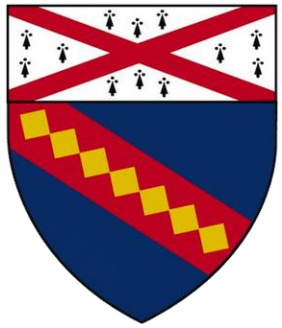
Methodological and Statistical Considerations in Implementation Science

Denise Esserman, PhD
Professor of Biostatistics
Yale School of Public Health



With support and materials provided by Peter Peduzzi, PhD and James Dziura, PhD

Study designs and statistical considerations

All design and statistical considerations are driven by the **research question and aims**.

PICO (T)

- Population
- Intervention
- Comparator
- Outcome
- Time

“In many ways the design of a study is more important than the analysis. A badly designed study can never be retrieved, whereas a poorly analysed one can usually be reanalysed. Consideration of design is also important because the design of a study will govern how the data are to be analysed.”

https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/13-study-design-and-choosing-statisti

Research Hypothesis:

Usually describes what you expect to find, e.g., what the investigator suspects the relationship truly is among the variables being studies

Closely related to the alternative hypothesis (used for sample size calculations)

Statistical Hypothesis (to be tested)

Usually referred to as the null hypothesis (what study trying to disprove)

NOTE: A clear research hypothesis tells the statistician which "null hypothesis" you would like to test and how to frame the "alternative hypothesis" of interest

Sample Size and Statistical Power Considerations

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β
Not rejected	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β

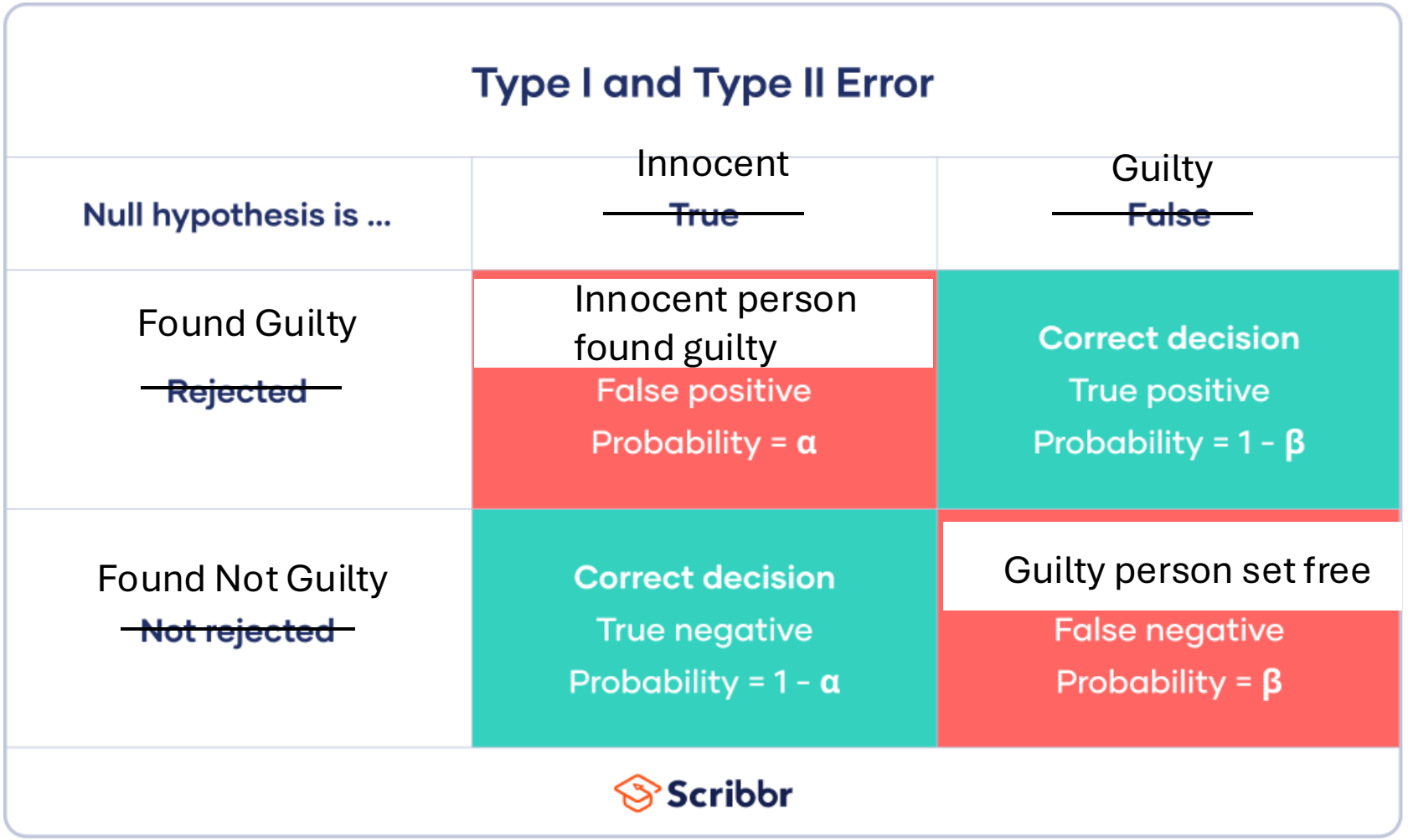
https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β **POWER**
Not rejected	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β

https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

Seriousness of the Errors



General Comments

- Sample size is one of most frequent calculations that practicing statisticians and epidemiologists do, but it is not often taught
- Particularly important for study design
- Has budget implications

General Comments

- Inadequate sample size reason for many inconclusive studies
  - There's a difference but can't detect it
- Sample size = best guess based on assumptions that need to be carefully examined
- Need right order of magnitude → don't want to target 100 patients when 1000 needed
  - Ethical consideration
- Sample size and analysis must be consistent with primary endpoint
  - Hypotheses, endpoints, sample size and analysis need to match up – one flows from the other

General Considerations

- Studies on margin = uncertainties about assumptions such as event rates, losses, etc.
  - Need increased power: 90% instead of the usual 80%
  - Need increased sample size to account for uncertainties
- Need to be conservative because rarely observe what it is hypothesized
- Validated software recommended for sample size determination, freeware may not be validated
- Good strategy is to calculate sample size using different software not only for replication but also because results can sometimes differ among packages, particularly for complex trial designs

Sample Size Determination Depends on the Endpoint

- Not acceptable to base sample size for a study on a secondary endpoint just because it gives a lower sample size
  - Strategy will lead to an underpowered study of primary hypothesis
- Types of endpoints
  - Discrete/categorical measurement
  - Continuous measurement
  - Time to event
  - Repeated measurements over time

Discrete Categorical Endpoint

Binary: presence/absence of factor: 30-day mortality

Ordinal: severity of symptoms (mild, mod, severe)

Nominal: blood type

Continuous Endpoint

- Blood pressure, cholesterol level
- Generally, requires lower sample size than discrete endpoints because use contains more information
  - e.g., actual blood pressure value instead of presence/absence of HTN based on some cutpoint
- Continuous measures could be highly variable in small studies when a few people have large changes
  - A consideration is change > than some threshold, i.e., more than a minimally clinically important difference (MCID)

Time to Event Survival Analysis

- Survival analysis has two components
  - Occurrence of the event (e.g., death) and time to the event occurring
  - If the event does not occur, then patients are censored at their last follow-up time or the end of the study (assume censoring is unrelated to the outcome)
  - Censoring = no longer in the risk set for the outcome
- Survival endpoints use more information than just presence or absence of the event
- Information contained in events and not number of patients, i.e., no events = no information

