COHORT STUDIES

Learning objectives for this session:

1) Know when it is appropriate/feasible to use a cohort study design
2) Understand the issues to consider in selecting exposed and comparison groups
3) Differentiate between a retrospective and prospective cohort study design
4) Understand the strength and weaknesses of the prospective and retrospective cohort study designs compared to RCT

Outside Preparation:

A cohort study is a type of observational study. By observational, we mean that the investigators simply observe what happens to people over time.

The researcher does not determine exposure status.

This is unlike the randomized control trial design where exposure status is determined by the investigator(s). Instead, the exposure status is a characteristic of (or determined by) the subject (study participant).

Examples of exposures that are often examined in cohort studies are: cigarette smoking, BRCA genes, lead paint in house, and cholesterol levels. It is not possible to study these exposures using a randomized controlled trial design. Why? Because it would not be ethical to randomize half your study participants to these exposures. The ethical issues are different in observational studies than in RCTs because the investigator is not exposing the participant to a potentially risky substance

Concept of a ‘cohort’: A group of individuals that are all similar in some trait and move forward together as a unit.

In the context of epidemiologic studies, a cohort is a group of individuals sharing a common characteristic and observed over time.
Examples of different types of cohorts:

- **Birth cohort**: A group of people born during a particular period or year.
- **Inception cohort**: all individuals assembled at a given point based on some factor, e.g. where they live or work
- **Exposure cohort**: individuals assembled as a group based on some common exposure, e.g. radiation exposure during desert testing, asbestos exposure in the shipyards, etc.

There are different ways of selecting study populations for a cohort study:

**a) Cohort study beginning with exposed and nonexposed groups:**

Start with: Exposed and Not exposed

Then, follow up for: Disease and No disease

In this case, subjects are selected based on their exposure status and disease outcomes are compared to a group of similar subjects who are unexposed.

An example of this is as follows: Researchers identify a group of subjects who smoke and a comparable group of subjects who do not smoke. In this example, smoking is the exposure. The researchers then follow both groups for 20 years and record the number of head and neck cancers that develop in each group (cohort). They then do a statistical analysis to determine if there is an association between smoking and head and neck cancer.

**b) Cohort study beginning with a defined population:**

Start with: Targeted population

Then, follow up for: Exposed and Not exposed

Disease and No disease

Not randomly assigned
In this case, we select a defined population before any of its members become exposed or before their exposures are identified. Then, using biological tests or questionnaires, we can determine who is ‘exposed’ and who is ‘unexposed’ and follow-up the two groups for disease outcomes.

An example of this: Researchers recruit a group of adult subjects living in Framingham, MA. They then determine who has elevated blood cholesterol levels and who doesn’t. The researchers then follow these subjects for 20 years and determine who and how many develop cardiovascular disease. They then do a statistical analysis to determine if development of cardiovascular disease is more common in those with elevated cholesterol levels compared to those with normal cholesterol levels.

There are two types of cohort study designs:

1) **Prospective cohort study** (also known as concurrent, longitudinal or follow-up cohort study). This is where the Investigator identifies the original population at the beginning of the study and follows them up **in real time** until the study ends at a pre-specified time (20 years in the following figure):

![Prospective Cohort Study Diagram]

2) **Retrospective cohort study** (also known as historical or nonconcurrent cohort study): Investigator identifies a group of subjects with historical records of both exposure and outcome available in the past.
Note that it is also possible to conduct a mixed-design of prospective and retrospective, where exposure is ascertained from historical records in the past, and follow-up and measurement of outcome continue in real time into the future.

**ASSESSMENT OF EXPOSURE**

Goal: To have an accurate measure of the true exposure

This is not as easy as in a randomized trial design because the participant determines the exposure, not the investigator (e.g. cigarette smoking, workplace exposures – does he/she wear a mask at work? Illicit drug use - do the participants tell the truth?)

The exposure definition should be clear and precise

All study subjects should fulfill the exposure/non-exposure definitions. The questions to ask yourself are:

- Is information on exposure status available or attainable?
- If so, how do I get an accurate measurement of it in all study participants?

For example, if the exposure of interest is smoking:

- Does my questionnaire contain questions on smoking?
- Am I interested only in whether or not a person currently smokes, or do I also want to know how much and how often my study participants smoke?
- Do I need to know if a participant has smoked in the past?
- What is the best way to get this information in this study population?

