**Key learning objectives one should have established prior to EPI 560**

1. Define and identify different types of cohort studies
2. Define and identify potential sources of bias that arise in cohort studies
3. Define and identify approaches to control for confounding in the design and analysis of cohort studies

Definitions

* **Cohort**: subjects are selected for inclusion based on exposure status or common characteristic and followed forward in time to determine who gets the outcome of interest. Subjects in cohort studies should be at risk of the outcome at the beginning of follow-up. Cohorts are valuable for monitoring incidence of disease, identifying risk factors for disease development, monitoring survival from disease, and identifying factors for disease progression.
* **Exposed (or index) cohort**: refers to individuals who have been exposed to the putative causal event
* **Unexposed (or referent) cohort**: refers to individuals who have not been exposed
* **Open**: participants can enter and exit the study at any point during follow up
* **Fixed**: a set number of individuals is enrolled into the study and exposure status is ascertained at the beginning of follow up. Individuals cannot move between exposure groups during follow-up.
* **Closed**: a fixed cohort that experiences no loss to follow up and death of participants is attributable only to the outcome of interest.
* **Prospective**: the research study begins prior to follow-up period, meaning data is collected in real-time and temporality between exposure and outcome can be reasonably ascertained.
* **Retrospective**: the data has already been collected and the research study is undertaken after the follow-up period has ended. data has already been collected and relies on previously enacted data collection instruments and assumptions must be made about temporality between exposure and outcome.

Sources of Bias

Information Bias

* Measurement Error
* Misclassification
* Non-response

Selection bias

* Volunteer bias
* Loss to follow-up

Immortal Time Bias: Period of follow-up between ‘exposure categorization’ and ‘experienced exposure’ resulting in time when the outcome cannot occur as a result of exposure: ‘immortal time’ (ex. Where date of prescription defines exposure but the drug was not taken until later). Person-time is inflated, rate in exposed is underestimated, and rate ratios are biased towards the null.

**Controlling for potential confounders in the design or analytical stages of a cohort study**

Matching:

The process of making a study group and a comparison group comparable with respect to extraneous factors. Unexposed individuals are matched with exposed individuals on one or more confounders.

Matching factors can be potential confounders and/or known risk factors of the outcome. Matching on a given variable removes the ability to analyze that factor later on. Matching increases statistical efficiency (precision) and controls for confounding, but if you match on something that is not a confounder, this can introduce bias or lead to loss of precision.

Mantel-Haenszel Adjustments:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Risk Ratio*** | |  |  |  |  | | --- | --- | --- | --- | | Stratum i | E+ | E- | Total | | D+ | ai | bi | m1i | | D- | ci | di | m0i | | Total | n1i | n0i | ni |   *For G strata* |  |
| ***Odds Ratio*** | |  |  |  |  | | --- | --- | --- | --- | | Stratum i | E+ | E- | Total | | D+ | ai | bi | m1i | | D- | ci | di | m0i | | Total | n1i | n0i | ni |   *For G strata* |  |
| ***Rate Ratio*** | |  |  |  |  | | --- | --- | --- | --- | | Stratum i | E+ | E- | Total | | D+ | ai | bi | m1i | | Person-Time | PT1i | PT0i | Ti |   *For G strata* |  |

Standardization

Provides a means of summarizing epidemiologic measures across strata by taking a weighted average of the stratum-specific measures. Weights are based on the experience of the reference group and reflects the measure of frequency we would have seen in the exposed had they been unexposed.

\*This method does not work well when there are n=0 strata.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Risk/prevalence data*** | |  |  |  |  | | --- | --- | --- | --- | | Stratum i | E+ | E- | Total | | D+ | ai | bi | m1i | | D- | ci | di | m0i | | Total | n1i | n0i | ni |   *For G strata* |  |
|  |  |  |
| ***Rate data*** | |  |  |  |  | | --- | --- | --- | --- | | Stratum i | E+ | E- | Total | | D+ | ai | bi | m1i | | Person-Time | PT1i | PT0i | Ti |   *For G strata* |  |

Inverse Probability of Treatment Weighting

Create a pseudo population through weighting, where all study participants are represented twice. Once for what would happen if everyone were exposed and once for what would happen if everyone were unexposed.

