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

Lecture 2

Yuanjia Wang, Ph.D.

Department of Biostatistics, Mailman School of Public Health
Columbia University

& Division of Biostatistics, New York State Psychiatric Institute

Part I: Communication in Statistical Consulting and Collaboration

Five major consulting settings: Academics; Pharmaceutical industry; Government; High-techs; Business

No matter which setting, good communication skills are required.

Verbal communication includes the ability to listen and the skill to articulate ideas orally. Writing skills are important as the report containing the analysis results must be understood by the researcher.

Structure meeting with collaborators:

Background: Projects are often based on previous studies, in which case there may be an established or accepted analysis method. Obtaining a relevant reference from the collaborator, ascertain whether the established analysis method is reasonable and applicable to the problem.

Status: What is the status of the project? If the study is in the pre-experiment or planning stage, the statistician's contribution can be important (ensuring the planned experiment will produce reliable data for the subsequent analysis). If the data have already been collected, direct questions towards the collection process:

- What design was used? How reliable? Was the experiment performed in a controlled environment? Randomization? Comparison group? Enough sample size to support the objectives?

Introduction to Communication

Aims: What are the aims and hypotheses associated with the study? Are the objectives commensurate with the results that can be obtained from a statistical analysis?

Hypotheses may need to be reformulated for the statistical analysis to provide valid conclusions.

Understands the distinction between causality and association. Confirmatory versus exploratory.

Our expectation for the collaborator: articulate the project's importance, motivation and scientific premise.

Communicate what the collaborator expects from us: identify the problem's statistical aspects. Recognize non-statistical questions that statisticians cannot answer.

Introduction to Communication

The most important contribution a statistician can make in any project is to help develop a clear specification of the project's goals.

Seek to gain agreement among the collaborators on the project's primary aims and where the real problems lie. Sometimes, what may seem like minor differences in opinions among collaborators may, in fact, not be.

Get priorities set. What is the single most important objective of this project?

See HEAL case study.

Technical Knowledge of the collaborator

Part of the statistician's role is to educate the collaborator. Explanations given in the context of the project; provide an interpretation of the outcome and purpose of a statistical procedure, not the mathematical details. Be patient, but avoid getting stuck on details.

Level of sophistication: The statistical methods employed for analysis need to be appropriate for the problem and may require introducing more sophisticated approaches. However, the collaborator needs to be able to interpret the analysis results irrespective of the level of statistical sophistication.

Example: methods for missing data (last observation carried forward; linear mixed effects models; inverse probability weighting; multiple imputation). HEAL example.

Overall and Specific Issues to Communicate

Objectives: exploratory versus confirmatory. Ensure the experimental design will provide statistically valid results; address issues related to the design (sample size, randomization, control group, implementation); plan for exploratory analyses.

Methodology: to ensure that the statistical procedure is applied appropriately, specific issues will need to be addressed, such as:

- What is the data type of each variable?
- Are there outliers or missing values?
- Identify important variables for the study before the analysis.

Post-experiment: consider the intended use of the experiment results (decision making; adopt the intervention?). For example, does the study's outcome depend critically on obtaining significance for a particular hypothesis? What are the consequences of getting a non-significant result? Include summary of evidence (confidence intervals) other than p-value.

Statistical Significance vs Clinical Significance

- Statistical significance depends on the sample size. In a large study (e.g., big data), a clinically unimportant difference can be statistically significant ($p < 0.05$).
- If an intervention is resource intensive and has potential side effects, it may decide not to intervene when there is a lack of evidence on its true benefit.

Statistical Significance vs Clinical Significance

- But, absence of evidence is not evidence of absence[1,2].
  - "There is no evidence to suggest" that hospitalizing compared with not hospitalizing patients with acute shortness of breath reduces mortality.
  - "There is no evidence to suggest" is ambiguous: hypothesis not tested; tested but inconclusive due to lack of power; proven to have no benefit; or a close call with risk outweighs the benefit in some patients.
  - Deciding not to act may have detrimental consequences (e.g., masking during COVID-19).
  - Examine the magnitude of effect size (confidence intervals), seek evidence in other (observational) studies, consider the potential harm of not intervening, and prevalence of exposure.

Check out the complete list of BMJ statistics notes.

[1] BMJ Statistics Notes.

