Biostatistics Support for Investigators: Tips for Collaborating with a Statistician

Pediatric Biostatistics Core
April 9, 2021

Scott Gillespie, MS, MSPH
Janet Figueroa, MPH

What is collaborative biostatistics?

- Collaborative biostatistics is the creative application of applied statistical tools to areas of biology and medicine

- Two broad areas where collaborative biostatisticians add value to the research enterprise:
  1. Study design
     - Hypothesis refinement; conceptualization of complex relationships between variables; sample size and power; statistical design plans
  2. Data analysis
     - Interpretation and reporting of results; technical write-ups

- These collaborations lead to stronger grant proposals and manuscript submissions (many reviewers expect statistical collaborators)

Biostatistician Make-up

- An experienced biostatistician is competent in four areas:
  1. Technical and analytical
     - Familiar with modern statistical methods and software coding
  2. Broad subject knowledge
     - Working knowledge of the biomedical content
  3. Communication
     - Ability to understand and be understood
  4. Problem-solving
     - Synthesize critical study components to answer research questions

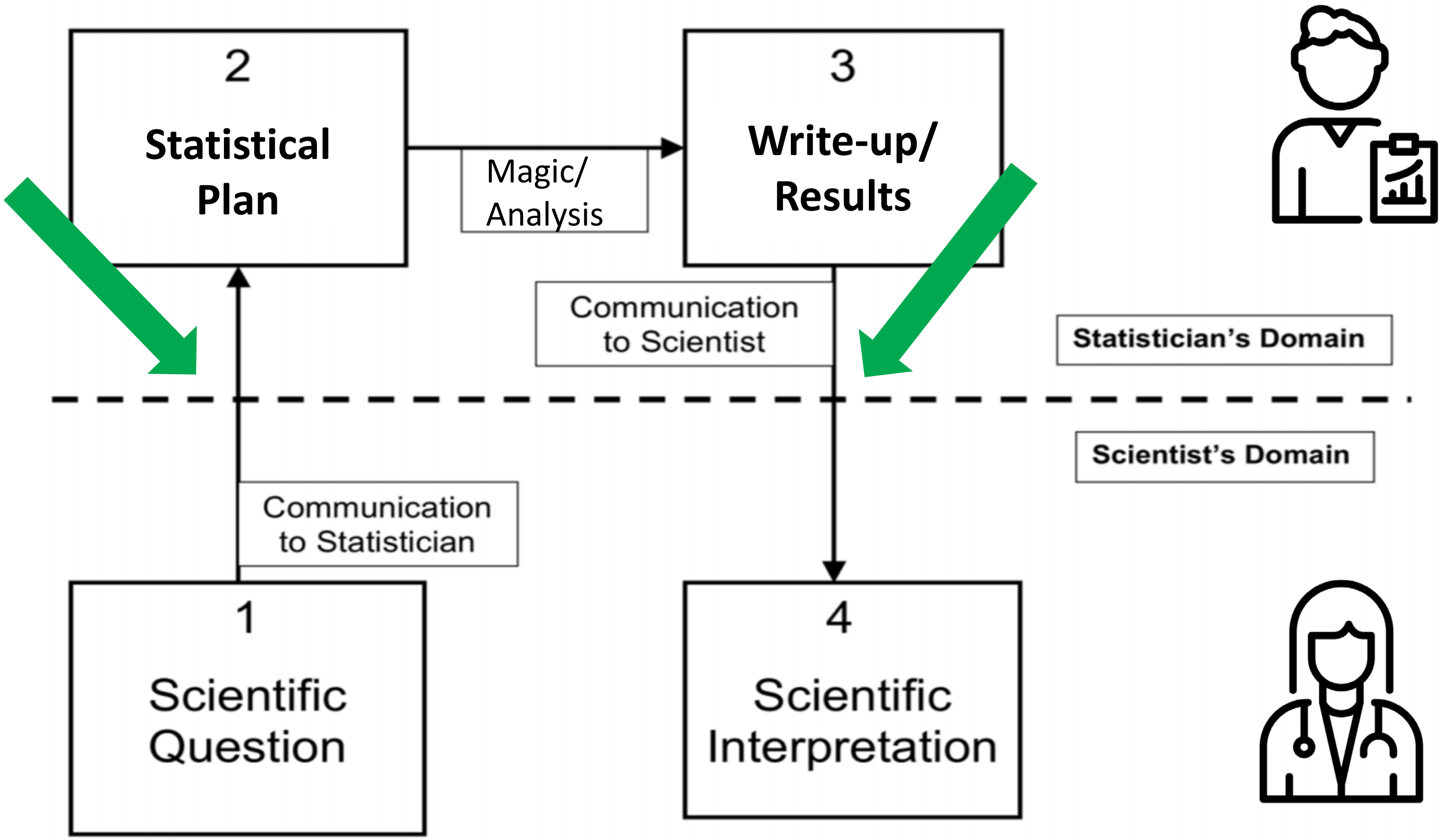
What information we may not know

- A statistician may lack specific subject knowledge for your study
  - You should not assume we are familiar with all acronyms, jargon, or instruments you propose to use (communication is key here)

- We are generally not database experts and may not be familiar with your data collection software
  - At Emory, we do see and can advise on, Redcap, Excel, some SQL databases

- We may not have experience with a niche method or analysis plan that is common for your field
  - Early contact, providing relevant papers, and table/figure mock-ups are helpful to educate your statistician on your data