Longitudinal Endpoints

- Repeated observations on each patient over time
- Repeated observations are correlated, and correlation needs to be considered in sample size
  - Sample size increases with increasing correlation for a fixed number of repeated observations
- Vulnerable to missing data
  - If expect a lot of missing data, may need to reconsider choice of endpoint and study design

Treatment Allocation Ratio

- Usually 1:1 = same number assigned to each treatment group because it gives maximum power
  - Ethical issue: maintains equipoise
- Can have an allocation ratio as extreme as 2:1 without big effect on power
- When to use unequal allocation?
  - Multiple experimental treatments compared with usual care – may want to assign fewer individuals to usual care
  - Desire more information about a new treatment relative to standard
  - Have an expensive treatment

Sample Size Specifications

- Treatment effect (clinically relevant)
  - Absolute metric: difference in means, proportions (Δ)
  - Relative metric: odds ratio, relative risk, hazard ratio

- Estimate of variability for absolute metrics
  - σ^2 estimated by SD^2, SD = standard deviation of a single measurement not the mean (standard error)
  - For proportions use p(1-p)

Inflations to Sample Size

- Losses and dropouts = inflate by 1/(1-loss rate)
  - When loss rate is high inflating by 1/(1-loss rate) can over-estimate N in survival studies
  - Inflating sample size for losses doesn't solve the missing data problem – still have missing data
- Nonadherence reduces treatment effect
- Multiplicity = Bonferroni correction to type I error
  - E.g., 3 treatment groups = 3 comparisons = $\alpha/3$
  - E.g., 2 primary outcomes = 2 treatment comparisons, one for each outcome = $\alpha/2$
  - Bonferroni is conservative – wouldn't necessarily use this in the analysis to control for multiplicity
- Interim looks for efficacy/futility = 3 to 5% inflation

Sensitivity Analyses

- To determine impact of assumptions made in determining sample size

- Calculate sample size under various scenarios for range of event rates, loss rates and effect sizes

Drop-ins and Dropouts

- Drop-ins from control to experimental (often termed crossovers) can dampen treatment effects if experimental is effective in an ITT analysis
- Dropouts have two meanings
  - Dropouts from treatment who need to be followed for outcomes – also called nonadherence rate
  - Dropouts from the study – results in censoring due to incomplete follow-up – typically referred to as lost to follow-up

What to do if target sample size can't be obtained

- Inability to recruit the target sample size is a common occurrence in clinical trials – patients disappear once a study begins → major reason for futility
- Administrative considerations – don't affect type I error
  - Reassess eligibility criteria and relax if appropriate – could affect event rates and increase N
  - Add more sites – can be costly
  - Reallocate funds to sites that can recruit more patients
  - Extend follow-up in survival studies to obtain more events – adds cost

What to do if target sample size can't be obtained

- Reduce power from 90% to 85% or 80%
- Reassess sample size assumptions for nuisance parameters which don't affect Type I error
  - Variability – lower reduces N
  - Event rates in the control arm – higher rates reduce N
  - In cluster trials, re-estimate the intracluster correlation coefficient (ICC) – lower ICC reduces N
- Re-estimation of sample size based on Δ affects Type I error – adaptive designs need to be pre-specified
- If despite all measures the target sample size cannot be obtained, the trial may be futile

Now that we have conducted the study, how do we analyze it?

Statistical Analysis Plan (SAP)

- Pre-planned analytic plan – before analysis is conducted
  - Goal is to conduct as unbiased analysis as possible
  - Analysis follows the study design
- Two general scenarios encountered in practice
  - Data already generated – by investigator/client, EHR, other large databases
    - Develop a plan but can refine it based on the inspection of the data
    - No control over how data were generated
  - Data to be generated from a study to be planned and executed
    - Pre-specified in a protocol – limited refinement of the plan, e.g., before data lock and unblinding in clinical trials
    - Can provide input about how data are generated via study design
- Creating an SAP requires a lot of data analysis experience
- Standalone, self-contained document

Elements of the SAP

- **General considerations**
  - Analysis population – who is included in/excluded from the analysis, e.g., intent to treat, modified ITT, per protocol, as treated
  - Unit of analysis
  - Software – R, SAS, Stata
- **Characterize the population via summary statistics, e.g., comparability of treatment groups**
  - No tests of significance for a RCT trial, proper randomization assures balance
  - No assurance of balance in an observational study - need to check for imbalances and consider adjusting for these
- **Analysis plan for each outcome**
  - Endpoint and how measured (e.g., time to event, instrument used)
  - Type I error and any control for multiplicity – multiple treatment groups, outcomes, subgroups, looks at the data; multiple genes being evaluated
  - Method of analysis – specify model and assumptions (e.g., PH, MAR, normality, etc.)
    - Testing model assumptions and alternative approaches if assumptions are violated (what ifs)
    - Sensitivity analyses, e.g., examination of MAR assumption, use of multiple imputation, etc.
  - Covariate adjustment variables (pre specified in a clinical trial, e.g., randomization strata)
  - Metric used for estimation of treatment effects, e.g., hazard ratio, LS means, odds ratio, risk ratio etc. + confidence intervals
  - Subgroup analyses – tests of heterogeneity (interactions)
  - References

What are the most important things to consider when choosing an analysis method?

- How the data were generated (design of the study)
  - Do we need to be concerned with clustering/nesting?
- N (individuals) and p (number of parameters)
- Types of data collected
- Goal of the research study
- One-sample vs. two-sample
- Sample size
  - Power/Precision
  - Determining appropriate tests
- Clinical trial
  - Follow intent to treat analysis

Analysis must match design

Examples:

- Individually randomized trial stratified by age and HIV status
  - ALL analyses need to adjust for age and HIV status
- Cluster randomized trial where students are nested within classrooms and classrooms nested within schools, where classrooms are randomized stratified by school.
  - ALL analyses need to account for clustering of individuals within schools and adjust for school (stratification).
- Stepped wedge cluster RCT
  - Analysis needs to account for within period and between period effects

