1) Researchers are investigating the number of falls experienced by elderly patients in different care settings. They visit three different nursing homes and, over a three-years period, record the total number of falls for each resident currently living in those homes. They also record the total number of resident-years of observation for each nursing home (summing the time each resident contributed to the study within the follow up). Moreover, some characteristics of the subjects are also measured, such as age, health status, presence of dementia and level of physical activity. Which kind of statistical tool could you use to describe the association between the outcome and the subjects' features ?

To analyze the association between the number of falls and the subjects' characteristics in this scenario, the most appropriate statistical tool is **Poisson regression** or **Negative Binomial regression**, specifically incorporating an **offset term** to account for differing observation times. The outcome variable is the total number of falls, which is a count variable (non-negative integers: 0, 1, 2, …). Because residents were followed for different lengths of time, we can not just compare the raw number of falls; we should model the rate of falls. To do this, we use a Poisson regression model where the expected count μ for an individual is modeled using the log link function:

$$\log(\mu) = \log(t) + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

$\log(t)$ (The Offset): t represents the "resident-years of observation" (the exposure time). By forcing the coefficient of $\log(t)$ to be exactly 1, the model effectively analyzes the incidence rate of falls (falls per resident-year) rather than just the absolute count.

$X_1, X_2, \ldots, X_k$: These are the predictors (age, health status, dementia, and physical activity level).

A strict assumption of Poisson regression is that the variance of the outcome equals its mean ($E(Y) = Var(Y)$). In epidemiological data this assumption is frequently violated; in this case for example some residents might be highly prone to falling (causing "clustering" of events).

The variance is often much larger than the mean (overdispersion). If overdispersion is present, Negative Binomial regression should be used instead. It includes an extra parameter to model this excess variance, ensuring p-values and confidence intervals remain accurate.

2) What is the difference in the following table between the "Crude" and the "Adjusted" Odds Ratio ? How do you interpret the change in the estimate with respect to the exposure variable (use of selective serotonin reuptake inhibitors) on the outcome (hepatocellular carcinoma) ?

TABLE 2 | Odds ratio and 95% confidence interval of hepatocellular carcinoma associated with selective serotonin reuptake inhibitors use by logistical regression model.

	Crude		Adjusted[†]	
Variable	**OR**	**(95% CI)**	**OR**	**(95% CI)**
Ever use of selective serotonin reuptake inhibitors (never use as a reference)	1.27	(1.13, 1.43)	0.92	(0.79, 1.08)

[†]*Initially, all variables were included in the univariable logistic regression model. Only those significantly associated with hepatocellular carcinoma in the univariable logistic regression model could be further included in the multivariable logistic regression model. Therefore, only metformin use, statin use, alcohol-related disease, chronic kidney disease, chronic liver disease, chronic obstructive pulmonary disease, diabetes mellitus, hyperlipidemia, and hypertension could be further included for adjustment.*

The shift from a significant increased risk (OR} = 1.27) to a non-significant, slightly lower risk (OR= 0.92) indicates a classic case of confounding. In the crude analysis, it falsely appeared that SSRI use was associated with a 27% increase in the odds of developing hepatocellular carcinoma. However, this association was not causal; it was driven by underlying differences in the patient populations. The adjustment variables listed in the legend are known risk factors or indicators for liver health. Therefore, is probable that patients who are prescribed SSRIs likely had a higher prevalence of these baseline comorbidities (e.g., chronic illness and severe lifestyle factors are heavily linked to higher rates of depression and anxiety, resulting in higher SSRI prescriptions). Therefore, the crude OR was simply picking up the risk of HCC caused by alcohol use and metabolic issues, not the SSRIs themselves.

3) Which is the statistical test used to compare two or more survival curves? Which are the assumptions behind this test ?

The standard and most widely used statistical test to compare two or more survival curves is the Log-Rank Test.

The log-rank test is a non-parametric test used to compare the survival distributions of two or more independent groups. It tests the null hypothesis that there is no difference between the survival curves of the populations being compared (i.e., the probability of an event occurring is identical at any given time point).

It compares the observed number of events in each group against the expected number of events if the null hypothesis were true, calculated at every distinct time point where an event occurs.

To safely use and interpret the log-rank test, the following assumptions must be met:

- Proportional Hazards Assumption : The log-rank test assumes that the relative risk of the event between the groups remains constant over time.

What it means: If Group A has twice the risk of dying compared to Group B at month 1, it must also have roughly twice the risk at month 12 and month 24.

Violations: If the hazard functions cross each other the proportional hazards assumption is violated. In this scenario, the log-rank test loses statistical power and can yield misleading results.

- Non-Informative (Random) Censoring: The reasons why individuals are censored (lost to follow-up, dropped out, or reached the end of the study without the event) must be unrelated to their underlying probability of experiencing the event.