General collaboration flow



2 Statistical Plan

Magic/ Analysis

3 Write-up/ Results

Communication to Scientist

Statistician's Domain

Scientist's Domain

Communication to Statistician

1 Scientific Question

4 Scientific Interpretation

Samsa, GP. "A day in the professional life of a collaborative biostatistician deconstructed: Implications for curriculum design. *Journal of Curriculum and Teaching*. 2018. Vol. 7(1): pp 20-31. doi:10.5430/jct.v7n1p20"

Statistical Considerations Checklist

1. Come to us early in your study process

When to contact your statistician

A. Study conceptualization 😆

B. Study is conceptualized, but needs polishing 😏

C. Data collection phase, before study starts 😐

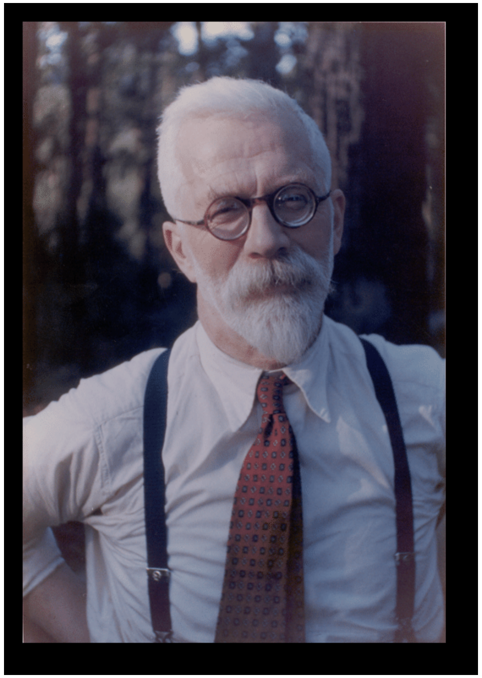
D. Data collected and needs analysis help 😟

E. Performed own analysis and needs checking 😬

F. Manuscript submitted, answering reviewer criticisms 😢

To call in the statistician after the experiment is done may be no more than asking them to perform a post-mortem examination: they may be able to say what the experiment died of.

-RA Fisher



Advantages to involving statisticians early

- Help you think through technical objectives of research study

- Reconstruct research questions into research hypotheses, and ultimately, into statistical hypotheses to inform analysis

- Help identify variable types, directions and anticipated strengths of relationships, moderators/mediators, and the role of nuisance characteristics (i.e., confounders)

- Ensure available data and planned analysis are appropriate for answering the research question(s)

Preferred Lead-Times

- Scientific Abstracts – At least 1 month
- Manuscript Preparation – At least 1 month
- Intramural Grant Applications – At least 6 weeks
- Extramural Grant Applications – At least 2 months
- The Pediatric Biostatistics Core generally will not shun you if you reach out after these lead times have passed
  - However, we may not have adequate time to advise in all phases of your study

Statistical Considerations Checklist

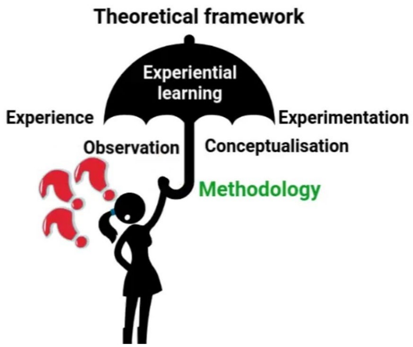
1. Come to us early in your study process

2. Have an idea of your research question and hypotheses

Steps for Research Question Development

1. Observation of a problem ✅
   - ”Why are things done this way?” or “What if we did…”

2. Conceptualize the theoretical framework ✅
   - A theoretical framework defines the key concepts in your research and proposes relationships between them:
     - What clinical characteristics will be measured (i.e., independent variables)
     - What outcome(s) are of interest (i.e., dependent variables)
     - What associations you may want to consider



Steps for Research Question Development

3. Literature review ✓
   - Electronic search of PUBMED, MEDLINE, etc. *using search terms identified in the theoretical framework stage* (e.g., ICS, CF, pseudomonas)
   - Identify weaknesses or holes in the literature and how your study could fill the gaps (e.g., Little confound consideration or non-representative samples)

4. Identify variables of interest (exposures, outcomes, confounds) ✓
   - At this point, we do not have to layout a prediction (i.e., hypothesis), but instead, have all information together to inform a quality research question

5. Formulate a well-informed research question ✓
   - “What is the effect of inhaled corticosteroids (ICS) on the incidence of lung infections in children with cystic fibrosis (CF)?”

Research Question to Research Hypothesis

- The research question presents a broad idea to be examined in a research study
  - “What is the effect of inhaled corticosteroids (ICS) on the incidence of lung infections in children with cystic fibrosis (CF)?”

- The research hypothesis *makes a prediction* and is a more focused attempt to empirically answer the research question
  - “Use of ICS will *increase the incidence of lung infections* in children with CF.”
  - A research hypothesis *is a bridge* connecting theory and observation

Statistical Considerations Checklist

1. Come to us early in your study process
2. Have an idea of your research question and hypotheses
3. Consider study feasibility and population, data availability

