From Bookshelf to Bedside: Applying Basic Statistical Concepts to Patient Care

N. Jim Rhodes, Pharm.D., BCPS
Infectious Disease Pharmacotherapy Fellow
Midwestern University, CCP
Northwestern Memorial Hospital
nrhode@midwestern.edu

Disclosures

- I have no actual or potential disclosures related to the content of this presentation.
- We will discuss proprietary resources and off-label use of FDA approved products as examples including:
  - Dronedarone
  - Cefepime
  - Piperacillin-tazobactam

Objectives

- At the end of this presentation, pharmacist participants should be able to:
  - Select an appropriate statistical test of a given hypothesis based on the level of measure and the distribution of the data.
  - Calculate the number need to treat or harm for a given intervention with a known absolute benefit or risk.
  - Interpret positive and negative study findings in the light of a given study's limitations.

Topic Overview

- Survey data indicate that PharmD only and residency-trained pharmacists struggle with the interpretation of biostatistical concepts
  - Discordance exists between self-assessment and performance
  - Positive attitude toward the subject and prior confidence were indicators of performance
    - Prior confidence was the only independent predictor
  - Additional reinforcement needed to support practice

Bookstaver PB, et al. *Ann Pharmacother.* 2012;7-8:991-9

In this session

- We will review:
  - Levels of measurement, hypothesis testing
  - Sample size, power, and statistical significance
  - Interpretation of relative risk, odds ratios, and numbers needed to treat or harm
  - Basic concepts related to correlation and regression
- Ultimate aims:
  - Provide lasting knowledge through application
  - Provide practical tools you can use in patient care

Where do you go…

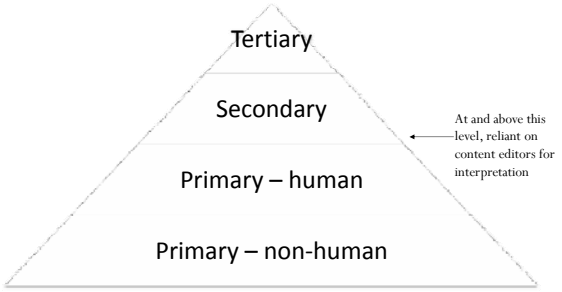
when you need data to answer a clinical question?

- Search primary literature expanding the search using MEDLINE/PubMed, Web of Science ™, EMBASE™
- Review tertiary resources to obtain any needed background, then proceed to secondary indexes to find informative / supportive primary literature
- "Google it" and use search results to find supportive primary or tertiary resources directly

Which of the following statements best describes you?

- I am **HIGHLY** confident when it comes to analyzing data and interpreting the literature.
- I am **VERY** confident when it comes to analyzing data and interpreting the literature.
- I am **SOMEWHAT** confident when it comes to analyzing data and interpreting the literature.
- I am **NOT VERY** confident when it comes to analyzing data and interpreting the literature.
- I am **NOT** confident when it comes to analyzing data and interpreting the literature.

Evidence-based practice resources



Adapted from Duke University Medical Center Library and the Health Sciences Library at the University of North Carolina at Chapel Hill. Accessed July, 2014. http://guides.mclibrary.duke.edu/ebmtutorial

Evidence-based practice resources

- Level of Evidence
  - 1: Systematic reviews of RCTs or individual RCTs
  - 2: Systematic reviews of cohort or individual cohort studies
  - 3: Systematic reviews of case-control or individual case-control studies
  - 4: Case-series and lower quality cohort or case-control data
  - 5: Expert opinion, mechanistic, based on bench research
- Grade of Evidence
  - A: Consistent level 1 data
  - B: Consistent level 2 or 3 data or extrapolations from level 1 data
  - C: Level 4 data or extrapolations from level 2 or 3 data
  - D: Level 5 data or inconsistent data at any level

Oxford Centre for Evidence-Based Medicine. Accessed July, 2014. http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/

Evidence-based practice resources

- Often, questions pharmacists receive center on optimal therapeutic options and adverse events
  - Efficacy
    - Generally, level 1 or 2 data needed to answer these question
      - Often, level 1 data does not exist, other levels and extrapolations needed
  - Toxicity
    - Generally, level 1 and 2 data insufficient or non-existent
      - Often, reliant on level 3 or 4 data and extrapolations from level 1 or 2