[2] EBM's Six Dangerous Words

Specific Issues to Communicate

Data management: how will data be collected, stored and provided in suitable format to the study statistician.

Data analysis error checking: The collaborator needs to be aware that the initial analysis stage will involve checking the data. The collaborator will need to provide any corrections.

Statistical analysis: Agree on the method used and the details of performing the computations.

Report writing: Whether there is an expectation of manuscript or presentation.

Time frame: a realistic time frame to allow for performing the analysis and completing the written report for the project.

Other Aspects of Communication

Tips on good communication:

- Learn something about the subject under consideration.
- Ask questions when something isn't clear (rephrase). Simple questions about seemingly minor details often bring misunderstandings of important issues to light.
- Be prepared to interrupt (politely) and redirect the collaborator toward relevant issues as necessary.
- Write memoranda that give your understanding of the problem; often bring to light still new avenues for improvement.

Other Aspects of Communication

Soft skills can be important when collaborating:

- Probing questions to understand the confusion
- For scientists in other fields or policymakers, it is not necessarily helpful to repeat the same understanding in our statistical/mathematical language. Focus on easier interpretations.

Developing good communication skills is an evolving process!

An Example of Miscommunication

FDA's debacle on miscommunicating effect of convalescent plasma (EUA issued on August 23, 2020)

F.D.A. 'Grossly Misrepresented' Blood Plasma Data, Scientists Say

Many experts — including a scientist who worked on the Mayo Clinic study — were bewildered about where a key statistic came from.

Mr. Trump called it a "tremendous" number. His health and human services secretary, Alex M. Azar II, a former pharmaceutical executive, said, "I don't want you to gloss over this number." And Dr. Stephen M. Hahn, the commissioner of the Food and Drug Administration, said 35 out of 100 Covid-19 patients "would have been saved because of the administration of plasma."

Dr. Hahn did not respond to repeated requests for clarification on his comments.

On Sunday night, an agency spokeswoman posted a chart on Twitter claiming that plasma "has shown to be beneficial" for 35 percent of patients — neglecting to mention that the figure was based on a subset of

Does evidence Support the Claim?

Mayo Clinic Study (observational study)

Effect of Convalescent Plasma on Mortality among Hospitalized Patients with COVID-19: Initial Three-Month Experience

59 **Objective:** To explore potential signals of efficacy of COVID-19 convalescent plasma.

60 **Design:** Open-label, Expanded Access Program (EAP) for the treatment of COVID-19
61 patients with human convalescent plasma.

62 **Setting:** Multicenter, including 2,807 acute care facilities in the US and territories.

63 **Participants:** Adult participants enrolled and transfused under the purview of the US
64 Convalescent Plasma EAP program between April 4 and July 4, 2020 who were
65 hospitalized with (or at risk of) severe or life threatening acute COVID-19 respiratory
66 syndrome.

72 **Results:** The 35,322 transfused patients had heterogeneous demographic and clinical
73 characteristics. This cohort included a high proportion of critically-ill patients, with 52.3%

....

84 was also observed in thirty-day mortality (p=0.021). The pooled relative risk of mortality
85 among patients transfused with high antibody level plasma units was 0.65 [0.47-0.92]
86 for 7 days and 0.77 [0.63-0.94] for 30 days compared to low antibody level plasma

The Issue: Relative risk vs Absolute Risk

Reported statistics in the Mayo Clinic study pre-print:

Supplemental Table 3. Crude Mortality (7 and 30 day) of patients with IgG transfused with COVID-10 Convalescent Plasma.

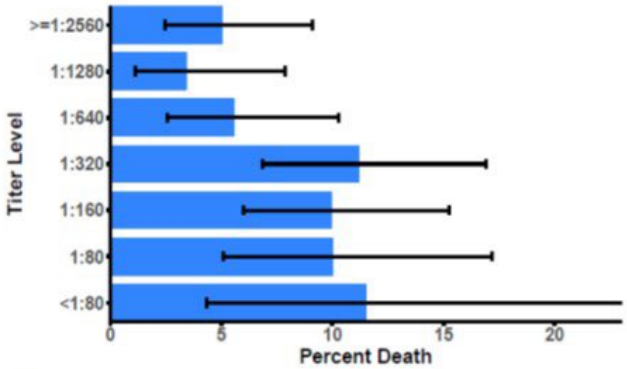
	Seven-day Mortality				Thirty-day Mortality			
	Sample, No	Events, No	Estimate, 95% CI	P-value	Sample, No	Events, No	Estimate, 95% CI	P-value
Ortho IgG				0.0483				0.0208
Low	561	77	13.7% (11.1%, 16.8%)		561	166	29.6% (26.0%, 33.5%)	
Medium	2,006	233	11.6% (10.3%, 13.1%)		2,006	549	27.4% (25.5%, 29.4%)	
High	515	46	8.9% (6.8%, 11.7%)		515	115	22.3% (18.9%, 26.1%)	