Study Details

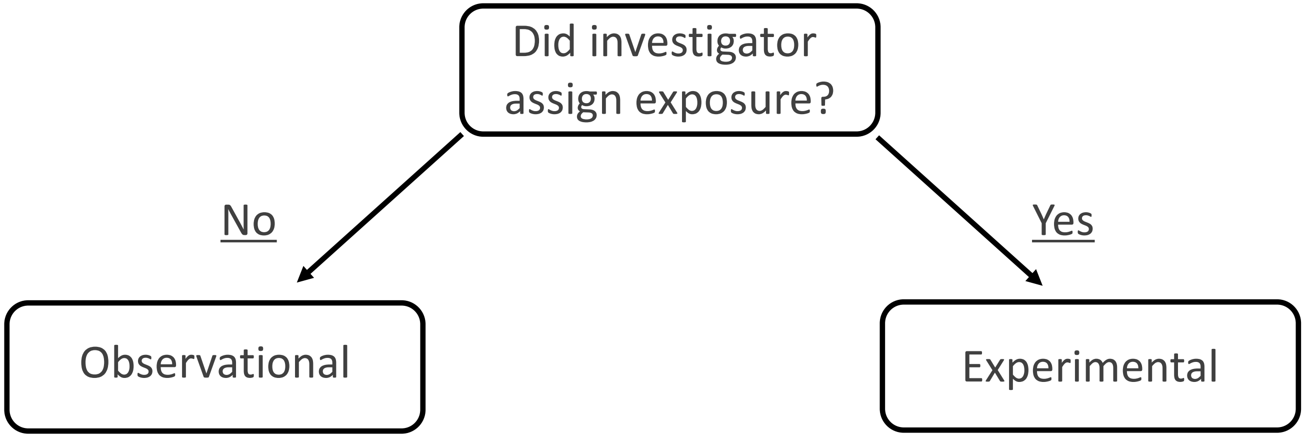
- Feasibility
  - A researcher should consider time, availability of subjects, facilities, finances
  - Must also evaluate their own experience and ethical considerations

- Study population
  - Determination of inclusion/exclusion criteria
  - Chart review, larger EHR data pull, prospective recruitment, registries
  - Willingness of the population to participate and attend visits (if prospective)

- Testability
  - Research questions consider relationships between exposures and outcome
    - Can variables needed to test the research question be reasonably gathered

Study Design Caveat

- Starting with the basics



Did investigator assign exposure?

No

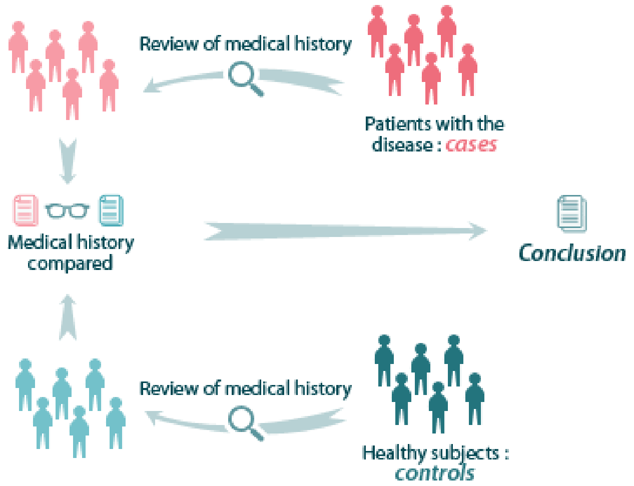
Yes

Observational

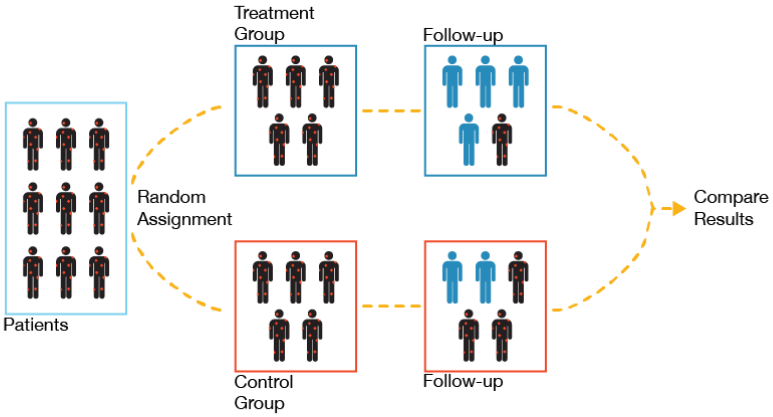
Experimental

Study Design Caveat

- Observational studies (epidemiological) – Do not involve any intervention or experiment but instead observe the natural relationships between exposures and outcomes