Adapted from Duke University Medical Center Library and the Health Sciences Library at the University of North Carolina at Chapel Hill. Accessed July, 2014. http://guides.mclibrary.duke.edu/ebmtutorial

The dilemma

- Many topics we need information on to treat patients (e.g. alternate or optimized dosing or adverse event data) lack randomized controlled trials
  - Critically analyzing available data essential
- Survey data suggest that even residency-trained pharmacists have low confidence with biostatistics
  - Predictors of higher confidence were
    - Positive attitude towards the subject
    - Higher confidence with the subject at baseline

Bookstaver PB, et al. *Ann Pharmacother.* 2012;7-8:991-9

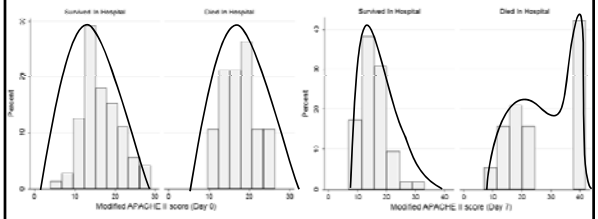
Levels of measurement

- Nominal (e.g. proportions achieving goal blood pressure)
  - Count and frequency
    - Categorical; often could retain ratio/interval form
    - Examples: gender, ethnicity, survival status, etc.
- Numeric (e.g. mean difference in blood pressure [mmHg])
  - Interval and ratio
    - Real (ratio) or arbitrary (interval) zero; equivalent difference in values
    - Examples: blood pressure, serum creatinine, glucose, time, etc.
- Ordinal (e.g. differences in stage of hypertension)
  - Categorical with magnitude
    - Differences between values not continuous or proportional
    - Examples: satisfaction scores, sedation scales, etc.

Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 1:86-9

Data distributions

- Normally distributed
- Non-normally distributed



Rhodes NJ, Nevrekar SN, McLaughlin MM, Wang SK, Crank CW, Qi C, Scheetz MH. K-1542. Evaluation Of Clinical Outcomes In Patients With Gram-negative Bloodstream Infections According To cefepime MIC [abstract]. ICAAC 2013. Denver, Colorado.

Hypothesis testing (part 1)

- What are we trying to accomplish?
  - Usually seek to compare 2 or more groups we observe
    - The fundamental question asked is:
      - Is the measure of central tendency (e.g. mean, median, etc.) or observed proportions different between my two populations?
    - Comparison takes the form of null and alternate hypotheses
      - The null hypothesis is the statement of no difference
        - » This is the actual hypothesis that we test
      - Even after running the test, you can still be wrong…
        - » Type I error: rejecting null hypothesis in error
        - » Type II error: rejecting the alternate hypothesis in error

Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 5:591-7

Hypothesis testing (part 1)

Statistical test result	Reality: Difference exists	Reality: No difference exists
Difference exists	True positive	Type I error
No difference exists	Type II error	True negative

Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 5:591-7

Hypothesis testing (part 1)

- Reliant on basic tenants of probability theory
  - What is a P-value?
    - The probability that the observed difference is due to chance
    - Multiple comparisons may yield erroneous results
      - Analogous to the probability of a Type I error occurring
  - Willingness to accept the alternate hypothesis depends on the amount of type I error you're willing to accept
  - In other words, are you willing to roll the dice?
    - If you roll enough, eventually will "hit"

Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 5:591-7

Hypothesis testing (part 1)

- Bivariate and univariate hypothesis tests
  - Assumptions:
    - Independent observations*, data must have the appropriate level of measure, data should conform to the distributional assumptions of test
  - Student's t-test and paired t-test*
    - Appropriate for interval or ratio data IF data normally distributed
      - Pre/post data appropriate for paired t-test (not independent)

Gaddis GM, Gaddis ML. *Ann Emerg Med.* 7:820-5

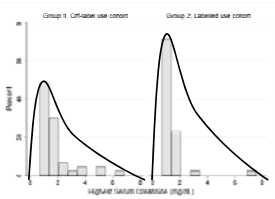
Hypothesis testing (part 1)

