# Collaborating with a biostatistician?

## *Best Practices and Tips*

Jiangxia Wang

Wilmer Biostatistics Center

Johns Hopkins Biostatistics Center

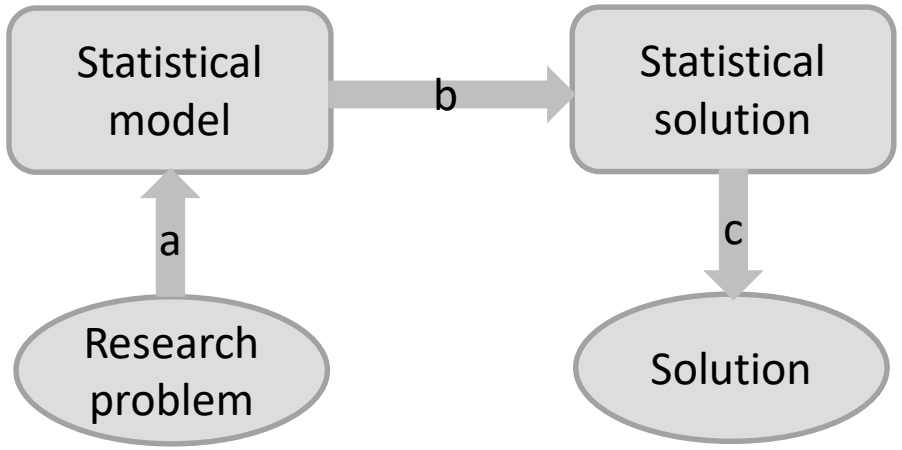
08/06/2019

# What is Statistical Collaboration?

" The collaboration of a statistician with another professional for the purpose of devising solutions to research problems"

Kirk RE. (1991) Statistical consulting in a university: dealing with people and other challenges. *American Statistician 45(1):28-34.*

# The Core Process of Statistical Collaboration




**a**: The researcher's problem will be translated into a more abstract statistical representation. It depends on how clearly and accurately the statistician have understood the problem.

**b**: The solution the statistician develops will be based on this abstraction and will reflect her training in statistics.

**c**: The statistical solution is translated back to the researcher. How the researcher understands and accepts the solution also depends on how the communication went.

# When Are Statisticians Contacted

- Study is a twinkle in the researcher's eye
- Study is more thought out, but needs some polishing to proceed
- Study design is defined, needs help with data collection considerations before study starts
- Data has been collected, needs help with analysis
- Data analysis has been performed by someone else, wants blessing
- Manuscript has been submitted, and needs help with reviewer's comments

# Advantages for Involving a Statistician Early

- Help clarify objectives of the research
- Formulate the research question as a statistical problem
- Help identify variables/measures that are important to the research objectives
- Important for the conclusions from the research to have a meaningful interpretation

# Possible Statistician's Roles

- Helper
  - Low-level of involvement in substantial aspects; acts like a technician
- Leader
  - Assumes responsibility for making sense of the client's data
  - Intellectually involved, client has passive role
- Data Blesser
  - Simple question, no opportunity to review
- Collaborator
  - Pool talents/expertise so sum of the parts is greater than the whole
  - Consultant is involved from the inception through report-writing
- Teaching
  - Important by-product of consulting

Kirk RE. (1991)

# Types of Statistical Assistance

- Study design
- Data collection
- Data analysis
- Manuscript/report/presentation preparation
- Peer review of manuscript
- Grant preparation
- Study review committee, like Data Safety and Monitoring Board for clinical trials.

# An Example

**Background**: A medical center developed and published a clinical prediction model to predict the risk of certain disease. A Johns Hopkins clinician is interested in collecting an independent cohort of patients from Johns Hopkins Hospital to validate the risk score. He sent the two published papers from the study group and wants the statistician's advice about his study.