Data showed a relative risk reduction of 35%, NOT an absolute risk difference of 35% (on a subgroup of $n = 1076$). Cannot say "35 out of 100 patients will be saved", while it is about 5 out of 100 patients will be saved (13.7%-8.9%=4.8%)!

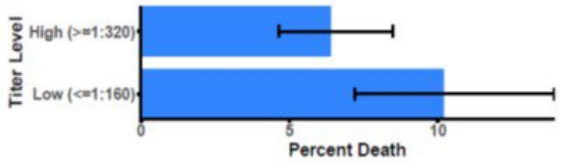
FDA's clarification

COVID-19 Convalescent Plasma Reduction in Death at 7 Days

Non-intubated patients treated within 72 h age 80 or less (n=1018)



Statistically significant 37% reduction in mortality in those treated with high titer convalescent plasma (p=.03)



High titer corresponds approximately to Ortho VITROS S/C level ≥ 12

www.fda.gov

1

A relative risk reduction of 37% (on a subgroup of patients), NOT an absolute risk difference of 35%.

FDA chief apologizes for overstating plasma effect on virus

Food and Drug Administration Commissioner Stephen Hahn is apologizing for overstating the life-saving benefits of using convalescent plasma to treat COVID-19 patients

By **MATTHEW PERRONE and DEB RIECHMANN Associated Press**

August 25, 2020, 1:58 PM • 5 min read

Further Evidence on Convalescent Plasma for Treating Severe Covid[3]

ORIGINAL ARTICLE

A Randomized Trial of Convalescent Plasma in Covid-19 Severe Pneumonia

- No significant difference between the convalescent plasma group and the placebo group in the distribution of ordinal clinical outcomes (ranging from recovery to death): odds ratio=0.83 (95% CI: 0.52 to 1.35; $p = 0.46$).
- Overall mortality 10.96% in the convalescent plasma group and 11.43% in the placebo group, risk difference $= -0.46\%$ (95% CI, -7.8% to 6.8%).
- Did not examine high titer.

[3] Simonovich et al. (2020). NEJM.

Further Evidence on Convalescent Plasma for Treating Severe Covid[4]

ORIGINAL ARTICLE

Early High-Titer Plasma Therapy to Prevent Severe Covid-19 in Older Adults

Romina Libster, M.D., Gonzalo Pérez Marc, M.D., Diego Wappner, M.D., Silvina Coviello, M.S., Alejandra Bianchi, Virginia Braem, Ignacio Esteban, M.D., Mauricio T. Caballero, M.D., Cristian Wood, M.D., Mabel Berrueta, M.D., Aníbal Rondan, M.D., Gabriela Lescano, M.D., et al., for the Fundación INFANT–COVID-19 Group*

RESULTS

A total of 160 patients underwent randomization. In the intention-to-treat population, severe respiratory disease developed in 13 of 80 patients (16%) who received convalescent plasma and 25 of 80 patients (31%) who received placebo (relative risk, 0.52; 95% confidence interval [CI], 0.29 to 0.94; P=0.03), with a relative risk reduction of 48%. A modified intention-to-treat analysis that excluded 6 patients who had a primary end-point event before infusion of convalescent plasma or placebo showed a larger effect size (relative risk, 0.40; 95% CI, 0.20 to 0.81). No solicited adverse events were observed.

[4] Libster et al. (2021). NEJM.

