Outcome-Dependent Sampling

Ashley I Naimi

August 2026

Contents

1 Sampling on the Outcome

Here we explore a set of tools generally referred to as **outcome-dependent sampling** (selection into the analytic cohort depends on the outcome value).

The canonical example in epidemiology is the case-control study, where we find people who experienced the event, and then sample only a fraction of the (typically much larger) group who did not. But this idea is far more general, and includes the (nested) case-cohort design, and a broader family of outcome-dependent sampling (ODS) designs. All share the same idea of evaluating the outcome to determine whom to collect (usually expensive) information on.

Why would we do this?:

1) **The outcome is rare.** If only a small fraction of a large cohort will ever become cases, then most of the information about the exposure-outcome relationship is concentrated in those people plus a modest comparison series. Collecting complete data on everyone usually returns very little additional precision.

2) **The exposure (or a key covariate) is expensive to measure.** Biomarker assays, genotyping, chart abstraction, and geocoded environmental linkage can cost hundreds of dollars *per person*. When outcome data already exist, e.g., in an established cohort, a registry, or an electronic health record system, measuring the expensive variable on a judiciously chosen subset can achieve nearly the same statistical efficiency as measuring it on everyone (Zhou et al., 2007; O'Brien et al., 2022).

3) **The dataset is huge.** Many claims and EHR datasets are now boasting trillions of observations, and while we typically will never build an analytic cohort using all these observations, we may still find that the number of records that end up in our analytic sample is so large, that the programs use for analysis take a long time to run. For instance, using the bootstrap for variance estimation is not always feasible in a huge analytic dataset.

The price we pay is that the sample no longer "looks like" the cohort. The outcome distribution in an outcome-dependent sample determined by design, so quantities we might have read directly from the data (e.g., risks and risk contrasts) are no longer directly estimable. The two questions that organize this entire lecture are:

- **Identification:** which cohort (target population) parameters can still be recovered from the biased sample, and under what assumptions?
- **Estimation:** how must the analysis change, in both the point estimate and the standard error, to account for the sampling design?

There are two general strategies for analyzing outcome-dependent samples. The first is to **re-weight** the sampled observations by the inverse of their sampling probabilities, restoring a "pseudo-population" that resembles the original cohort. The second is to **correct the likelihood** (or estimating function), conditioning each person's contribution on the fact that they were sampled. Classical case-control studies are a special case in which key design terms cancel entirely, and no correction is needed at all.

2 Motivating Data Example

We'll organize the first part of this lecture with a brief review (hopefully) of case-control studies. Consider 20,000 individuals free of the outcome at baseline sampled into a closed cohort. Say of these, 10,000 are exposed and 10,000 unexposed. All are followed for exactly 5 years with no loss to follow-up and no competing events.

By the end of follow-up, 1,813 of the exposed and 952 of the unexposed have experienced the outcome:

```
cohort_2x2 <- data.frame(
  exposure = c("Exposed", "Unexposed"),
  cases    = c(1813, 952),
  noncases = c(8187, 9048),
  total    = c(10000, 10000))
```

	Cases	Non-cases	Total
Exposed	1,813	8,187	10,000
Unexposed	952	9,048	10,000

The full-cohort two-by-two table. All 20,000 members of the hypothetical cohort, cross-classified by exposure and end-of-follow-up outcome status. With this complete cohort in hand, we can directly compute the odds ratio and risk ratio directly:

- The 5-year **risks** are $R_1 = 1{,}813/10{,}000 = 0.181$ in the exposed and $R_0 = 952/10{,}000 = 0.095$ in the unexposed, giving a **risk ratio** of $0.181/0.095 = 1.90$.
- The **odds** of the outcome are $1{,}813/8{,}187 = 0.221$ in the exposed and $952/9{,}048 = 0.105$ in the unexposed, giving a cohort **odds ratio** of $0.221/0.105 = 2.10$.

  We do not have person-time at risk here, but:

- If events occur uniformly over the 5 years, each case contributes on average 2.5 years at risk, so the exposed accrue $10{,}000 \times 5 - 1{,}813 \times 2.5 = 45{,}468$ person-years and the unexposed $47{,}620$. The **rates** are 0.0399 and 0.0200 per person-year, and the **rate ratio** is 2.00.

Now suppose the exposure had to be measured by an assay costing $500 per person, which means measuring everyone will cost $10 million. A case-control study that keeps all 2,765 cases and samples an equal number of controls will only need 5,530 assays, which will cost about $2.8 million (72% cost reduction). The designs in this lecture are about engineering that reduction without giving up a valid answer.

3 Case-Control Studies

The case-control study reverses the direction of a cohort study's data collection. Rather than classifying people by exposure and following them forward to observe outcomes, we identify people by their *outcome* status: **cases**, who experienced the event during the study period, and **controls**, sampled from those eligible to be compared with the cases, and then look *backward* to ascertain their exposure. For this reason the design is often called *retrospective*, though the label is about the sampling logic rather than calendar time: a case-control study can be (and, at its best, is) conducted inside a fully prospective cohort, sometimes referred to as the **study base** or source population (Wacholder et al., 1992a; Lash et al., 2021).

What to read: Lash & Rothman. Case-Control Studies. In: *Modern Epidemiology*, 4th Ed, Chapter 8.

In the context of a case-control study, it's useful to distinguish between three objects:

- the **source population** (say here, our 20,000-person cohort), in which the scientific question and the target parameter live;

- the **observed case-control sample**: all (or nearly all) of the cases, plus a modest series of controls, which is what we actually collect exposure data on; and
- the **target parameter**: a contrast defined in the source population (a risk ratio, rate ratio, or odds ratio), *not* a property of the sample.

To formalize the design, let Y denote the outcome indicator (1 = case by end of follow-up), A the binary exposure, and S an indicator that a person is *sampled* into the case-control study. The defining feature of the design is that sampling depends on the outcome:

$$\pi_1 = P(S = 1 \mid Y = 1), \qquad \pi_0 = P(S = 1 \mid Y = 0),$$

typically with π_1 near 1 (we take all the cases we can find) and π_0 small (a few controls per case). Critically, we assume sampling depends *only* on outcome status: given Y, whether a person is sampled has nothing to do with their exposure,

$$P(S = 1 \mid Y, A) = P(S = 1 \mid Y).$$

Violations of this assumption are one route to selection bias, and much of the craft of control selection, e.g., the principles of study base, deconfounding, and comparable accuracy (Wacholder et al., 1992a,b), is about making it plausible.

3.1 What Can and Cannot Be Estimated

What do the sampled data let us estimate directly? Within the sampled cases, the observed exposure prevalence estimates $P(A = 1 \mid Y = 1)$ in the source population; within the sampled controls, it estimates the exposure distribution of whatever group the controls were sampled from. In other words, case-control data can be used to compute **exposure distributions conditional on outcome**, in contrast to what we usually work with (outcome distributions (risks) conditional on exposure).

The quantities that are *not* directly estimable are just as important. For instance, risks and the risk difference are not directly computable in a case-control study because $P(Y = 1 \mid A = a)$ cannot be computed from the data. The ratio of cases to controls in the data was chosen by us. If we sample

one control per case, half the sample are cases regardless of whether the true 5-year risk is 14% (as here) or 0.1%. In fact, any feature of the the outcome distribution itself is not computable from case-control data. Any statement of the form "X% of our study participants experienced the outcome" describes a design decision and not the population of interest.

Note: The Sampled Outcome Distribution Is Not the Cohort's

This is the first and most common mistake in analyzing outcome-dependent samples: treating the case fraction in the data as if it were the risk in the population. It shows up in subtle forms. A predictive model's intercept, and therefore all of its predicted probabilities, fit to case-control data will reflect the designed case fraction, not the population risk. Similarly, the positive predictive value of a screening test cannot be estimated from a case-control sample, because predictive values depend directly on prevalence, which the design has erased. Absolute measures can be *restored* if external information on the population risk or prevalence is brought in (Greenland, 2004), or if the sampling fractions themselves are known. But nothing in the case-control data alone contains that information.

