Types of Designs #1:
DESIGNING CLINICAL RESEARCH
Biomedical studies are often conducted to “demonstrate” or confirm or establish a relationship between an exposure or explanatory factor and an outcome or response variable. The demonstration is accomplished by comparing the outcomes or responses from different levels of the explanatory factor or exposure. Different ways to show case the relationship form different “designs”.
The simplest form of designed experiments is the “completely randomized design” where treatments are randomly assigned to the experimental units – regardless of their characteristics. This design is most useful when the experimental units are relatively homogeneous with respect to known confounders.
A confounder is a factor which may be related to the treatment or the outcome even if the factor itself may not be under investigation. A study may involve one or several confounders. In a clinical trial, the primary outcome could be SBP reduction and the baseline SBP is a potential confounder. Patients’ age may be another one. In theory, values of confounders may have been balanced out between study groups because patients were randomized. But it is not guaranteed; especially if the sample size is not very large.
If confounder or confounders are known, heterogeneous experimental units are divided into homogeneous “block”; and randomizations of treatments are carried out within each block. The result would be a “randomized complete block design”.
A Simple Example:

An experiment on the effect of Vitamin C on the prevention of colds could be simply conducted as follows. A number of $n$ children (the sample size) are randomized; half were each given a 1,000-mg tablet of Vitamin C daily during the test period and form the “experimental group”. The remaining half, who made up the “control group” received “placebo” – an identical tablet containing no Vitamin C – also on a daily basis. At the end, the “Number of colds per child” could be chosen as the outcome/response variable, and the means of the two groups are compared.

Some other factors might affect the numbers of colds contracted by a child: age, gender, etc... Let say we focus on gender.
THE CHOICES

- We could perform complete randomization - disregard the gender of the child, and put Gender into the analysis as a covariate; or
- We could randomize boys and girls separately; at the end the proportions of boys in the two groups are similar and there would be no need for adjustment.
- The first approach is a complete randomized design; the second is a randomized complete block design.
- Similarly, we could block using “age groups”.
The term “treatment” may also mean different things; a treatment could be a factor or it could be multifactor. For example, let consider two different aspects of a drug regimen: Dose (Low, High) and Administration mode (say, one tablet a day or two tablets every other day). We could combine these two aspects to form 4 combinations; then treating them as 4 treatments and apply a complete randomize design. We call it a (balanced) Factorial Design; the analysis is similar to that of a randomized complete block design.
THE ROLE OF STUDY DESIGN

In a “standard” experimental design, a linear model for a continuous response/outcome is:

\[
Y = \begin{bmatrix} \text{Overall} \\ \text{Constant} \end{bmatrix} + \begin{bmatrix} \text{Treatment} \\ \text{Effect} \end{bmatrix} + \begin{bmatrix} \text{Experimental} \\ \text{Error} \end{bmatrix}
\]

The last component, ‘experimental error”, includes not only error specific to the experimental process but also includes “subject effect” (age, gender, etc…). Sometimes these subject effects are large making it difficult to assess “treatment effect”. 
Blocking (to turn a completely randomized design into a randomized complete block design) would help. But it would only help to “reduce” subject effects, not to “eliminate” them: subjects in the same block are only similar, not identical – unless we have “blocks of size one”. And that the basic idea of “Cross-over Designs”, a very popular form in biomedical research.
In the most simple cross-over design, subjects are randomly divided into two groups (often of equal sign); subjects in both groups/series take both treatments (experimental treatment and placebo/control) but in different “orders”.

Group 1: Period #1 (Treatment) – washout – Period #2 (Placebo)
Group 2: Period #1 (Placebo) – washout – Period #2 (Treatment)

Of course, “order effects” and “carry-over effects” are possible. And the cross-over designs are not always suitable. They are commonly used when treatment effects are not permanent; for example some treatments of rheumatism.
DESIGNING CLINICAL RESEARCH:
NUTS & BOLTS OF THE BASICS
How data will be collected? (It’s the complex issue of when to do what):

- Cross-section Design,
- Case-Control Design (retrospective),
- Cohort Design (prospective); Clinical Trials (one-arm, two-arm; open-label or randomized; double or triple blind)
CROSS-SECTIONAL DESIGNS
In a cross sectional study, investigators draw a sample, randomly, from the population, then make all measurements for all variables on a single occasion- or within a very short period of time – without a follow up. They study distributions the variables within that sample; sometimes designating predictor and outcome variables based on “biological plausibility”, then correlating one to the other.
<table>
<thead>
<tr>
<th></th>
<th>Factor Present</th>
<th>Factor Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Disease</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Take One Sample
Example:

In the National Health and Nutrition Examination Survey (NHANES), a sample designed to represent the U.S. population is interviewed and examined. These surveys have been carried out periodically and all data are available for public use. They make up a major source of information about the health and the habits of the U.S. population; one could obtain estimates such as prevalence of smoking or a disease.
In addition to studying distributions and obtaining parameter estimates, cross-sectional studies can also be used for examining associations. For examples, a cross-sectional finding in NHANES III is an association between childhood obesity and hours watching television. The choice of which variables to label as predictors and which as outcomes depends on the cause-and-effect hypotheses of the investigator rather than on the study design.
Serial Surveys:

These form a special case. A series of cross-sectional studies of a single population observed at several points in time – the case of those NHANES – is sometimes used to draw (informal) inferences about changing patterns of population characteristics over time.
**Strengths:**

A major strength of cross-sectional studies is that there is no waiting time for the outcome to occur, and no loss to follow-up. This makes them fast and inexpensive. And the obvious strength of their sizes. A cross-sectional study, because of its low cost, could be included as the first step in a cohort study or an experiment.
Weaknesses:

The major weakness of cross-sectional studies is the difficulty of establishing causal relationships from “observational” data collected in a cross-sectional time frame.

Cross-sectional studies are also “impractical” for the study of rare diseases if the sample was collected from the general population; we might need 10,000 subjects or more to find just one case of a rare disease. What would happen to statistical power?
CASE-CONTROL STUDIES
<table>
<thead>
<tr>
<th></th>
<th>Factor Present</th>
<th>Factor Absent</th>
<th>Sample 1: Cases</th>
<th>Sample 2: Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Present</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Disease</td>
<td></td>
<td></td>
<td>Sample 1: Cases</td>
<td>Sample 2: Controls</td>
</tr>
</tbody>
</table>

These are “retrospective”; obtaining past data from cases and from controls (people without the disease). The research focus is the disease.
Case-control studies “began” as epidemiologic studies to try to identify risk factors for diseases. Therefore, the term “cases” refer to those with the disease under investigation. However, the term has become more generic; case-control design can also be used to look at other outcomes, such “disability” among those who already have a disease. In that case, “controls” are those with the disease but not disability. The “case” in a case-control study maybe a patient who has had a good but rare outcome; say, recovery from a usually fatal disease.
Some investigators and scientists refer to case-control studies as “confirmatory observational studies” and to cross-sectional studies as “exploratory observational studies”. Both are observational; without interventions. Case-control studies are “retrospective”; obtaining “past” data from cases and from controls. The research focus is the disease (‘status” under investigation).
 STEPS IN THE DESIGN

- Select the sample of cases
- Select the sample of controls
- Measure predictor variable & potential confounders and effect modifiers
Case-control studies cannot yield estimates of the incidence or prevalence of a disease because the proportion of the study subjects who have the disease is determined by how many cases and how many controls the investigator chooses to sample, rather than by their proportions in the population.
What case-control studies do provide is descriptive information on the characteristics of the cases and, more important, an estimate of the strength of the association between each predictor variable and the presence or absence of the disease. These estimates are in the form of the odds ratio which approximates the Relative Risk if the disease is relatively rare.
**Strength:** Efficiency for Rare Outcomes

A major strength of case-control studies is their rapid, high yield of information from relatively few subjects. For rare diseases or outcomes, cross-sectional design is impractical; it requires a study size that no investigator could afford. Cohort design, which is normally larger, is also impractical because it requires a follow-up time longer than most investigators could afford.
**Strength:** Generating Hypotheses

The retrospective approach of case-control studies, and their ability to investigate a large numbers of possible predictors make them useful for generating hypotheses about, say, the causes of new outbreak of disease so that a more thorough investigation or investigations could follow.
**Weaknesses:**

Case-control studies have great strengths but they also have major limitations; among them:

1. Accuracy of exposure histories.
2. Appropriateness of controls.
3. Unlike cohort designs, we can only study one disease/outcome at a time.
**Sampling Bias:**

The data collection in a case-control study begins with the cases. But how do we know if these cases are representative of all patients who developed the disease; those who are undiagnosed, or died, are not included. And some included might be misdiagnosed (easier problem here). The more difficult decisions faced by investigators of a case-control study, however, relates to the more open-ended task of selecting controls. The ones included might be inappropriate and confound study conclusion.
Hospital- or Clinic-based Controls

One strategy to compensate for the possible selection bias caused by obtaining cases from a hospital or clinic is to select controls from the same facility. However, the risk factor of interest might be related to causes for which those control seek care; if so, prevalence of the risk factor in the control group would be falsely high, biasing the study results toward “the null”. That’s why, some studies use two control groups.
Matching & Multiple-matching:

Matching, and multiple matching, is a relatively simple method of ensuring that cases and controls are comparable with respect to major factors that are related to the disease but are not interest to the investigators. Examples are gender and Age group. Matching does have its adverse consequences, especially when “modifiable factors”, such as income and cholesterol levels, are matched.
Population-based Cases:

Population-base case-control studies are now possible for many diseases, like cancers, because of a rapid increase in the creation of maintenance of “disease registries”. Cases obtained from disease registries are generally more representative of the population of patients. When information on the cases and controls can come from the same sources, the design has the potential for eliminating sampling bias. Later, we will cover such a form, the “nested case-control design”.