Another example - Workplace exposures to chemicals:

- Which chemical(s) am I interested in?
- How do I measure exposure to this chemical – biological measurement, questionnaire?
- How do I assess previous history of exposure or length of time of exposure to the chemical?

Accurate assessments of exposure are generally easier in a prospective study than in a retrospective study because you can design the tools to measure the exposure and make the measurements yourself.

The exposure should be assessed blind to outcome (preferably before you know what the outcome will be). This is not an issue in prospective cohort studies because the outcome has not yet occurred at the time exposure is assessed. But this is a serious issue in retrospective cohort studies because you have to be vigilant to keep the people doing the assessments blind to the outcome (which has already occurred).

Sources of exposure information:

Objective Data: Lab tests (e.g., blood cholesterol levels)
Medical records (often incomplete)
Environmental tests – water, soil

Subjective Data: Questionnaires (e.g. surveys on smoking, alcohol, drug use, depression, etc.) Keep in mind that self-report of exposures are often subject to inaccuracies.

ASSESSMENT OF OUTCOME

The Outcome should also be clearly and precisely defined.

All study subjects with the outcome should actually fulfill the outcome criteria (sounds obvious - but not always followed)

Some potential sources of outcome information are:

State Cancer Registries
State Death Registries
Death Certificates (notoriously inaccurate / incomplete)
National Death Index
Medical Records – if attending doctor
Priest, family → Confirm with doctor and medical records (if possible)
Study measures (laboratory tests, questionnaire responses, physical measurements)

Similar to an RCT, the outcome should be assessed blind to the exposure status. This is true for both prospective and retrospective cohort studies.
COMPARABILITY OF STUDY GROUPS or SELECTION OF A COMPARISON GROUP

If you start with a defined study population and determine who in that study population is exposed, then your comparison group is the remaining subjects who are unexposed. However, if you start with an exposed group, then you need to make sure you select an appropriate comparison group.

The major principle for selecting a comparison group is that the groups being compared should be as similar as possible with respect to all other factors that may be related to the disease except for the exposure of interest. Therefore if there is really no association between exposure and disease, then the disease rates in the two groups will be essentially the same.

It is also important to ensure that the information collected from the exposed and unexposed groups are comparable.

Since cohort studies are not randomized, they are more vulnerable to bias and confounding (We will see more on this in future lectures).

Remember in Lecture 2 (RCT) you learned that:

*Overall goal of randomized treatment allocation is to have study arms that are identical with respect to all determinants of the outcome under study except the treatment being studied. Thus, any difference in outcome must be due to treatment.*

The goal is similar for a cohort study:

You want the study groups (exposed vs. non-exposed) to be similar for all other (non-study) determinants of the outcome except the exposure being studied.

An important difference between the cohort study vs. the randomized clinical trial is that the exposed and unexposed groups may not be comparable in terms of other important determinants of the outcome.

Remember that the process of randomization tends to balance the arms of a clinical trial for determinants of the outcome (when the numbers of subjects randomized is large).

But in a cohort study we do not have randomization to help us make the groups comparable with respect to other determinants of the outcome - large numbers of subjects does not balance arms in an observational study.

Since many exposures are found in the same people (smokers tend to drink more alcohol than non-smokers), the correlation of determinants of the outcome can lead to a problem called “confounding”. e.g., you think higher rates of breast cancer in people who drink alcohol are due to the alcohol, but you have to factor out the effect of smoking, which is more common among those who drink.
Briefly, confounding is a mixing of effects, as in the example above where the effect that you think is due to alcohol is really due to smoking. You will learn more about confounding in a future lecture.

The pros and cons for a **prospective** cohort study are as follows:

**Pros:**
- Good for rare exposures
- Can assess multiple outcomes
- Can generate incidence data
- Generally no ethical issues

**Cons:**
- Bad for rare outcomes
- Subjects can change exposure status (smokers quit and non smokers start smoking)
- Takes time
- Relatively expensive
- Comparability of study groups
- Loss to follow-up
- Are all subjects really free of the outcome at the start of the study?