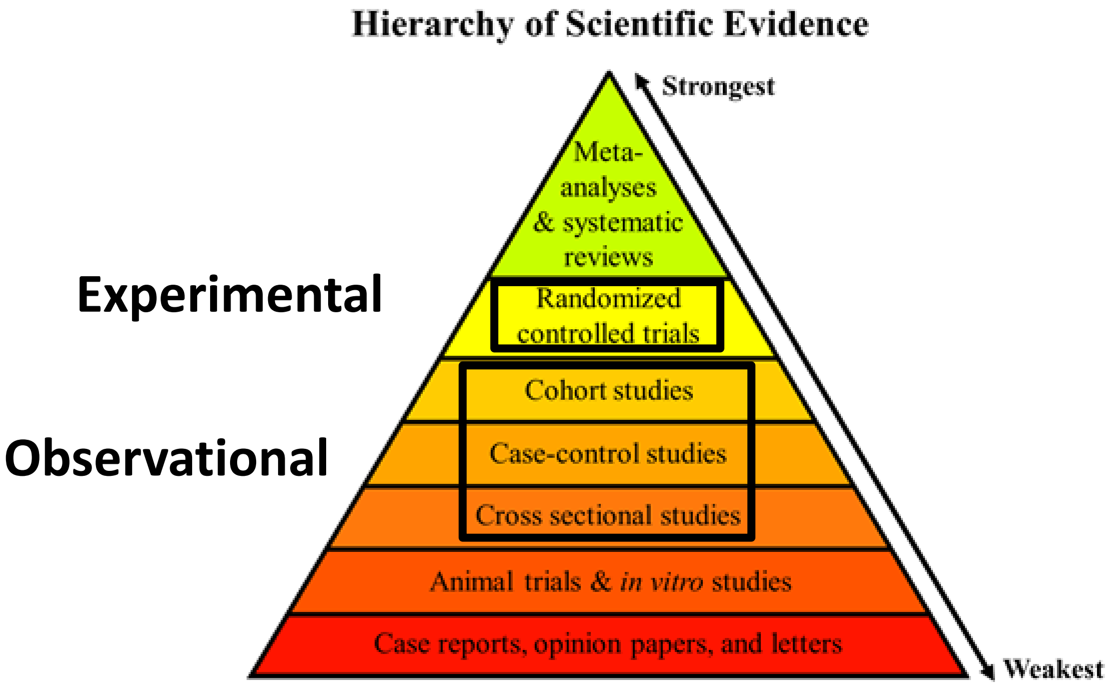

Review of medical history
Patients with the disease : cases
Medical history compared
Conclusion
Review of medical history
Healthy subjects : controls

- Experimental (interventional) studies – Entail manipulation of the study factor (exposure) and randomization of subjects to exposure (i.e., treatment) groups



Study Design Caveat

- **Observational studies**
  - Surveys
  - Cohort studies
  - Cross-sectional studies
  - Case-control studies
- **Experimental studies**
  - Randomized trials (RCTs)
  - Non-randomized studies
- **Systematic reviews**
- **Qualitative research**
- **Animal/lab trials**



Hierarchy of Scientific Evidence

Strongest

Meta-analyses & systematic reviews

Experimental

Randomized controlled trials

Observational

Cohort studies

Case-control studies

Cross sectional studies

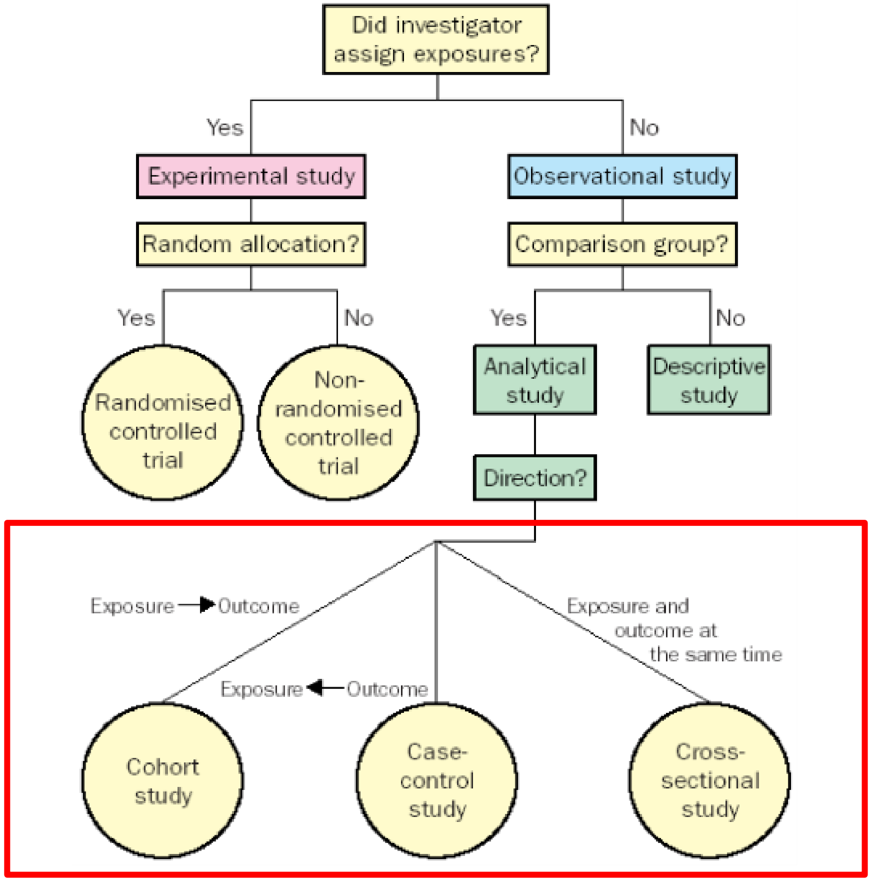
Animal trials & *in vitro* studies

Case reports, opinion papers, and letters

Weakest

Primary Types of Observational Studies

- Most studies carried out by new researchers are observational
  - Case-Control Study
  - Cross-Sectional
  - Cohort Study
    - Retrospective
    - Prospective



Statistical Considerations Checklist

1. Come to us early in your study process
2. Have an idea of your research question and hypotheses
3. Consider study feasibility and population, data availability
4. Conceptualize data collection, formatting and layout

Data formatting and layout

- Most data our group encounter arrives in Excel files that originate from:

  1. REDCap (CHOA or Emory)
     - Retrospective chart review or prospective studies
     - Clearly formats and standardizes data

  2. CHOA EHR/EPIC data pulls
     - Retrospective studies, but too big for individual review
     - Some formats and standards, not as clean as REDCap

  3. Retrospective or prospective data collected without REDCap or data standardization (least preferred)

Data standards – Which is correct?

Patient	Age
A	1
B	2
C	3

Patient	Age	Exclude
A	1	
B	2	1
C	3	

✓

Patient ID	Dose	% of Doses
1	1 mg	10
2	1	10%
3	0.001 g	0.1
4	1 mg/day	100%
5	1 mg/kg weight/ day	1

Patient ID	Dose in mg	% of Doses
1	1	10%
2	1	10%
3	1	10%
4	1	100%
5	1	100%

✓

Data standards – Which is correct?