Two or more Control Groups:

Selection of controls can be very tricky, especially might not be representative of patients; for example, hospital-based cases. In those studies, it is advisable to use more than one control groups; for example a hospital-based and a population-based control groups. The former might be biased toward the Null. On the other hand, be prepare to deal with “multiple decision problem”.
**Measurement Bias:** The Need for Blinding

Besides selection bias, case-control studies might be biased due to measurement error caused by their retrospective approach: the “recall bias” especially when it occurs to a different extent in cases and in controls. A necessary solution is the need for blinding: both observers (interviewers, for example) and study subjects could be both blinded to the case-control status of each subject and to the risk factor being studied (not an easy task at all!).
Case-crossover Studies:

A variant of the case-control design, useful for the short-term effects of varying or intermittent exposures, is the "case-crossover design". As with the "regular" case-control studies, case-crossover studies are retrospective studies that begin with a group of cases. However, in regular case-control studies, the exposures of the cases are compared with exposures of a group of controls. Each case in case-crossover studies serves as his/her own control. Exposures of the cases at or right before outcome time are compared with exposures of those same cases at one/more other points in time.
Relative Risk & Odds Ratio
One of the most often used ratios in epidemiological studies is the *Relative Risk*, a concept for the comparison of two groups or populations with respect to a certain unwanted event (disease or death). The traditional method of expressing it in prospective studies is simply the ratio of the incidence rates:

\[
\text{Relative Risk} = \frac{\text{Disease Incidence in Group 1}}{\text{Disease Incidence in Group 2}}
\]
Usually, group 2 is under standard conditions - such as non-exposure to a certain risk factor - against which group 1 (exposed) is measured. A relative risk which is greater than 1.0 indicates harmful effects whereas a relative risk which is less than 1.0 indicates beneficial effects. For example, if group 1 consists of smokers and group 2 non-smokers, then we have a *relative risk due to smoking*. 

$$ \text{Relative Risk} = \frac{\text{Disease Incidence in Group 1}}{\text{Disease Incidence in Group 2}} $$
The relative risk, also called risk ratio, is an important index in epidemiological studies because in such studies it is often useful to measure the increased risk (if any) of incurring a particular disease if a certain factor is present. In cohort studies such an index is readily obtained by observing the experience of groups of subjects with and without the factor as shown above. In a case-control study the data do not present an immediate way to estimate this parameter; we do not know the risk of the control group.
Suppose that each subject in a large study, at a particular time, is classified as positive or negative according to some risk factor, and as having or not having a certain disease under investigation. For any such categorization the population may be enumerated in a 2x2 table, as follows:

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed (+)</td>
<td>Yes (+)</td>
<td>A+B</td>
</tr>
<tr>
<td>Un-exposed (-)</td>
<td>No (-)</td>
<td>C+D</td>
</tr>
<tr>
<td>Total</td>
<td>A+C</td>
<td>B+D</td>
</tr>
</tbody>
</table>

The entries A, B, C and D in the table are sizes of the four combinations of disease presence-and-absence and factor presence-and-absence and the number N at the lower right corner of the table is the total population size. The relative risk is

\[
RR = \frac{A}{A + B} \div \frac{C}{C + D} = \frac{A(C + D)}{C(A + B)}
\]
In many situations, the number of subjects classified as disease positive is very small as compared to the number classified as disease negative, that is,

\[ RR = \frac{A}{A + B} + \frac{C}{C + D} = \frac{A(C + D)}{C(A + B)} \]

\[ C + D \cong D \]
\[ A + B \cong B \]
\[ RR \cong \frac{AD}{BC} = \frac{A/B}{C/D} = \frac{A/C}{B/D} \]
The resulting ratio, \( \frac{AD}{BC} \), is an approximate relative risk, but it is often referred to as odds ratio because

- A/B and C/D are the odds in favor of having disease from groups with or without the factor;
- A/C and B/D are the odds in favor of having exposed to the factors from groups with or without the disease.
The two odds, $A/C$ and $B/D$, can be easily estimated in case-control studies, by using sample frequencies, $a/c$ and $b/d$. 