**Example**

1. **Determine if there is confounding by calculating the crude and stratum-specific risk ratios.**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| |  |  |  | | --- | --- | --- | | **Crude** | **E+** | **E-** | | **D+** | 220 | 59 | | **D-** | 1780 | 4050 | | **Total** | 2000 | 4100 | | |  |  |  | | --- | --- | --- | | **C=1** | **E+** | **E-** | | **D+** | 200 | 10 | | **D-** | 800 | 90 | | **Total** | 1000 | 4200 | | |  |  |  | | --- | --- | --- | | **C=0** | **E+** | **E-** | | **D+** | 20 | 40 | | **D-** | 980 | 3960 | | **Total** | 1000 | 4000 | |
|  |  |  |

*\*Crude measure of association differs from stratum-specific; therefore, C is confounding the association between E and D.*

1. **Calculate the probabilities of being exposed and unexposed in each stratum of C**
2. **Calculate the reciprocal of each probability. These values are the weights.**

11/10, 11, 5, and 5/4

1. **Multiply the observed cell counts by the corresponding weights.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **C=1** | **E+** | **E-** | | **D+** | 200\*(11/10) = 220 | 10\*11 = 110 | | **D-** | 800\*(11/10) = 880 | 90\*11 = 990 | | **Total** | 1100 | 1100 |   RR = 2.0 | |  |  |  | | --- | --- | --- | | **C=0** | **E+** | **E-** | | **D+** | 20\*5 = 100 | 40\*(5/4 = 50 | | **D-** | 980\*5 = 4900 | 3960\*(5/4) = 4950 | | **Total** | 5000 | 5000 |   RR = 2.0 |

1. **Create new crude table using the above pseudo population & calculate crude RR.**

|  |  |  |
| --- | --- | --- |
| **Crude** | **E+** | **E-** |
| **D+** | 320 | 160 |
| **D-** | 5780 | 5940 |
| **Total** | 6100 | 6100 |

RRIPTW = 2.0

No confounding

**Key points covered in prior iterations of 560**

**Michael Kramer EPI 560**

*Designing cohort studies for causal inference*

* **Exchangeability**: the experience of the exposed cohort (or exposed person-time) is exchangeable for the experience of the unexposed cohort (or person-time)
* **Positivity:** non-zero probability of ‘assignment to’ any level of exposure within every strata of covariates needed to assure exchangeability
* **Consistency:** well-defined and consistently experienced version of the exposure

**Key ingredients at baseline:**

1. Well-defined target population
2. All individuals are at risk of incidence of disease
3. Conditioning on exposure-assignment mechanism for exchangeability
   1. Prior knowledge inform threats to exchangeability through confounders
   2. To strengthen exchangeability between exposed & unexposed cohorts by:
      1. Restrict to single strata of confounders
      2. Matching
      3. Measuring all confounders
4. Every type of person (strata of confounders) has some chance of being either exposed or unexposed [positivity]
5. Exposure is well defined [consistency]
   1. Define whether the exposure varies across these dimensions: Timing, duration, intensity/frequency
   2. Categorizing people’s experience as ‘exposed’ or ‘unexposed’: time-varying exposure levels, induction period. Is there a causal threshold? If exposure must accumulate to specific level to be causal, then perhaps ‘exposure’ is defined at the time when that threshold has been met?

**Classifying person-time**

Induction period: distinguish timing of exposure from time of exposure-risk, exposure (acute or chronic) may not lead to immediate disease; thus, failure to account for induction period can lead to inappropriate assignment of ‘exposed time at risk’ and biased estimates of incidence/measures of association. To handle induction period: all induction-period time can be classified as ‘unexposed’ even among those who have exposure (exact induction period may not be known) OR omit all follow-up during induction (reduced precision)

A single person may contribute both exposed and unexposed person time:

* Time-varying exposure
* Movement through induction period
* Accrual of chronic exposure to threshold level

**Define & distinguish external and internal validity**

***Internal validity***: when the estimate from the sample is unbiased for the true sample or source average treatment effect (SATE). Internal validity is specific to the study & source population; whereas, external validity requires specification of the target population.

