P9185: Statistical Practices and Research for Interdisciplinary Sciences (**SPRIS**)

Lecture 1

Yuanjia Wang, Ph.D.

Department of Biostatistics, Mailman School of Public Health
Columbia University

& Division of Biostatistics, New York State Psychiatric Institute

History and Overview of the Course

- History and rationale
  - Launched in 2019 to address a gap in our doctoral training: applied analysis skills, effective collaboration or communication, team science
  - Scope expanded: emerging exciting topics; methodological research motivated from real world problems
- *Syllabus*: didactic lectures, case studies, emerging topics, current critical issues in the field, class projects and presentations

Introduction to the Course: (Bio)Statistics

What defines statistics/biostatistics as a field? A collection of tools (ANOVA, chi-squared, t-tests, regression, machine learning techniques etc)? Can AI replace statisticians (e.g., *the Automatic Statistician*)?

[1] *COPSS Awards and Fisher Lecture, JSM 2017*.

Introduction to the Course: (Bio)Statistics

What defines statistics/biostatistics as a field? A collection of tools (ANOVA, chi-squared, t-tests, regression, machine learning techniques etc)? Can AI replace statisticians (e.g., *the Automatic Statistician*)?

As Professor Robert Kass explained[1]:

Two Fundamental Tenets of the Statistical Paradigm

1. Statistical models are used to express knowledge and uncertainty about a signal in the presence of noise, via inductive reasoning and inference from data.

2. Statistical methods may be analyzed to determine how well they are likely to perform.

[1] *COPSS Awards and Fisher Lecture, JSM 2017*.

Big Picture of Statistical Inference (Standard View)[2]

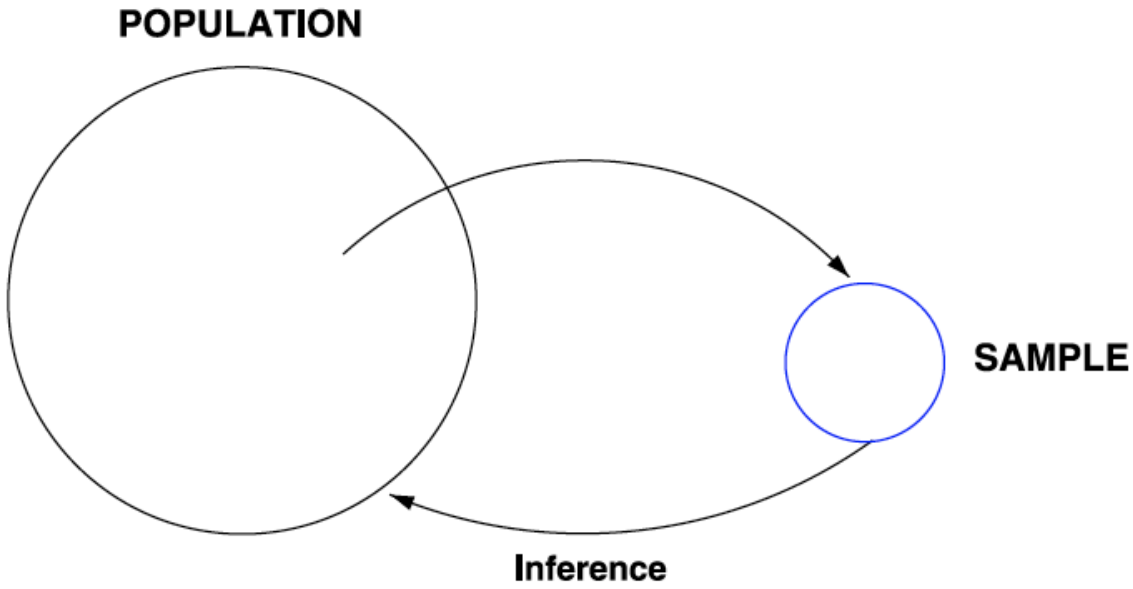


FIG. 3. *The big picture of statistical inference according to the standard conception. Here, a random sample is pictured as a sample from a finite population.*

[2]Kass. Statistical Inference: The Big Picture. *Statistical Science*. 2011, Vol. 26, No. 1, 1–9

Big Picture of Statistical Inference (Pragmatic View)[2]