FDA Re-issued the EUA on Convalescent Plasma

Recommending the treatment with high-titer in a subgroup[5]:

Following the August 23, 2020, authorization, additional studies, including randomized, controlled trials, have provided data to further inform the safety and efficacy of COVID-19 convalescent plasma, and further characterize product attributes and patient populations for its use. For the December 28, 2021 authorization, FDA reviewed additional studies including several randomized controlled trials and observational studies, which reported on the use of COVID-19 convalescent plasma in both the inpatient and outpatient settings. Based on assessment of these data, transfusion of COVID-19 convalescent plasma in hospitalized immunocompetent patients is unlikely to be associated with clinical benefit and the known and potential benefits do not outweigh the known and potential risks in this population. However, evidence supports a potential clinical benefit of transfusion of COVID-19 convalescent plasma with high titers of anti-SARS-CoV-2 antibodies to treat COVID-19 in patients with immunosuppressive disease or receiving immunosuppressive treatment. Based on the totality of the scientific evidence available, it is reasonable to believe that the known and potential benefits of COVID-19 convalescent plasma with high titers of anti-SARS-CoV-2 antibodies, when used under the conditions described in this authorization, outweigh its known and potential risks for

[5] FDA EUA on Convalescent Plasma

An Example of Difficulties in Communicating Risks

Why most people who now die with Covid in England have had a vaccination

David Spiegelhalter and Anthony Masters

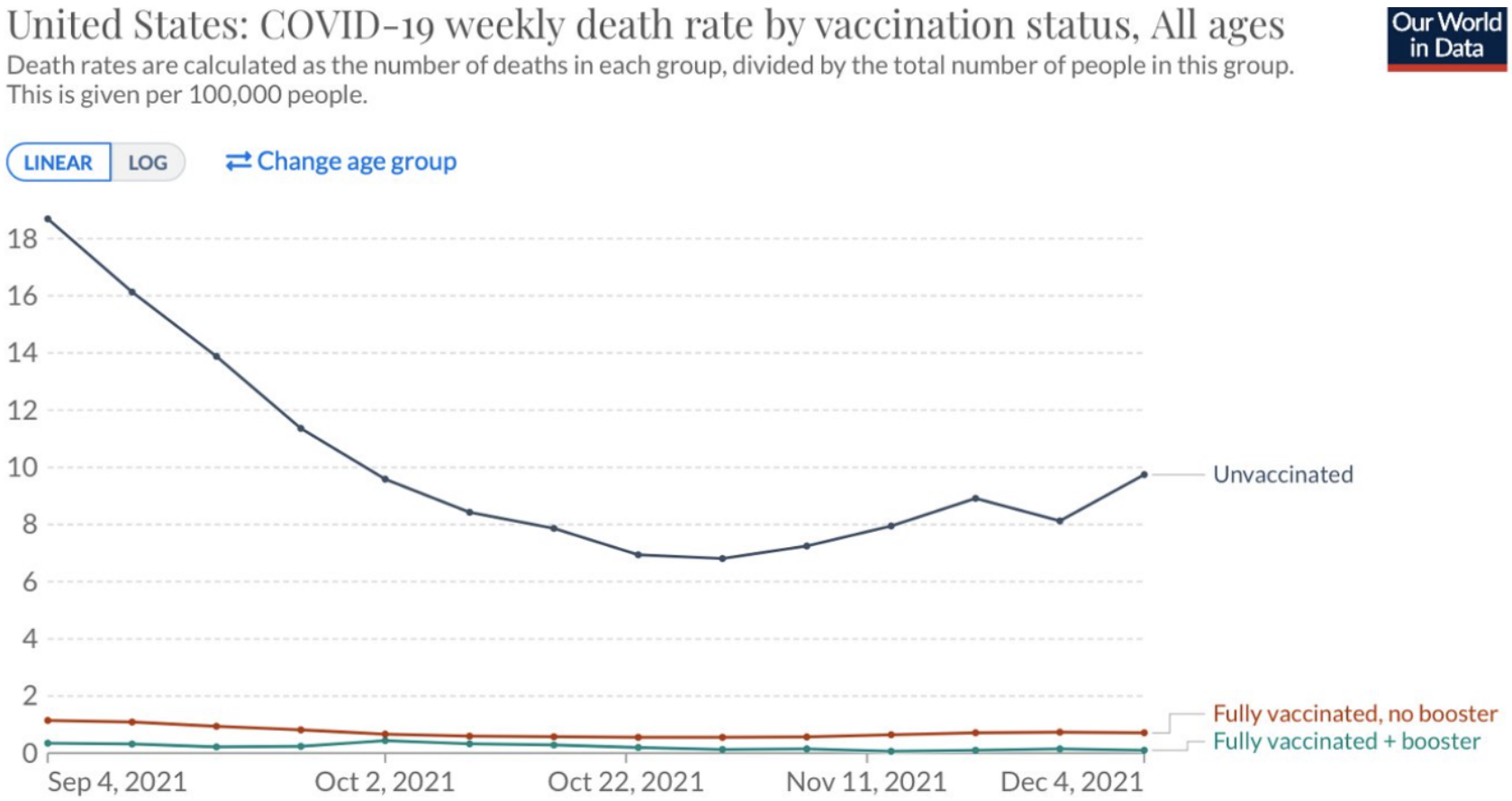
Don't think of this as a bad sign, it's exactly what's expected from an effective but imperfect jab

