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

Lecture 3

Yuanjia Wang, Ph.D.

Department of Biostatistics, Mailman School of Public Health
Columbia University

& Division of Biostatistics, New York State Psychiatric Institute

THE DEPARTMENT OF BIOSTATISTICS

Columbia University MAILMAN SCHOOL OF PUBLIC HEALTH

Statistical Credibility of Experiments

Quality of experimental results depends on the scientific plausibility (conceptual model/causal model/theoretical model) of how the results could have occurred and statistical credibility of the observed results.

Scientific plausibility assesses the scientific underpinning of the proposed conceptual model of interest.

Statistical credibility attempts to assess the chance that the observed result is ”close” to the true underlying phenomenon (both sampling and non-sampling variability).

Scientific Plausibility

Scientific conceptual model plausibility depends on

1. the completeness of the description of the conceptual model of the experimental process (all steps and not just one ”big black box”)
2. quantify the model and proper measurements of the important parameters
3. the logic of the conceptual model (e.g., internal consistency and no miracles) and the parsimony of its assumptions and steps
4. the support which accepted past evidence (knowledge) provides for the conceptual model

The strength of the conceptual model plausibility is not independent of the observed evidence. The model plausibility may be judged as low if the observed evidence does not support the conceptual model.

Strength of Statistical Credibility

Statistical credibility/strength relates to

1. the care utilized in the formulation of the problem of interest, the design of the experiment, and its implementation
2. the attention given to controlling the variability (e.g., prescreening, blocking, covariates, and sufficient sample size)
3. the protection taken against possible sources of bias (e.g., randomization, double-blind, reduce bias in missingness)
4. the basis for the statistical inferences (e.g., analysis procedures, control of multiplicities)

The strength of the statistical credibility is independent of the observed results. From the statistical credibility viewpoint an experiment can be successful even if its results do not show support for the hypothesized treatment effect.

Three Types of Studies

- Preliminary study/pilot study: a small-scale test to provide insight on the feasibility of a part of the proposed full-scale experiment.
- Confirmatory study: a formal attempt at "independently" validating a result/hypothesis (typically a result of a previous pilot experiment) such that it has a high degree of statistical credibility.
  - Focused objectives (hypothesis)
  - Detailed protocol documenting designs of the experiment
  - Performs pre-specified analyses (can be confirmatory or exploratory nature)
- Exploratory study: a guided search for substantial evidence or indication of a treatment effect (e.g., subgroup analysis, test for moderation effect, mediation effect).

Research Design Considerations

Internal validity: the linkage between what is observed and how it can be interpreted. Research design determines internal validity.

External validity: relies on the synthesis of research design with external assumptions that extend its scope. Has implications on the interpretation (e.g., local population versus extended population).

Need to consider sampling process and measurement process.

- Coverage: the extent to which all possible types of subject are included or eligible for inclusion
- Sampling bias/error: the extent to which subjects are different with respect to the true expected values for phenomena under study
- Measurement error/response error: the extent to which multiple observations of the same phenomenon for the same subject differ.

Measurement Process

The measurement process: the manner by which data are obtained and the form in which they are expressed.

Relevance of measurements: provides information on the underlying construct.

Reliability: yields consistent values for the same phenomena in the sense of minimal error.

- Precision: multiple observations of the same phenomenon for the same subject by the same observer are similar.
- Objectivity: multiple observations of the same phenomenon for the same subject by different observers are similar. training observers/raters
- Constancy: the extent to which the phenomenon under study is stable in value within the same subject (i.e., tends not to fluctuate due to intrinsic biological variability, etc).
- Congruence: multiple measures of multiple aspects for the same phenomenon within the same subject are in general agreement.

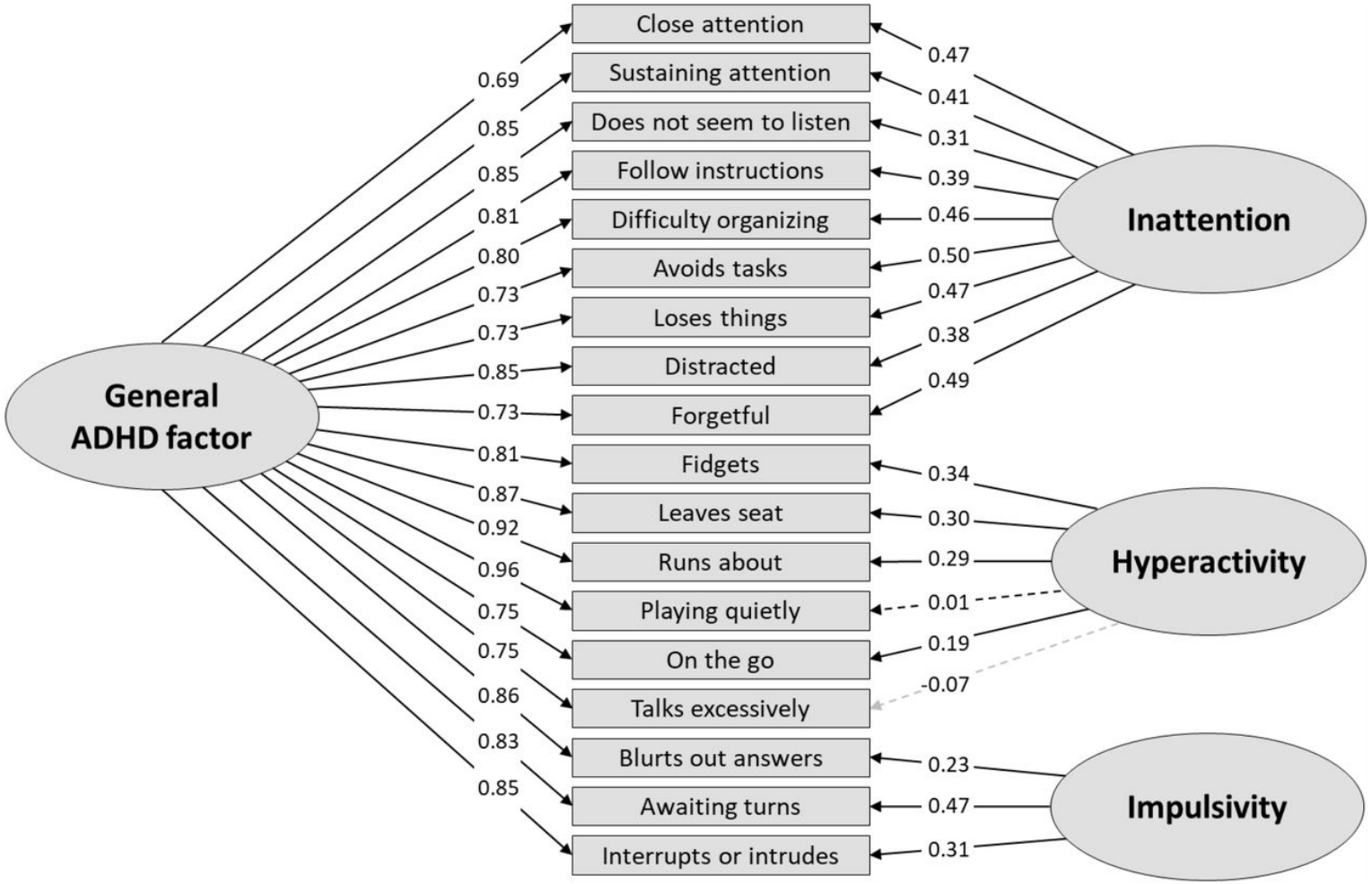
Validity of Measurement Process

A measurement process is considered valid if it is relevant, reliable, and yields values that provide a convincing basis for evaluating the research questions.