So what survives? It turns out that the exposure **odds ratio** contains important information within it that can be used to recover information about the outcome distribution within exposed and unexposed groups.

Within the case-control sample we can compute the odds of being exposed among cases, divided by the odds of being exposed among controls. The reason this quantity has a special role is that it is *invariant to the outcome-dependent sampling*: it equals the corresponding contrast in the source population, with the sampling fractions cancelling exactly.

Let's return to the cohort two-by-two table and denote the four cohort cell counts by N_{ay}, for exposure level a and outcome level y; e.g., $N_{11} = 1{,}813$ is the number of exposed cases. Suppose we sample cases with probability π_1 and controls with probability π_0. Because sampling ignores exposure given outcome, the *expected* cell counts in the case-control sample are just the cohort counts scaled by the relevant sampling fraction:

sampled cases:	$\pi_1 N_{11}$ exposed,	$\pi_1 N_{01}$ unexposed,
sampled controls:	$\pi_0 N_{10}$ exposed,	$\pi_0 N_{00}$ unexposed.

The exposure odds ratio computed from the sample is the familiar cross-

product ratio of these four cells:

$$\widehat{OR}_{\text{caco}} = \frac{\pi_1 N_{11} \times \pi_0 N_{00}}{\pi_0 N_{10} \times \pi_1 N_{01}} = \frac{N_{11} N_{00}}{N_{10} N_{01}}.$$

Both π_a's cancel *exactly*, without our ever needing to know their values. The case-control exposure odds ratio equals the cross-product ratio of the full cohort table, even though we never observed the full cohort table (Cornfield, 1951; Miettinen, 1976).

Note two important features here: First, the cancellation required $P(S = 1 \mid Y, A) = P(S = 1 \mid Y)$ (the probability of being selected into the case-control sample is independent of the exposure). If exposed cases were sampled differently from unexposed cases, the π_a's would not cancel. Second, notice that the result says the case-control odds ratio equals the *cohort* cross-product ratio of whatever table the controls represent. We expand on this latter issue next.

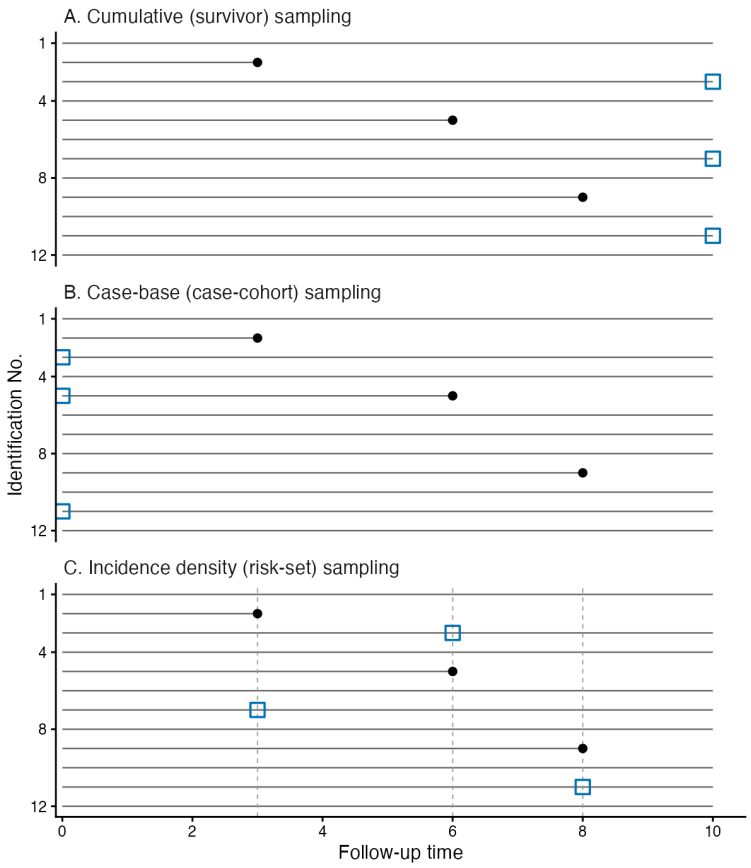

Figure 1: Three ways to sample controls from the same 12-person cohort, in which three participants (2, 5, and 9) become cases (solid dots) at times 3, 6, and 8. Open squares mark sampled controls, positioned at the time their control status is defined. Panel A (cumulative): controls are drawn at the end of follow-up from those who never became cases. Panel B (case-base): controls are drawn at baseline from the full cohort, regardless of what happens later – note that participant 5, a future case, is legitimately sampled as a control. Panel C (incidence density): at each case's event time (dashed lines), a control is drawn from those still at risk at that moment. All three schemes keep the same cases; they differ only in the comparison series, and consequently in which cohort parameter the exposure odds ratio identifies.

3.2 Cumulative Sampling Identifies the Cohort Odds Ratio

The scheme implicitly assumed in the algebra above, and in most textbook introductions, is **cumulative sampling** (sometimes "survivor sampling"): wait until the end of the follow-up period, then sample controls from among those who never became cases. The control series then represents the *non-cases*, N_{10} and N_{00}, which is exactly the situation of the cancellation argument. The sample exposure odds ratio therefore estimates the cohort **odds ratio**, $[R_1/(1-R_1)]\,/[R_0/(1-R_0)]$.

In our motivating cohort, suppose we keep all 2,765 cases ($\pi_1 = 1$) and sample 2,765 of the 17,235 non-cases ($\pi_0 = 0.160$). In expectation we obtain 1,313 exposed and 1,452 unexposed controls, so:

$$\widehat{OR}_{\text{caco}} = \frac{1{,}813/1{,}313}{952/1{,}452} = \frac{1.381}{0.656} = 2.10,$$

which reproduces the cohort odds ratio of 2.10 (in expectation), despite using only 16% of the non-cases.

3.3 Recovering the Cohort Risk Ratio

Case-control odds ratios are often interpreted as risk ratios, but this is not always justified. There are three ways in which this will be the case:

The rare-disease approximation holds. When the outcome is rare in both exposure groups, $1-R_1 \approx 1-R_0 \approx 1$, and the cohort odds ratio approximately equals the cohort risk ratio (Knol et al., 2012). But note critically that this has nothing to do with the fact that we did a case-control study; it is not a feature of the case-control odds ratio. It is just a simple fact about the relationship between the cohort odds ratio and the cohort risk ratio.

An exact conversion using the baseline risk. The odds ratio and risk ratio are deterministically linked through the baseline risk. Writing R_0 for the baseline (unexposed) risk, note that one converts to the other exactly:

$$RR = \frac{OR}{(1-R_0) + R_0 \times OR}.$$

Under cumulative sampling, the case-control odds ratio returns the cohort odds ratio (denoted OR), but the baseline risk R_0 must come from *outside* the case-control data, e.g., from surveillance data, from the source cohort itself

if the study is nested, or from the known sampling fractions. No rare-disease approximation is needed, just one piece of absolute-occurrence information that outcome-dependent sampling altered (Greenland, 2004).

Change the control-sampling scheme. This is a more elegant route and is based entirely on how controls are sampled. Instead of sampling controls from the non-cases at the end of follow-up, we sample them from the entire cohort at baseline, before anyone's outcome is known.