**External validity:** when the true sample ATE is unbiased for the target population ATE. A target population is necessary for meaningful inference and decision making beyond the sample at hand. The target population should be well defined (clearly specified) and informed by the research question.

**Defining Validity**

**ATE**: Average treatment effect (ATE) in target population:

**SATE** Average treatment effect in source population:

Average association in study population:

**Internal bias (interval validity)** = [SATE in source population] – [Association in study population]

**External bias (external validity)** = [ATE in target population] – [SATE in source population]

Target Validity = Internal Validity + External Validity = [ATE in target population] – [Association in study population]

**Identifiability assumptions for target ATE given observed sample association**

1. Study participants are **exchangeable** with target population, possibly conditional on covariates, W:
2. Every target population strata of W has non-zero probability of selection into sample: for all w
3. Equal distribution of version of treatment in the study sample and target population
4. No interference in the study population and the target population

S: Indicator of membership in sample (e.g. selection = 1)

Y(a): Potential outcome

w: Covariates related to selection and Y(a)

Sources of external bias:

**1. Lack of exchangeability**

Estimands are the target ATE and SATE. Each is the average of individual counterfactual causal effects (e.g. doomed, immune, harmed, helped); each causal type might be explained by presence/absence of other component causes (e.g. effect modifiers, interactions); if sample has different distribution of causal types (or different distribution of effect modifiers) from target, the average treatment effect will be different

**2. Different treatment version (ex. Protocol vs. cohort)**

**3. Non-positivity (eligibility/exclusion criteria may exclude some strata)**

Estimators for target population ATE

If W=w are strata of covariates that violate exchangeability between sample and target population (e.g. modifiers of effect that are differently distributed)

G-formula (non-parametric standardization)

Target ATE = [sample association for each strata of W = w] x [Distribution of strata W in the target population]

**Dana Flanders EPI 739**

***Additional biases in cohort studies***

Index Event Bias

* People are selected for the study because they’ve already had some event (unknown factor U), resulting in collider bias. There is no way to adjust for this type of bias.
* In the example below, researchers would like to measure the effect of obesity on mortality for individuals with end stage renal disease. Because some people could have died prior to having an ESRD diagnosis, selection into the study has to be contingent on surviving until diagnosis. These factors that facilitated entry into the study can bias the effect of obesity on mortality towards the null, and all confounders related to ESRD and mortality should be accounted for.
  + Explanation: collider bias from conditioning on an index event (ESRD) / survival
    - U = dietary choices, exercise habits, etc.

ESRD

U

Death

Obesity

Flanders, W. D., Eldridge, R. C., & McClellan, W. (2014). A nearly unavoidable mechanism for collider bias with index-event studies. *Epidemiology*, *25*(5), 762-764.

\*Other examples in Modern Epidemiology 4 – pages 154 – 156.

Ghost Time Bias

Refers to the failure to stop collecting person-time when it should be stopped. For example, women appear to have a much lower mortality rate over time compared to men. However, researchers were using the national death index and some deaths were not reported. Women (vs. men) were more likely to have unreported deaths because they are more likely to change their last names.

* e.g., comparing mortality rates for women vs men:

IDR

Women

vs. Men

1.0 –

Time

|

30+ yrs old

What we should have observed

Allowing post-eligibility exposure events to define a cohort

* When eligibility starts after the exposure starts
  + Eligibility = service in the army 1964 – 1972, excluding those with officer paygrade
  + Pay grade was determined by last discharge
  + If someone served in Vietnam until the last day of 1972 (extreme example), but had an increase in pay grade that day, they would not be included in the study
  + Researchers should have used pay grade at first enrollment to determine eligibility
* How to avoid this: defined cohort and once someone is included (eligible) for the study, they cannot later become ineligible (e.g., ineligible for receiving an increase in pay grade)