- face validity: based on intuitive grounds, background knowledge
- construct validity: based on substantive theoretical ground
- consensual validity: agreement among experts
- convergent validity: association with other measures that are accepted as valid
- predictive validity: association with correct decision about the research objectives of the study

Validity of Measurement Process

Figure. Bi-factor model for the ADHD rating scale[1]



[1] Arildskov et al. (2022)

Examples of Research Designs

- Observational Studies
  - Case-control
  - Cross-sectional studies
  - Longitudinal studies
- Sample Surveys
- Clinical Trials
- Designed Experiments/Mechanistic Experiments

Examples of Research Designs

In an observational study, the investigator generally takes a passive role and simply observes a particular cohort of subjects and collects the responses.
Examples: natural history study of disease; epidemiological study of health risk factors; case-control retrospective design; prospective cohort study.
Challenges: confounding bias.

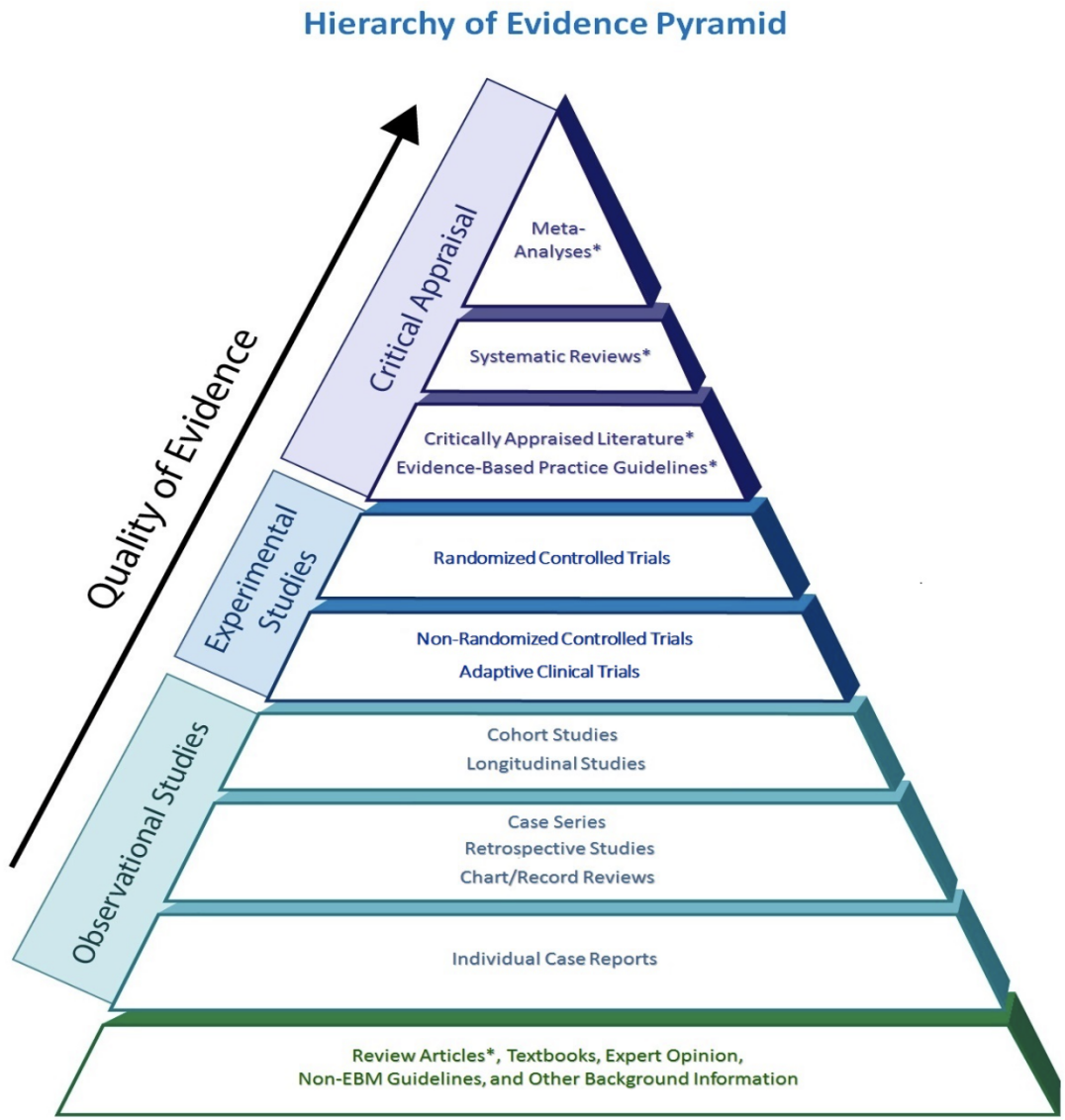
Survey sampling: take samples to infer characteristics about a population (simple random sample, stratification, cluster, multistage etc).
Challenges: maximize the response rate and improve the quality of the data collected.

Examples of Research Designs

Clinical Trials: used to test intervention effect.
Components of phase III trials: control group, randomization, written protocol. High internal validity.
Challenges: may lack generalizability/external validity.

Designed Experiments: biological experiments, animal studies, and cell studies.
Principles: control, randomization, replication. Randomized complete block design, split-plot, incomplete block, factorial design.
Challenges: standardization of research protocol; control variability; batch effects.