Types of Outcomes to Consider

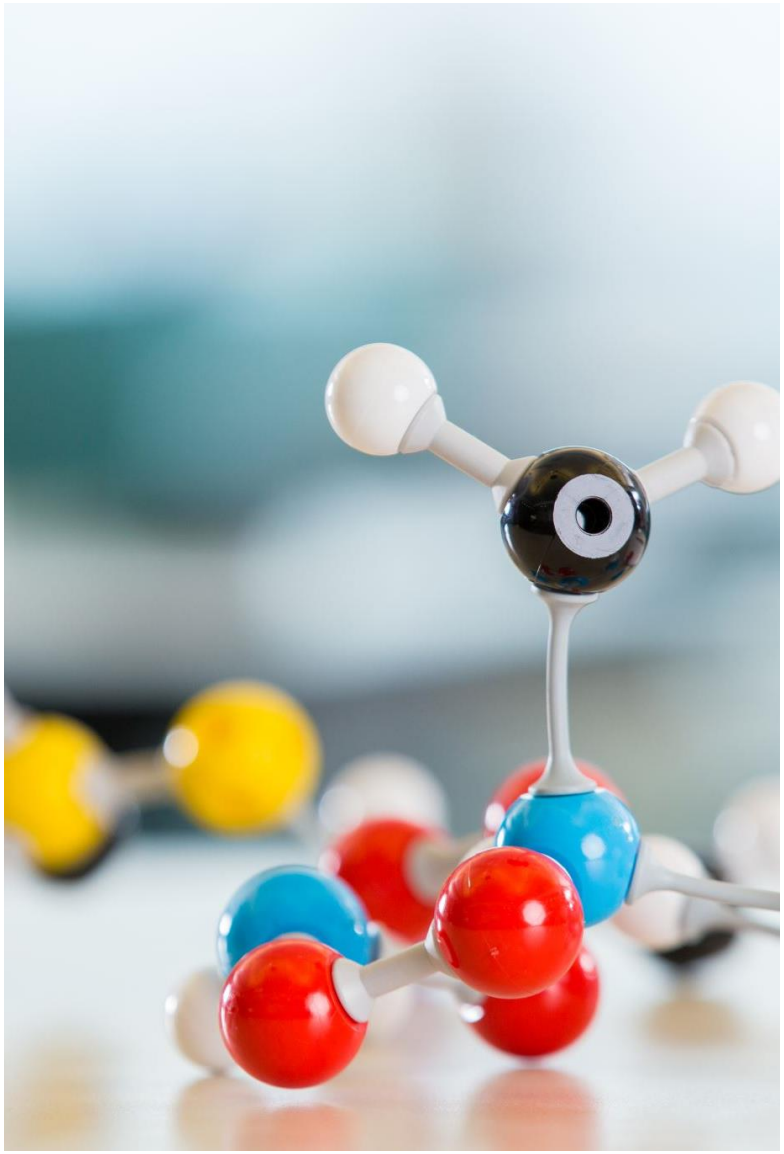
- Continuous Data
- Binary Data
- Count Data
- Time to event data
- Rates
- Ordinal
- Repeated Measures

Example: Different Analysis Techniques for CONTINUOUS DATA

Example: blood pressure, BMI, cholesterol, QOL

- One time point – simple randomization
    - T-test for independent groups: estimate difference in means
    - Assumptions: normality, equality of variances, independence
- One time point – stratified randomization
    - Linear regression, adjust for stratification (possibly baseline value): estimate least squares mean difference
    - Assumptions: normality, independence, heteroscedasticity, linearity
- Repeated measures over time
    - Linear Mixed Model
        - Estimate average treatment effect over time
        - Estimate least squares mean difference at specified time point
    - Generalized estimating equations (GEE): estimate marginal mean differences
    - Assumptions: similar to linear model

HOW IMPORTANT IS IT TO CHECK MODEL ASSUMPTIONS?



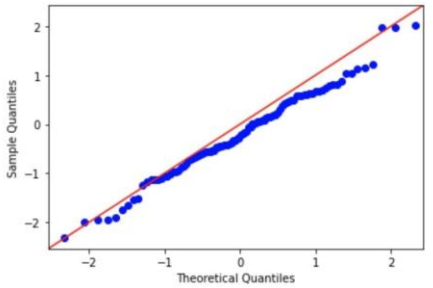
Reasons Assumptions are NOT checked

- Unaware of the assumption of the statistical test being used
- Unaware of standard approaches to test the assumptions and evaluate violations
- Unaware how to remedy violations and/or select a new test if violations cannot be remedied
- Believe the test is robust enough and thus assumptions do not need testing

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6467598/

Parametric versus Non-parametric tests

- Parametric tests make assumptions about population parameters (e.g., mean, standard deviation, proportion), while a non-parametric tests does not assume anything about the underlying distribution
- Some assumption of parametric test
  - Normality
    - Can use plots (e.g., Q-Q plot) to assess normality
    - Test of normality
  - Independence
  - Equal variance



- Advantages and Disadvantages non-parametric approaches
  - More powerful if assumptions violated (less powerful if they have not)
  - Small sample sizes okay
  - Used for different types of data (e.g., ordinal, nominal, interval)
  - Can handle outliers

https://builtin.com/data-science/parametric-vs-nonparametric

Missingness Mechanisms

Need to understand what assumptions the models make about missing data

01

Missing Completely at Random (MCAR) – the likelihood of missing data is unrelated to observed variables or unobserved values

Example: participant unintentionally skips a question; accidently drop test tube

02

Missing at Random (MAR) – likelihood of missing data is related to observed variables but not to unobserved values

Example: Younger participants report using condoms more than older participants in an HIV prevention trial

03

Missing Not at Random (MNAR) – likelihood of missing data is related to unobserved values

Example: Participants who engage in risky behavior intentionally withhold condom use information

https://ydata.ai/resources/understanding-missing-data-mechanisms

QUESTIONS?

CASE EXAMPLE: WHAT IF?

Working with HIV clinics to adopt addiction treatment using implementation facilitation (WHAT-IF?)[1-4]

Background[1]

- Tobacco, alcohol and opioid misuse are associated with substantial morbidity and mortality among people with HIV (PWH).
- Evidence-based counseling and medications for addiction, but infrequently offered in HIV clinics
  - Both patients and clinician hesitancy (due lack of knowledge) are barriers to adoption
- **Type 3** effectiveness-implementation study: address implementation challenges.
  1. Conduct formative evaluation of barriers to and facilitators of implementing addiction treatment for PWH
  2. Evaluate impact of Implementation Facilitation (IF) on promoting adoption of addiction treatment and clinical outcomes

Stepped-Wedge Design[1]

- Conducted in 4 clinics in the northeast United States
- Open cohort design

Clinic 4	Control	Control	Control	Control	IF	Evaluation	Maintenance
Clinic 3	Control	Control	Control	IF	Evaluation	Maintenance	Maintenance
Clinic 2	Control	Control	IF	Evaluation	Maintenance	Maintenance	Maintenance
Clinic 1	Control	IF	Evaluation	Maintenance	Maintenance	Maintenance	Maintenance
Time Point	**Baseline**	**6 months**	**12 months**	**18 months**	**24 months**	**30 months**	**36 months**
	July 2017	**Jan 2018**	**July 2018**	**Jan 2019**	**July 2019**	**Jan 2020**	**July 2020**

IF= Implementation Facilitation

- Guided by Promoting Action on Research Implementation in Health Services Research (PARiHS) framework.

Stepped-Wedge Design Rationale[1]

Clinic 4	Control	Control	Control	Control	IF	Evaluation	Maintenance
Clinic 3	Control	Control	Control	IF	Evaluation	Maintenance	Maintenance
Clinic 2	Control	Control	IF	Evaluation	Maintenance	Maintenance	Maintenance
Clinic 1	Control	IF	Evaluation	Maintenance	Maintenance	Maintenance	Maintenance
Time Point	Baseline	6 months	12 months	18 months	24 months	30 months	36 months
	July 2017	Jan 2018	July 2018	Jan 2019	July 2019	Jan 2020	July 2020

IF= Implementation Facilitation

1) Allows all sites to receive the Intervention
2) Facilitates conduct of the study related to logistical and personnel challenges with a small team of investigators leading the IF
3) Allows for consideration of temporal trends, which is particularly relevant
given the heightened focus on the opioid epidemic

Outcomes

- **Primary: Provision of Addiction Treatment**
  - Assessed using electronic health record (EHR) data
- Secondary:
  - Implementation outcomes:
    - Clinician, staff, and organizational readiness to provide addiction treatments
    - Models of care
  - Patient level effectiveness outcomes
    - Antiretroviral regimen receipt
    - HIV viral suppression
    - VACS Index 2.0 scores
    - Retention in HIV Care

Patient Participants

- All patients with HIV receiving care in the participating clinics from July 26, 2016 through July 25, 2020
  - Receiving care = scheduled visit at clinic during time period of interest
  - Eligible to enter cohort at any point during study period (open cohort)
- Patients considered to have substance use (opioid, alcohol, tobacco) if documented in problem list, encounter reason, or ICD code

Sample Size[1,5]

- The primary goal of this study was to assess the impact of IF on provision of addiction treatment among eligible patients by the clinics over time (baseline vs. evaluation vs. maintenance periods).
- Informed by prior work, we anticipated an 11% absolute increase during the evaluation phase and 19% absolute increase during the maintenance phase from baseline.
- A parallel group design, unadjusted for clustering and repeated measures would require a sample size of 592 (296 in each the control and intervention arms) to detect the estimated effect size assuming 90% power and Type I error of 0.05.
- To account for the stepped wedge design, we made the following assumptions:
  - Each clinic would provide a minimum of 300 addiction treatment eligible patients
  - Intracluster correlation, $\rho = 0.001$
  - Number of steps $\kappa = 4$
  - Number of baseline measurements, $b = 1$
  - Number of measurements taken after each step, $t = 1$
- This yielded a derived design effect of 0.63 and assuming a cross-sectional design required a sample size of **375 across the four clinics**.

