NOTE: When you are studying for the second exam, make sure you review the material from
the first exam. It will be covered heavily again!!

1. The main take-home messages in this lecture deal with understanding the conceptual
designs, statistical metrics, biases, and advantages/disadvantages of case-control versus
cohort studies. Review all of these topics from the first exam.

2. The Doll & Hill case-control study of the association of smoking with lung cancer only
studied male cases and controls. Why? To maximize internal validity – the likelihood that
the study subjects are as similar to each other as possible. What’s the tradeoff?
Reduced generalizability – If smoking has different biological effects in women, the study
will not reveal them and we can’t use the study results to draw conclusions about that half of
the population. These two concepts are generally at odds with each other in the design of
any epidemiological study.

3. The odds ratio (OR) of smoking being associated with lung cancer increased when the
case-control data were stratified by daily cigarette use. This phenomenon is referred to as a
dose-response relationship.

4. The case-control study seemed very conclusive, but there were other possible explanations
for the observed results, such as chance (statistics are usually used to minimize this effect),
selection bias (only hospitalized cases and controls were used), measurement bias
(exposure-suspicion, a major problem with case-control studies), confounding (see Topic 3),
and reverse causation (the idea that a developing lung cancer makes someone more
inclined to smoke).

5. To put another nail in the coffin (if you will) of smoking’s association with lung cancer, a
prospective cohort study was conducted. This type of study can reduce the concerns of
reverse causation and exposure-suspicion bias described above.

6. The prospective cohort study confirmed the dose-response relationship of smoking
cigarettes and increased risks of lung cancer.

7. Smoking was also found to be associated with cardiovascular disease (CVD), but with a
much lower relative risk. At this point, it is important to remember the relative incidences of
lung cancer and CVD. CVD is much more common in the general population, so if you
do the math correctly, you’ll find that the population attributable risk of CVD due to smoking
is much higher than the PAR of lung cancer. Remember to consider this interplay between
high RR & low incidence versus low RR & high incidence. If you do, you will understand
why PAR% is such a useful value to calculate!

Topic 2. Statistical reporting and interpretation.
1. This topic is basically a review of several concepts from M1 statistics. You are probably not
going to be asked to calculate any statistics, but you must be able to interpret p-values,
confidence intervals and study power. These concepts are frequently tested on Step I.

2. When evaluating a new diagnostic test for a disease, we first compare it to a “gold standard”
test that we consider to represent the “truth” and we can construct a 2x2 table to begin
estimating the new test’s quality. In this table, α error represents the false positive rate, and
β error represents the false negative rate. The inverses of these errors represent the
specificity and sensitivity of the test, respectively.
### Table: Test/study conclusion (-) vs. Test/study conclusion (+)

<table>
<thead>
<tr>
<th>“Truth”/Disease (-)</th>
<th>Test/study conclusion (-)</th>
<th>Test/study conclusion (+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The null hypothesis</td>
<td>Correct result!</td>
<td>False positive!</td>
</tr>
<tr>
<td></td>
<td>Probability = 1-$$\alpha$$</td>
<td>Probability = $$\alpha$$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>“Truth”/Disease (+)</th>
<th>False negative!</th>
<th>Correct result!</th>
</tr>
</thead>
<tbody>
<tr>
<td>The alternative hypothesis</td>
<td>Probability = $$\beta$$</td>
<td>Probability = 1-$$\beta$$</td>
</tr>
</tbody>
</table>

### 3. $$\alpha$$ error is defined as the probability of seeing a result when the null hypothesis is actually true. This relates directly to the statistical p-value of the study.

### 4. $$\beta$$ error is defined as the chance of finding a null result when there really is an association between variables. This error is not directly calculated, but we can estimate this with the power of the study or test. It is important to understand the definition of power: it is the likelihood that a study, as it was designed, can detect a specified association between an exposure and an outcome. Consider these ways of increasing the power of a study, which hopefully make sense:

- If we reduce the variability of study subjects, making the population more homogenous, the power of the study increases. Remember how Doll and Hill studied the effects of smoking only in men? By doing so, they increased the power of their study for finding a significant relationship between smoking and lung cancer.
- Improve the measurement methods.
- Increase the sample size. Increasing sample size is the best (but usually the most expensive) way to increase the power of a study.
- Matched pair analysis. This technique is pretty much the same as reducing variability.
- Relaxing our statistical standards. If we set $$\alpha$$=0.05, it will be easier to detect a statistically significant result than if our $$\alpha$$=0.01 (see below), however, this method is often frowned upon.