Pyramid of Evidence Hierarchy for Evidence-Based Medicine (EBM)[2]



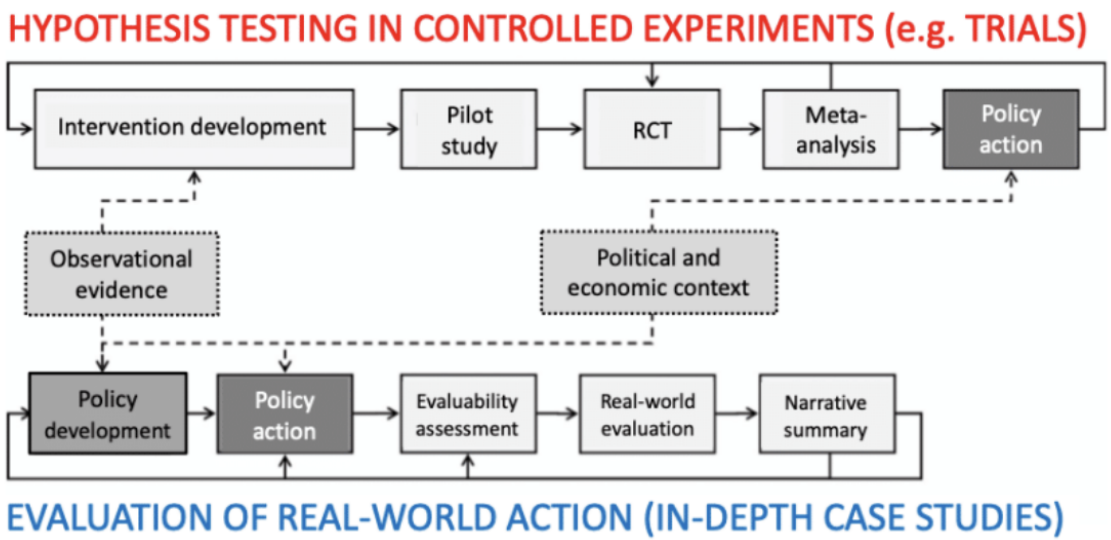
[2] Wagoner, B. et al. (2004). http://library.downstate.edu/EBM2/2100.htm

Evidence-based Medicine+ (EMB+)

EBM+: Mechanistic experiments, evidence other than RCTs[3]

Table 2	A suggested hierarchy of evidence for mechanistic evidence (drawing partly on previous publications)[12 14]
Level 1 (strongest)	Necessary and sufficient conditions for causality (eg, multiple features of the causal chain) supported by multiple independent studies, confirmed by multiple independent research groups using accepted best research methods. No high-quality disconfirming studies found.
Level 2	Indicators of causality (eg, more than one feature of the causal chain) supported, especially if by more than one method and confirmed by independent studies. No high-quality disconfirming studies found.
Level 3	Suggestion of causality (eg, one feature of the causal chain) supported by multiple independent studies. No high-quality disconfirming studies found.
Level 4	Sparse evidence supporting feature(s) of the causal chain. Disconfirming studies found but these are not definitive.
Level 5 (weakest)	No supporting studies. High-quality disconfirming studies found.

Combine trials with RWD (Greenhalgh et al. 2022):



Using Epidemiological Data and Mechanistic Study to Test SARS-CoV-2 New Variant of Concern (VOC)

Yuanjia Wang, Department of Biostatistics, Columbia University

Using Epidemiological Data and Mechanistic Study to Test SARS-CoV-2 New Variant of Concern (VOC)

Several new variants of concern (VOCs) were discovered in the UK (alpha), South Africa (beta), and Brazil (gamma) up to early 2021.

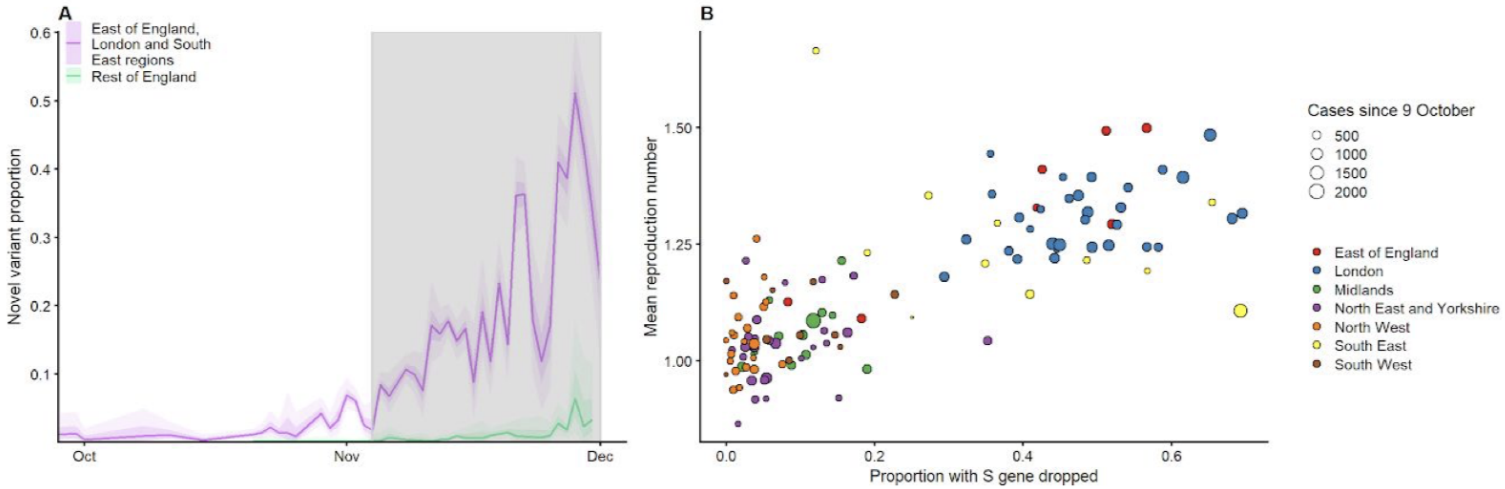
- SARS-CoV-2 genome is constantly mutating: does it matter?
- >12,000 mutations have been identified
- Most do not qualify as a new 'strain'
- Most have no proven implications for human transmission or pathogenesis

Why are new variants alpha, beta, gamma, delta, and omicron designated as VOCs? Can we prove that they are more transmissible, have more (or less) pathogenesis, or weaken the effect of vaccines?

Alpha Variant B.1.1.7.

Study: *Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England*. (2020). Pre-print.

- BY117 was discovered in November in the UK.
- Higher case rate correlates with the prevalence of the variant in the regions across the UK.



- Correlation $\neq$ Causation!