Implementation facilitation (IF)[1]

- Adapted existing IF manual used to promote mental health treatment in primary care
- Included bundle of activities designed to promote addiction treatment in HIV clinics – tailored to site specific needs
- Informed by survey data collected prior to IF
  - Formative evaluation of barriers and facilitators
    - Site visit, focus groups, review of work flow, EHR documentation
  - Academic detailing
- Tracked conduct and participation in IF activities using tracking log

Analysis[3]

- Intent-to-Treat Approach based on time clinics were scheduled to cross over from control to facilitation
- Descriptive Statistics
  - Characterize baseline characteristics of clinic populations
- Generalized estimating equations (GEE)
  - Included study phase, site and time
  - Generated odds ratios and 95% confidence intervals
- Sensitivity analyses
  - Included all patients regardless of substance use disorder (given may not be uniformly captured)
  - Excluded final period due to COVID-19 pandemic starting
- Two-sided $p<0.05$ for statistical significance
- SAS software version 9.4

Results – Population[3]

3647 (range: 366-1548) patients engaged in care across the 4 clinics

- Mean age was 49 (SD=12) years
- 50% were Black
- 22% were Hispanic
- 39% were Female

- 3% (N=121) had opioid use disorder
- 3% (N=126) had alcohol use disorder
- 12% (N=420) had tobacco use disorder

Impact of Facilitation on Provision of Addiction Treatment[3]

Table 2. Provision of Medications for Addiction Treatment Among Treatment-Eligible Patients Across All Sites by Study Period, Results From Generalized Estimating Equation

	Provision of MOUD[a]		Provision of MAUD		Provision of MTUD	
Study period	Patients, No. (%) [95% CI]	*P* value	Patients, No. (%) [95% CI]	*P* value	Patients, No. (%) [95% CI]	*P* value
Control	243 (27) [22-32]	Reference	251 (8) [5-12]	Reference	810 (33) [30-36]	Reference
Intervention	117 (28) [22-35]	.55	122 (13) [8-21]	.09	444 (41) [37-46]	.001
Evaluation	135 (28) [22-35]	.59	112 (13) [8-21]	.11	471 (40) [36-45]	.005
Maintenance	198 (29) [22-36]	.48	180 (17) [12-24]	.009	643 (38) [34-41]	.047

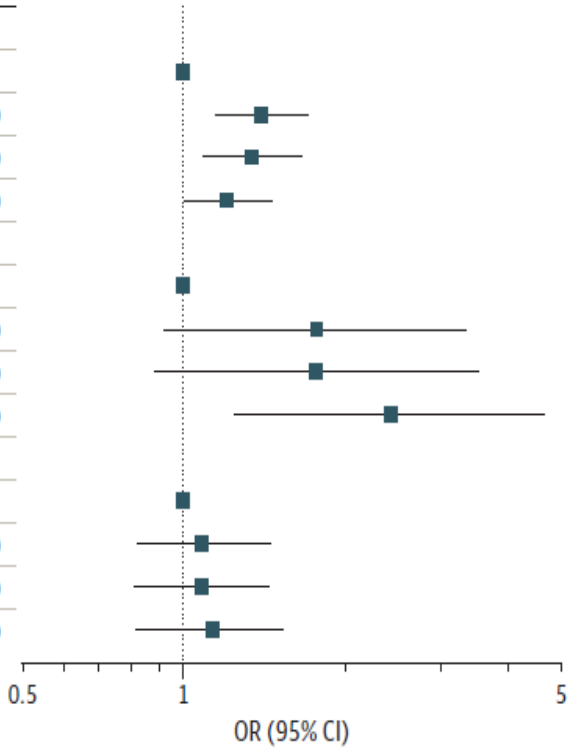
Abbreviations: MAUD, medications for alcohol use disorder; MOUD, medications for opioid use disorder; MTUD, medications for tobacco use disorder.

[a] MOUD exclusively included buprenorphine products.

Impact of Facilitation on Provision of Addiction Treatment[3]

Figure 1. Provision of Medications for Addiction Treatment Among Treatment-Eligible Patients Across Sites by Study Period

WHAT-IF phase	OR (95% CI)	*P* value
Tobacco use disorder		
Control	1 [Reference]	
Intervention	1.40 (1.14-1.71)	.001
Evaluation	1.35 (1.09-1.66)	.005
Maintenance	1.21 (1.00-1.46)	.047
Alcohol use disorder		
Control	1 [Reference]	
Intervention	1.76 (0.92-3.37)	.09
Evaluation	1.77 (0.88-3.55)	.11
Maintenance	2.43 (1.25-4.71)	.009
Opioid use disorder		
Control	1 [Reference]	
Intervention	1.09 (0.82-1.46)	.55
Evaluation	1.08 (0.81-1.45)	.59
Maintenance	1.12 (0.82-1.54)	.48



OR indicates odds ratio; WHAT-IF, Working with HIV Clinics to adopt Addiction Treatment using Implementation Facilitation.

Limitations/Strengths

- Use of electronic health record data (strength and weakness)
- Did not distinguish between treatment initiation and continuation
- Did not focus on medications obtained elsewhere (e.g., opioid treatment programs)
- COVID
- Did not evaluate impact of IF on quality of care
- Limited generalizability – 4 clinics in northeast US
  - Strength: Included both academic and community affiliated HOV clinics with varying level of resources and infrastructure
- Cannot attribute gaps in provision of addiction treatment to physicians or patients given way outcome measured