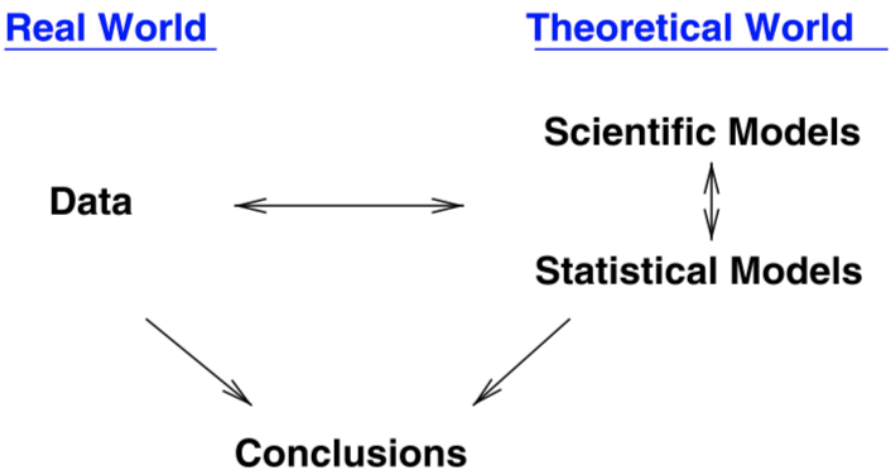


FIG. 1. *The big picture of statistical inference. Statistical procedures are abstractly defined in terms of mathematics but are used, in conjunction with scientific models and methods, to explain observable phenomena. This picture emphasizes the hypothetical link between variation in data and its description using statistical models.*

Important step: abstraction (leap of faith) from scientific model to statistical model

[2]Kass. Statistical Inference: The Big Picture. *Statistical Science*. 2011, Vol. 26, No. 1, 1–9

Big Picture of Statistical Inference (Pragmatic View)[2]