### 5. If we assume that the null hypothesis is true (e.g. drug X does not really cause result Y), the p-value is the chance that we would still observe a result that makes the null hypothesis look false (e.g. it looks like drug X caused result Y, but this was just a random fluke). Therefore, the lower the p-value, the less likely it is that our observed results were due solely to chance. A very low p-value makes us more confident in rejecting the null hypothesis. Make sure you understand this definition.

### 6. When we design a study, we decide in advance what $$\alpha$$ level is acceptable for us to reject the null hypothesis. Typical values often selected for $$\alpha$$ are 0.05 and 0.01. The lower the $$\alpha$$ value, the lower the p-value needs to be before we reject the null hypothesis. If we set $$\alpha$$ at 0.05 and then measure a p-value of 0.044, we reject the null hypothesis and claim that our observed results were not due to chance. What this really means is that, if we were to repeat the identical experiment with the same subjects 1000 times, 956 times we would observe a result favoring the association, and only 44 times we would observe a result refuting the association.

- Do not make the mistake of assuming that the p-value tells us how many people responded a certain way in the trial. Tests questions will often try to trick you with this idea.
7. **Confidence intervals (CI)** are often used to express how tightly clustered a group of data points are distributed about a mean value for the population.

- For example, if the mean diastolic blood pressure of a population is expressed as 75 mmHg with a 95% confidence interval of 65 mmHg to 85 mmHg, what does this mean?
- *It does not* mean that 95% of the population has a diastolic blood pressure between 65 mmHg and 85 mmHg. *Again, don’t be fooled.* It means that we are 95% sure that the true mean diastolic BP for the population is between 65 mmHg and 85 mmHg.
- What if we consider relative risks instead of a linear variable? For example, we want to see if hypertensives taking hydrochlorothiazide (HCTZ) have a decreased risk of having an MI. Our “null hypothesis” is that HCTZ does not have an effect (RR=1) and we set our $\alpha$ at 0.05. Our study finds a RR of MI of 0.8 among HCTZ users versus people who do not take HCTZ. Does this mean we can publish the study? Not yet. Let’s say our study had a fairly low power and our 95% CI is 0.68-1.4. First of all, what does this mean? It means that our p-value is >0.05, since the RR=1 is included in our confidence interval. Remember that the null hypothesis is RR=1, so we cannot exclude the null hypothesis with 95% certainty. Any time the CI includes a RR of 1, we cannot exclude the null hypothesis. Let’s repeat the study with 10x as many subjects. Now we find a RR of 0.8, with a 95% CI of 0.74-0.93. Let’s publish the study! Our statistics allow us to exclude the null hypothesis of RR=1, because our CI does not include this value. The p-value this time must be <0.05 and therefore the results are significant.

**Topic 3. Confounding and effect modification.**

1. **Effect modification (EM)**

- Effect modification implies that there exists a relationship between exposure and disease that is different among different subgroups of the study population. Most of the time effect modification is due to a biological difference inherent to the subgroups. *An example would be if males metabolize a cholesterol-lowering drug differently from females, resulting in different cholesterol lowering effects in the two genders.* — In this example the “effect modifier” of the relationship between the drug and the observed outcome (cholesterol lowering) is gender.
- In order to identify EM, first stratify your data by subgroups, then draw 2x2 tables and estimate the odds ratio (or RR) of each stratum. If they are substantially different from each other (a difference of >20% is typical), the stratified variable is an EM.
- Report the stratified ORs (or RRs) separately. It is inappropriate to use the crude OR.

2. **Confounding**

- Confounding refers to the false association between a disease and an exposure caused by a third variable (the confounding variable) that: (1) is associated with the disease in the absence of the exposure; (2) is associated with the exposure but not as a result of being exposed; and (3) is not necessarily a cause of the disease. If you don’t take the confounding variable into effect when you analyze your study’s data, your crude OR will be very misleading. *An example of confounding would be to report that more men have heart disease than women. While this may be true, in a given birth cohort, more men smoke than women. However, smoking is an independent predictor of heart disease. Smoking is therefore a confounder on the effect of gender on heart disease risk. Note that smoking is not a biological difference between the genders, i.e., it is not an effect modifier.*
- In order to identify confounding, first take your crude data, draw a 2x2 table and calculate a crude OR. Next, stratify your data by the presumed confounding variable, draw new 2x2 tables, and calculate stratum-specific ORs. If these ORs are similar to
each other but different from the crude OR, you can call the stratified variable a confounder.