This is called "case-base sampling" or "inclusive sampling" (when embedded in an existing cohort, it is the control-sampling scheme of the *case-cohort* design of the next section). Under case-base sampling with fraction π_b, the expected control counts are $\pi_b N_{1+}$ exposed and $\pi_b N_{0+}$ unexposed, where N_{a+} is the total number of cohort members with exposure a, cases included. Then, the case control odds ratio is:

$$\frac{N_{11}/(\pi_b N_{1+})}{N_{01}/(\pi_b N_{0+})} = \frac{N_{11}/N_{1+}}{N_{01}/N_{0+}} = \frac{R_1}{R_0} = RR,$$

or the **risk ratio itself**, exactly, with no rare-disease approximation and no external information (Miettinen, 1976; Rothman et al., 2008). The trick is that a baseline sample of the cohort estimates the exposure distribution of the *denominators* of the risks, rather than of the non-cases. In our numbers, 2,765 controls drawn from the 20,000-person baseline cohort yield, in expectation, 1,382.5 exposed and 1,382.5 unexposed controls, so that:

$$\frac{1{,}813/1{,}382.5}{952/1{,}382.5} = \frac{1{,}813}{952} = 1.90.$$

Two caveats travel with this result: some sampled controls will later become cases, and complete follow-up of the cohort's case count is assumed, i.e., no losses to follow-up and no competing events (O'Brien et al., 2022).

3.4 Recovering the Cohort Rate Ratio

Under incidence density or "risk-set" sampling, controls are sampled longitudinally throughout follow-up: whenever a case occurs, one or more controls are drawn from the people still at risk **at that moment** (Panel C of the figure). A person's chance of being sampled as a control is then proportional to the amount of *time* they spend at risk, so the exposure distribution among the controls estimates the exposure distribution of the cohort's **person-time**. Writing

PT_1 and PT_0 for the exposed and unexposed person-time, the expected control series is proportional to (PT_1, PT_0), and the sample cross-product becomes:

$$\frac{N_{11}/PT_1}{N_{01}/PT_0} = \frac{\text{rate}_1}{\text{rate}_0} = \text{rate ratio},$$

the cohort **incidence rate ratio**, again exactly and again with no rare-disease assumption (Miettinen, 1976; Rothman et al., 2008). In our motivating cohort, this returns $0.0399/0.0200 = 2.00$. The requirements are that controls be sampled in proportion to person-time at risk, which risk-set sampling accomplishes automatically, and that exposure be defined *as of the sampling time*. Note also that under density sampling a person may be sampled as a control at one time and later become a case, and may even be sampled as a control more than once; both are correct features of the scheme, since the person-time denominators they represent really do include that time.

When each control is *matched* to its case's event time, and the matching retained in the analysis, this design is the **nested case-control study**, whose conditional-logistic analysis estimates the rate (hazard) ratio of a proportional hazards model; we return to that machinery in Week 11 (Lash et al., 2021; O'Brien et al., 2022).

Note: "The" Odds Ratio from a Case-Control Study

The phrase *case-control odds ratio* describes an estimator (a cross-product ratio computed from the sampled two-by-two table), not an estimand. Depending on the control-sampling scheme, the identical arithmetic estimates a cohort odds ratio (2.10 in our example), a cohort risk ratio (1.90), or a cohort rate ratio (2.00). The key message: a case-control odds ratio need not be an "odds ratio" that we'd expect in a cohort study.

4 Case-Cohort Studies

The case-base sampling idea, i.e., "draw the comparison series from the baseline cohort, blind to future outcomes", becomes even more powerful when we embed it in an existing, well-characterized cohort and carry the full time-to-event information along. This is the **case-cohort design** (Prentice, 1986; O'Brien et al., 2022).

What to read: O'Brien et al. The Case for Case-Cohort. *Epidemiology* 2022;33(3):354–361

Suppose we have a large prospective cohort, e.g., tens of thousands of

participants, enrolled with baseline questionnaires and banked biological specimens, and outcomes accruing over years of follow-up (think of the Nurses' Health Study or the Sister Study). Say that a new hypothesis arises: does a serum biomarker, measurable in the banked baseline blood at $150 per assay, predict incident breast cancer?

Assaying the entire cohort would cost millions and consume irreplaceable specimen volume. The outcome information, meanwhile, is already collected for everyone. The question is which specimens should we assay?

4.1 The Design and the Subcohort

A case-cohort sample consists of two, possibly overlapping, pieces (Prentice, 1986; O'Brien et al., 2022):

1) the **subcohort**: a random sample of the full cohort, drawn *at baseline* and hence without any regard to future case status (i.e., cases and non-cases are sampled). We denote the subcohort sampling fraction by q; a subcohort might be, say, a 10% random sample of the cohort;

2) **all the cases**: everyone in the cohort who develops the outcome during follow-up, whether or not they happen to fall in the subcohort.

The expensive covariate is then measured *only* for subcohort members and cases. Because the subcohort is drawn at random from the baseline cohort, some of its members will, by chance, become cases: this overlap is expected and handled explicitly by the analysis.

Note the direct lineage from the previous section: the subcohort is exactly a case-base control series, but instead of reducing each person to a single exposed/unexposed count, we retain their full event-time and censoring information, and can thus deploy more advanced methods (censoring, competing risks) than a two-by-two table.

To make the construction concrete, return to the motivating cohort of 20,000, but keep the time axis rather than collapsing data into a two-by-two table. Everyone enters at time zero and is followed for up to 5 years. Each of the 2,765 cases has an event time T_i recorded somewhere in $(0, 5]$, averaging 2.5 years under the uniform-occurrence device from the motivating example, while the 17,235 non-cases are followed for the full 5 years; the cohort as a

whole therefore accrues $45{,}468 + 47{,}620 = 93{,}088$ person-years. Suppose the assay budget supports a 10% subcohort, $q = 0.10$. The case-cohort sample can be assembled as follows:

1) **Draw the subcohort at baseline.** A random 10% of the 20,000 yields 2,000 subcohort members: in expectation, 1,000 exposed and 1,000 unexposed, together accruing one-tenth of the cohort's person-time (about 9,309 person-years). Because the draw is blind to the future, the subcohort contains, in expectation, $0.10 \times 2{,}765 \approx 277$ people who will go on to become cases, while the remaining ≈ 1,723 remain non-cases throughout follow-up.

2) **Add all the cases.** The 2,765 cases join the sample, bringing their event times. About 277 of them are already in the subcohort, so the union contains $2{,}000 + 2{,}765 - 277 = 4{,}488$ *distinct* people (i.e., nobody appears twice). The ≈ 2,488 cases *outside* the subcohort contribute their event time and their assayed exposure, but their earlier follow-up will not be used as comparison person-time, because the subcohort already represents it.

We then use time-varying weights to appropriately weight the subcohort's person-time. Scaling up by $1/q = 10$ reconstructs the cohort's 93,088 person-years of comparison time, but each of the 2,765 events must be counted exactly once; and the analysis must prevent the ≈ 277 subcohort cases from being counted both as scaled-up subcohort events *and* as sampled cases.