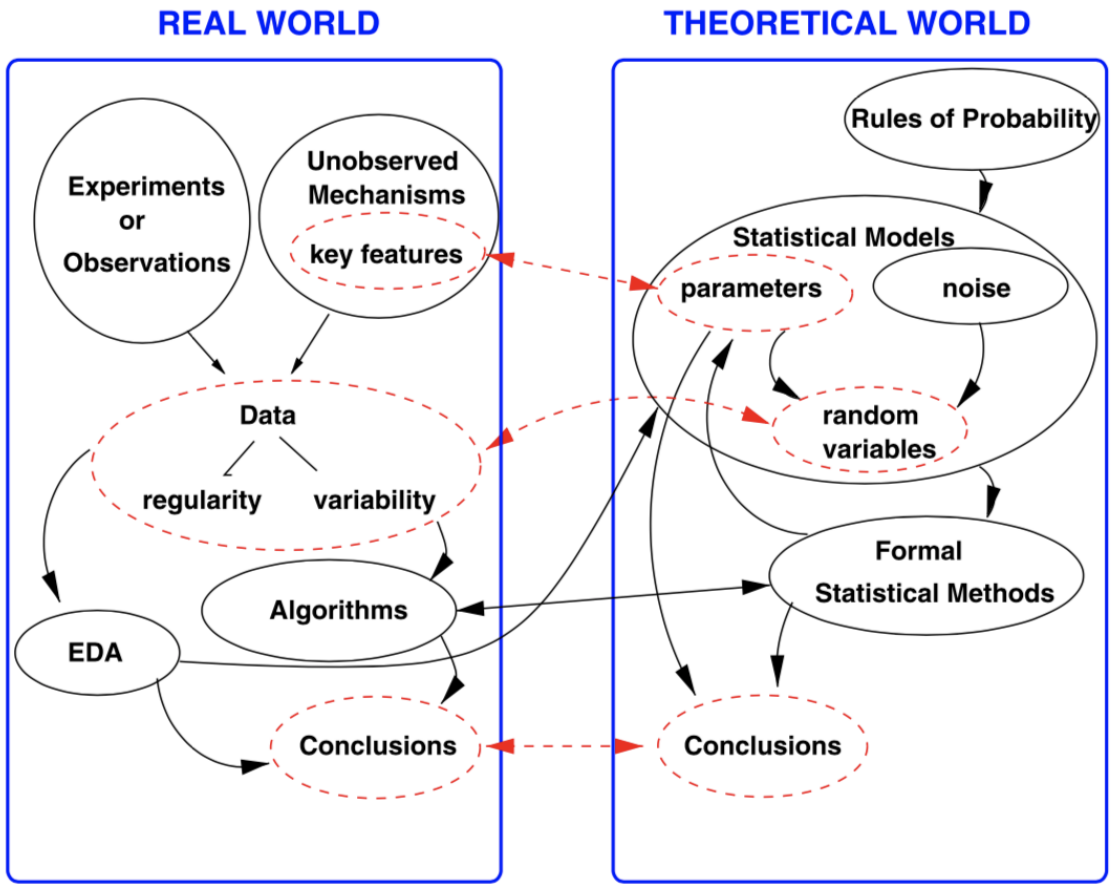


Figure 4. A more elaborate big picture, reflecting in greater detail the process of statistical inference.

[2]Kass. Statistical Inference: The Big Picture. *Statistical Science*. 2011, Vol. 26, No. 1, 1–9

Ten Simple Rules for Effective Statistical Practice[3]:

- Rule 1: Statistical Methods Should Enable Data to Answer Scientific Questions
- Rule 2: Signals Always Come with Noise
- Rule 3: Plan Ahead, Really Ahead
- Rule 4: Worry about Data Quality
- Rule 5: Statistical Analysis Is More Than a Set of Computations

[3]Kass RE, Caffo BS, Davidian M, Meng X-L, Yu B, Reid N (2016) Ten Simple Rules for Effective Statistical Practice. PLoS Comput Biol 12(6): e1004961.

Ten Simple Rules for Effective Statistical Practice[3]:

- Rule 6: Keep it Simple
- Rule 7: Provide Assessments of Variability
- Rule 8: Check Your Assumptions
- Rule 9: When Possible, Replicate!
- Rule 10: Make Your Analysis Reproducible

[3]Kass RE, Caffo BS, Davidian M, Meng X-L, Yu B, Reid N (2016) Ten Simple Rules for Effective Statistical Practice. PLoS Comput Biol 12(6): e1004961.

Introduction to the Course (SPRIS): Statistical Practice

Professor John Tukey:
"When we ask for the true sources of significant innovations in statistics, we find two sources, each indispensable:

- a desire for mathematically neat and tidy accounts of what to do in more or less idealized situations;
- a slow, pervasive disquiet about either the practical functioning of familiar techniques or the absence of ways to approach important questions."

"If we are able to preserve an irreplaceable source for future significant innovations, we must make it natural for a significant fraction of our sharpest minds *to learn to consult and to continue consulting, at least to a degree, throughout their career*. This means that our Ph.D. students will need streamlined training in statistical consultation, too."

Emphasis placed on exploratory data analysis (as opposed to confirmatory data analysis).

Introduction to the Course (SPRIS): Research

Biostatistics research focuses on health data science: answering important biomedical research questions in a scientific manner not only with well-trodden theory-driven hypotheses but as well with data generated by systems which exist today as never before.

Statistics offers probabilistic descriptions and mathematical models of the mechanisms by which the observed data are generated (generative models). Using statistical language and rigorous inferential tools, important queries can be posed to investigate scientific questions and draw conclusions while accounting for uncertainty and rule out chance findings.

This data-driven approach complements the traditional hypothesis-driven science and offers new insights using data.

Introduction to the Course (SPRIS): Interdisciplinary Sciences

Why interdisciplinary sciences?

Biomedical research problems are increasingly complex. "*Solving the puzzle of complex diseases, from obesity to cancer, will require a holistic understanding of the interplay between factors such as genetics, diet, infectious agents, environment, behavior, and social structures*"[4]. Elias Zerhouni, former director of the U.S. National Institutes of Health (NIH).

The most successful research teams of the future will likely involve unexpected combinations of experts, such as "*biological scientists, engineers, mathematicians, physical scientists, computer scientists, and others*".

[4] Zerhouni, E. (2003), "The NIH Roadmap," Science, 302, 63–72.