✓

Patient	Steroid
1	dexamethasone
2	dexamethasone
3	dexamethasone
4	prednisone
5	prednisone
6	prednisone
7	prednisone

Patient	Steroid
1	decdron
2	dex
3	dexamethasone
4	oral pred
5	oralpred
6	orapred
7	ORAPRED

Patient ID	Dose in mg
1	1
2	2
3	EXCLUde
4	1
5	1 (dose was delayed)

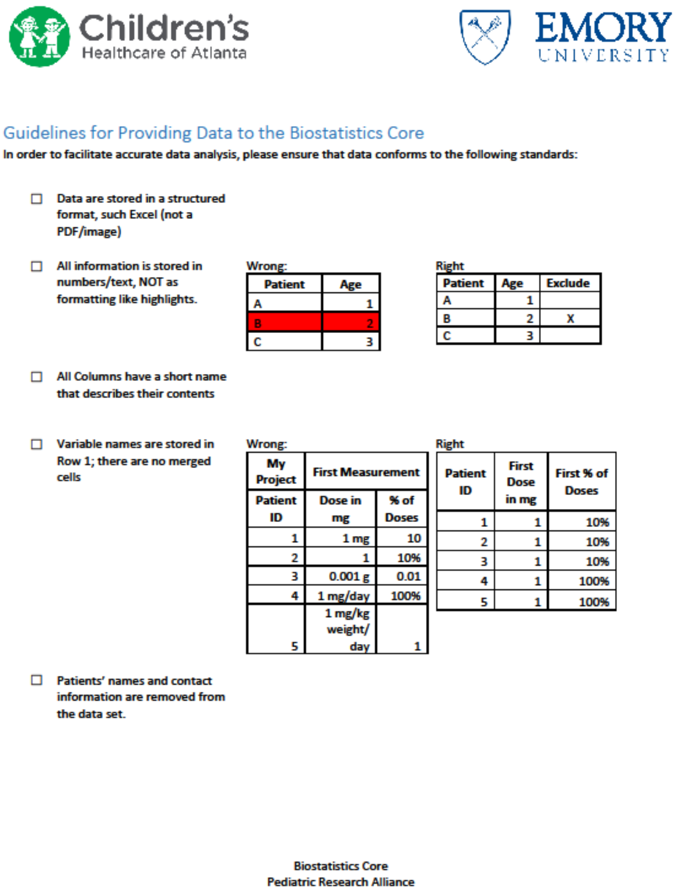
ID	Dose in mg	Exclude	Delayed
1	1		
2	2		
3		1	
4	1		
5	1		1

✓

- Clean data expedite and inform analysis

Suggested Data Guidelines

- https://www.pedsresearch.org/uploads/blog/doc/Biostatistics_Core_Data_Guidelines.pdf



Children's
Healthcare of Atlanta

EMORY
UNIVERSITY

Guidelines for Providing Data to the Biostatistics Core

In order to facilitate accurate data analysis, please ensure that data conforms to the following standards:

□ Data are stored in a structured format, such Excel (not a PDF/image)

□ All information is stored in numbers/text, NOT as formatting like highlights.

Wrong:

Patient	Age
A	1
B	2
C	3

Right

Patient	Age	Exclude
A	1	
B	2	X
C	3	

□ All Columns have a short name that describes their contents

□ Variable names are stored in Row 1; there are no merged cells

Wrong:

My Project	First Measurement	
Patient ID	Dose in mg	% of Doses
1	1 mg	10
2	1	10%
3	0.001 g	0.01
4	1 mg/day	100%
5	1 mg/kg weight/ day	1

Right

Patient ID	First Dose in mg	First % of Doses
1	1	10%
2	1	10%
3	1	10%
4	1	100%
5	1	100%

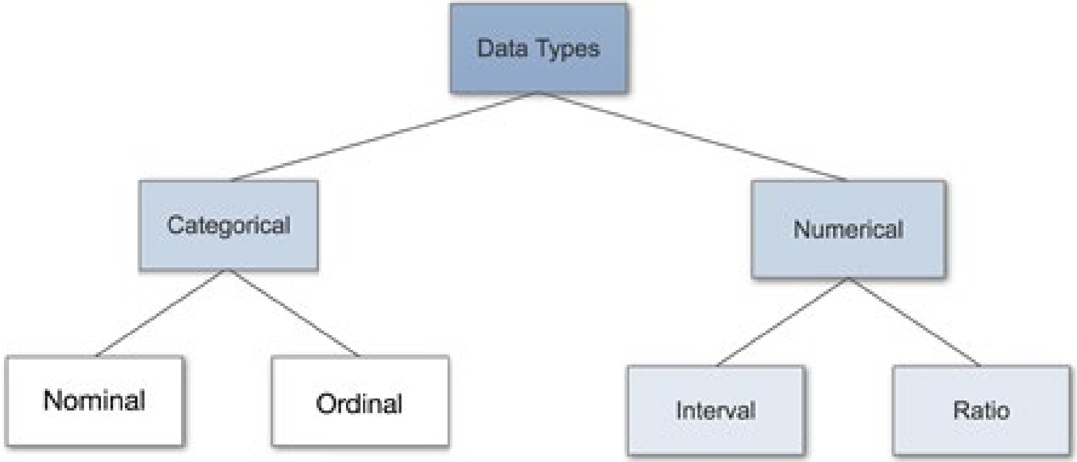
□ Patients' names and contact information are removed from the data set.