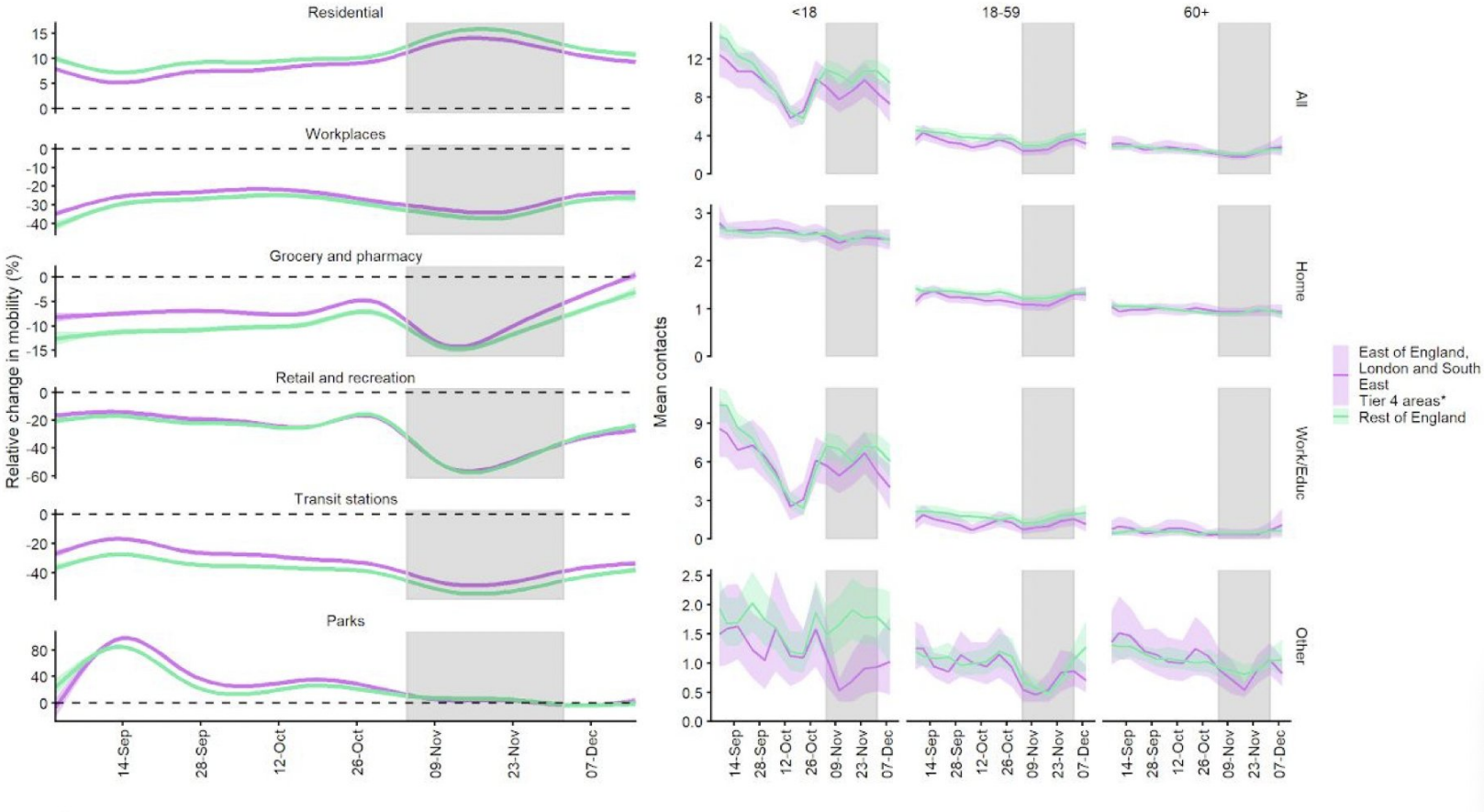
Potential Confounders

Founder effect:

- Variant of concern (VOC) happens to transmit at a more densely connected network of hosts, superspreader events.
- VOC happens to transmit at regions of high viral activity where social distancing or other interventions are not practiced.

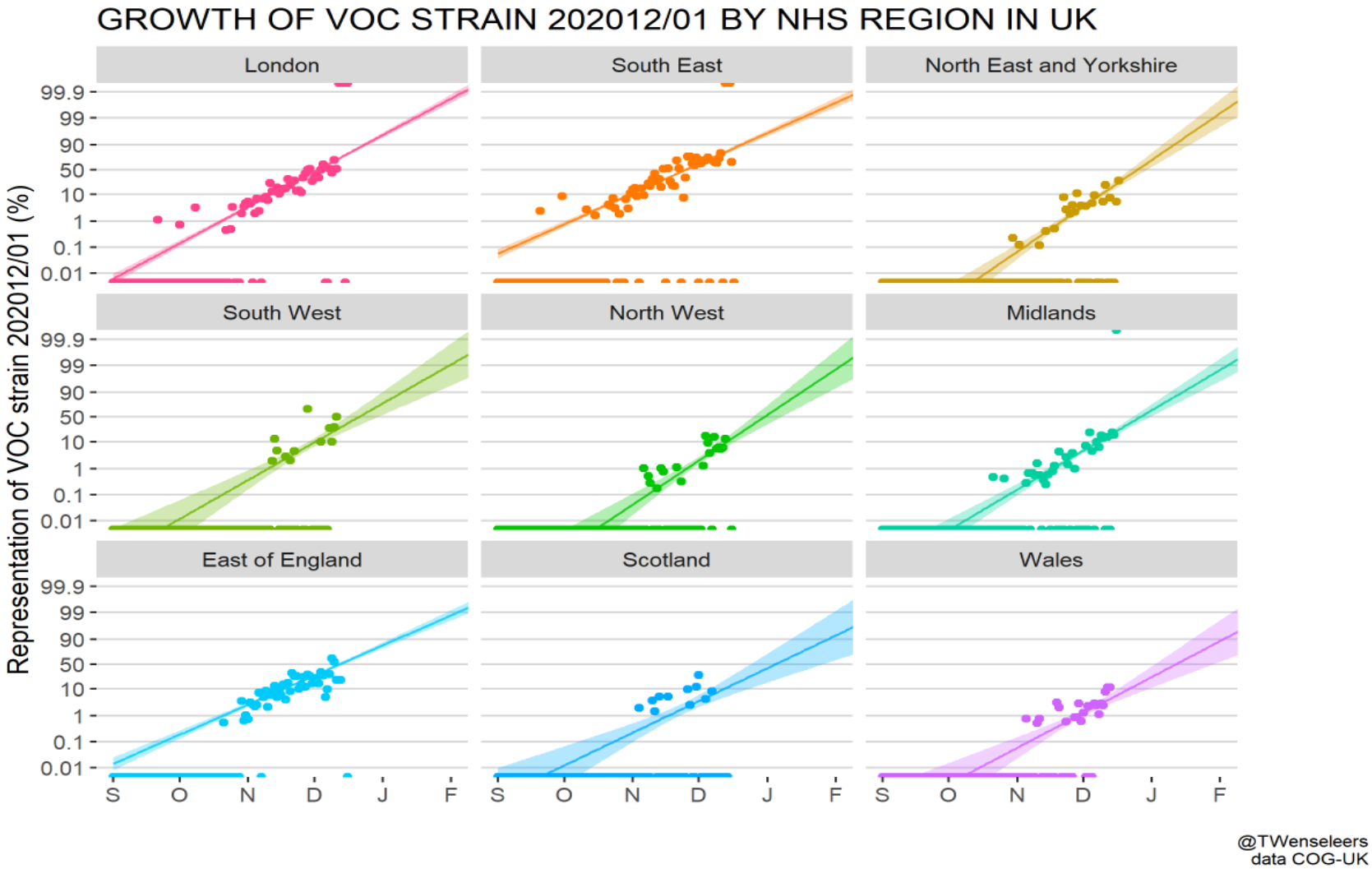
Potential Confounders

No increased mobility or social contacts:



Potential Confounders

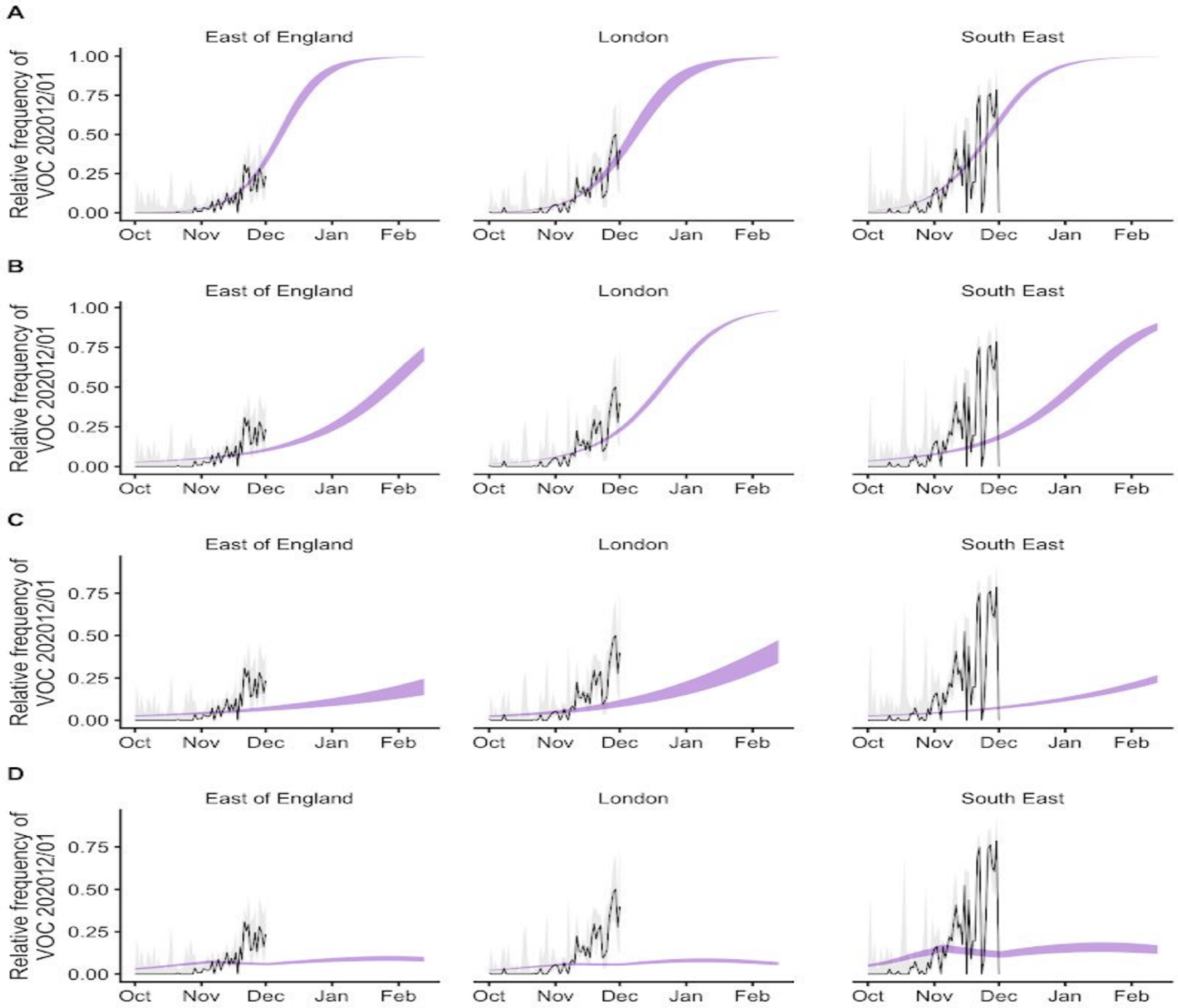
Also, growth rates similar across regions



It is unlikely to be the founder effect. VOC causes true

Biological Mechanism of Increased Infection Rate

A: increased transmissibility; B: immune escape; C: increased susceptibility among children; D: shorter generation time



Increased Transmissiblity

- Transmissiblity increases by 56% (R_t increases between 0.39 to 0.93)
- Contact tracing data: secondary attack rate of the VOC increased to 15% compared to 10% for other variants in the same period (10/5/2020 to 12/1/2020).
- Limitation: not on the same cohort.

Explore Biological Mechanism of Increased Transmissibility

- Indirect evidence: PCR C_t value decreases by around 2 for the alpha variant (fewer repeats of genome reads needed to detect a variant).
- Viral load inferred from PCR C_t value suggests an increase of 0.5 in median log10 viral loads.
- Biological experiment: measure shedding of infectious virus by patients infected with VOC and compare with other variants.

Lessons Learned from VOC B.117

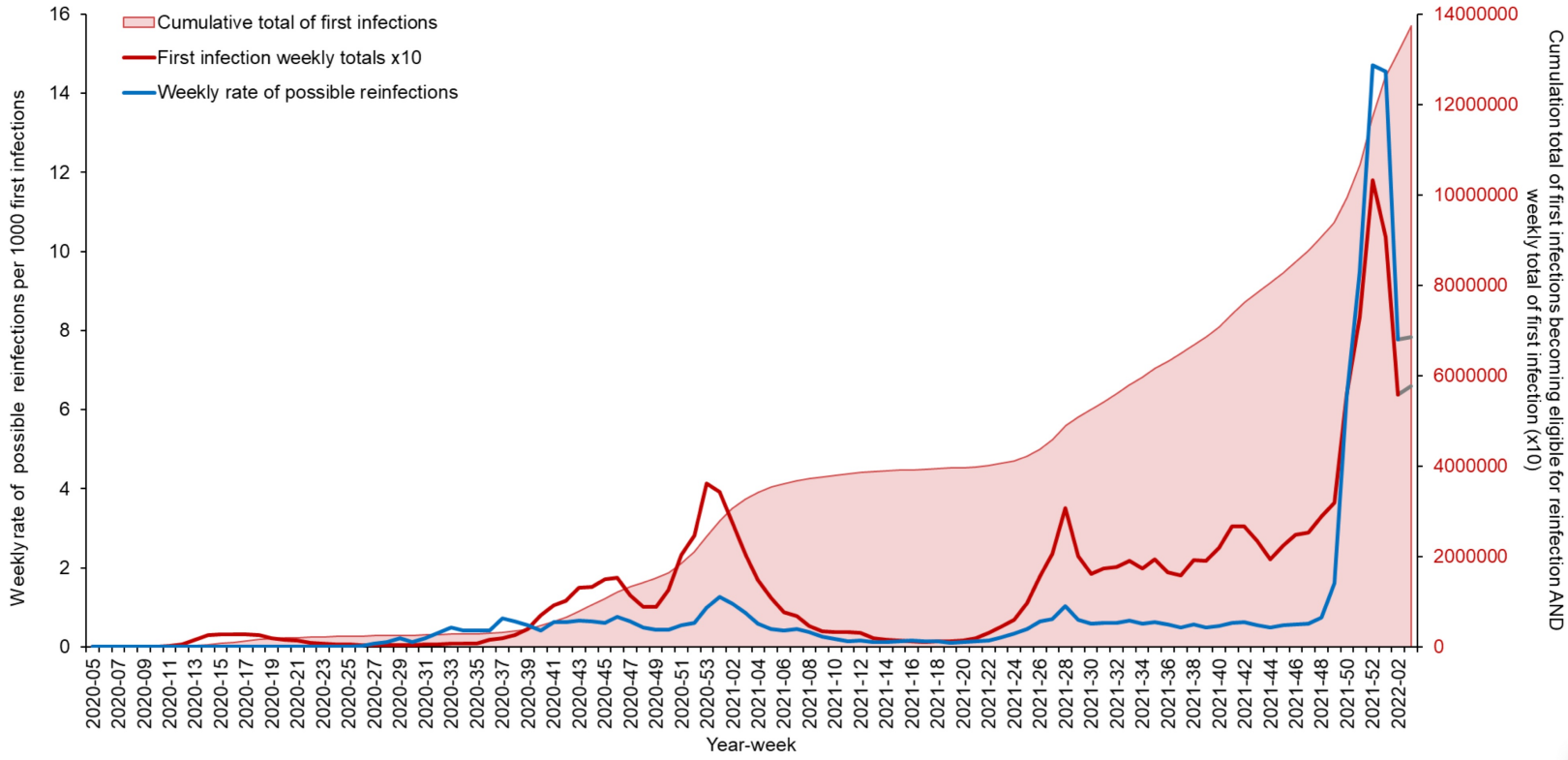
- Warned the rest of the world. Give time to others to prepare and react.
- Real-time surveillance (sequencing), testing, and contact tracing data are valuable (no good system in the US, but recently ramped up)!
- Use epidemiological data, biological experiments, mechanistic models, and evidence totality to draw inference.
- Rapid rollout of vaccines is critical.