Overall Objectives of the Course (SPRIS)

- Biostatistics is a interdisciplinary field, intersects statistical, biological, public health, and medical sciences. Emphasis on medical and public health applications.
- Previous courses covered components of statistical analytics (e.g., various regression techniques, model building, experimental designs)
- Transfer learned components in real world contexts
- Identify novel statistical research problems with important public health and medical applications and advance the field of statistics
- Prepare students to be effective research collaborators and analytic partners to enter and lead interdisciplinary research efforts
- Gain important skills for data-driven decision making

Overall Objectives of the Course (SPRIS)

First and foremost, biostatisticians need strong disciplinary training in the theory and methods of statistics (all your other courses).

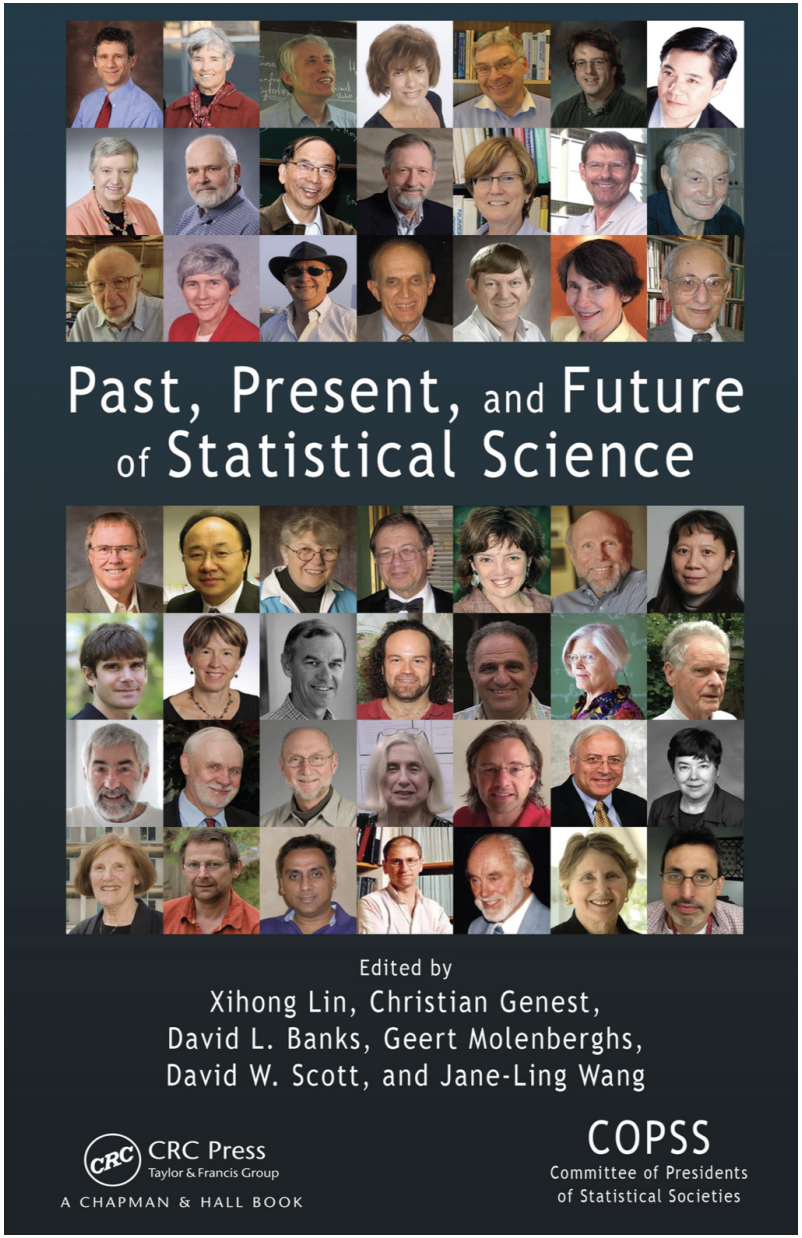
Biomedical research is diverse and complex in nature, requiring a fast-evolving and growing set of tools. No class can teach a student everything they would need to know for any specific project.

Therefore, the best skills are to learn how to listen actively, think statistically, communicate effectively, and acquire new tools and skills on the go.