- Bivariate and univariate hypothesis tests
  - Wilcoxon rank-sum (Mann-Whitney U)
    - Appropriate for interval or ratio data IF data NOT normally distributed
  - Chi-square and Fisher's Exact
    - Appropriate for frequency (count) data for nominal measures
      - Use Fisher's Exact for 2x2 tables with cell counts < 5 and for larger tables with >20% of cells with counts < 5

Gaddis GM, Gaddis ML. *Ann Emerg Med.* 7:820-5

Active Learning Question 1

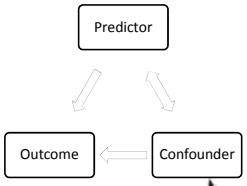
- While reviewing the findings of a retrospective study of dronedarone use, you note that the authors report the mean (SD) highest serum creatinine of patients receiving the drug at their center was 1.59 (1.17) mg/dL. Considering the distribution of the observations, which of the following options pairs the MOST appropriate statistical test and rationale to use to compare treatment groups?
  - Student's t-test as the data are continuous and normally distributed
  - Wilcoxon Rank-Sum test as the data are continuous and skewed
  - Pearson's Chi-square test as the data are frequencies with more than 5 observations in each stratum
  - Fisher's Exact test as the data are frequencies with fewer than 5 observations in each stratum



Rhodes NJ, McLaughlin MM, Fotis M, Scheetz MH. Unpublished data. 2014

Once we have stats…

- We have an idea of that the measures of central tendency in the data may differ, but the job isn't done yet…
  - Ultimately, our goal is to make inferences from the data
    - Even if there are imbalances in study groups for predictor or outcome variables, unadjusted differences may still be informative
    - But accounting for imbalances should improve understanding of relationships in the data



Effect size and measures of association

- What is an "effect size"?
  - We often discuss the results of prospective randomized trials in terms of risk reduction (relative and absolute)
    - Absolute risk reduction (ARR) or increase
      - The unadjusted difference in risk between the experimental group and the control group (may also compare other groups within a cohort)
      - ARR = $R_{exp} - R_{cont}$
    - Relative risk reduction (RRR) or increase
      - The unadjusted ratio of risk between the experimental group and the control group (may also compare other groups within a cohort)
      - RRR = R_{exp} / R_{cont}

Adapted from Duke University Medical Center Library and the Health Sciences Library at the University of North Carolina at Chapel Hill. Accessed July, 2014. http://guides.mclibrary.duke.edu/ebmtutorial

Effect size and measures of association

- An effect size can be relative or absolute
  - Comparisons should be made using similar measures
    - Meta-analyses and systematic reviews compare effect sizes
  - A simpler approach:
    - Number needed to treat (risk decrease)
      - NNT = 1 / ARR
        - » $NNT_{peri\text{-}operativeBB\text{-}died}$ = 1 / 0.0357 = 28 patients
    - Number needed to harm (risk increase)
      - NNH = 1 / ARI
        - » $NNT_{peri\text{-}operativeBB\text{-}stroke}$ = 1 / 0.00565 = 177 patients

Guay J, Ochroch EA. *J Cardiothorac Vasc Anesth.* 2013 Oct;5:834-44
Adapted from Duke University Medical Center Library and the Health Sciences Library at the University of North Carolina at Chapel Hill. Accessed July, 2014. http://guides.mclibrary.duke.edu/ebmtutorial

Effect size and measures of association

- Often, odds ratios will be reported rather than the relative or absolute risk change
  - Cohort studies and case-control studies
  - Odds ratios require addition interpretation

Parameter	Survived In Hospital n=72 (79.1%)	Died In Hospital n=19 (20.9%)	Univariate OR for mortality (95% CI)
Cefepime MIC ≥ 4 mg/L	25 (34.7%)	11 (57.9%)	2.58 (0.92-7.25)
Modified APACHE II ≥ 16.5	34 (47.2%)	13 (68.4%)	2.42 (0.83-7.10)

  - In this case:
    - Greater than 2 fold increase in mortality above cutoff values

Rhodes NJ, Nevrekar SN, McLaughlin MM, Wang SK, Crank CW, Qi C, Scheetz MH. K-1542. Evaluation Of Clinical Outcomes In Patients With Gram-negative Bloodstream Infections According To cefepime MIC [abstract]. ICAAC 2013. Denver, Colorado.