The figure below draws out what this weighting scheme looks like in practice:

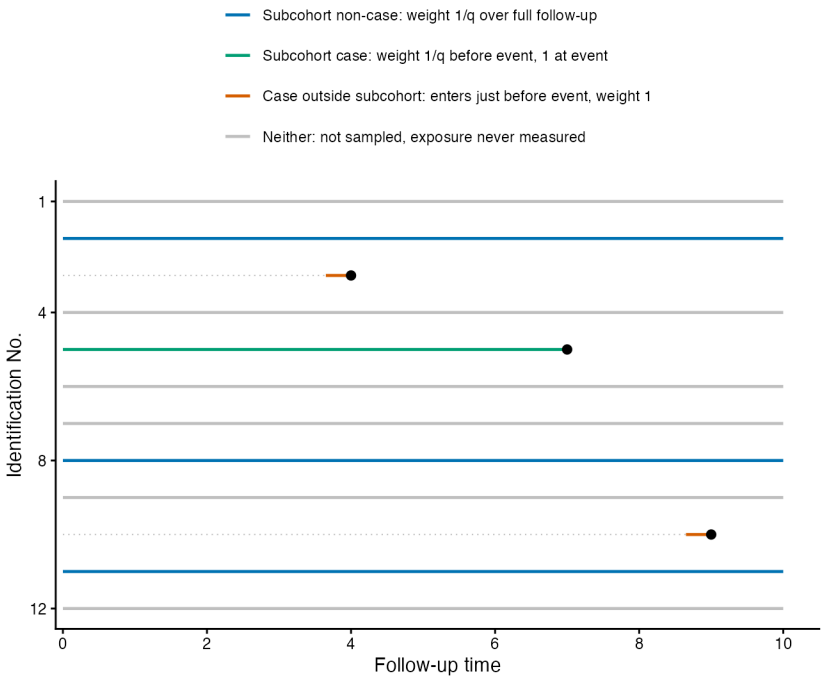

Figure 2: A case-cohort sample drawn from the 12-person cohort, with a 4-person subcohort (participants 2, 5, 8, 11) and three cases (solid dots: participants 3, 5, and 10). Line color encodes each person's role and analysis weight. Subcohort non-cases (blue) contribute all their follow-up at weight 1/q. Participant 5, a subcohort member who becomes a case (green), contributes follow-up at weight 1/q until just before the event, then weight 1 at the event itself. Cases outside the subcohort (orange) enter the analysis only just before their event, at weight 1; their earlier follow-up (dotted) is not used. The remaining cohort members (grey) are never sampled and their exposure is never measured.

4.2 Why Case-Cohort Designs?

Why prefer this over the nested case-control design of the previous section? Flexibility O'Brien et al. (2022). **First,** one subcohort can be used for many outcomes. Because the subcohort was drawn without reference to *any* outcome, the same subcohort, with its expensive covariates already measured, can serve as the comparison group for breast cancer today, ovarian cancer next year, and stroke after that, and only the new cases need assaying each time. A nested case-control study's controls, by contrast, are selected (and often matched) specifically for one case series and one outcome. Each new outcome usually requires a new control series, and a control group required to remain free of an ever-growing list of diseases becomes progressively less representative of the source population's exposure distribution. **Second,** the subcohort is a standalone random sample, which means that cross-sectional and case-independent questions (e.g., how does the biomarker vary with age, race, or neighborhood?) can be answered directly in the subcohort, because it is a miniature of the full cohort. **Third,** New measurements can be layered onto the same subcohort (e.g., genotyping added to an earlier biomarker study), and the design remains compatible with pooling and meta-analysis.

When is the case-cohort design *not* the best tool?: with a single pre-specified rare outcome, staggered entry, and strong concerns about specimen degradation or batch effects (where matching cases and controls on storage time and analytic batch is protective), a nested case-control design can be somewhat more statistically efficient and logistically safer (Langholz and Thomas, 1990; Wacholder, 1991; O'Brien et al., 2022). The efficiency differences for a single question are usually modest; the flexibility differences across a cohort's scientific lifetime are not.

4.3 Estimation: Weighting the Sample Back to the Cohort

Had we measured the biomarker on everyone, the working analysis (in this course, always some form of regression) might be a Cox proportional hazards model,

$$h(t \mid X) = h_0(t)\exp(\beta' X),$$

where X contains the exposure of interest and adjustment covariates, $h_0(t)$ is an unspecified baseline hazard (a nuisance function, in Week 2's language), and $\exp(\beta_j)$ is the hazard ratio for a one-unit difference in X_j; we treat the Cox model's own mechanics, and Week 2's caveats about hazard ratios as causal estimands, in Weeks 12 and 13. The target parameter β is defined by the *full-cohort* model. The estimation problem is that our sample over-represents cases (all of them are in) and under-represents everyone else (only a fraction q).

The modern way to see the solution is through **inverse probability of sampling weights** (O'Brien et al., 2022; Barlow, 1994). Each sampled person is weighted by the reciprocal of the probability that a person like them entered the sample, creating a pseudo-population that reconstructs the full cohort:

- a **subcohort member** stands in for $1/q$ cohort members (a 10% subcohort gives each member weight 10) over their entire follow-up;
- a **case** must count, at their event time, exactly once: cases were sampled with probability 1, so at the moment of the event the weight is 1.

The subtlety, and it is the key to understanding the whole analysis, concerns cases *inside* the subcohort. If subcohort members all carried weight $1/q$ throughout, then the subcohort's cases would already be weighted up to

represent *all* the cohort's cases, in expectation; adding the sampled cases from outside the subcohort at any positive weight would then double-count events. The resolution (O'Brien et al., 2022) is to make the weights *time-varying*: a subcohort case carries weight $1/q$ from baseline until just before their event, i.e., while serving as part of the comparison person-time, and weight 1 *at* the event, exactly like every other case. A case outside the subcohort contributes nothing until just before their event (t minus a vanishingly small ϵ), then enters with weight 1. In this scheme every case's event is counted exactly once, at weight 1, and all comparison person-time is counted at weight $1/q$; this is the structure drawn in the figure above.

With the weights defined, estimation can proceed by the **weighted partial likelihood** (historically called a *pseudo-likelihood*) (Prentice, 1986; Self and Prentice, 1988; Barlow, 1994). We'll see this more in Week 13, but the Cox partial likelihood compares, at each observed event time t_j, the case's covariates to those of everyone still at risk. The weighted score equation solved by the case-cohort estimator is:

$$\sum_{j:\,\text{events}} \left\{ X_{(j)} - \frac{\sum_{i \in \mathcal{R}(t_j)} w_i(t_j)\, X_i \exp(\beta' X_i)}{\sum_{i \in \mathcal{R}(t_j)} w_i(t_j) \exp(\beta' X_i)} \right\} = 0,$$

where $X_{(j)}$ is the covariate vector of the case at the j-th event time, $\mathcal{R}(t_j)$ is the *sampled* risk set at t_j, and $w_i(t_j)$ are the weights just described. In words: at each event, the estimator asks "how does this case's exposure compare to the weighted average exposure of the people who were also at risk at this moment?", precisely what the full-cohort Cox model asks, except that the risk-set average is now *estimated from the weighted subcohort* rather than computed from the entire cohort.

To see the equation's moving parts, take a single event time and a single binary exposure (so $X = A$ and β is a scalar). Suppose the case is exposed, so $X_{(j)} = 1$, and that the sampled risk set at that moment holds three subcohort members — one exposed and two unexposed, each at weight $1/q = 10$ — plus the case itself at weight 1. Exposed people enter both sums as $w_i e^{\beta}$; unexposed people contribute only w_i to the denominator, since $e^{\beta \times 0} = 1$ and $X_i = 0$ erases them from the numerator. This event's term is therefore:

$$X_{(j)} - \frac{10\, e^{\beta} + 1 \cdot e^{\beta}}{10\, e^{\beta} + 1 \cdot e^{\beta} + 10 + 10} = 1 - \frac{11\, e^{\beta}}{11\, e^{\beta} + 20},$$

that is, the case's own exposure minus the weighted share of the at-risk "person-equivalents" that are exposed (11 of 31, when $\beta = 0$).

The full score is just one such term per event, each computed with the weights in force at that moment (a subcohort case counts at weight 10 in other people's risk sets, but at weight 1 in the term for its own event), and $\hat{\beta}$ is the value that makes the terms sum to zero. That one-dimensional root-finding is exactly what `coxph()` performs when handed the weighted records.