Be prepared to play collaborative and leadership roles in interdisciplinary science, and to excel in team research initiatives.

Cover examples of collaboration as well as conducting original statistical dissertation research motivated from real world problems.

Career Pathways and Advices



Contents

Introduction to Statistical Consulting and Collaboration

[5]Morton (2005). Statistical collaboration to impact policy decisions. Statistics in Medicine, 24(4), 493-501.

Introduction to Statistical Consulting and Collaboration

What is statistical consulting versus collaboration[5]?

2. COLLABORATING VERSUS CONSULTING

Statistical analysis in health services research, or indeed any policy area, is an iterative process. Formulating research questions, and refining them based on data quality and availability requires the integral involvement of the statistician. Communicating between the realm of data and analysis, the statistician's traditional domain, and the content area, the policy researcher's traditional domain, is essential. I suggest that in order to facilitate this communication and thus have a larger impact, a statistician needs to collaborate rather than consult. Webster's Dictionary defines *to collaborate* as 'to work jointly with others or together especially in an intellectual endeavor' [1]. The key component in this definition is that you cannot collaborate alone, or even by spending the majority of your time working alone or at a distance. The synonyms listed in Roget's Thesaurus include 'cooperation, contribution, synergy, symbiosis, joint effort, team work' [2]. These synonyms stress the synergy of working together, the whole being more than the sum of the parts.

In contrast, *a consultant* is defined as 'one who gives professional advice or services' [1] with 'adviser, counselor, expert' [2] as synonyms. The statistical consultant's involvement on a project tends to be restricted. His or her interaction with other project members may be short-term, intermittent, and/or limited to a particular technical issue.

[5]Morton (2005). Statistical collaboration to impact policy decisions. Statistics in Medicine, 24(4), 493-501.

Introduction to Statistical Consulting and Collaboration

Ten steps of consulting and collaboration[6]:

1. Gaining Interest in Consulting
2. Strengthening Technical Skills
3. Understanding and Interacting with Collaborators
4. Developing Relationship Skills
5. Practicing Research Centered Collaboration
6. Recognizing the Importance of Collaboration to Research Practice
7. Collaborating to Improve your Own Practice
8. Taking the Lead in Research
9. Taking the Lead Among Statisticians
10. Leading the Way for Statisticians in Research

[6]Love-Myers et al. (2015). JSM 2015.

Introduction to Statistical Consulting and Collaboration

Ten steps of consulting and collaboration:

Introduction to Statistical Consulting and Collaboration

Practically, in academic medical centers:

- statistical consulting: conduct simple requested analysis, explain minor technical details; usually simple questions resolved within a short period (0.5-1 hour); may not require repeated scheduled meetings
- statistical collaboration: answers research questions which requires long-term partnership with the biostatistician, preferably from the initial stage, may require repeated meetings to define questions, design experiments, perform analysis, draft manuscripts, respond to reviewer's comments and so on.

Introduction to Statistical Consulting and Collaboration

Practically, in academic medical centers:

- statistical consulting: conduct simple requested analysis, explain minor technical details; usually simple questions resolved within a short period (0.5-1 hour); may not require repeated scheduled meetings
- statistical collaboration: answers research questions which requires long-term partnership with the biostatistician, preferably from the initial stage, may require repeated meetings to define questions, design experiments, perform analysis, draft manuscripts, respond to reviewer's comments and so on.

Examples: *Biostatistics, Epidemiology, and Research Design (BERD)* at the Department of Biostatistics

Mental Health Data Science at the Department of Psychiatry and New York State Psychiatric Institute

Classroom Learning vs Real World Collaboration

Difference between classroom environment and consulting or collaboration environment:

- In courses, the subject matter and questions are well defined. Methods for solving problems are generally contained in course materials. Dealing with clear problem with a precise optimal solution in ideal settings.
- In consulting/collaboration, neither the questions nor the answers are unambiguous. Statistical collaborator's job: play a major role in the transition from a vague research idea to a reasonably quantified problem.
- In practice, many alternatives may be possible and our most difficult task may be to develop and explore these with other researchers. Achieving a reasonable compromise between goals is the objective.
- Non-statistical skills are needed (conducting a consulting session, interpersonal relations, written and oral communication skill etc).