- In order to report your data when confounding variables are present, you need to adjust for the confounder(s). Examples include direct and indirect age adjustments of populations (review those from Exam #1), the Mantel-Haenzsel OR adjustment (you DO NOT have to know how to do this) and multiple linear and logistic regression (review this- see below).

- Matched-pair analysis. Sometimes a matched-pair analysis is used to control for confounders by having a matched control for each case in a study. In simple terms, this means that two nearly identical people are chosen but one has the disease and the other doesn’t. Then you ask each member of the matched pair if they also had the exposure of interest. Does this sound like a type of study you already know? Each pair is really a case-control study with 2 people in it! Consider the example in Question #5 of the Exam II Practice Questions on the website, which is looking at the association of eating peanut butter and developing peptic ulcer disease (PUD) in 10 matched pairs:

<table>
<thead>
<tr>
<th>Pair</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUD (+):</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Controls:</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

When data are represented this way, consider the top row of the table to be your cases and the bottom row to be your controls. The data could also be in a vertical table too. Take a moment to orient yourself to the data if you don’t understand the table. Finally, you could also draw 10 2x2 tables to represent the same data.

How do we use this type of data? First, go through the table and identify the discordant pairs: those are the pairs with one + and one – in them. You can disregard the concordant pairs, where both the case and control are exposed or unexposed. For each discordant pair with the case exposed (+ in the top row), add 1 to the numerator of a fraction. For each discordant pair with the control exposed (+ in the bottom row), add 1 to the denominator. For the above data, your fraction would look like this:

\[
\frac{0 + 0 + 1 + 0 + 0 + 0 + 1 + 1 + 1 + 1}{0 + 0 + 0 + 1 + 1 + 0 + 0 + 0 + 0 + 0} = \frac{4}{2} = 2
\]

This method is really a modification of what is known as the MH adjustment formula, but I have simplified it for you here.

**Topic 4. Multivariable Modeling.**

1. Multiple linear regression

- For a continuous variable, such as blood pressure, we can derive a formula to predict “normal” blood pressures in different populations, taking into account predictors or confounders such as age, gender, race, etc.

- Let’s say that at age 0 (hypothetically), the average systolic BP should be 30 mmHg (variable \(\alpha\)). This would be estimated by graphing a large population’s blood pressure by age and finding the y-intercept on the graph.

- Let’s say that for every year that we age (variable \(X_1\)) above 0, systolic BP increases by a factor of 1.5 (variable \(\beta_1\)).

- Let’s say that men tend to have a systolic BP 10mmHg higher than women do. (variable \(X_2\) for gender with male=1 and female=0, \(\beta_2\) is 10).

- So our systolic BP estimation, in mmHg, is calculated by:
  \[\alpha + \beta_1X_1 + \beta_2X_2\]
  and a 20-year-old woman would be expected to be at:
30 + (1.5)(20) + (10)(0) = 60. Obviously these are not real numbers we’re using...
and a 60-year-old man would be expected to be at:
30 + (1.5)(60) + (10)(1) = 130.

- This is multiple linear regression, a method for formulaically representing and controlling
  for every known predicting and confounding variable in a measurement. The systolic BP
  formula should have dozens of modifiers in it.
- When reporting a result like this, $X_1$ is the “most important” variable, the one you want to
  study. For example, for systolic BP the most significant variable predicting outcome is
  age, and all of the other $\beta$ and $X$ values are adjusted confounders/predictors.

2. Multiple logistic regression
- The purpose is the same as multiple linear regression, except logistic regression is used
  to adjust the probability of a binary or dichotomous outcome by predictors and
  confounders. A binary outcome has two possible states, such as alive/dead, disease +/-, etc.
- The formula is similar to the one above:
  \[
  \ln(\text{odds of outcome}) = \alpha + \beta_1X_1 + \beta_2X_2...
  \]
  where the variables are the same as the ones above.
- How is this different? This time we are calculating an odds ratio to predict the binary
  outcome. The OR will then be >1, favoring the outcome, or <1, not favoring it.