WHAT IF? References

1. Edelman EJ, Dziura J, Esserman D, Porter E, Becker WC, Chan PA, Cornman DH, Rebick G, Yager J, Morford K, Muvvala SB, Fiellin DA. Working with HIV clinics to adopt addiction treatment using implementation facilitation (WHAT-IF?): Rationale and design for a hybrid type 3 effectiveness-implementation study. Contemp Clin Trials. 2020 Nov;98:106156. doi: 10.1016/j.cct.2020.106156. Epub 2020 Sep 23. PMID: 32976995; PMCID: PMC7511156.
2. Muvvala, Srinivas B. MD, MPH; Gan, Geliang MPH; Morford, Kenneth L. MD; Dziura, James PhD; Esserman, Denise PhD; Porter, Elizabeth MBA; Chan, Philip A. MD; Cornman, Deborah H. PhD; Reynolds, Jesse MS; Yager, Jessica E. MD; Fiellin, David A. MD; Edelman, E. Jennifer MD, MHS. Facilitation and Preferred Models for Delivering Substance Use Disorder Treatment in HIV Clinics: Results From a Multisite Randomized Trial. Journal of Addiction Medicine 17(6):p e388-e391, 11/12 2023. | DOI: 10.1097/ADM.0000000000001192
3. Edelman EJ, Gan G, Dziura J, Esserman D, Porter E, Becker WC, Chan PA, Cornman DH, Helfrich CD, Reynolds J, Yager JE, Morford KL, Muvvala SB, Fiellin DA. Effect of Implementation Facilitation to Promote Adoption of Medications for Addiction Treatment in US HIV Clinics: A Randomized Clinical Trial. JAMA Netw Open. 2022 Oct 3;5(10):e2236904. doi: 10.1001/jamanetworkopen.2022.36904. PMID: 36251291; PMCID: PMC9577676.
4. Edelman EJ, Gan G, Dziura J, Esserman D, Morford KL, Porter E, Chan PA, Cornman DH, Oldfield BJ, Yager JE, Muvvala SB, Fiellin DA. Readiness to Provide Medications for Addiction Treatment in HIV Clinics: A Multisite Mixed-Methods Formative Evaluation. J Acquir Immune Defic Syndr. 2021 Jul 1;87(3):959-970. doi: 10.1097/QAI.0000000000002666. PMID: 33675619; PMCID: PMC8192340.
5. W. Woertman, E. de Hoop, M. Moerbeek, S.U. Zuidema, D.L. Gerritsen, S. Teerenstra, Stepped wedge designs could reduce the required sample size in cluster randomized trials, J. Clin. Epidemiol. 66 (7) (2013) 752–758.

QUESTIONS?

CASE EXAMPLE
STRIDE (Strategies to Reduce Injuries and Develop Confidence in Elders) Trial[1-6]

Overview of Trial Design[1]

- Pragmatic, multisite cluster-randomized, parallel group superiority trial
  - 10 US healthcare systems; 86 practices (patients nested within practices)
  - Unit of randomization was the practice
- Testing a multifactorial, individually-tailored intervention implemented by a nurse Falls Care Manager (FCM) in primary care setting
- Control condition: enhanced usual care
- Used stratified, covariate constrained randomization
  - Stratification by health care system
  - Balanced on practice size, geography (urban, rural), and race/ethnicity (non-white vs. white)
- Primary outcome: Time to first serious fall-related injury (adjudicated)
- Secondary outcomes: All fall injuries and all falls, well-being measures

Inclusion and Exclusion Criteria

INCLUSION CRITERIA

Identified as being at increased risk of falls

Answered yes to one of the following:

1. Have you fallen and hurt yourself in the past year?
2. Have you fallen two or more times in the past year?
3. Are you afraid that you might fall because of balance or walking problems?

Age 70 or older.*

Able to provide telephone consent or proxy consent with patient assent.

EXCLUSION CRITERIA

Patient enrolled in hospice.

Patient resides in nursing home.

Patient not capable of providing informed consent and does not have an available proxy.

Patient does not speak English or Spanish.

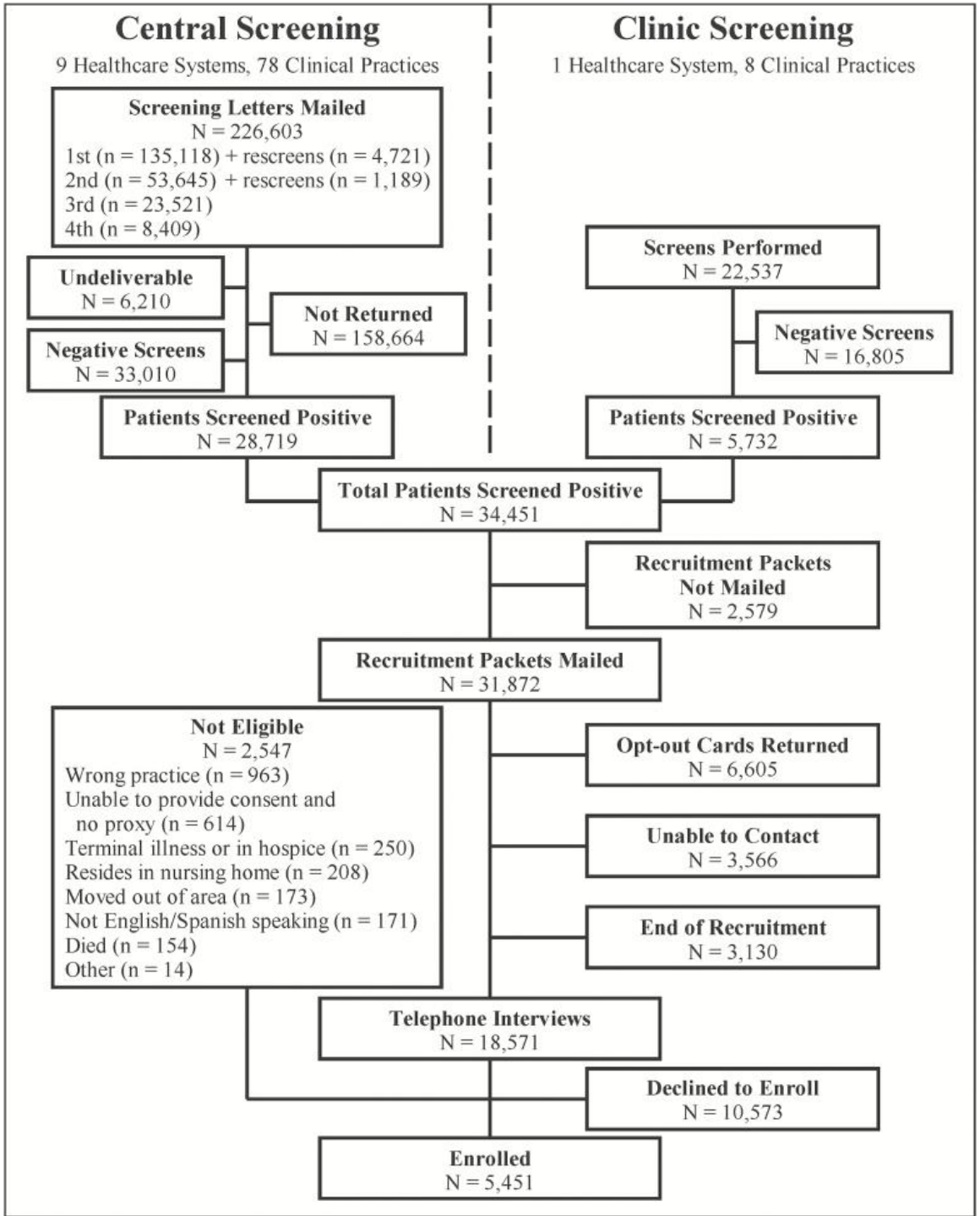
Screening and Recruitment[4]

Primary recruitment strategy was centralized screening through Yale Recruitment and Assessment Center (RAC) (9 out of 10 sites)

- Clinics provided RAC with names and addresses of age-eligible patients at each practice
- Updated biannually

Clinic Screening was used by other site

- Clinic provided names and addresses of those screened positive
- Regularly transmitted to RAC



Challenges with the Design and Conduct

- Estimation of the sample size for the initial trial
- Delivery of the intervention
  - FCM access
- Slow recruitment
  - Changes to inclusion criteria
  - Sample size re-estimation
  - Extended follow-up
- Issues with ascertainment of the primary outcome
  - Use of electronic health data
  - Ascertainment bias concerns
- Planning for the primary analysis
  - Appropriate contingencies