6

- As of 6/25/2021, in Public Health England's technical briefing 43% (50 of 117) who have died after catching the new strain had BOTH vaccinations, with the majority (60%) having received at least one dose.
- Vaccine does not work? No! Expected from an effective but imperfect vaccine. Considering a high vaccination rate (e.g., 93% of those aged 65-69). Had the vaccine been ineffective, the percentage would have been even higher (close to the vaccination coverage rate of 93%).

[6] Guardian COVID headline

Better Representation of Data



Case Study 1: Healing Emotion After Loss (HEAL) Trial

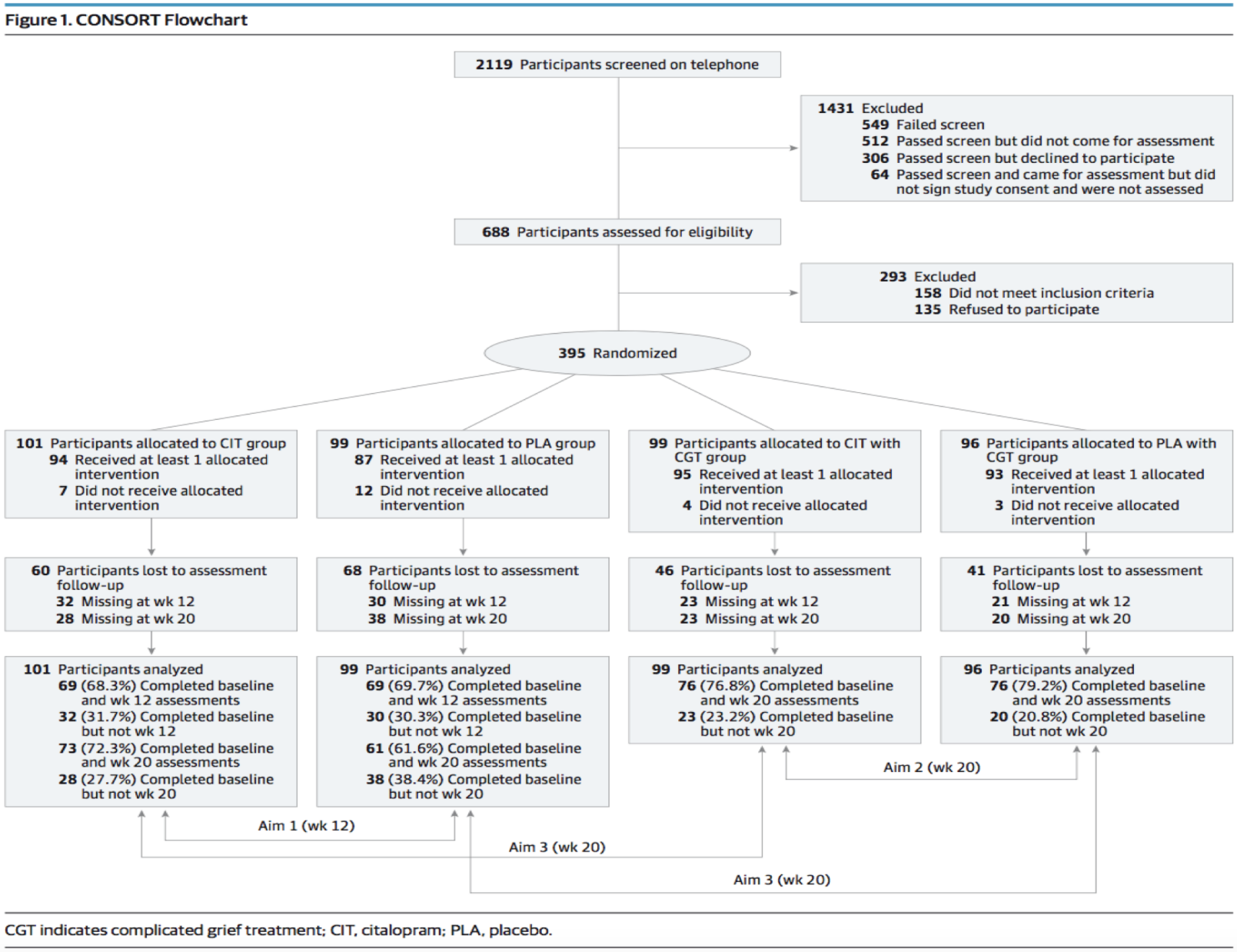
Analysis Plan, Data Collection and Measurements