Steps of Consulting/Collaboration

Essential Steps of Collaboration Sessions

1. identification of relevant aspects of the problem context (background)
2. definition of the scientific goals (hypothesis/assumptions; not necessarily hypothesis testing)
3. determination of the actions to be taken (proposals)
4. discussion of various aspects of the collaboration relationship and who will do what when.

Identification of the Context

Both research context and the collaborator's overall situation, e.g., ongoing study? data collection complete? design a study? other collaborators? data provenance? time constraint? similar research done before? attempts have been made for the current problem?

Tips: avoid answering any questions on statistics asked initially by the client/collaborator (e.g., should I use repeated measures ANOVA or MANOVA) because frequently the statistical test proposes may not the one most pertinent to the problem. Avoid providing "the right answer to the wrong question".

Statistical topics to explore in the session: (a) objectives of the experiment, hypothesis formulation and specification; (b) experimental plan and design; (c) implementation, data management; (d) analysis; (e) interpretation and inference.

Identification of the Context

Frequently reflect back to the collaborator his or her own emerging understanding of the client's situation.
"The 75 subjects were divided into three groups and observed under seven different conditions."

Ask probing questions: Did you randomize the assignment of treatment to subjects?

If working in a substantive field about which you know little, it is essential that the collaborator learn about this limitation. It is crucial that you know your limitations.

Determination of Goals

Issues may emerge when the statistician set out to clarify, formulate, and (perhaps) reformulate the collaborator's research questions.

For example, the collaborator's data may not appropriately address the research objectives or when a client's theory is not clearly thought through to define appropriate research objective (e.g., causation versus correlation; within-subject effect versus population effect).

Help the collaborator explore gap between the desired goals and the resources available, how to redesign the study, or which research questions can be addressed with the existing data set.

Given that the data set is appropriate, formulate clear research objectives. Ask questions such as, "What would you hope to answer by performing that analysis?", or "How will you know if you have answered that question?"

Determination of Analysis Plan

Depends on the step 1, i.e., the clearer the problem is, the clearer and more creative solution can be.

Statistician role: discuss alternatives regarding the quality of the data and the design, assumptions of various methods and how the data relate to them, purpose of the analysis (planning future studies or publication or decision making), resources available (time, money, computing, and expertise), the collaborator's understanding of statistics, and whether the collaborator will be presenting the analysis on his or her own.

Important: Communicate level of commitment and level of service, co-authorship, or the right to use the data for methods research.

Other Issues

Deal with status difference?

Constantly evaluate your own motivation and expectations: For example, choose the role of scientist (take the appropriate level of responsibility for the study's outcome) or statistical service provider only be concerned about the short-term outcomes?

Consider organizational structure, career aspirations, job responsibilities.

Easier said than done. No substitute for practice, observation of experts, and supervised consulting experiences. Prepare yourself for future consulting/collaborating roles.

Training videos on effective consulting and collaboration (ASA statistical consulting section).

Summary: Process of Statistical Consulting and Collaboration

(i) represent the problem; (ii) determine the solution strategy; (iii) execute the strategy; and (iv) evaluate the results:

1. Understand the problem and the context
2. Understand the variables and data structure
3. Exploratory Data Analysis (EDA)
4. Model specification: (a) Analytic Plan
5. Model specification: (b) Model selection
6. Assess model validity
7. Sensitivity analysis
8. Assess whether the research question was answered
9. Communicate results and conclusions effectively to the investigator in non-technical language

An iterative process.

Overview of Collaboration/Consulting Environments

Five Major Collaboration/Consulting Environments

Academics

Government

Pharmaceutical industry

Business

Technology Companies

Academic Setting

In addition to Biostatistics and Statistics departments, there are consulting divisions in other parts of university (School of Medicine, Social Sciences, Cancer Centers).

Consulting groups/divisions specialize in certain disciplines (e.g., cancer, psychiatry).

Roles: original research + providing consulting service and internal support for researchers across a broad range of disciplines + educate next generation statistical leaders, collaborators and consultants

Government

The U.S. government and state agencies are some of the main sources of employment for statisticians.