Effect size and measures of association

- Comparison of RR and OR
  - RR
    - Fraction with outcome of exposed / Fraction with outcome not exposed
    - Result of analysis no different when CI inclusive of "1" (ratio)
      - Example: relative risk = 0.92 (95% CI: 0.85 – 1.10)
    - If CI of risk difference includes "0" then no different
      - Example: risk difference = 0.25 (95% CI: -0.92 – 0.55)
  - OR
    - Fraction exposed with outcome / Fraction unexposed with outcome
    - Result of analysis no different when CI inclusive of "1" (ratio)
    - If CI of probability difference includes "0" then no different

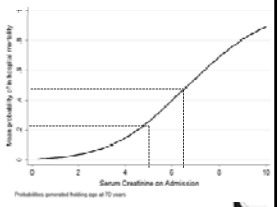
Adapted from Duke University Medical Center Library and the Health Sciences Library at the University of North Carolina at Chapel Hill. Accessed July, 2014. http://guides.mclibrary.duke.edu/ebmtutorial

Limitations on generalizability

- Unadjusted / unstandardized analyses are limited when confounders present
- Biases may exist throughout study design and implementation
  - Selection – sample reflects population of interest?
  - Measurement – differential recording of predictors?
  - Follow-up – differential detection of outcomes?
  - Many others possible depending on design!

Active Learning Question 2

- While reviewing the findings of the same study of dronedarone, you note that the authors report a mean probability of in-hospital death of at least 25% among patients with an admission serum creatinine of at least 5 mg/dL holding age constant at 70 years. Patients with an admission creatinine of at least 6.75 mg/dL have a predicted risk of in hospital mortality of at least 50% holding age constant at 70 years.
  - What is the risk ratio between the two "groups" as creatinine increases from 5 to 6.75 mg/dL?
  - What is the absolute risk increase associated with having a 1.75 mg/dL higher creatinine?
  - The NNH?



Rhodes NJ, McLaughlin MM, Fotis MA, Scheetz MH. Unpublished data. 2014

Translating to practice

- Why are many studies instructive or hypothesis generating but not necessarily practice changing?
  - Often the study design limits casual associations
    - RCTs limit bias and may show causality of treatment on outcome
    - Cohort and case control studies subject to confounding and biases
  - Often the population of interest is small or very specific
    - Difficult to study, rely of patterns in broader population
  - Even more frequently sample size limits power
    - The relative frequency of the outcome impacts power greatly

Translating to practice

- Recall type I error
  - Generally set an upper limit on false positives (1 in 20 times = 0.05)
    - Example: If you observe a difference in prolonged QTc between two groups of 27.9% v. 0% ($P < 0.001$), the difference may be due to chance 1 in 1000 times
- Recall type II error
  - Desired upper limit on false negatives (1 in 5 times = 0.20) **varies**
    - Example: If you observe a difference in mortality between two groups (N = 79) of 6.98% v. 0 % ($P = 0.25$); the similarity may be due to chance 1 in 2 times
- Power and sample size
  - An adequately powered study should be able to detect a (known or probable) effect size at least 80% of the time
    - Reliant on prior studies or theory to drive estimated sample size needed

Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 5:591-7

Translating to practice

- Considerations for power
  - A rare event is studied relative to the general population
    - Case-control design may be preferred to maximize power
  - Very small differences may be found to be statistically significantly different if large sample sizes obtained
    - Careful assessment of target sample sizes needed
  - Clinically meaningful differences should help drive assessments of power (1 – beta) and sample size needs
    - Again, literature or practice will be helpful to set targets

Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 5:591-7

Translating to practice

- Questions to ask
  - Once the level of evidence is classified, methods are understood, and adequate sample / power assessed:
    - Was the difference clinically meaningful?
    - Were subjects in study similar to patients I care for?
    - Was the follow-up adequate to detect outcomes?
    - Were exposures comparable to my practice / population?
    - What was the effect size? The NNH/NNT?
    - Would a larger sample possibly change results?