This is the sense in which the case-cohort estimator "relates to the estimator we would have obtained from the complete cohort": it solves the same equation with the full-cohort risk-set averages replaced by unbiased sampled estimates of them, and it converges to the same β as the full-cohort estimator, at the cost of extra sampling variability (Prentice, 1986; Self and Prentice, 1988).[1]

Two practical notes. First, the same weighting logic is not Cox-specific: weighted logistic, Poisson, or additive-hazards regressions of the sort we build in Weeks 6 and 11 accommodate case-cohort data identically, i.e., weight the observations, then fit (O'Brien et al., 2022; Ding et al., 2017). Second, the framework extends immediately to **stratified** designs: if the subcohort oversamples a subgroup (sampling fraction q_A in group A, q_B in group B), or if only a fraction of cases is sampled, the weights simply become the reciprocal of each person's own, known, sampling probability (O'Brien et al., 2022). A useful design check before modeling: build a weighted frequency table of the case-cohort sample; the weighted person-time, covariate distributions, and case counts should approximately reproduce the full cohort's (O'Brien et al., 2022).

[1] Several asymptotically equivalent variations exist, differing in exactly how the risk set is constructed: Prentice's original proposal includes each non-subcohort case in the risk set just before its own event time (the version above, and the one implemented in standard software); Self and Prentice's version uses the subcohort alone; and later authors proposed more efficient variants that reuse all cases throughout, as reviewed in Ding et al. (2017). The differences matter little in practice, and the weighted-partial-likelihood framing covers the design variants an applied analyst will actually encounter.

4.4 Variance Estimation for the Case-Cohort Design

Fitting the weighted model and reading off the standard errors printed by default is **wrong**, and understanding why is the second key lesson of the design.

A model-based (information-based) variance treats the data as if they were $\sum_i w_i$ independent observations: a subcohort member with weight 10 is counted as if she were 10 people. The weighted pseudo-population is *larger* than the actual sample, so the naive variance is generally too small, and confidence intervals too narrow. Moreover, the same subcohort members appear in the comparison group at (nearly) every event time, so the terms of the score function are correlated across risk sets in a way full-cohort theory does not

anticipate. The result is that a naive analysis reports the precision that the *full cohort* analysis would have earned, when in truth we measured the exposure on a fraction of it.

The additional variability has a specific source: **the sampling of the subcohort itself**. Had a different 10% random subcohort been drawn, the estimated risk-set averages, and hence $\hat{\beta}$, would differ. A valid variance estimator must add this design-based, between-possible-subcohorts variability to the usual model-based uncertainty. The practical solution, advocated by O'Brien et al. (2022) and standard since Barlow (1994), is the **robust (sandwich) variance estimator**, which we treat in its own right in Week 7; for now the logic suffices:

- For each sampled individual, compute their **influence** on $\hat{\beta}$: the amount $\hat{\beta}$ would shift if that individual were removed (the `dfbeta` residual of Week 7's sandwich machinery).
- Estimate the variance of $\hat{\beta}$ by the empirical variability of these influence contributions, aggregated *by individual*, so that a subcohort case's multiple weighted records (pre-event and at-event) are summed into one influence per person before squaring.

Because each person, not each weighted record, is the independent unit, this estimator correctly captures both the outcome-model uncertainty and the subcohort-sampling uncertainty. It is what SAS's PHREG (via `COVSANDWICH`), Stata, and R's `survival` package (via `cluster(id)` or `robust = TRUE`, or the purpose-built `cch()` function) compute (Barlow et al., 1999; Therneau and Li, 1999; O'Brien et al., 2022).[2]

Week 11's lab turns this weighting-and-variance pipeline into practice, once the regression toolkit is in hand. Lab 2, for its part, works the first half of this lecture: starting from the cohort you built in Lab 1, it draws all three control-sampling schemes and verifies, against the fully enumerated cohort, that each scheme's exposure odds ratio recovers its own cohort parameter — the odds ratio, the risk ratio, and the rate ratio in turn.

[2] One refinement you may encounter: because the subcohort is drawn *without replacement* from a finite cohort, the plain sandwich estimator is slightly conservative, and finite-population corrections (subtracting a term proportional to q) exist, as do the original asymptotic variance formulas of Prentice (1986) and Self and Prentice (1988). For an applied analysis, the robust sandwich estimator, with individuals as the clustering unit, is the recommended default: it is simple, general, at worst mildly conservative, and it extends unchanged to the stratified and generalized designs.

4.5 Implementation Summary

An applied analyst's checklist for a case-cohort analysis:

1) **Construct the sample.** Draw a random subcohort of fraction q from the baseline cohort (optionally stratified, with known fractions per stratum);

identify all cases over follow-up; measure the expensive covariates on the union of the two.

2) **Define inclusion indicators and weights.** Record, for every sampled person, subcohort membership and case status, and assign weights: $1/q$ for subcohort person-time (using each person's own stratum-specific q if stratified), weight 1 for every case at their event, with subcohort cases switching from $1/q$ to 1 just before the event, and non-subcohort cases entering just before their event.
3) **Check the weighted sample.** Weighted person-time, case counts, and available covariate distributions should approximately match the full cohort.
4) **Fit the weighted regression.** Weighted Cox partial likelihood for hazard ratios (or a weighted GLM for risk/rate models), with the weights from step 2.
5) **Compute the robust variance.** Sandwich estimator with the *individual* as the clustering unit (`cluster(id)`); never the model-based variance.
6) **Report** the estimates, robust standard errors, confidence intervals, and, always, the sampling design itself: the subcohort fraction, stratification, and case ascertainment, since without these the analysis cannot be reproduced or even interpreted.

5 Outcome-Dependent Sampling

Step back and look at the two designs so far through one lens. In both, we chose whose expensive data to collect *using the outcome*: case-control sampling conditions on a binary end-of-follow-up status; case-cohort sampling conditions on whether the failure event occurred during follow-up. Both concentrate measurement where the information is, on the (rare) cases, while a cheap random sample covers the denominators.

Outcome-dependent sampling (ODS) is the general principle of which these are the two oldest special cases: *observe the expensive exposure with a probability that depends on the observed value of the outcome*, deliberately oversampling the outcome values that carry the most information about the exposure-outcome relationship (Ding et al., 2017). Once stated this way, the idea plainly does not require a binary outcome:

- With a **continuous outcome**, the informative observations are typically those

in the *tails* of the outcome distribution: if the exposure is related to the outcome, individuals with unusually high or low outcome values are disproportionately informative about the slope, just as, in a case-control study, the (rare) cases were disproportionately informative. The corresponding ODS design supplements a simple random sample with additional draws from the extremes (Zhou et al., 2002, 2007).

What to read: Zhou et al. An efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology* 2007;18(4):461–8

- With a **censored failure time**, we can go beyond the case-cohort design's case/non-case dichotomy and let selection depend on *when* the event occurred: sample all or most of the unusually early failures (and perhaps the unusually late ones), plus a random sample of everyone (Ding et al., 2014, 2017).
- With **longitudinal binary responses**, the informative subjects are those whose outcome *varies* over follow-up, and sampling can depend on a summary of each person's whole response vector (Schildcrout and Heagerty, 2011).

Two running examples make this concrete. Ding et al. (2014) studied effects of environmental contaminants on time-to-pregnancy in the Norwegian Mother and Child (MoBa) cohort: with an assay budget covering only a fraction of the eligible women, they measured the exposure on a small random sample *plus* supplemental samples drawn from the most informative regions of the time-to-pregnancy distribution. Schildcrout and Heagerty (2011) planned a substudy of the Childhood Asthma Management Program continuation cohort (CAMPCS), asking whether an interleukin-10 polymorphism modifies the decline of asthma symptoms through adolescence: symptom outcomes (whether a physician had been contacted about symptoms since the last visit) were already recorded quarterly for everyone, genotyping the stored blood was the expensive step, and 60% of children *never* contacted a physician over the whole of follow-up, contributing very little information about symptom dynamics. In both cases the cohort's outcome data were already complete; the design question was purely *whose specimens to run through the lab*.