Biostatistics Core
Pediatric Research Alliance

Data Types

- Primary exposure variable
  - Generally, primary exposures are discrete (e.g., prematurity yes/no)
  - These discrete, independent variables lump patients into mutually exclusive groups for statistical testing

- Outcome variable(s)
  - Outcome variables may be discrete or continuous
  - For discrete outcomes, data may be ordered or unordered (nominal)



Data Types and Statistical Tests

- Hypotheses and data types determine the appropriate statistical tests for a study

Exposure	Outcome				
	2 Categories	>2 Categories	Ordered Categories	Normal Continuous	Non-Normal Continuous
2 Categories	Chi-square test of independence	Chi-square test of independence	Chi-square test for trend	Two-sample t-test	Mann-Whitney Test
>2 Categories	Chi-square test of independence	Chi-square test of independence	Chi-square test for trend	One-way ANOVA	Kruskal-Wallis Test
Ordered Categories	Chi-square test for trend	Chi-square test for trend	Spearman correlation	Spearman correlation	Spearman Correlation
Normal Continuous	Logistic regression	Nominal logistic regression	Ordinal logistic regression	Pearson correlation and linear regression	Spearman correlation and generalized linear regression
Non-Normal Continuous	Logistic regression	Nominal logistic regression	Ordinal logistic regression	Spearman correlation and linear regression	Spearman correlation and generalized linear regression

Statistical Considerations Checklist

1. Come to us early in your study process

2. Have an idea of your research question and hypotheses

3. Consider study feasibility and population, data availability

4. Conceptualize data collection, formatting and layout

5. Statistical hypotheses and sample size

Sample size and statistical hypotheses

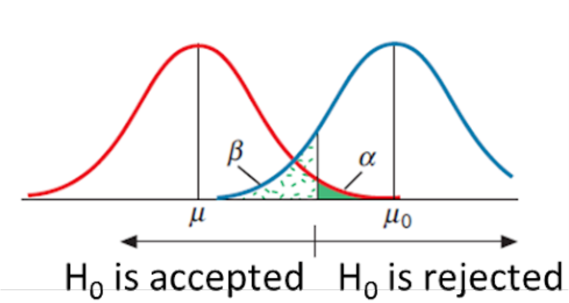
- We have discussed research hypotheses thus far, but need to convert these questions to statistical hypotheses for analysis

- Thinking back to our example: "What is the effect of inhaled corticosteroids (ICS) on the incidence of lung infections in children with cystic fibrosis (CF)"
  - We have two groups, based on our primary exposure: ICS+ and ICS-
  - The outcome may be an *incidence proportion (i.e. of lung infections)*

- Converting this information into statistical hypotheses, we would have:
  - **Null:** $p_{ICS+} = p_{ICS-}$
  - **Alternative:** $p_{ICS+} > p_{ICS-}$ or

    $p_{ICS+} < p_{ICS-}$ or

    $p_{ICS+} \neq p_{ICS-}$ *(two-sided)*



Power and Sample Size

- Power is the probability of **correctly** rejecting a null hypothesis when there is indeed a difference to reject

- Statistical power is important
  - If a study is too small, you will not be able to answer the research question and waste time, resources, and money (worst case)

  - If a study is too large, the research question may be answerable, but you will still waste valuable time and resources (better, but not ideal)

- Depending on study design, most grants should have some discussion of power

Power and Sample Size

- When discussing power with your statistician, it is helpful to have:
  1. Statistical hypotheses and an idea of directionality
  2. Rough estimates of means/proportions and variances for the outcome you are trying to study (i.e., effect size, *d*)
  3. Proportional breakdowns of your exposure groups
  4. Some idea of how many subjects you can realistically consent