Adapted from Duke University Medical Center Library and the Health Sciences Library at the University of North Carolina at Chapel Hill. Accessed July, 2014. http://guides.mclibrary.duke.edu/ebmtutorial

Active Learning Question 3

- As you continue to review the dronedarone study, you note that the author's final model of in hospital mortality does not include treatment group. The unadjusted difference in mortality between off-label versus labelled use groups was 3 (7%) v. 0 (0%) with an OR (95% CI) of 6.83 (0.34-137; P = 0.2). Which of the following statements BEST describes why treatment group was not found to be an independent predictor of in hospital mortality?
  - A 7% absolute mortality difference is not clinically significant
  - A 6-fold increased odds of mortality is not clinically significant
  - The study is underpowered to determine the predictive power of treatment group on outcomes
  - The study is underpowered to determine the predictive power of treatment group on mortality

	In hospital mortality = Yes	In hospital mortality = No	Total
Off-label use = Yes	3 (7%)	40 (93%)	43
Off-label use = No	0 (0%)	39 (100%)	39
Total	3	79	82

Rhodes NJ, McLaughlin MM, Fotis MA, Scheetz MH. Unpublished data. 2014

Drawing inferences

- Descriptive statistics
  - Tell you about the sample
  - Measure central tendencies / distributions within data
- Inferential statistics
  - Provide a sense of whether the differences between study groups differ systematically beyond what we would expect merely due to chance alone
  - Differentiate "signal" from "noise"
  - Foundation of hypothesis testing

Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 5:591-7

Drawing inferences

- Often, we have a clinical data point and wish to predict future success or failure using that data
  - To what extent is the data point related to the outcome of interest?
  - Can we predict the outcome reliably from the data?
- These questions are inferential and translational
  - Fundamentally seek to quantify relationships in the data
  - Can utilize two tools to answer these questions
    - Correlation
    - Regression

Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 12:1462-8

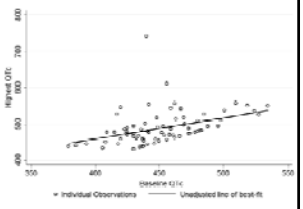
Correlation

- The first in a related set of concepts that can tell us:
  - Is there a relationship between X and Y
  - How strong is the relationship between X and Y?
- Measures of correlation:
  - Correlation coefficient "r" gives the relationship
    - + 1 is a "perfect" positive relationship
    - – 1 is a "perfect" inverse relationship
  - Coefficient of determination "r^2" gives the strength
    - An r^2 of 0.9 (i.e. $r^2 = 0.9$) implies that 90% of the variability in Y is related to the variability in X
  - Mere correlation does not imply causation though!

Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 12:1462-8

Correlation

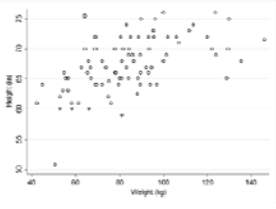
- Example:
  - Dronedarone off-label versus labelled use
  - Correlate QTc highest v. baseline
    - $r^2 = 0.1595$
    - That is, baseline QTc explains ~16% of the total variability in the highest QTc
    - Correlation coefficient ("r") = 0.52
    - That is, there is a positive association between baseline and highest observed QTc values



Rhodes NJ, McLaughlin MM, Fotis MA, Scheetz MH. Unpublished data. 2014

Regression

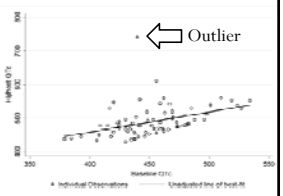
- Regression is usually an iterative process
  - Question may be straightforward:
    - "Can I predict height from body weight?"
  - Often, relationship not quite as simple
    - Can consider multiple variables simultaneously
    - Useful when dealing with confounders



Rhodes NJ, McLaughlin MM, Fotis MA, Scheetz MH. Unpublished data. 2014

Regression

- Regression can be perilous
  - Must consider if interactions or collinearity between predictor and outcome
    - Additional steps required
  - Likewise, how should "influential" outliers be handled in the data?
  - Authors should present visual checks of data to help the reader understand if outliers present and overall model fit



Rhodes NJ, McLaughlin MM, Fotis MA, Scheetz MH. Unpublished data. 2014

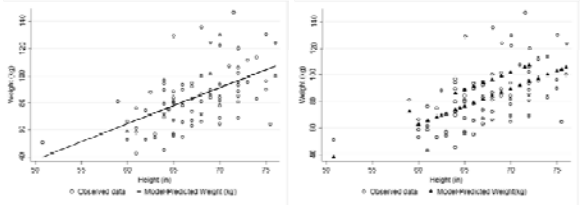
Regression

Univariate regression

- "One variable" at a time

Multivariate regression

- Multiple variables simultaneously



Triangles = Diabetes status, the upper triangles corresponding to a positive history of diabetes