Statistical Analysis Plan

- formulate goals precisely and quantify goals
- specify variables precisely
- specify classes of candidate models
- describe blocking, the realm of generalizability
- specify estimation procedure (precision, bias); consider alternative means of exposing true uncertainty
- describe plan for data gathering in detail; describe randomization procedure in detail
- describe data documentation procedures in detail
- specify ways that data will be checked
- specify how data will be explored, summarized and analyzed

Reporting guidelines: CONSORT for randomized trials; STROBE for observational studies; TRIPOD for prediction models.

Consort and data acquisition process: participant recruitment, screening, consent, and data collection.



Data Collection

Data errors and problems may include entry errors, missing values, duplicates, outliers, and data inconsistencies and discrepancies, any of which may affect the validity, reproducibility, and, thus, the quality of studies.

In large-scale studies, budgets may be allocated for personnel with distinct roles, including principal investigators, study coordinators, data collectors, database managers, and statisticians.

Statisticians may need to play multiple roles oversee quality control over the whole study flow (study design, data acquisition, preprocessing, and analysis).

Example: status tracking in HEAL

Participant Recruitment, Screening, Consent

Multi-site study: Examine subject characteristics in each site to determine whether a study site effect was present (important to reduce selection bias, especially for epidemiological studies, need to collect relevant covariates for adjustment; for RCT has implications on generalizability and test of moderation).

Check whether ineligible participants were recruited (e.g., for studies of eating disorders subjects BMI$\leq$17.5).
HEAL inclusion criteria

1. Age 18-95
2. ICG$\geq$30 and Clinical confirmation of CG as the most important clinical problem

HEAL exclusion criteria

- Lifetime bipolar, psychotic disorder, or dementia
- Current substance use disorder
- MoCA score $\leq$21
- Active homicidal/suicidal ideation (acute risk)
- Prior failure of citalopram or escitalopram
- Concurrent psychotherapy OR antidepressant use

Researcher Training

In HEAL, therapists and independent evaluators at each study center were initially trained by experienced principal investigators with hands-on practice of measurements and treatment of CG. A treatment manual of procedures for treatment and measurements was provided.

Consensus calls were held to discuss and share field experiences regarding participant recruitment. Training for independent evaluators (IEs) includes multiple IEs rating training subject videos.

Status tracking data collected (e.g., 224 - completed 16 sessions in 24 weeks, 13 sessions in 20 weeks; started 9/23/10, finished 2/28/11)

Statisticians: analyze the reliability of IE ratings (e.g., kappa coefficient for dichotomous outcomes and intra-class correlation [ICC] for continuous outcomes).

Data Collection

Actual data collection required careful attention to the administration of the study questionnaire(s) by interviewers and completion by participants (clinician-administered interview vs patient self-report).

An active research area: analysis of patient-reported outcomes. NIH Patient-Reported Outcomes Measurement Information System (PROMIS) program. Involve psychometric analysis (internal/external validity).

Variable coding and a data code book. The data codebook describes the content of the dataset and includes original survey questions and skip patterns; variable definitions; variable name, type, label, and values; code for missing data; and other characteristics of each variable.