Sample size estimation

- Determined sample size for unclustered design using log-rank test and then inflated for clustering and interim monitoring
- Target sample size of 6000 participants was selected based on following assumptions
  - 18-month recruitment and 36-month trial with 18-month minimum follow-up
  - Type I error 5% two-sided and 90% power
  - Uniform accrural
  - Equal allocation
  - No adjustment for non-adherence – accounted for in estimate of treatment effect
  - All patient's followed to end of trial
  - 7% death rate without experiencing fall (semi-competing risk)
  - 3% inflation for interim monitoring for efficacy and futility
  - 53% inflation for design effect of clustering, based on 86 practices enrolling 70 participants and an ICC of 0.0076 estimated from a prior trial using a different analysis technique
    - No methods available to estimate ICC for time to event (also an issue for monitoring)
  - Needed 844 events to detect a 20% reduction for intervention relative to control

Primary Outcome Ascertainment

Primary outcome: First serious fall-related injury (adjudicated)

Initial plan

- Use self-report **or** electronic health record (EHR) **or** Centers for Medicare and Medicaid (CMS) to detect a serious fall injury

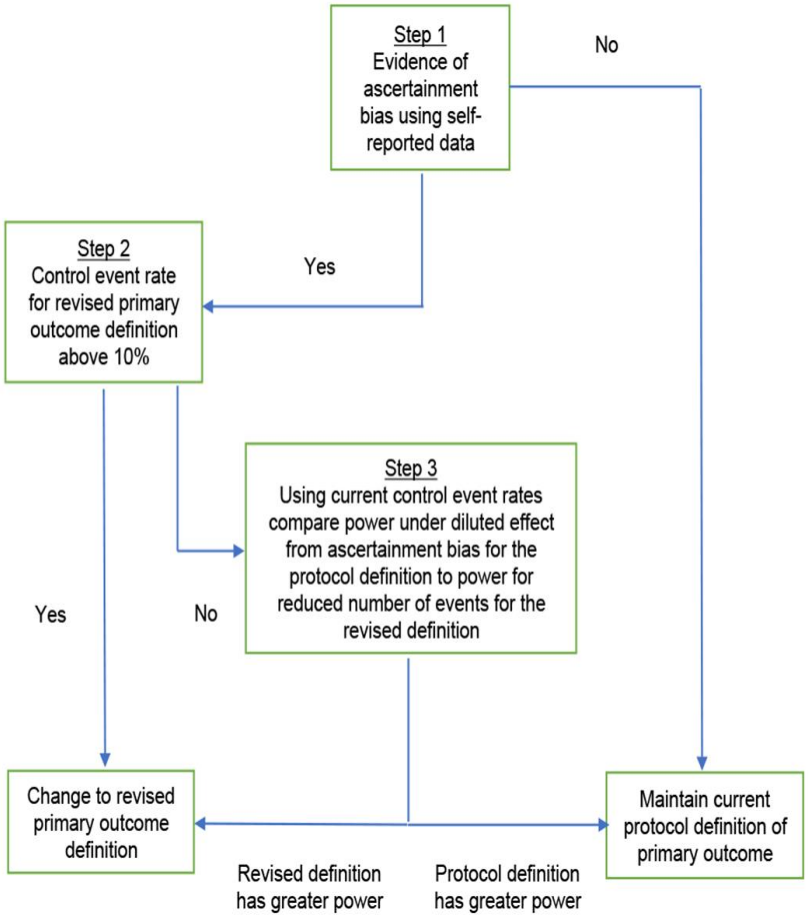
CHALLENGES

- Quality of EHR was not consistent across health care systems
- Not all individuals have CMS data
  - ~40% Medicare Advantage
- Variability in receipt of the data across sites and from data different sources
- Independent sources of data for adjudication

RESULT: Self report became primary source of outcome ascertainment

Ascertainment bias[5-6]

- Ascertainment bias: potential for unblinded observers (e.g., FCM) to influence ascertainment of outcomes.
- Prespecified interim monitoring plan included decision algorithm for revising primary outcome definition
- Original definition:
  - Serious fall injury requiring medical attention
    - Type 1: fracture other than thoracic/lumbar vertebral, joint dislocation cut requiring closure
    - Type 2: head injury, sprain or strain, bruising or swelling, other
  - Revised definition
    - Excluded type 2 injuries that did not necessarily require an overnight hospitalization (most subject to bias)
- Found evidence of bias and changed to revised definition



Trial Monitoring

- Monitor ICC
  - No standard methods for survival data
  - Impact of timing of practices and capturing of events
    - Only could use self-report data – huge lag in adjudication process
  - Difficult to make decisions about trial extensions without knowing

- Recruitment
  - Lowered age criteria from 75+ to 70+
  - Sample size re-estimation
    - Due to slower than expected enrollment in initial first 6-months, DSMB approved extension of recruitment period to 20 months and minimum follow-up of 20 months for total of 40-month trial duration
    - Reduced sample size requirements to 5,322 (same number of events needed 844)
- Note: after recruitment ended the follow-up was further extended from 40 month to 44 months because of lower than project event rates in control group

Final Analyses[2]

- Used principle of intent-to-treat, i.e., practices/participants will be analyzed according to original treatment assignment regardless of adherence to protocol.
- Analysis of primary outcome:
  - Used a survival model that incorporates semi-competing risk (death) and clustering (multi-state model)
    - Adjusted for randomization of practices by healthcare system and the covariates used for the constrained randomization (practice size, location, race/ethnicity)
  - Participants lost to follow-up without prior serious fall injury censored at data last seen
  - Sensitivity analyses: adjusted for pre-specified set of baseline covariates (age, sex, race/ethnicity, education, number of chronic conditions, number of positive screening items)
  - Standard approaches (e.g., test of proportional hazard assumption) used to examine fit of models
  - Contingency: If multi-state model did not converge, fit a model developed by Zhou[7]

Figure S1. CONSORT Diagram for the STRIDE Study Showing the Enrollment and Randomization of Primary Care Practices, and their Disposition Through the Course of the Study