- Census Bureau: survey sampling methodology, data visualization, time series analysis, and data mining
- Food and Drug Administration (FDA): statistical reviews in the regulating process of drugs and devices (also food and cosmetics)
- National Institute of Health (NIH)
- Centers for Disease Control and Prevention (CDC)
- Environmental Protection Agency (EPA)
- Departments of Health in each state
- National Security Agency (NSA)

Join the American Statistical Association (ASA) *Health Policy Statistics Section (HPSS)* or *Government Statistics Section (GSS)* to learn more.

Pharmaceutical Industry

Drug development is a complex and lengthy process that can take 7 to 15 years for a single drug at a cost that may reach tens of millions of U.S. dollars. Three processes:

- Discovery and decision
- Preclinical studies
- Clinical studies

Discovery and decision process starts with the discovery of a new compound or of a new potential application of an existing compound: in vitro testing in test tubes; and in vivo testing on cells.

If adequate results, then enter preclinical studies stage.

Preclinical Stage

In the preclinical stage, the initial toxicology of the compound is studied in animals.

Perform initial formulation of the drug development and specific or comprehensive pharmacological studies in animals.

The evidence of potential safety and effectiveness of the drug is assessed. Provide reasonable confidence that the drug dosage will not be fatal and can be tolerated by humans. Extensive tests done on animals suggesting sufficient evidence that the drug will be of benefit to human subjects.

To proceed further, a U.S.-based company needs to file a Notice of Claimed Investigational New Drug Exemption.

Clinical Stage

Phase I: Establish the initial safety information about the effect of the drug on humans (e.g., range of acceptable dosages and the pharmacokinetics). Small study of health volunteers (4-20 subjects).

Phase II: Patients who will potentially benefit from the new drug. Test effective dose ranges and initial effects of the drug. Up to several hundred patients.

Phase III: Assessment of safety, efficacy, and optimum dosage designed with controls and treatment groups (hundreds or even several thousand patients). Successful results obtained will lead to submission of a New Drug Application (contains the results from all three stages) reviewed by the FDA.

Postmarket Activities

Phase IV: Outcomes research analysis and follow-up studies to examine the long-term effects of the drug.

Ensure all claims made by the company about the new drug can be substantiated by ”clinical evidence.” All reported adverse effects must be investigated by the company and, in some cases, the drug may need to be withdrawn from the market.

Current research using electronic health records (EHRs) and *FDA Adverse Event Reporting System (FAERS)* to investigate drug effectiveness and safety.

Statistician's Roles

- Participate in the development plan for studying a drug.
- Study design and protocol development. Randomization schemes.
- Data cleaning and database construction format.
- Analysis plan and program development for analysis.
- Report preparation. Produce tables and figures.
- Integrate clinical study results, safety and efficacy reports.
- Communication and NDA defense to the FDA review panel.
- Publication support and consultation with other company personnel.

Technology Companies

Long history of supporting statistical research in telecommunication (e.g., Bell labs and AT&T). Unique features of research in high-tech industry:

- Extremely large databases available to statisticians (network testing, package transmission and delays, recommendation systems, AB testing/personalized advertising). Many tasks are research-oriented rather than production-oriented.
- Statistical expertise: statistical tests/modeling, data mining, machine learning, big data techniques

Business Sectors

Consulting companies: market research, survey design and analysis, financial analysis, database management, private consultants, expert witness for legal cases, news media.

Join the ASA *Statistical Consulting Section* to learn more.

Case Study: Statisticians' Responses to the COVID-19 Pandemic

Rise to the Challenges of the COVID-19 Pandemic

Modeling disease transmission, study prognostic factors, design and analysis of vaccine/treatment trials, advising local health departments.

- Societal responses and behavioral changes are major tools for preventing outbreak before a vaccine is widely administered.
- Behavioral changes can be shaped by public view towards risks.
- Communicating risks with the public
  - Low mortality risk? Consider intensity and excess mortality.
  - Only reporting the low absolute risks to certain individuals (e.g., younger adults) does not convey their risks to vulnerable populations.
  - Vaccine coverage versus efficacy
  - Professor Jeffrey Morris's COVID data science blog

Examples of the *HPSS responses* and our own modeling efforts during 2020. An interview with *Dr. Dean Follman* (2021). In our department *Biostatistics in Action*.

Case Study: COVID-19 Forecast Model