Other VOCs? Immune escape?
Implications for next steps (JAMA opinion piece, January 28, 2021)

Immune Escape: Omicron Variant

UK Study on Omicron Reinfections:

Figure 13 (a): The weekly rate of possible COVID-19 reinfections with cumulation of first infections becoming eligible for reinfection and weekly total of first infection* (England only to week 3 2022, provisional early data$^{\Delta}$)



Highlights the importance of vaccinating the world.

Clinical Trials

Clinical Trials

Clinical trial: a controlled experiment with a clinical event as an outcome measure, done in a clinical setting, and involving persons with a specific disease or health condition.

A randomized controlled trial (RCT) is a clinical trial in which participants are randomly assigned to separate groups that compare treatments.

Randomization: the process of assigning clinical trial participants to treatment groups. Randomization gives each participant a known (usually equal) chance of being assigned to any group. Successful randomization requires that group assignments cannot be predicted in advance.

Considerations in Clinical Trials

Purpose of randomization:

- Alleviate systematic difference (or bias) between the groups due to factors other than the intervention, i.e., guarantees no unmeasured confounding. Thus, the observed difference is due to intervention.
- Achieve comparability (similar baseline distribution) between the groups (sometimes no guarantee for small sample sizes)

RCTs are considered the gold standard of study designs because the potential for bias (selection into treatment groups) is avoided. However, external validity may be poor due to stringent inclusion/exclusion criteria. In some settings, it is unethical to use RCT (e.g., smoking and lung cancer).

Types of randomization

Simple Randomization

Permuted Block Randomization

Stratified Block Randomization

Dynamic (adaptive) Random Allocation

Simple Randomization

Coin Tossing for each trial participant

Computer generated random sequences

Example:
Two Groups (equal probability): AABABAAABAABAAA……
Two groups (2:1 ratio): BBAABABBABAABAA……..

Permuted Block Randomization

Used for small studies to maintain reasonably good balance among groups

In a two group design, blocks having equal numbers of As and Bs (A = intervention and B = control, for example) are used, with the order of treatments within the block being randomly permuted.

Permuted Block Randomization

With a block size of 4 for two groups(A,B), there are 6 possible permutations and they can be coded as: 1=AABB, 2=ABAB, 3=ABBA, 4=BAAB, 5=BABA, 6=BBAA

Each number in the random number sequence in turn selects the next block, determining the next four participant allocations.
For example, the sequence 612614.... will produce BBAA AABB ABAB BBAA AABB BAAB.

In practice, using a single block size of two or four is too small since researchers may guess the assignment and risk selection bias. Mixing block sizes with the size kept unknown to the investigator. Maintains concealment. Simple randomization determines which block size to use next.

Stratified Block randomization

Stratified block randomization can further restrict chance imbalances to ensure the treatment groups are similar for selected prognostic variables or other patient factors. A set of permuted blocks is generated for each combination of prognostic factors.

Typical examples of such factors are age group, severity of condition, and treatment center. Stratification simply means having separate block randomisation schemes for each combination of characteristics ('stratum')

For example, in a study where you expect treatment effect to differ with age and sex you may have four strata: male over 65, male under 65, female over 65 and female under 65.

The analysis will include stratification variable as a covariate.

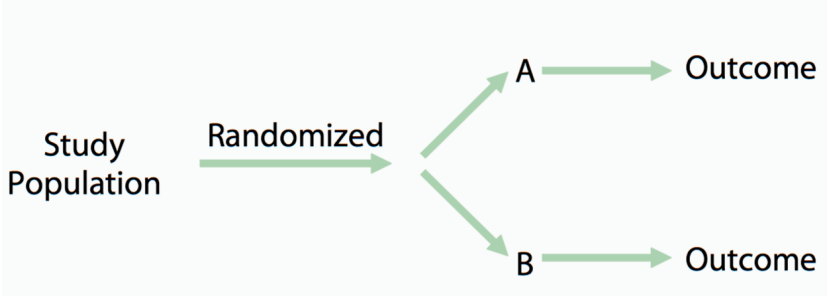
Adaptive Randomization

Simple and block randomization are defined, and allocation sequences set up, before the start of the trial.

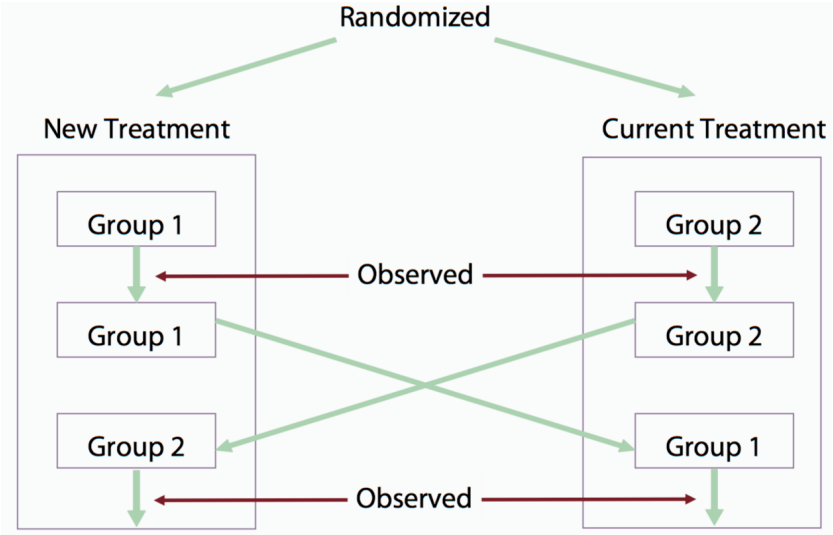
Adaptive randomization: allocate patients to treatment group by checking the allocation of similar patients already randomized, and allocating the next treatment group adaptively to best balance the treatment groups across all stratification variables. Biased coin randomization and minimisation are two such methods.