5.1 The Design

We describe the two-component ODS design in the failure-time setting of Ding et al. (2014; as reviewed in Ding et al., 2017); the continuous-outcome and

longitudinal versions share its anatomy. Partition the range of observed event times into K mutually exclusive strata $\mathcal{A}_k = (a_{k-1}, a_k]$ using pre-specified cut-points. The ODS sample then consists of:

1) a **simple random sample (SRS)** of size n_0 from the full cohort, drawn without regard to anything, exactly like a subcohort; and
2) **supplemental samples**: from each time-stratum $\mathcal{A}_k$, an additional random sample of size n_k drawn *from the cases whose events fell in that stratum*, with the allocation $(n_1, \ldots, n_K)$ chosen to oversample the informative regions (in the MoBa study, the extremes).

The expensive exposure is measured only for these $n_0 + n_1 + \cdots + n_K$ individuals. The case-cohort design is the special case with a single stratum containing all cases sampled with probability one; the extra generality is that selection may now depend on the *timing*, not just the occurrence, of the event, and that not all cases need be sampled, which matters when the outcome is not rare and even the cases are too numerous to afford (Ding et al., 2017).

In the longitudinal binary setting of Schildcrout and Heagerty (2011), the same anatomy appears with the time-strata replaced by **response-vector strata**. With Y_{ij} the binary outcome for subject i at visit $j = 1, \ldots, n_i$, classify each subject by their total $\sum_j Y_{ij}$ into three strata: *never* ($\sum_j Y_{ij} = 0$), *sometimes* ($0 < \sum_j Y_{ij} < n_i$), and *always* ($\sum_j Y_{ij} = n_i$). The design samples subject i with a probability determined by their stratum:

$$P(S_i = 1 \mid Y_{i1}, \ldots, Y_{in_i}) = \pi(g_i), \qquad g_i \in \{\text{never}, \text{sometimes}, \text{always}\},$$

with the $\pi(\cdot)$'s fixed by the investigator. In the CAMPCS substudy, the most efficient designs took *all* of the "sometimes" children, those whose symptoms varied, and hence who carry the information about change over time, and few or none of the others: relative to random sampling of the same number of children, this cut standard errors for the genotype-by-time interaction by roughly a third, equivalent to a much larger study at no additional assay cost (Schildcrout and Heagerty, 2011).

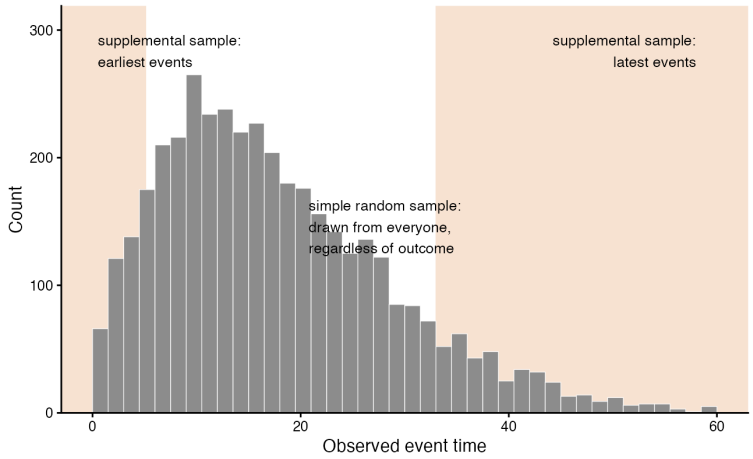

Figure 3: Anatomy of a two-component outcome-dependent sampling design for a failure-time outcome. The histogram is the cohort's distribution of observed event times. The expensive exposure is measured for a simple random sample drawn from the entire cohort (providing an unbiased picture of the population and its denominators) plus supplemental samples drawn deliberately from the shaded, most informative regions – here, the earliest and latest events. Subjects in the shaded strata are consequently over-represented in the analytic sample relative to the cohort, and the analysis must correct for exactly this over-representation.

5.2 Why the Naive Analysis Fails

Suppose we simply pooled the SRS and the supplemental samples and fit the ordinary regression model, e.g., the Cox model, or the logistic model for the longitudinal outcomes, as if the data were a random sample. What goes wrong?

The sample's outcome distribution is distorted by construction: subjects with informative outcomes appear at many times their cohort frequency. A likelihood built from $P(\text{outcome} \mid \text{exposure})$ evaluated on such a sample is simply the wrong likelihood: it describes data drawn at random, and ours were not. The fitted coefficients converge to values that blend the true regression parameters with the sampling design, and the reported standard errors describe the wrong experiment.

It is worth being precise about why this may be surprising. In the classical case-control setting, a famous result of Prentice and Pyke (1979) says that ordinary logistic regression applied to case-control data as if prospectively sampled gives valid estimates of every coefficient *except the intercept*, which absorbs the sampling fractions (this is why "run the logistic regression and ignore the design" is the standard, and correct, analysis of a simple case-control study; only absolute risk predictions are off, which returns us to the note earlier in this lecture). That result is a small miracle specific to its circumstances: a *binary* outcome, the *logistic* link, and sampling that depends on the outcome *only through that binary value*. None of these hold under general ODS: with a continuous or failure-time outcome there is no intercept in which the design

can quietly hide, and with sampling based on response-vector summaries, the distortion contaminates the covariate coefficients themselves, including the exposure effects we care about (Schildcrout and Heagerty, 2011; Ding et al., 2017). Under general outcome-dependent sampling, an unadjusted analysis is biased, not just miscalibrated.

So the design must enter the analysis. As throughout this lecture, there are two principled ways in.

5.3 Estimation Strategy 1: Inverse Probability Weighting

The first strategy is by now familiar: weight each sampled subject by the inverse of their known sampling probability, $w_i = 1/\pi(g_i)$, and solve the weighted version of the estimating equations we would have used on the full cohort (Cai et al., 2001; Yu et al., 2015). The "sometimes" child sampled with certainty gets weight 1; a "never" child sampled at $\pi = 0.05$ stands in for 20 such children. In expectation, the weighted sample score equals the full-cohort score, which is exactly the argument that justified the case-cohort weights, so the weighted estimator targets the same β as a full-cohort analysis. Yu et al. (2015) develop precisely this weighted pseudo-score approach for the failure-time ODS design under the additive hazards model, with the weights built from the known design quantities: the stratum sizes in the cohort and the numbers sampled from each.

The requirements are minimal and worth stating: the sampling probabilities must be *known* (they are, when we designed the sampling) or estimable from cohort counts (sampled fraction of each stratum), and every stratum with nonzero cohort representation must have a nonzero sampling probability; the outcome and stratum labels must be available for the whole cohort, which is what lets us compute the stratum-specific sampled fractions in the first place. Its weaknesses are also familiar from the case-cohort discussion, but amplified: when a design deliberately samples few subjects from an uninformative stratum, those few carry enormous weights, and estimators in which a handful of observations carry extreme weight are noisy. This is not a mere theoretical worry: in the simulation studies of Schildcrout and Heagerty (2011), standard errors from weighted estimation were, in the least favorable settings, more than 50% larger than those from the likelihood-based alternative of the next subsection, precisely because the few sampled "never" children carried very

large weights.