**Questions:**

1. What advantages does the clinician gain by contacting a statistician at this stage?

2. What type of roles do you see the statistician will play in this project?

3. Imagine the clinician contacted the statistician after collecting the data, what can potentially go wrong?

# What Statisticians May Not Know

- May have little or no subject knowledge- don't assume that they are familiar with certain variables or instruments/acronyms
- May not be experienced with your database software
- May not be a good database programmer
- May not be familiar with statistical methods specific to your field

# Reproducible Research Components

- Data
  - Database Software
  - Data Documentation
  - Variable Names/Value labels
  - Datafile Version Control
  - Transmission of Data Files

- Documents
  - Protocols, analysis plans, reports, meeting notes, emails
  - Data management and Analysis programs

- Individual work folder for each study that stores data and documents

# Communication – Subject Matters

Project 1 email folder:

| Subject | Received* |
|---|---|
| IRB re-submission-PLEASE READ | Fri 3/22/2019 9:59 AM |
| eIRB: Agree to Participate | Fri 3/22/2019 9:58 AM |
| Re: regression model | Thu 3/21/2019 7:34 AM |
| Re: regression model | Mon 3/18/2019 7:31 AM |
| Re: figure edits | Wed 3/13/2019 7:33 AM |
| Re: figure edits | Tue 3/12/2019 9:59 AM |
| re-submission to IRB | Sat 3/2/2019 8:21 AM |
| figure edits | Sat 3/2/2019 7:48 AM |
| Re: Re: | Fri 3/1/2019 10:05 AM |
| | |
| Re: Re: | Wed 2/27/2019 4:58 PM |
| Re: Re: | Mon 2/25/2019 9:10 AM |
| Re: Re: | Wed 2/20/2019 10:55 AM |
| | Wed 2/20/2019 8:45 AM |
| Re: Figures | Mon 2/18/2019 3:10 PM |
| Fw: Bloomberg Project Task Order | Mon 2/18/2019 3:10 PM |
| Re: Figures | Sun 2/17/2019 10:06 AM |
| Re: Figures | Tue 2/12/2019 10:53 AM |
| Figures | Sun 2/10/2019 9:21 AM |

Project 2 email folder:

| Subject | Received |
|---|---|
| Re: Analyses using TNFa-R1 and CD25 | Mon 8/20/2018 1:23 PM |
| Re: Analyses using TNFa-R1 and CD25 | Mon 8/20/2018 12:29 PM |
| Re: Analyses using TNFa-R1 and CD25 | Fri 8/17/2018 2:24 PM |
| Re: Analyses using TNFa-R1 and CD25 | Fri 8/17/2018 12:56 PM |
| Analyses using TNFa-R1 and CD25 | Thu 8/16/2018 5:01 PM |
| Re: raw vs. log-transformed variables for power calculations | Thu 8/16/2018 4:10 PM |
| Re: raw vs. log-transformed variables for power calculations | Thu 8/16/2018 4:02 PM |
| Re: raw vs. log-transformed variables for power calculations | Thu 8/16/2018 3:54 PM |
| Re: raw vs. log-transformed variables for power calculations | Thu 8/16/2018 3:28 PM |
| Re: raw vs. log-transformed variables for power calculations | Thu 8/16/2018 3:14 PM |
| Re: raw vs. log-transformed variables for power calculations | Thu 8/16/2018 2:43 PM |
| Re: raw vs. log-transformed variables for power calculations | Thu 8/16/2018 12:40 PM |
| Re: raw vs. log-transformed variables for power calculations | Thu 8/16/2018 11:41 AM |
| Re: raw vs. log-transformed variables for power calculations | Wed 8/15/2018 2:34 PM |
| raw vs. log-transformed variables for power calculations | Wed 8/15/2018 12:57 PM |
| Re: TIME SENSITIVE: additional power calculations | Tue 8/14/2018 11:19 AM |
| Re: TIME SENSITIVE: additional power calculations | Mon 8/13/2018 8:25 PM |
| Re: TIME SENSITIVE: additional power calculations | Mon 8/13/2018 4:53 PM |
| TIME SENSITIVE: additional power calculations | Mon 8/13/2018 12:38 PM |

Which one works better if you need to go back and look for information?

# Dataset Distribution

- Be careful about HIPAA!
- A dataset containing Patient Medical Information (PMI) cannot be e-mailed unless it is encrypted
- Best bet:  only distribute de-identified datasets
  - Redcap will create one for you automatically
- If someone e-mails a statistician an unencrypted dataset with PMI, she might be obligated to report them.
- PMI includes dates and ages if >90
- Consider up-to-date JH technology, such as Safe Desktop, OneDrive, for file distribution

# Main Points

- Develop a collaboration ***early***
- Both you and the statistician should ***be involved*** in that collaboration
- Useful data is ***well-documented*** data
- For written communication, such as emails, reports, use descriptive names that can be interpreted in the future
- You and the statistician can save each other time