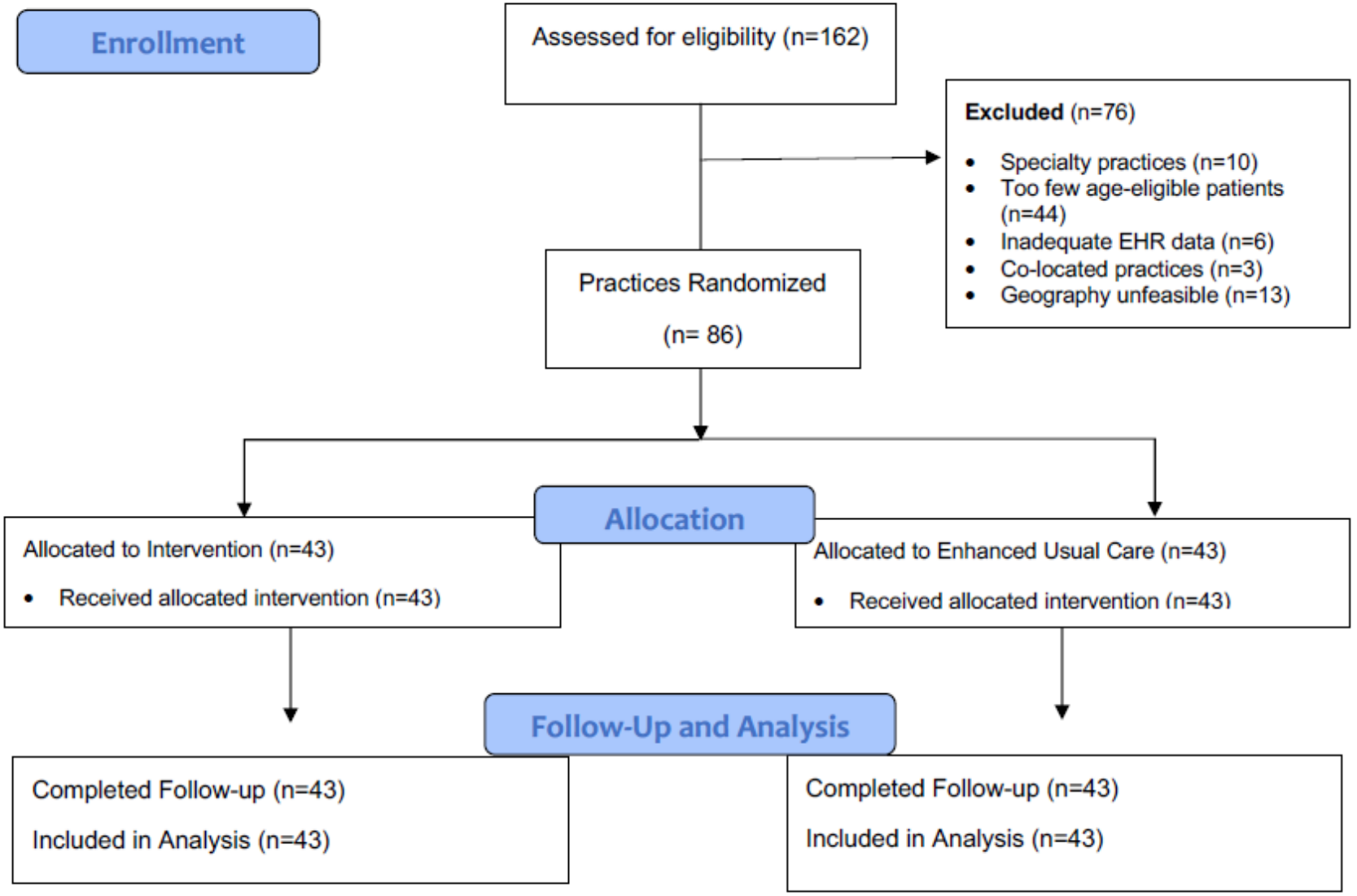
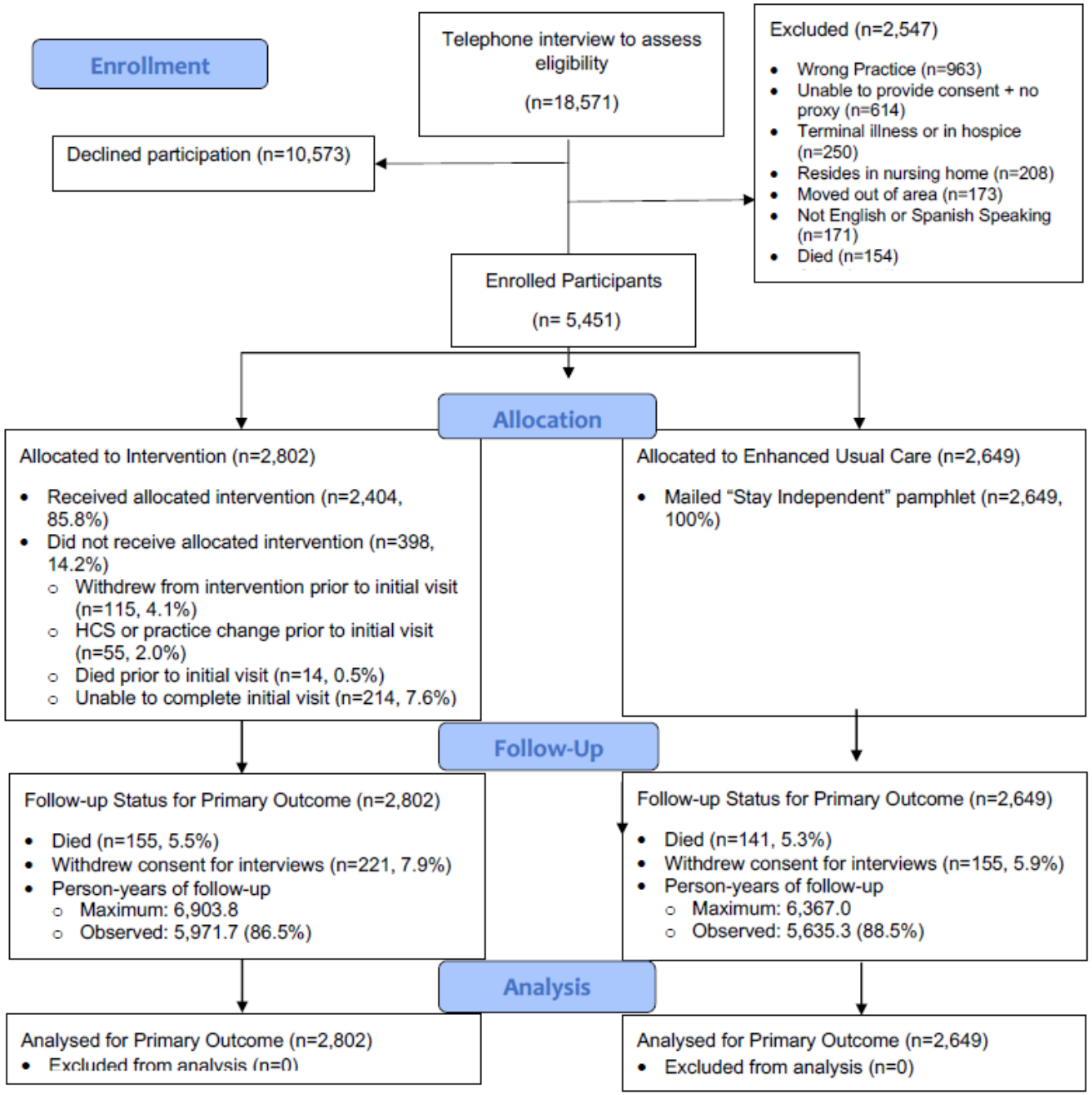
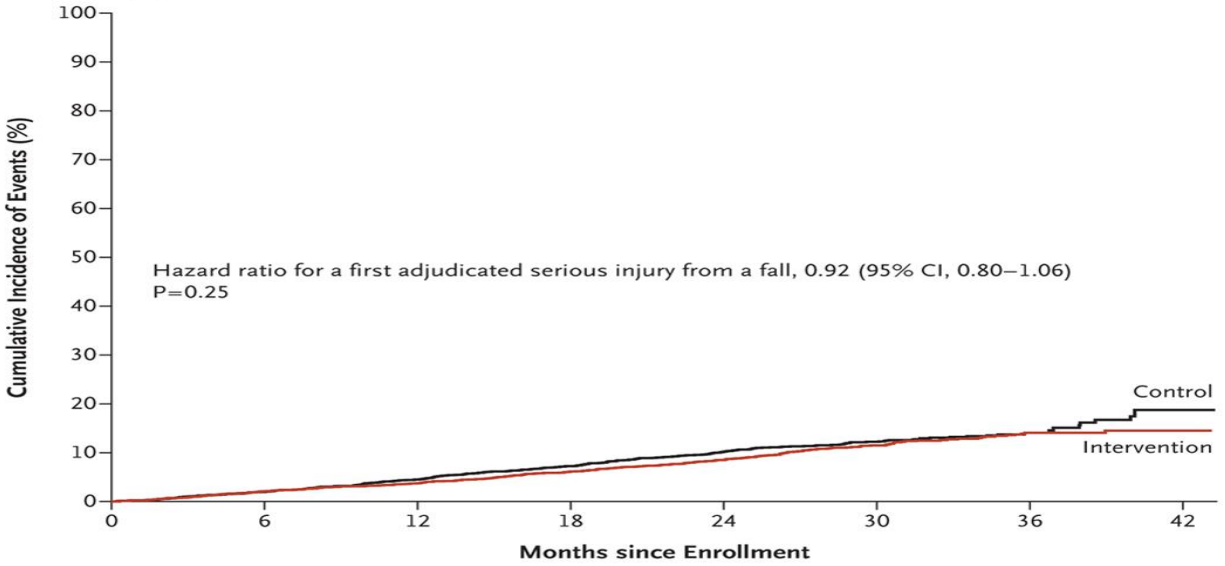