Clinical Trial Designs

Parallel group design:



Crossover design:



Clinical Trial Designs

Factorial design:

		Treatment B: +	Treatment B: −
Treatment A	+	Both A and B	A only
	−	B only	Neither A nor B

Example: HEAL

Potential Bias and Error in RCT Analysis

Sources of bias:

- Missing data, dropouts
- Deliberate exclusions
- Handling noncompliance

Error: Multiple comparisons

Reasons of Missing Data

- Subject dropped out and refused further follow-up
- Subject stopped drug or otherwise did not comply with protocol and investigators ceased follow-up. Not good practice!
- Subject died or experienced a major medical event that prevented continuation in the study
- Subject did not return–lost to follow-up
- Subject missed a visit
- Subject refused a procedure
- Data not recorded

Concerns of Missing Data

Missingness may be associated with unobserved outcomes and intervention (missing not at random, MNAR)

We usually don't know the form of this association, but if we fail to account for the (true) association, we may bias our results.

Can explore extent of bias through additional modeling (e.g., modeling how missing data depends on unobserved outcomes) and sensitivity analysis.

Missing Data and Prognosis

For most missing data, it is possible that missingness is related to prognosis

- Subject feels worse, doesn't want to come to the clinics
- Subject feels much better, no longer interested in study
- Subject feels study treatment not helping, drops out
- Subject intolerant to side effects

Thus, missing data raise concerns about biased results.

Can't be sure of the direction of the bias; can't be sure there is bias; can't rule it out

Distinguish treatment drop out versus assessment drop out.

Strategy to Deal with Drop Out

Excluding patients with missing values can bias results, increase Type I error (false positives)

Collecting and analyzing outcome data on non-compliant patients may dilute results, increase Type II error (false negatives)

General principle: we can compensate for dilution with sample size increases, but can't compensate for potential bias of unknown magnitude or direction

Advice collaborators to continue collect assessments even after a patient drops out of treatment!

Intention-To-Treat (ITT) Principle

ITT for analyzing RCTs: All randomized patients should be included in the (primary) analysis in their assigned treatment groups, even if they have stopped taking assigned treatments.

ITT Principle and Missing Data

ITT: Analyze all randomized patients in groups to which randomized

Implication of ITT principle for design: collect all required data on all patients, regardless of compliance with treatment. Thus, avoid missing data.

What to do when data are unavailable?

Modified ITT

1. All randomized eligible patients...
2. All randomized eligible patients who received any of their assigned treatment...
3. All randomized patients for whom the primary outcome is known... complete-case analysis----may be biased

Modified ITT 1

Exclude randomized subjects who turn out to be ineligible?

- probably won't bias results–unless level of eligibility depends on treatment and/or outcome
- greater chance of bias if eligibility assessed after data are revealed and study is unblinded

Modified ITT 2

Exclude randomized subjects who never started assigned treatment?

- probably won't bias results if study is double-blind
- in unblinded studies (e.g., surgery vs drug), refusals of assigned treatment may result in bias if "refusers" are excluded

Modified ITT 3

Exclude randomized patients who become lost-to-follow-up so outcome is unknown?

- possibility of bias, but can't analyze what we don't have
- model-based analyses may be informative if assumptions are reasonable
- sensitivity analysis important, since we cannot verify assumptions

Noncompliance

There will always be people who don't take medical treatment as prescribed

There will always be noncompliant subjects in clinical trials

We evaluate data from a trial in which some subjects do not adhere to their assigned treatment regimen under ITT principle.

More sophisticated analysis: per-protocol analysis.

Dealing with Missing Data

Analysis of data when some are missing requires assumptions

The assumptions are not always obvious, may not be checked

When a substantial proportion of data is missing, different analyses may lead to different conclusions

- robustness of findings should be assessed

When few data are missing, approach to analysis probably won't matter

Common Practices

Ignore those with missing data; analyze only those who completed study (completers analysis)

For those who drop out, analyze as though their last observation was their final observation (last observation carried forward)

For those who drop out, predict what their final observation would be on the basis of outcomes for others with similar characteristics; or predict based on their own previous observations (mixed effects model, multiple imputation)

Assumptions for Common Practices

Analyze only subjects with complete data (completers only analysis)

- Assumption: those who dropped out would have shown the same effects as those who remained in study (missing completely at random; MCAR)
- However, those who drop out may be different from those who remain in study

Last observation carried forward

- Assumption: those who dropped out would not have improved or worsened had they remained in study
- However, dropout may relate to perception of getting worse or better

Mixed effects model or multiple imputation

- Assumption: available data will permit unbiased estimation of missing outcome data (missing at random; MAR)
- However, we can only predict outcome or missingness using data that are measured; unmeasured variables may be more important in predicting outcome or missingness (missing not at random; MNAR)

sensitivity analysis

Sensitivity Analysis

Analyze data under different assumptions to see how much the inference changes

Such analyses are essential to understanding the potential impact of the assumptions required by the selected analysis

If all analyses lead to the same conclusion, will be more comfortable that results are not biased in important ways

Useful to pre-specify sensitivity analyses and consider what outcomes might either confirm or cast doubt on results of primary analysis

Sensitivity Analysis Under "Worse Case Scenario"

Simplest type of sensitivity analysis

- Assume all on investigational drug were treatment failures, all on control group were successes
- If drug still appears significantly better than control, even under this extreme assumption, very robust result
- Note: if more than a tiny fraction of data are missing, this approach is unlikely to provide persuasive evidence of benefit. May miss a drug effect.

Other Approaches of Sensitivity Analysis

Different ways to model possible outcomes

Different assumptions can be made

- Missing at random (predict based on available data; Inverse probability weighting; Use of linear mixed effects models)
- Nonignorable missing (must create model for missing mechanism; pattern mixture model)

Simple analyses (e.g., completers only) can also be considered sensitivity analyses

If different analyses all suggest the same result, can be more comfortable with conclusions

Multiplicity Issues

Multiplicity Issues

Multiplicity refers to the multiple judgements and inferences we make from data: hypothesis tests; confidence intervals

Multiplicity leads to concern about inflation of Type I error, or false positives

There are many types of multiplicity to deal with

- Multiple endpoints
- Multiple subgroups
- Repeated testing over time rate of improvement can avoid this

Avoid ”data-driven” testing: avoid identify comparisons that look “interesting” and perform significance tests for these results

Situations that Multiple Testing May Arise

Multiple treatment arms: A, B, C

Subgroups: gender, age, co-morbidity, previous drug use...

Site groupings: country, type of clinic...