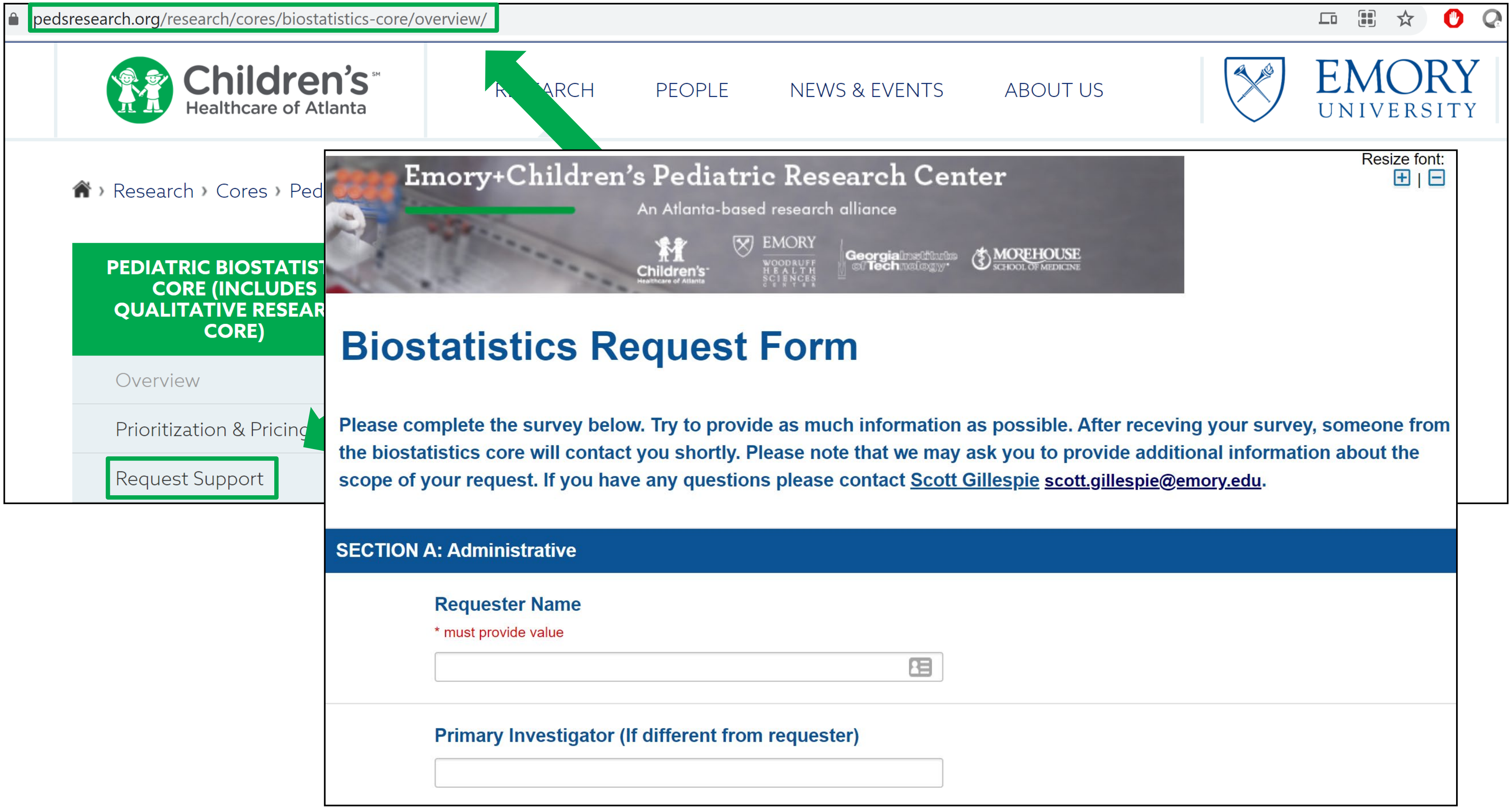
- An example:
  - Alternative hypothesis: $p_{ICS+} \neq p_{ICS-}$
  - Assuming lung infection outcome: p_{ICS+} = 35%; p_{ICS-} = 20%
  - 40% ICS+ / 60% ICS- split in study sample
  - What sample size do I need to detect this difference?
    - **N=125 ICS+ / N=186 ICS- (N=311 total)**

Statistical Considerations Checklist

1. Come to us early in your study process
2. Have an idea of your research question and hypotheses
3. Consider study feasibility and population, data availability
4. Conceptualize data collection, formatting and layout
5. Statistical hypotheses and sample size
6. How to reach us and funding considerations

Pediatric Biostatistics Core

- Internet search - “Emory Pediatric Biostatistics Core”


pedsresearch.org/research/cores/biostatistics-core/overview/
Children’s Healthcare of Atlanta
RESEARCH
PEOPLE
NEWS & EVENTS
ABOUT US
EMORY UNIVERSITY
Research › Cores › Ped
PEDIATRIC BIOSTATIS
CORE (INCLUDES
QUALITATIVE RESEAR
CORE)
Overview
Prioritization & Pricing
Request Support
Emory+Children’s Pediatric Research Center
An Atlanta-based research alliance
Children’s Healthcare of Atlanta
EMORY WOODRUFF HEALTH SCIENCES CENTER
Georgia Institute of Technology
MOREHOUSE SCHOOL OF MEDICINE
Resize font:
⊞ | ⊟

Biostatistics Request Form

Please complete the survey below. Try to provide as much information as possible. After receving your survey, someone from the biostatistics core will contact you shortly. Please note that we may ask you to provide additional information about the scope of your request. If you have any questions please contact Scott Gillespie scott.gillespie@emory.edu.

SECTION A: Administrative

Requester Name

* must provide value

Primary Investigator (If different from requester)