What it means: A patient who drops out at year 2 must have the same future risk of experiencing the event as a patient who remains in the study at year 2.

Violations: If patients drop out of a clinical trial because they are feeling too sick from side effects (meaning they are closer to death/event), this is informative censoring and will bias the results.

- Independence of Survival Times: The survival times of individual participants must be independent of one another.

What it means: The probability of one participant experiencing the event cannot influence or be related to the probability of another participant experiencing it.

4) Which is the dependent variable in the Cox regression model ?

In a Cox proportional hazards regression model, the dependent variable is the hazard rate, or more specifically, the hazard function over time, denoted as h(t).Because the Cox model is used for survival analysis (time-to-event data), the dependent variable simultaneously incorporates two distinct pieces of information: Time (t): The continuous time elapsed until an individual experiences the event of interest or is censored. Status (D): A binary indicator variable showing whether the event actually occurred (1) or if the subject was censored (0).

Mathematically, the dependent variable is modeled on a logarithmic scale as the hazard at time t, given a set of predictors $(X_1, X_2, \dots, X_k)$:

$$\log\big(h(t)\big) = \log\big(h_0(t)\big) + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

5) What does "censoring" means in survival analysis ? Provide some examples.

In survival analysis, censoring occurs when we have some information about a subject's time-to-event, but we do not know the exact time the event of interest happened. Instead of discarding these incomplete pieces of data, survival analysis techniques (like Kaplan-Meier curves and Cox regression) incorporate them because they still provide vital information: we know the subject survived at least up until the moment they were censored.

The Right-Censoring (The Most Common):

This happens when the study ends, or a participant leaves, before the event occurs. We only know that the true survival time is greater than the recorded observation time. There are three common reasons for right-censoring:

- Administrative End of Study: The study concludes as planned, but the patient is still alive or has not experienced the event.

Example: A 5-year study tracks cardiovascular mortality. Patient A survives the entire 5 years without a heart attack. Their data is right-censored at exactly 5 years.

- Loss to Follow-up: The patient moves away, drops out, or researchers lose contact with them.

Example: In a study tracking cancer recurrence, a patient moves to another country after 18 months of symptom-free follow-up. We know they were disease-free for 18 months, but we don't know what happened after. Their data is right-censored at month 18.

- Competing Risks / Death from an Unrelated Cause: The patient experiences a different event that makes observing the primary event as the first one impossible.

Example: A study is tracking the time until a specific hip implant fails. Two years into the study, the patient unfortunately dies in a car accident. Since the hip implant never failed while they were alive, their data is right-censored at year 2.

6) A research team is studying the effectiveness of a new cancer treatment, "Thera-C," on patients with a specific type of aggressive lymphoma. The study design is as follows:

Study Population: Patients diagnosed with this specific lymphoma between January 1, 2020, and December 31, 2022, at a single hospital. They are observed from date of diagnosis to date of death or administrative censoring, fixed at 31 December, 2024. Treatment Groups:

- Group A (Thera-C): Patients who received at least one dose of Thera-C after their initial diagnosis during the follow up.
- Group B (Standard Care): Patients who did not receive Thera-C and were treated with the standard of care along all the follow up.
- Outcome: Overall survival, measured in months from the date of initial diagnosis.

Analysis: Kaplan-Meier survival curve and log-rank test are used to compare the survival distributions of Group A vs Group B. The results show that Group A has a significantly better survival than Group B. The researchers conclude that Thera-C is highly effective in prolonging the lives of patients with this lymphoma. Critique this study design and its conclusion. What potential bias is present, and how does it affect the results?

This study design suffers from a major, classic methodological flaw in epidemiology and survival analysis known as immortal time bias (specifically, a type of selection bias or time-varying confounding). Because of this bias, the researchers' conclusion that Thera-C is highly effective is highly suspect and likely completely invalid. Immortal time refers to a span of follow-up during which the study outcome (death) cannot physically occur. In this study design, patients are assigned to Group A (Thera-C) if they received *at least one dose of the drug at any point during the follow-up*. Look closely at how the timeline unfolds for a typical patient in Group A:

They are diagnosed -> They live for 6 months on standard care -> At month 6, they receive their first dose of Thera-C.

By definition, for those first 6 months, that patient had to survive in order to live long enough to receive the treatment. If they had died at month 3, they would have been classified into Group B (Standard Care). Group A is mathematically guaranteed to live longer: Group A artificially pools together patients who survived long enough to get the treatment. Their survival clock starts at diagnosis, but they are credited with "survival time" before they even touched the drug. Group B is penalized: Group B inadvertently absorbs all the patients who had aggressive disease and died quickly before they ever had a chance to receive Thera-C. This bias heavily favors the treatment group. It creates a false appearance of a treatment benefit (or severely exaggerates a true modest benefit). Even if Thera-C were a completely useless sugar pill, Group A would still show a significantly better survival curve and a significant log-rank test simply because the dead-on-arrival or early-dying patients were systematically forced into Group B. Beyond immortal time bias, there are other severe issues with this study:

Confounding by Indication: This is an observational study, not a randomized controlled trial (RCT). Why did some patients get Thera-C while others got Standard Care? It is highly probable that younger, healthier patients with better organ function were selected to receive the new

treatment, while frail, older, or sicker patients were kept on standard care. The better survival in Group A might just reflect that they were healthier to begin with.