5.4 Estimation Strategy 2: Correcting the Likelihood

The second strategy fixes the likelihood rather than the sample. If subject i is in our data *because* they were sampled, then the correct probability of their data is not $P(y_i \mid x_i)$ but the probability *conditional on having been sampled*:

$$L_i = P(y_i \mid x_i, S_i = 1) = \frac{P(S_i = 1 \mid y_i)\, P(y_i \mid x_i; \theta)}{P(S_i = 1 \mid x_i; \theta)},$$

which is just Bayes' theorem, with the numerator using the design's known sampling rule (given the outcome, sampling ignores x) and the denominator averaging that rule over the outcomes the model says subject i *might* have had:

$$P(S_i = 1 \mid x_i; \theta) = \sum_g \pi(g)\, P(\text{outcome in stratum } g \mid x_i; \theta).$$

Maximizing $\sum_i \log L_i$ gives the **ascertainment-corrected maximum likelihood (ACML)** estimator of Schildcrout and Heagerty (2011). Read the correction as a penalty for predictability: each subject's ordinary likelihood contribution is divided by their probability, according to the model, of being ascertained at all. A subject whose covariates already made them very likely to be sampled (say, covariates that strongly predict "sometimes" status) contributes less new information than a subject whose being sampled was a surprise, and the denominator prices this in. Note also what the correction does *not* require: only the *relative* sizes of the $\pi(g)$'s matter (multiplying all of them by a constant cancels from the ratio), so the correction is driven by the design's oversampling pattern, not by absolute sampling rates.

For the three-stratum longitudinal design, the correction term is computationally cheap in a way that looks almost accidental but is worth appreciating. The denominator needs $P(\text{never} \mid x_i; \theta)$, $P(\text{always} \mid x_i; \theta)$, and $P(\text{sometimes} \mid x_i; \theta)$, and the first two are simply the model's likelihood evaluated at two "pseudo-outcomes": the all-zeros and all-ones response vectors, with "sometimes" available as one minus their sum. So ACML can be implemented with three calls to any standard longitudinal likelihood routine per subject: one at the observed data and two at the pseudo-outcomes (Schildcrout and Heagerty, 2011).

The same conditioning idea drives estimation for the failure-time ODS design, with one added wrinkle. In the likelihood of Ding et al. (2014), an SRS subject contributes their ordinary density; a supplemental subject from stratum $\mathcal{A}_k$ contributes their density *divided by the population probability of being an observed case in* $\mathcal{A}_k$, i.e., the probability that a random cohort member would have qualified for that stratum, again "the probability of being ascertained." The wrinkle is that under the semiparametric Cox model this denominator involves the unspecified baseline hazard and the censoring distribution, i.e., nuisance functions. The solution is to *estimate* the nuisance components nonparametrically from the SRS component, which, being a genuine random sample, identifies them, plug them into the corrected likelihood, and maximize the result over β; Ding et al. (2014) call this an *estimated maximum semiparametric empirical likelihood* estimator, and show it is consistent, asymptotically normal, and more efficient than both simple random sampling with the same assay budget and the comparable case-cohort analysis (Ding et al., 2014, 2017). The technical scaffolding is beyond what we need in this course; the structure to retain is that it is the *same* Bayes-theorem correction as ACML, with nuisance functions estimated from the design's random-sample component.

Why this works: both strategies succeed for the same underlying reason, and seeing it unifies the whole lecture. The design guarantees that, *given the outcome (and design strata), sampling is independent of the expensive exposure*, and the sampling probabilities are known by design. Weighting uses that knowledge to undo the distortion in the *sample*; likelihood correction uses it to build the distortion into the *model*. The assumptions for identification are: (i) sampling depends only on the outcome quantities that define the design strata, never additionally on the unmeasured exposure; (ii) the sampling probabilities (or the cohort stratum counts) are known; (iii) the cohort itself represents the target population; and, for the likelihood route, (iv) the outcome model is correctly specified, since the correction term is computed from it.

5.5 Variance Estimation for Outcome-Dependent Samples

The two strategies come with different variance stories, and keeping them straight is the difference between honest and dishonest confidence intervals.

For the weighted estimators, everything said about the case-cohort variance applies with full force: the usual model-based variance from the weighted fit

pretends each weighted subject is w_i independent people, understating uncertainty exactly when the design is most aggressive (large weights). A **sandwich (robust) variance estimator** is mandatory, with the subject as the independent unit, so that in the longitudinal setting all of a subject's repeated observations aggregate into a single influence contribution before squaring (Cai et al., 2001; Yu et al., 2015). In the sandwich, the "bread" (the derivative of the weighted score) and the "meat" (the empirical variance of subject-level weighted score contributions) are both computed from the sample, while the weights entering them are known constants from the design. When sampling probabilities are *estimated* from cohort counts rather than fixed a priori, plugging estimated probabilities into the weights is standard and, somewhat counterintuitively, tends to make the plain sandwich slightly conservative, so it remains a safe default.

For the likelihood-corrected estimators, the situation is friendlier, and this is one of the arguments for the approach. Because the ascertainment-corrected likelihood is a *genuine* likelihood for the data actually observed, i.e., it is the correct conditional probability of the sample, standard maximum-likelihood machinery applies to it directly: the inverse of the observed information from the *corrected* likelihood is a valid variance estimator, and Wald, score, and likelihood-ratio inference all behave as usual, provided the outcome model is correctly specified (Schildcrout and Heagerty, 2011). The design quantities $\pi(g)$ enter the information matrix as known constants; nothing about them needs to be estimated. What the analyst must *not* do is take the information matrix from an *uncorrected* fit, i.e., the corrected point estimate is only half of the procedure; the corrected curvature is the other half. In practice, software maximizing the ACML objective returns the corrected information automatically, and a sandwich variance built on the corrected score adds protection against model misspecification at little cost. For the failure-time version, Ding et al. (2014) supply the analogous plug-in variance estimator for their profile-likelihood procedure, with the same division of labor: design quantities known, nuisance functions estimated from the SRS, curvature taken from the corrected objective.

Which strategy should an applied analyst prefer? Schildcrout and Heagerty (2011) compare them head-to-head: when cluster sizes are equal and weights are moderate, the two perform similarly; when cluster sizes vary or some

strata are sampled at very low rates, ACML is substantially more efficient. Weighting, for its part, asks less of the model (the weighted score is unbiased even if secondary features of the outcome model are misspecified), is trivial to implement with standard weighted-regression software, and generalizes across model types without new derivations. A reasonable house rule for this course: *use the likelihood correction when a well-specified likelihood is available and the design's weights would be extreme; use weighting when robustness and implementation simplicity dominate, and always with the sandwich variance*. What is never reasonable is the naive analysis in either direction: unweighted-and-uncorrected point estimates, or corrected point estimates with uncorrected standard errors.

Note: Design-Based Knowledge Is Real Knowledge

A recurring student instinct is to treat the sampling probabilities as one more thing to be estimated, or worse, to ignore. Resist it in both directions. When *you* designed the sampling, the probabilities $\pi(g)$, or the sampling fractions q, π_1, π_0, are *known constants*, and the analysis should use them as such: they appear in the weights, in the ascertainment correction, and in the variance formulas, with no uncertainty attached. Ignoring known sampling probabilities throws away the very information that makes the biased sample interpretable; it is the analytic equivalent of shredding the study protocol. Conversely, when analyzing someone else's outcome-dependent sample, e.g., an existing case-control study, or a substudy whose sampling rule was informal, the first analytic task is archaeological: reconstruct how people got into the data. If the sampling rule cannot be recovered, at least approximately, neither weighting nor likelihood correction can be applied, and the estimand quietly changes from a cohort parameter to "an association in whatever sample this is."

5.6 Planning an ODS Study

One more practical idea from Schildcrout and Heagerty (2011) deserves attention, because it inverts the usual order of study planning. In a *prospective* study we compute power from guessed parameters. In a *retrospective* ODS substudy, we are in a far stronger position: the outcomes and the cheap covariates are already in hand for the whole cohort, and only the expensive exposure is missing. Schildcrout and Heagerty (2011) exploit this with a **comparative feasibility analysis**: posit a model for the exposure's distribution (and candidate effect sizes), use an EM algorithm to simulate the missing exposure *conditional*

on the observed outcomes and covariates, and then rehearse each candidate design, i.e., each choice of $[\pi(\text{never}), \pi(\text{sometimes}), \pi(\text{always})]$, on the simulated data, comparing the standard errors each would deliver *before spending a single assay dollar*. In the CAMPCS application, this rehearsal correctly anticipated the standard errors of the eventual analysis, and identified designs that concentrated sampling on the response-variable children as uniformly best. The general lesson generalizes well beyond this method: when the outcome data already exist, the design of the expensive phase should be *simulated, not guessed*.