Pediatric Biostatistics Core

Scott Gillespie, MS, MSPH
scott.gillespie@emory.edu
404-727-4113, 404-785-6869

Chao Zhang, PhD
chao.zhang2@emory.edu

Traci Leong, PhD
tleong@emory.edu
404-727-9169

Amanda Thomas, MSPH
amanda.thomas@emory.edu
404-712-9571

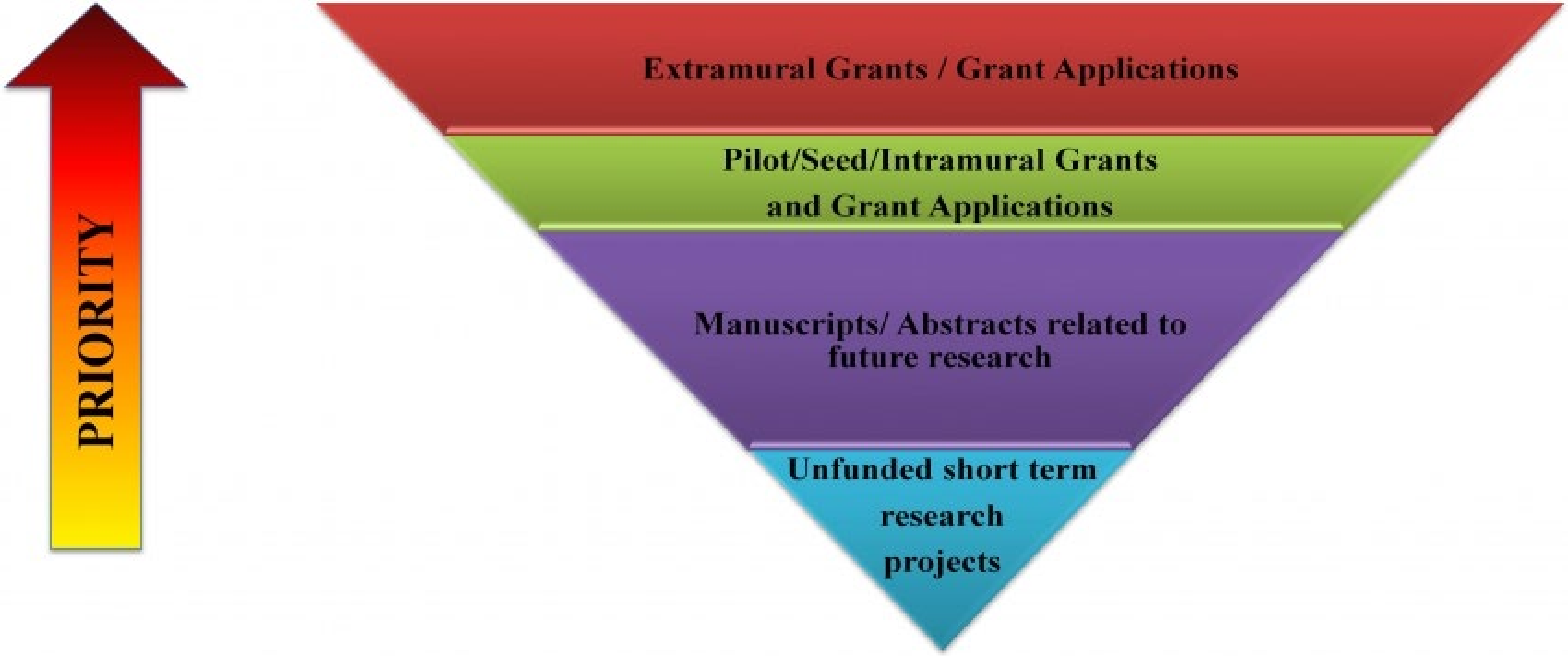
Michael Scott Kelleman, MSPH
michael.kelleman@emory.edu
404-727-5882

Janet Figueroa, MPH
janet.figueroa@emory.edu
404-785-7507

Yijin Xiang, MPH
yijin.xiang@emory.edu

Anna Wood, MPH
anna.wood@emory.edu

Core priorities



Extramural Grants / Grant Applications

Pilot/Seed/Intramural Grants and Grant Applications

Manuscripts/ Abstracts related to future research

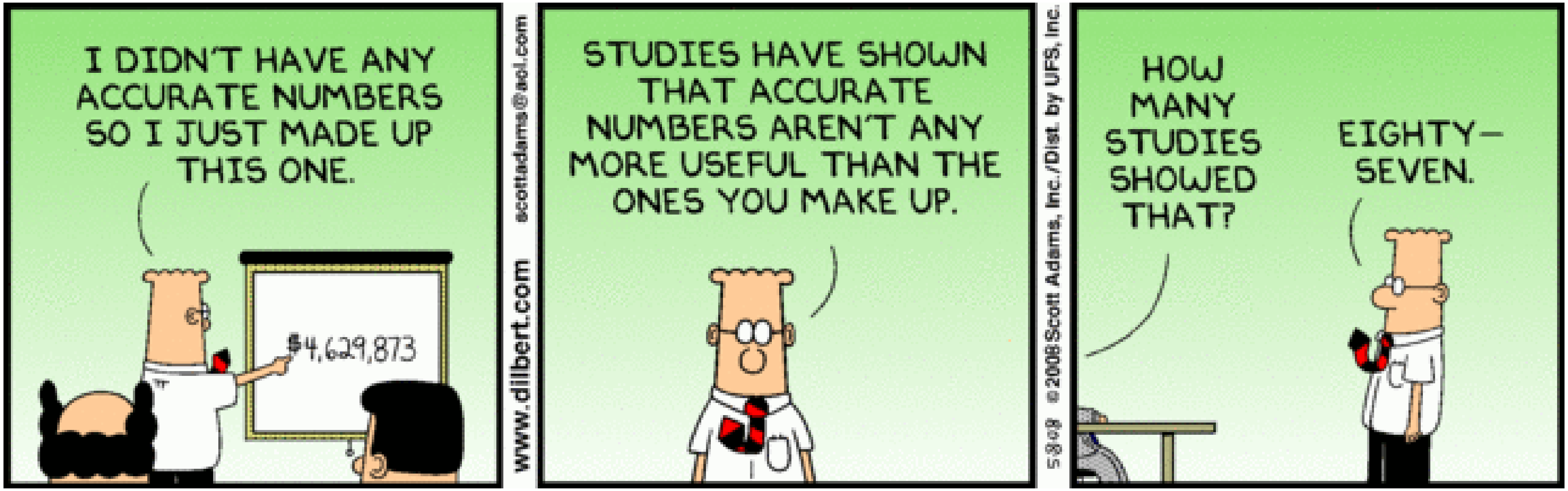
Unfunded short term research projects

Funding the core

- All new studies from pediatric investigators are entitled to some amount of free support, depending on project
  - 4 hours for a manuscript or abstract
  - 8-16 hours for grant applications

- Once the subsidized period is exhausted, additional biostatistics time must be funded using:
  - Awarded grants
  - Senior investigator or departmental funds
  - Other funding mechanisms

- If funding is an issue, speak with us early, so a reasonable plan can be devised


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.
www.dilbert.com scottadams@aol.com
5-8-08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.

Biostatistics Support for Investigators: Tips for Collaborating with a Statistician

Pediatric Biostatistics Core
April 9, 2021

Scott Gillespie, MS, MSPH
Janet Figueroa, MPH