The pros and cons of a **retrospective** cohort study are as follows:

**Pros:**
- Good for rare exposures
- Relatively short time to complete the study
- Relatively inexpensive
- Can assess multiple outcomes
- Can generate incidence data
- Generally no ethical issues

**Cons:**
- Bad for rare outcomes
- Imperfect information in medical records
- Loss to follow up
- Comparability of study arms
- Subjects can change exposure status
• Are all subjects really free of the outcome at the start of the study?

ESTIMATING RISK FROM COHORT STUDIES

In the RCT lecture, you learned how to calculate a relative risk. Again, the relative risk (RR) is defined as the probability (risk) of developing a disease in the exposed group compared to the probability of developing the disease in the unexposed group. In a cohort study, the RR is calculated the same way as in the RCT:

\[
\text{Relative risk} = \frac{\text{Incidence in exposed}}{\text{Incidence in unexposed}} = \frac{a}{a + b} / \frac{c}{c + d}
\]

However, an alternative approach is to use the concept of odds. The odds of an event is defined as the ratio of the number of ways the event can occur to the number of ways the event cannot occur.

\[
\text{Odds} = \frac{P}{1 - P}, \text{where } P = \text{probability of an event}
\]

Consider the Cohort Study design:

<table>
<thead>
<tr>
<th>First Select…</th>
<th>Then Follow to See Whether…</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disease develops</td>
</tr>
<tr>
<td>Exposed</td>
<td>a</td>
</tr>
<tr>
<td>Not exposed</td>
<td>c</td>
</tr>
</tbody>
</table>

What is the Probability (or risk) that the disease will develop in an exposed person?

Probability (risk, incidence) = \( \frac{a}{a + b} \)

What are the odds that the disease will develop in an exposed person?

\[
\text{Odds} = \frac{a}{b}
\]

Similarly, among unexposed persons, the probability that the disease will develop is \( \frac{c}{c + d} \) and the odds of the disease developing in unexposed persons are \( \frac{c}{d} \).
In a cohort study, to answer the question of whether there is an association between the exposure and the disease, we can either use the relative risk (defined above) or we can use the **odds ratio** (also called the **relative odds**).

In a cohort study, the ratio of the odds of development of disease in exposed persons to the odds of development of disease in unexposed persons is:

\[
\frac{a/b}{c/d} = \frac{a \times d}{b \times c}
\]

Odds Ratio measures the strength of the association, or the size of the effect of the exposure on the disease (not its statistical significance)

- OR >1 suggests the risk factor promotes the outcome
- OR <1 suggests the risk factor prevents the outcome
- OR =1 suggests no effect of the risk factor on the outcome
- OR = 1 is the **null value**
- Note: OR has no units

**RISK DIFFERENCE (ATTRIBUTABLE RISK)**

The **Attributable Risk** is another measure of association between exposure and outcome. Specifically, the Attributable Risk compares the incidence of disease (outcome) in the exposed and unexposed groups by subtracting the incidence in the unexposed from the incidence in the exposed to give the "excess" incidence due to exposure.

For Example:

Incidence of lung cancer in smokers = 140 /100,000 persons at risk per year

Incidence of lung cancer in non-smokers = 10 /100,000 persons at risk per year

Attributable Risk (also known as Risk Difference or Excess Risk)

\[
= 140 /100,000 - 10 /100,000 \text{ persons at risk per year}
\]

\[
= 130/100,000 \text{ persons at risk per year}
\]

Interpretation: 130 cases of lung cancer per 100,000 smokers per year are attributable to smoking or...130 cases of lung cancer could be prevented by preventing smoking.
Relative Risk and Attributable Risk give two different types of information:

Incidence Rates per 100,000 Population per Year

<table>
<thead>
<tr>
<th>Disease</th>
<th>Smokers</th>
<th>Non-Smokers</th>
<th>RR</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Cancer</td>
<td>140</td>
<td>10</td>
<td>14</td>
<td>130</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>669</td>
<td>413</td>
<td>1.6</td>
<td>256</td>
</tr>
</tbody>
</table>

Interpretation: Smoking has a stronger association with lung cancer (RR=14) than it does with cardiovascular disease (RR=1.6). Nevertheless, preventing smoking will save more lives due to reduced incidence of cardiovascular disease (256/100,000 smokers per year) than it will due to reductions in the incidence of lung cancer. This is because cardiovascular disease is much more common in the population than lung cancer.