Quality Control in Data Preprocessing

Focus on clinical data. Specialized QC steps for genomic and neuroimaging data.

Data Review. As a front-end process, trained researchers' data review of forms and questionnaires is critical to reduce errors and evaluate data integrity. Study-specific ranges, skip patterns, and dependence among variables (not necessarily statistical dependence). Example: Columbia Suicide Severity Scale.

Data Entry and Verification. Random checking. Export data from online systems to analytic forms (SAS, SPSS, CSV, etc). Medication use forms are error-prone.

Data Cleaning. Outlier detection, logic, and consistency checks. Cross-checks on the same variable measured on repeated occasions using different questions. For example, in electronic health records, a subject documented as being diagnosed with type 2 diabetes (T2D) according to ICD coding but with no T2D medications or insulin treatment may not be a true T2D patient.

Missing Data Checks

Missing data is a common issue for most studies. How to handle missing data is an active area of research (more on this later).

At the data collection stage, check for incomplete data sent by study coordinators, data entry errors, or interruptions. HEAL example:

Table 1: Assessment Instrument Timetable

Week Treatment Visit[c]	**MA1**	**1** C/P	**2** C/P	**3** C	**4** C/P	**5** C	**6** C/P	**7** C	**8** C/P	**9** C	**10** C	**11** C	**12** C/P	**13** C	**14** C	**15** C	**16** C/P	**20** C	**MA2** C/P	**Follow up** C/P
Interview																				
SCID-IV	X																			
Struct. Clinical Interview for CG	X	X											X				X	X	X	X
CSS [d]	X				X				X				X				X	X	X	X
Clinician Suicide Assessment Checklist		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		
SIGH-A	X				X				X				X				X	X	X	X
CGI-I and S (Clinician Rated)[e]		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		
CGI-I & S (IE Rated)[f]	X				X				X				X				X	X	X	X
Acceptability Questionnaire (Clinician Rated)		X																		
Self Report																				
ICG	X				X				X				X				X	X	X	X
GRAQ	X								X				X				X	X	X	X
TBQ	X				X				X				X				X	X	X	X
QIDS-SR [e]	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
IES	X				X				X				X				X	X	X	X
CSQ									X										X	X

Exploratory Data Analysis

Exploratory Data Analysis (EDA)

Goals of EDA:

- describing features of distributions to answer substantive questions (e.g., about group differences)
- describing features of distributions to investigate underlying formal assumptions (e.g., t-test)
- looking for patterns in data (e.g., clustering)
- from univariate to multiviarate and high-dimensional (e.g., PCA).

Simple example: compare two group means using parallel boxplots to compare the distribution of a quantitative response variable between two independent groups

Graphical Display

- Distribution: Histogram, Boxplot, Forest plot for meta-analysis
- Y versus X: Scatter Plot, Spaghetti Plot, Heatmap
- Multivariate: Contour Plot, Trellis Plot, Clustering, Correlation Matrix
- Diagnostic: Q-Q Plot, Residual-vs.-Fitted
- Recent advancements: Network plots, Multidimensional scaling, Neighbor Embedding (t-SNE)

Examples of EDA

Spaghetti plot

Subject-specific trajectories

Network plot

Pre-processing of electronic health records

t-SNE plot

Case Study 2: RAISE

NIMH Landmark Study, Featured Story (75th Anniversary, 2023)

RAISE-ing the Standard of Care for Schizophrenia: The Rapid Adoption of Coordinated Specialty Care in the United States

The Recovery After an Initial Schizophrenia Episode research initiative, launched by the National Institute of Mental Health (NIMH) to test the effectiveness of coordinated specialty care to treat first-episode psychosis, has transformed the mental health landscape in the United States and helped thousands of people with schizophrenia achieve better outcomes.

Read the 75th Anniversary Feature Story

Case Study 2: RAISE

Recovery After an Initial Schizophrenic Episode (RAISE)

RAISE-IES Duration of Untreated Psychosis Study

FINAL DELIVERABLES:

- Case Study Reports
- Statistical Analysis
- Database
- Recruitment and Engagement Materials
- Process and Procedure Manuals
- Barriers Summary Reports
- Financial Model Summary Reports
- Clinical Impact Reports

December 31st, 2013