7) Which kind of study design can produce data that are suitable for survival analysis methods ?

Data suitable for survival analysis requires tracking the time until an event occurs for individual subjects. Therefore, the study design must be longitudinal (following subjects over a period of time) rather than cross-sectional.

The primary study designs that produce data suitable for survival analysis include: Population based (Cohort) Studies: this is the most common observational design for survival analysis. A group of individuals (the cohort) who are initially free of the outcome of interest are identified, their baseline exposures are measured, and they are followed forward in time to see who develops the outcome. The data collection could be prospective: participants are recruited in the present and followed into the future and time-to-event is recorded precisely as it happens, or the data collection could be retrospective (historical): the cohort is identified using historical records (e.g., medical charts or occupational registries from 2015). Researchers trace the data forward from that historical baseline to a later point (e.g., 2025) to extract the time-to-event or censoring status.

Randomized Controlled Trials (RCTs): In an RCT, participants are randomly assigned to an intervention group (e.g., a new drug) or a control group (e.g., a placebo) and followed prospectively over time. Survival analysis is heavily used here to evaluate outcomes like overall survival, progression-free survival, or time to disease recurrence.

8) What is the difference between "Marginal" and "Conditional" causal effect of a treatment ? Which are the statistical tools useful to estimate a marginal effect ?

In causal inference and epidemiology, the distinction between marginal and conditional causal effects depends on the population to which the causal effect applies.

Marginal Causal Effect (Population-Averaged): The marginal effect is the average causal effect of a treatment across the entire population. It answers the broad public health question: "What would happen to the overall rate of disease if we treated every single person in the population, compared to if we treated nobody?"

Target: The whole population.

Intuition: It averages out individual-level differences (confounders) across the entire group.

Analogy: A randomized controlled trial (RCT) naturally estimates a marginal effect because randomization balances all covariates across the treatment groups, allowing to directly compare the overall group averages.

Conditional Causal Effect (Stratum-Specific / Subject-Specific): The conditional effect is the causal effect of a treatment within a specific subgroup of people who share the exact same characteristics (covariates), or for a specific individual. It answers the clinical question: "For a 65-year-old male smoker with diabetes, what is the causal effect of this treatment?"

Target: Specific strata of covariates (e.g., age, sex, comorbidities).

Intuition: It measures the treatment effect while holding individual characteristics strictly constant.

Optional: The "Collapsibility" Catch

In linear regression models, marginal and conditional effects are mathematically identical (in absence of interactions). However, for non-linear models commonly used in epidemiology (like Logistic Regression for Odds Ratios or Cox Proportional Hazards for Hazard Ratios), the measures are non-collapsible. This means that even in the absence of confounding, the conditional Odds Ratio/Hazard Ratio will naturally differ from (and usually be further from the null than) the marginal Odds Ratio/Hazard Ratio.

When working with observational data, we can not simply compare the treated and untreated groups to find the marginal effect because of confounding. To estimate a marginal causal effect while adjusting for confounding, we must use tools that simulate an RCT by balancing covariates across the entire population. The primary statistical tools include:

Propensity Score-Based Methods: the propensity score is the probability of a subject receiving the treatment given their baseline covariates. It can be used in two main ways to find marginal effects:

Inverse Probability of Treatment Weighting (IPTW): This is the most popular tool for marginal effects. It assigns a weight to each individual proportional to the inverse of their probability of receiving the treatment they actually got. This creates a "pseudo-population" where the distribution of confounders is completely identical between the treated and untreated groups. A simple, unadjusted model (like a weighted t-test or weighted logistic regression) run on this pseudo-population yields the marginal causal effect.

Propensity Score Matching (Full Population): Matching treated individuals to untreated individuals with similar propensity scores. If done across the entire sample, it balances covariates globally to estimate a marginal effect (specifically, the Average Treatment Effect on the Treated, or ATT, which is a marginal effect restricted to the treated subpopulation).

Optional: G-Methods (G-Computation / Standardization): This is a sort of simulation-based approach. First, we fit a multivariable regression model predicting the outcome based on treatment and all confounders. Then, we use that model to predict two counterfactual scenarios for every single individual in the dataset: what their outcome would be if they were treated, and what it would be if they were untreated. Finally, we average those individual predictions across the entire dataset and subtract them to find the true marginal causal effect.