Figure S2. CONSORT Diagram Showing Flow of Participants Through the Study.



A First Adjudicated Serious Injury from a Fall



No. at Risk (cumulative no. of events)								
Control	2649 (0)	2457 (50)	2307 (113)	2146 (179)	1816 (248)	924 (279)	398 (294)	5 (301)
Intervention	2802 (0)	2566 (56)	2423 (98)	2251 (158)	1951 (215)	1054 (267)	437 (290)	3 (291)

B First Participant-Reported Injury from a Fall

No. at Risk (cumulative no. of events)								
Control	2649 (0)	2194 (333)	1810 (650)	1494 (898)	1156 (1091)	553 (1182)	220 (1224)	3 (1238)
Intervention	2802 (0)	2320 (308)	1968 (582)	1667 (802)	1300 (1005)	648 (1142)	245 (1202)	2 (1211)

Conclusions[2]

- Intervention less effective (did not reach 20% reduction)
  - Unexpected finding since previous trials showed benefits with respect to individual components
- Potential reasons
  - Adherence to intervention – potential difficulties implementing recommendations
  - No resources directly provided by the trial (real-world implementation) only referral
  - No monitoring of adherence – may have fallen below threshold
  - Participants chose recommendations – potentially valuable recommendations not implemented
  - Less effective approaches may have been chosen to address risk factors
  - A non-trivial (14.2%) proportion received no intervention

Summarization of conducting trial

- Challenging: what is easy on paper is not easy in practice
- Often required to make substantive changes to protocol to deal with issues within a trial
- Not all data are the same or can be obtained in the same time frame
- Sometimes serious concerns don’t arise until after the trial is underway or due to changes in the protocol (e.g., ascertainment bias issues).

STRIDE References

1. Bhasin S, Gill TM, Reuben DB, Latham NK, Gurwitz JH, Dykes P, McMahon S, Storer TW, Duncan PW, Ganz DA, Basaria S, Miller ME, Travison TG, Greene EJ, Dziura J, Esserman D, Allore H, Carnie MB, Fagan M, Hanson C, Baker D, Greenspan SL, Alexander N, Ko F, Siu AL, Volpi E, Wu AW, Rich J, Waring SC, Wallace R, Casteel C, Magaziner J, Charpentier P, Lu C, Araujo K, Rajeevan H, Margolis S, Eder R, McGloin JM, Skokos E, Wiggins J, Garber L, Clauser SB, Correa-De-Araujo R, Peduzzi P. Strategies to Reduce Injuries and Develop Confidence in Elders (STRIDE): A Cluster-Randomized Pragmatic Trial of a Multifactorial Fall Injury Prevention Strategy: Design and Methods. J Gerontol A Biol Sci Med Sci. 2018 Jul 9;73(8):1053-1061. doi: 10.1093/gerona/glx190. PMID: 29045582; PMCID: PMC6037050.

2. Bhasin S, Gill TM, Reuben DB, Latham NK, Ganz DA, Greene EJ, Dziura J, Basaria S, Gurwitz JH, Dykes PC, McMahon S, Storer TW, Gazarian P, Miller ME, Travison TG, Esserman D, Carnie MB, Goehring L, Fagan M, Greenspan SL, Alexander N, Wiggins J, Ko F, Siu AL, Volpi E, Wu AW, Rich J, Waring SC, Wallace RB, Casteel C, Resnick NM, Magaziner J, Charpentier P, Lu C, Araujo K, Rajeevan H, Meng C, Allore H, Brawley BF, Eder R, McGloin JM, Skokos EA, Duncan PW, Baker D, Boult C, Correa-de-Araujo R, Peduzzi P; STRIDE Trial Investigators. A Randomized Trial of a Multifactorial Strategy to Prevent Serious Fall Injuries. N Engl J Med. 2020 Jul 9;383(2):129-140. doi: 10.1056/NEJMoa2002183. PMID: 32640131; PMCID: PMC7421468.

3. Ganz DA, Yuan AH, Greene EJ, Latham NK, Araujo K, Siu AL, Magaziner J, Gurwitz JH, Wu AW, Alexander NB, Wallace RB, Greenspan SL, Rich J, Volpi E, Waring SC, Dykes PC, Ko F, Resnick NM, McMahon SK, Basaria S, Wang R, Lu C, Esserman D, Dziura J, Miller ME, Travison TG, Peduzzi P, Bhasin S, Reuben DB, Gill TM. Effect of the STRIDE fall injury prevention intervention on falls, fall injuries, and health-related quality of life. J Am Geriatr Soc. 2022 Nov;70(11):3221-3229. doi: 10.1111/jgs.17964. Epub 2022 Aug 6. PMID: 35932279; PMCID: PMC9669115.

4. Gill TM, McGloin JM, Latham NK, Charpentier PA, Araujo KL, Skokos EA, Lu C, Shelton A, Bhasin S, Bianco LM, Carnie MB, Covinsky KE, Dykes P, Esserman DA, Ganz DA, Gurwitz JH, Hanson C, Nyquist LV, Reuben DB, Wallace RB, Greene EJ. Screening, Recruitment, and Baseline Characteristics for the Strategies to Reduce Injuries and Develop Confidence in Elders (STRIDE) Study. J Gerontol A Biol Sci Med Sci. 2018 Oct 8;73(11):1495-1501. doi: 10.1093/gerona/gly076. PMID: 30020415; PMCID: PMC6175032.

5. Esserman DA, Gill TM, Miller ME, Greene EJ, Dziura JD, Travison TG, Meng C, Peduzzi PN. A case study of ascertainment bias for the primary outcome in the Strategies to Reduce Injuries and Develop Confidence in Elders (STRIDE) trial. Clin Trials. 2021 Apr;18(2):207-214. doi: 10.1177/1740774520980070. Epub 2021 Mar 7. PMID: 33678038; PMCID: PMC8009806.

6. Greene EJ, Peduzzi P, Dziura J, Meng C, Miller ME, Travison TG, Esserman D. Estimation of ascertainment bias and its effect on power in clinical trials with time-to-event outcomes. Stat Med. 2021 Feb 28;40(5):1306-1320. doi: 10.1002/sim.8842. Epub 2020 Dec 14. PMID: 33316841; PMCID: PMC9007163.

7. Zhou B, Fine J, Latouche A, Labopin M (2011). Competing risks regression for clustered data. Biostatistics.13(3):371-383.

QUESTIONS?