3. Propensity scores
- When the outcome being studied is very rare, it is not appropriate to use logistic
  regression. For example, a rare side effect of a drug might occur in only 1 in 100,000
  people. That means it might be difficult to assemble enough patients using the drug who
  have this side effect to power an effective study.
- What if we do the opposite – look for factors that are associated with the choice to give
  patients one treatment versus another? A propensity score is used to predict how likely
  someone is to receive a certain treatment. It is not used to predict the outcome, just the
  treatment choice.
- Our study can then compare patients with similar propensity scores, and we have thus
  reduced the effect of many potentially confounding variables from our investigation.
- You will not need to calculate propensity scores, but you must understand them.

Topic 5. Clinical Trials.
1. The challenge in designing a “good” clinical trial lies in having everything about the
   participants standardized in advance, including who will be eligible, who won’t be eligible,
   how randomization and blinding will be accomplished, how they will be treated, how they will
   be tested/evaluated, and what outcomes will be assessed.

2. Remember the things that made a study have good power. They come back to us here: are
   there enough subjects in the study (taking drop-outs and loss-to-follow-up into account)?

3. Generalizability vs. Internal Validity: the more alike the study subjects are to each other,
   the more internally valid. For example, if you select only women for your study, then the
   results are more likely to be accurate for women, but necessarily generalizable to all people.
   Remember that these two terms are always at odds with each other. Internal validity tries to
   answer the question, “Was the observed outcome truly due to the intervention?” while
   generalizability tries to answer the question, “Will another population respond to the
   intervention in the same way?” Related terms are efficacy and effectiveness. From
   pharmacology remember that efficacy is a measure of biological effect; effectiveness refers
   to how a population under standard circumstances will respond. For example, if a drug is
very efficacious but has lots of side effects, it may not be very effective because patients
don’t comply with their medication. Cholestyramine is a good drug for lowering LDL
cholesterol - good efficacy. However, because cholestyramine gives you bloating, diarrhea,
and/or constipation, patients do not like to take it - not very effective at reducing the risk of
cardiovascular disease!

4. Selecting controls. **Historical Controls:** people who received a different intervention in the
past. Their environments, standard quality of care, and life expectancy were likely different
from today’s study subjects. **Contemporaneous Controls:** present day controls who are in
a different place. Bias and confounders can creep into these control populations, especially
if things like quality of care, diagnostic methods, population differences, referrals, or
methods of data collection vary between locations. **Concurrent Controls:** people who are
at the same place and same time as the active-intervention study participants. **Randomized**
controls are always better than **non-randomized** controls.

5. Types of randomization. Remember that every clinical trial published has a table of
“baseline characteristics” to show that randomization was successful.
- **Complete Randomization:** every participant enrolled in the study has an identically
  fixed, random chance of being either a control or intervention patient. In a small study, it
  is possible to have too many patients randomly wind up in one arm and too few in the
  other.
- **Allocation Concealment:** whoever is enrolling patients in a study does not know the
  formula for randomization, so they cannot predict whether the “next” person to enroll will
  get control or interventional treatment. Note that allocation concealment is not the same
  as blinding.
- **Restricted Randomization:** allocation of patients to control or intervention arms is not
  completely random. Can be multiple types:
  - **Stratification:** If you know that there is a confounding variable that can mess
    with the outcomes, then separate all participants by that variable (e.g., smoking)
    and randomize each group separately. That way, an equal number of [smokers]
    will wind up in the control and intervention arms of the trial.
  - **Blocking:** A “block size” of patients is selected, for example 6. We know that for
    every 6 patients who enroll, 3 will be controls and 3 will receive intervention. If
    there are 6 envelopes with assignments, there will be 3 for each arm. The 6
    patients will randomly be handed the 6 envelopes.
  - **Minimization** or dynamic balancing: A computer helps us to “randomize”
    participants in such a way that the control and intervention arms have equal
    numbers of patients. Let’s say we want to enroll 500 participants in a study with
    50% controls. The first patient to enroll will be randomly assigned to one arm, for
    example the control arm. The second patient to enroll will be “randomized” by the
    computer, but with a 250/249 (slightly more than 50/50) chance of being placed
    in the intervention arm. Let’s say that she is also placed into the control group.
    The third patient now gets a 250/248 chance of being in the intervention arm.
    The chances of “randomization” dynamically change as each patient enrolls to try
    to weight assignments toward the smaller arm.