6 The Three Designs Side by Side

Design	What is sampled?	Why use it?	Main inferential issue
Case-control	All (or most) cases; controls sampled conditional on binary case status (from non-cases, the baseline cohort, or risk sets)	Efficient study of rare outcomes; exposure measured on a small fraction of the source population	Absolute risks not identified; which cohort ratio (odds, risk, rate) the sample odds ratio equals is set by the control-sampling scheme
Case-cohort	All cases, plus a random subcohort drawn from the baseline cohort blind to outcome	Avoid measuring expensive covariates on the full cohort; one subcohort reusable across many outcomes	Weighted (pseudo-likelihood) estimation with time-varying weights; robust sandwich variance to capture subcohort sampling
General outcome-dependent sampling	A simple random sample, plus supplemental samples drawn with probabilities depending on the outcome value (timing, extremity, or response pattern)	Concentrate expensive measurements on the most informative subjects; handles continuous, failure-time, and longitudinal outcomes	Naive regression biased in all coefficients; correct by inverse probability weighting or ascertainment-corrected likelihood, with design-aware variances

The week in one table. The three designs form a progression: sampling on a binary outcome, then on failure status within an existing cohort, then on general features of the outcome. Down the rows, the designs gain flexibility and efficiency; the corresponding analyses take on more of the correction burden, from "nothing to correct if you only want the ratio," to weights, to full likelihood corrections.

7 Conclusion

This week extended Week 2's central discipline, i.e., know your estimand, and know what your data must contain to reach it, to studies in which the data's

very presence depends on the outcome. Three points to carry forward.

First, **outcome-dependent sampling changes the analysis, not the target.** The estimand remains a parameter of the source cohort: a risk ratio, rate ratio, odds ratio, or regression coefficient defined exactly as if we had enumerated everyone. The sample is a device for reaching that target at a fraction of the cost, and every valid analysis of such a sample either exploits an invariance (the odds ratio's cancellation), reweights the sample back into a pseudo-cohort, or corrects the likelihood for the ascertainment.

Second, **the design decisions are load-bearing and must be recorded.** Which control-sampling scheme was used; the subcohort fraction; the stratum-specific sampling probabilities: these numbers are not administrative trivia but constants that appear inside the estimators and their variances. A biased sample plus a known sampling rule is a valid study; a biased sample with a forgotten sampling rule is just a biased sample.

Third, **standard errors are where naive analyses die quietly.** Point estimates from weighted analyses can look entirely sensible while their model-based standard errors overstate precision, sometimes dramatically, because weighted records are not independent people and sampled subcohorts vary. The sandwich/robust machinery that fixes this is not specific to these designs, and Week 7 develops it in general; this week supplied the canonical examples of why it exists.

Lab 2 puts the case-control schemes into practice: starting from the cohort you built in Lab 1, you will draw cumulative, case-base, and incidence-density control samples and verify — against the fully enumerated cohort — that the same exposure odds ratio recovers the cohort odds ratio, the risk ratio, and the rate ratio in turn. In Week 4 we begin assembling the course's central machine, regression itself, and in Week 11 we will return to precisely the designs of this week — including the case-cohort weighting and variance machinery above — equipped with the regression toolkit needed to analyze them in full.

References

William E Barlow. Robust variance estimation for the case-cohort design. *Biometrics*, 50:1064–1072, 1994.

William E Barlow, Laura Ichikawa, Dan Rosner, and Shizue Izumi. Analysis of case-cohort designs. *J Clin Epidemiol*, 52:1165–1172, 1999.

Jianwen Cai, Bahjat Qaqish, and Haibo Zhou. Marginal analysis for cluster-based case-control studies. *Sankhyā, Series B*, 63:326–337, 2001.

Jerome Cornfield. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst*, 11: 1269–1275, 1951.

Jieli Ding, Haibo Zhou, Yanyan Liu, Jianwen Cai, and Matthew P Longnecker. Estimating effect of environmental contaminants on women's subfecundity for the MoBa study data with an outcome-dependent sampling scheme. *Biostatistics*, 15:636–650, 2014.

Jieli Ding, Tsui-Shan Lu, Jianwen Cai, and Haibo Zhou. Recent progresses in outcome-dependent sampling with failure time data. *Lifetime Data Anal*, 23 (1):57–82, 2017. DOI: 10.1007/s10985-015-9355-7.

Sander Greenland. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol*, 160:301–305, 2004.

M. J. Knol, S. Le Cessie, A. Algra, J. P. Vandenbroucke, and R. H. Groenwold. Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. *Can Med Assoc J*, 184, 2012. DOI: 10.1503/cmaj.101715. URL https://doi.org/10.1503/cmaj.101715.

Bryan Langholz and Duncan C Thomas. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *Am J Epidemiol*, 131:169–176, 1990.

Timothy L Lash, Tyler J VanderWeele, Sebastien Haneuse, and Kenneth J Rothman. *Modern Epidemiology*. Wolters Kluwer, Philadelphia, PA, 4th edition, 2021.

Olli Miettinen. Estimability and estimation in case-referent studies. *Am J Epidemiol*, 103(2):226–235, 1976.

Katie M O'Brien, Kaitlyn G Lawrence, and Alexander P Keil. The case for case-cohort: An applied epidemiologist's guide to reframing case-cohort studies to improve usability and flexibility. *Epidemiology*, 33(3):354–361, 2022. DOI: 10.1097/EDE.0000000000001469.

Ross L Prentice. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73:1–11, 1986.

Ross L Prentice and Ronald Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.

K. J. Rothman, S. Greenland, and T.L. Lash. *Modern Epidemiology*. Wolters Kluwer, Philadelphia, PA, 3rd edition, 2008.

Jonathan S Schildcrout and Patrick J Heagerty. Outcome dependent sampling from existing cohorts with longitudinal binary response data: study planning and analysis. *Biometrics*, 67(4):1583–1593, 2011. DOI: 10.1111/j.1541-0420.2011.01582.x.

Steven G Self and Ross L Prentice. Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann Stat*, 16:64–81, 1988.

Terry M Therneau and Hongzhe Li. Computing the Cox model for case cohort designs. *Lifetime Data Anal*, 5:99–112, 1999.

S Wacholder, J K McLaughlin, D T Silverman, and J S Mandel. Selection of controls in case-control studies. i. principles. *Am J Epidemiol*, 135(9):1019–1028, 1992a.

S Wacholder, D T Silverman, J K McLaughlin, and J S Mandel. Selection of controls in case-control studies. ii. types of controls. *Am J Epidemiol*, 135 (9):1029–1041, 1992b.

Sholom Wacholder. Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology*, 2:155–158, 1991.

Jichang Yu, Yanyan Liu, Dale P Sandler, and Haibo Zhou. Statistical inference for the additive hazards model under outcome-dependent sampling. *Can J Stat*, 43(3):436–453, 2015.

Haibo Zhou, Mark A Weaver, Jing Qin, Matthew P Longnecker, and Mei-Cheng Wang. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, 58:413–421, 2002.

Haibo Zhou, Jianwei Chen, Tiina H Rissanen, Susan A Korrick, Howard Hu, Jukka T Salonen, and Matthew P Longnecker. An efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology*, 18(4):461–468, 2007. DOI: 10.1097/EDE.0b013e31806462d3.