Covariates accounted for in analysis

Repeated testing over time

Multiple endpoints

- different outcome: mortality, progression, response
- different ways of addressing the same outcome: different statistical tests

survival endpoints, e.g. PFS

Handle Multiple Comparisons

1. Ignore the problem; report all interesting results: Common bad practice and will be criticized by statistical reviewers (especially for high profile journals).
2. Perform all desired tests at the nominal level and warn reader that no accounting has been taken for multiple testing. Leave it to the readers for interpretation. Not ideal.
3. Limit to only one test. Not realistic.
4. Adjust the p-values/confidence interval widths in some statistically valid way. Better practice.

Perform Only One Test

Single (pre-specified) primary hypothesis

Single (pre-specified) analysis

No consideration of data in subgroups

Not really practical. Common practice:

- pre-specify the primary hypothesis and analysis, consider all other analyses "exploratory"
- perform pre-planned subgroup analysis

Multiple Comparisons Adjustment Procedures

Divide desired α by the number of comparisons (Bonferroni)
Bonferroni-type stepwise procedures
Control false discovery rate Confirmatory trials don't do FDR control

Multivariate testing for overall heterogeneity (e.g., omnibus ANOVA test), followed by pairwise tests

Resampling-based adjustments. Important and practical tool.

Importance of a Control Group

Before-after comparison is not sufficient to establish a treatment effect:

- In order to assess effect of new treatment, must have a comparison group
- Changes from baseline could be due to factors other than intervention
  - Natural variation in disease course
  - Patient expectations/psychological effects (placebo effect)
  - Regression to the mean
- Cannot assume investigational treatment is the cause of observed changes without a control group

Some Final Comments about Analysis of RCTs

There are many pitfalls in the analysis and interpretation of clinical trial data

Awareness of these pitfalls will prevent errors in drawing conclusions

For some issues, no consensus on optimal approach (e.g., best deign for subgroup analysis)

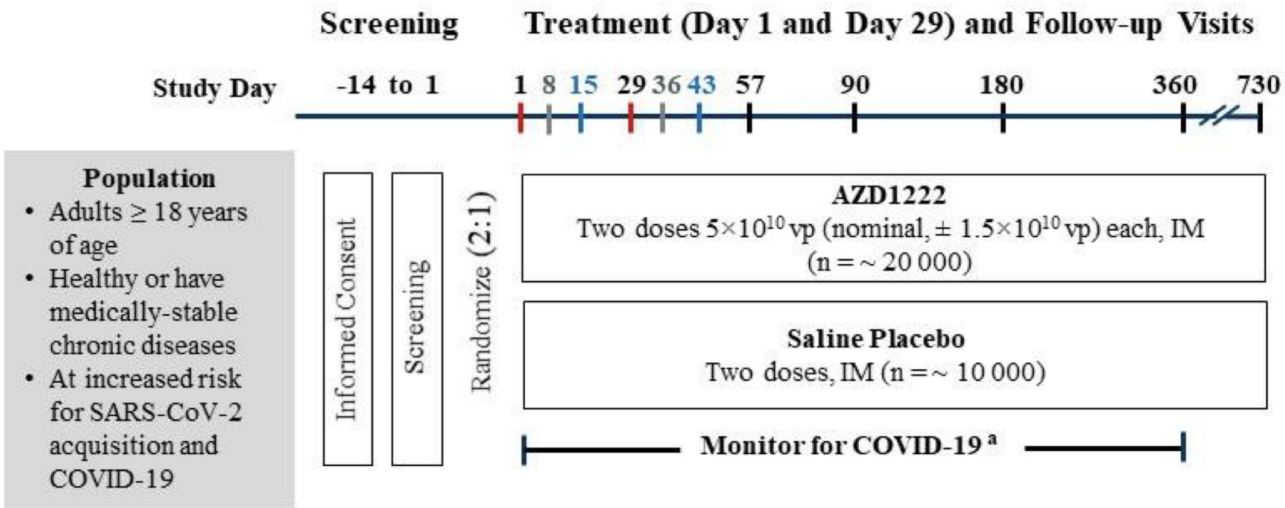
Statistical rules are best integrated with clinical judgements

Example: Vaccine RCTs for COVID-19

Vaccine Trials for COVID-19

- Trial protocols published: Moderna, Pfizer, AstraZeneca
- Similar designs (below is from AstraZeneca)

Figure 1 Study Design



[a] Participants who present with qualifying symptoms will be tested for SARS-CoV-2 and if positive, will complete illness visits.

Red bars (Day 1 and Day 29): Administration of study intervention.
Gray bars (Day 8 and Day 36): Visits will be telephone contacts, not study site visits.
Blue bars (Day 15 and Day 43): Visits will only be for participants in the substudy. The first participants randomized in each age group, including 1 500 participants 18 to 55 years of age, 750 participants 56 to 69 years of age, and 750 participants ≥ 70 years of age, will also participate in a substudy assessing the reactogenicity and immunogenicity of AZD1222.
COVID-19 = coronavirus disease 2019; IM = intramuscular; SARS-CoV-2 = severe acute respiratory syndrome-coronavirus-2; vp = viral particles.

Primary Endpoint

SARS-CoV-2 infection

- Informative for controlling COVID spread
- Not very clinically relevant

Symptomatic COVID

- Clinically relevant and moderate number of cases expected
- Most cases can be mild and less clinically relevant

Severe COVID

- Most clinically relevant (hospitalizations and deaths)
- Few cases to be expected. Need large sample size

Primary Endpoint

FDA guideline:
Either symptomatic COVID or COVID infection is acceptable.

Should consider powering efficacy trials for formal hypothesis testing on a severe COVID endpoint or evaluate as a secondary endpoint.

Vaccine Efficacy (VE)

VE: percent reduction in relative risk comparing vaccine group vs placebo group

$$VE = 1 - \frac{\text{Risk in vaccine}}{\text{Risk in placebo}}.$$

Relative effect, not absolute effect (90% efficacy $\neq$ 10% chance of having a risk outcome.

For approval, FDA guideline specifies:

- A point estimate of VE for the primary endpoint $\geq 50\%$
- Lower bound of CI $> 30\%$
- Overall type I error control for one-sided test at 2.5%

Safety outcomes are evaluated as well.

RCT Results

Example of RCT design and analysis: AZ study protocol, statistical analysis plan, results.

Summary of three major COVID-19 vaccine trial results:

Trial	VE COVID 95% CI	Cases Txt:Plb	VE Severe 95% CI	Severe Txt:Plb
AZ	74.0% (65.3%, 80.5%)	203 73:130	100%	8 0:8
Pfizer	94.6% (90.0%, 97.9%)	170 8:162	88.9% (19.8%, 99.7%)	10 1:9
Moderna	94.1% (89.1%, 97.0%)	196 11:185	100.0 (86.9%, 100.0%)	30 0:30

Highly positive! Vaccine administration became an urgent issue.

Emerging Topic on Causal Inference:
Self-Matched Learning for the Analysis of Electronic Health Records (EHRs)