6. The controls in clinical trials frequently perform better than random folks in the general
population. Why does this happen?
- **Volunteers** for clinical trials tend to care more about their well-being, take good care of
  themselves, and comply with medications.
- **Eligibility** criteria, if too stringent, select for people who are different from the general
  population. There should **always** be an explanation somewhere in the paper of why
certain groups are excluded from the study.
Placebo Effect: Even though someone received no intervention, it made them feel better.

Hawthorne Effect: People who know they are being observed perform better, i.e., they are more prone to “do their best,” especially in complying with treatment instructions.

Regression Toward the Mean: Everything in life ultimately slides toward mediocrity. Unusually good or bad performance on an initial examination will usually normalize toward the population mean upon repeated testing.

7. A new bias! Will Rogers Effect or Stage Migration Bias. If we design a new staging system for a disease, there will be apparent improvement at all stages. Remember Dr. Gayed’s explanation: If you take the dumbest people in Oklahoma, and move them to California, you will simultaneously increase the average IQ of each state.

8. Subgroup Analysis. SA is the term for looking at small mini-populations within a study population for additional outcomes. Generally, if this is part of the study design it can be done. If the authors of a crappy study perform subgroup analysis after the study is over, it is called “data dredging” or “data mining,” i.e., trying to find something meaningful without a preformed hypothesis. In other words, the authors of the failed study are trying to “put lipstick on a pig.”

9. Blinding. Patients should not know whether they are receiving control or interventional treatment (single-blind). The treating physician should not know which treatment each patient is receiving (double-blind). Whoever is analyzing the data should not know which treatment each participant received while analyzing their outcomes (triple-blind).

10. How do we handle dropouts or crossovers from one arm to the other?
   - Intention-to-Treat Analysis. Considered the best protocol, and certainly best for effectiveness assessment. Once a patient is randomized, he or she is always analyzed with that group. This applies even if a patient in the intervention arm stops taking the study drug or is accidentally treated as a control.
   - Per-Protocol Analysis. If patients accidentally or intentionally stop treatment or cross from one arm to the other, they are excluded from analysis. This maintains randomization, but reduces the study power if there are lots of dropouts.
   - As-Treated Analysis. If patients accidentally or intentionally stop treatment or cross from one arm to the other, they are analyzed with the group that they joined. As-treated analysis violates randomization and is generally considered “bad.” Of course, as-treated analysis is better for assessing the efficacy of a treatment but could ruin the effectiveness assessment.

11. Jadad and Delphi Scales - assessing study quality. Jadad assigns a score of 0-5 based on whether a study adequately describes randomization, blinding, and withdrawals/dropouts. Delphi is a 9-point scale that assesses a study’s randomization, allocation concealment, baseline characteristics, eligibility criteria, triple blinding, measurements of primary outcomes, and intention-to-treat analysis.

12. Non-Inferiority Trials. Once an effective treatment has been established, any “new” treatments that come along must be shown to be as good as the existing treatment or maybe almost as good, but with fewer side effects. As an example, think about something like statin drugs. The first statin was lovastatin. Large clinical trials were conducted as the drug first came to market which showed how this drug could lower LDL levels and thus reduce the risk of cardiovascular disease. In time, as “newer” statins were developed, it is conceivable that non-inferiority trials would have been conducted for each new statin to show that it had close to the same effects as lovastatin, and perhaps was even better.
12. **Decisions to stop a trial.** A clinical trial has several conflicting duties:
   ✷ The health and well-being of current trial participants.
   ✷ Future patients who could benefit from productive results.
   ✷ Getting enough evidence to change current practices.
   ✷ Getting enough evidence to withstand criticism.

So, if an intervention is clearly harming participants receiving the new intervention, it is appropriate to stop the trial early. If the new intervention is clearly benefiting participants, it is also appropriate to stop so that the controls can receive the intervention. The job of the Data Safety Monitoring Board is to assess when there is sufficient evidence to stop the trial for one reason or the other.