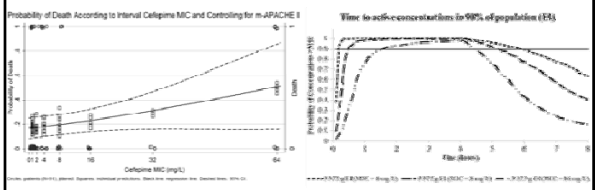
Rhodes NJ, McLaughlin MM, Fotis MA, Scheetz MH. Unpublished data. 2014

Dealing with confounders

- Confounders present a challenge in clinical studies
  - Unmeasured
    - Typically a limitation of retrospective studies
  - Measured
    - Several approaches to handle known confounders
      - Exclusion criteria
      - Criteria matching on confounder
      - Stratification by known confounder
      - Adjustment for confounder statistically

Dealing with confounders

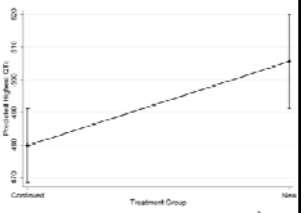
Adjustment **Stratification**



Rhodes NJ, Nevrekar SN, McLaughlin MM, Wang SK, Crank CW, Qi C, Scheetz MH. K-1542. Evaluation Of Clinical Outcomes In Patients With Gram-negative Bloodstream Infections According To cefepime MIC [abstract]. ICAAC 2013. Denver, Colorado.
Rhodes NJ, MacVane SH, Kutti JL, Scheetz MH. *Clin Infect Dis*. 2014. doi: 10.1093/cid/ciu402.

Active Learning Question 4

- Though the authors failed to detect a significant difference in mortality according to off-label and labelled use groups for dronedarone, they did observe greater increases in QTc among patients who were initiated on therapy in hospital adjusting for baseline QTc and multiple confounders. Which of the following statements BEST describes an inference that can be drawn?
  - Baseline QTc was negatively associated with the highest QTc observed
  - Initiation of dronedarone in hospital was positively associated with the highest QTc observed
  - Baseline QTc poorly predicts the highest QTc observed in the final model
  - Initiation of dronedarone in hospital poorly predicts the highest QTc observed in the final model



Rhodes NJ, McLaughlin MM, Fotis MA, Scheetz MH. Unpublished data. 2014

Summary

- Ideally, therapeutic decisions are based on high quality Level 1A evidence
  - In reality, many questions encountered in practice will not have systematic review or RCT level data to support
- Understanding appropriate biostatistical methods and correct interpretation of results essential
  - Limitations increase beyond Level 1evidence
  - Often, results can be instructive even if the strength of associations are less firm compared to Level 1 evidence

Summary

- When reviewing study data, always ask:
  - Were study groups comparable (fair comparison)?
    - If not, how were confounders handled?
  - Were exposures comparable to my practice / population?
    - If not, external validity may be in question.
  - Was the follow-up adequate to detect outcomes?
    - Even RCTs are subject to biases if methods not sound.

Summary

- What was the effect size?
  - Was the study likely powered to observe that effect?
- Was the difference clinically meaningful?
  - If many trials, look for consistency of effect.
- Were subjects in study similar to patients I care for?
  - If not, external validity may be in question.
- What was the strength of association between predictors and outcomes?
  - Low power or small samples may limit predictions

Additional resources

- Duke University Medical Center Library and the Health Sciences Library at the University of North Carolina at Chapel Hill. Available at: http://guides.mclibrary.duke.edu/ebmtutorial. Accessed July, 31, 2014.
- Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 1:86-9
- Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 3:309-15
- Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 5:591-7
- Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 7:820-5
- Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 9:1054-9
- Gaddis GM, Gaddis ML. *Ann Emerg Med.* 1990. 12:1462-8
- Rosner B. *Fundamentals of biostatistics.* 7th ed. Boston, MA: Brooks/Cole, Cengage Learning; 2010.