Methodological and Statistical Issues in Research Proposals

Thomas Travison, Dae Kim,
Fah Vasunilashorn, Rich Jones
NIDUS/CEDARTREE

5th Annual Delirium Boot Camp
November 3, 2017
The Inn at Longwood Medical

1. Pilot studies
2. Measurement validation studies
3. Power and sample size

Part 1

Common design and methodological issues in clinical trials pilot and feasibility studies

Pilot study: definition

A small-scale study or experiment intended to inform the design of, or the decision as to whether to conduct, a larger study.

Typically focuses specific attention on aspects of research methodology including choice of measurements, suitability of research environment, participant availability, and resource allocations.

Not intended to develop target 'effect size,' but *may* be used to *inform* definition of minimal clinically important differences

Pilot v. “Pilot”

2. Conduct good-faith effort to understand options:

Three other existing programs incorporating [redacted] *had more than 10 patients per year* [**identifying criterion for minimally reasonable change in design**];

these demonstrated total potential enrollees numbering [redacted] [**providing quantitative information supporting conclusions**]

3. State conclusions clearly

Based on this new pilot study, we concluded there are no additional feasible programs from which we can expand enrollment.

Purpose of a pilot study

To provide preliminary information about **feasibility** of doing a definitive study / trial

- Are participants truly available?
  - How many must be screened to enroll?
- Can measurements be done?
- Is the environment suitable?
- Is resourcing adequate (in particular: time)?

Purposes of a pilot study

To provide preliminary information about **measurement variation**

- What is the degree of natural (biological) variation in endpoints?
- Are there stratifying factors that are critical to consider at design time? Can acknowledging these help us overcome variation?
- Are there potential confounders?

Purposes of a pilot study

To provide information about **outcomes performance characteristics**

- Are measurements affiliated with gold standards or reference standards?
- Do measurements display intra- and inter-rater reliability?

Purposes of a pilot study

To provide information about **sample selection**

- Are there particular subpopulations ill-suited to enrollment? Should certain populations be over-represented?
- Is there evidence that efficacy / effectiveness or safety may vary across subpopulations?
- Are there potential confounders?

Purposes of a pilot study

To look for suggestive evidence of **efficacy/effectiveness**

- Is there evidence suggesting a favorable treatment effect?

Pilot v. “Pilot”

J Gerontol A Biol Sci Med Sci. 2007 Nov;62(11):1237-43.

Designing clinical trials of interventions for mobility disability: results from the lifestyle interventions and independence for elders pilot (LIFE-P) trial.

Espeland MA[1], Gill TM, Guralnik J, Miller ME, Fielding R, Newman AB, Pahor M; Lifestyle Interventions and Independence for Elders Study Group.

Author information

Abstract

BACKGROUND: Clinical trials to assess interventions for mobility disability are critically needed; however, data for efficiently designing such trials are lacking.

METHODS: Results are described from a pilot clinical trial in which 424 volunteers aged 70-89 years were randomly assigned to one of two interventions-physical activity or a healthy aging education program-and followed for a planned minimum of 12 months. We evaluated the longitudinal distributions of four standardized outcomes to contrast how they may serve as primary outcomes of future clinical trials: ability to walk 400 meters, ability to walk 4 meters in < or =10 seconds, a physical performance battery, and a questionnaire focused on physical function.

RESULTS: Changes in all four outcomes were interrelated over time. The ability to walk 400 meters as a dichotomous outcome provided the smallest sample size projections (i.e., appeared to be the most efficient outcome). It loaded most heavily on the underlying latent variable in structural equation modeling with a weight of 80%. A 4-year trial based on the outcome of the 400-meter walk is projected to require N = 962-2234 to detect an intervention effect of 30%-20% with 90% power.

CONCLUSIONS: Future clinical trials of interventions designed to influence mobility disability may have greater efficiency if they adopt the ability to complete a 400-meter walk as their primary outcome.

Pilot v. “Pilot”

Heart Lung. 2017 Jul - Aug;46(4):234-238. doi: 10.1016/j.hrtlng.2017.05.002. Epub 2017 Jun 9.

Delirium prevention in critically ill adults through an automated reorientation intervention - A pilot randomized controlled trial.

Munro CL[1], Cairns P[2], Ji M[2], Calero K[3], Anderson WM[3], Liang Z[2].

Author information

1 University of South Florida College of Nursing, 12901 Bruce B. Downs Blvd, MDC 22, Tampa, FL 33612-4766, USA. Electronic address: cmunro2@health.usf.edu.

2 University of South Florida College of Nursing, 12901 Bruce B. Downs Blvd, MDC 22, Tampa, FL 33612-4766, USA.

3 University of South Florida Morsani College of Medicine, 12901 Bruce B. Downs Blvd, MDC 19, Tampa, FL 33612-4766, USA.

Abstract

OBJECTIVES: Explore the effect of an automated reorientation intervention on ICU delirium in a prospective randomized controlled trial.

BACKGROUND: Delirium is common in ICU patients, and negatively affects outcomes. Few prevention strategies have been tested.

METHODS: Thirty ICU patients were randomized to 3 groups. Ten received hourly recorded messages in a family member's voice during waking hours over 3 ICU days, 10 received the same messages in a non-family voice, and 10 (control) did not receive any automated reorientation messages. The primary outcome was delirium free days during the intervention period (evaluated by CAM-ICU). Groups were compared by Fisher's Exact Test.

RESULTS: The family voice group had more delirium free days than the non-family voice group, and significantly more delirium free days ($p = 0.0437$) than the control group.

CONCLUSIONS: Reorientation through automated, scripted messages reduced incidence of delirium. Using identical scripted messages, family voice was more effective than non-family voice.

Pilot v. “Pilot”

Unpublished substudy undertaken to demonstrate feasibility of enrollment plan and satisfy a reviewer critique:

Reviewer Question: Are there additional participants / programs that could be used to broaden the study population?

1. **Identify resource and/or operational constraints essential to design; state clearly. Do not compromise or you will pay later:**

   Pilot work demonstrated that over 80% of eligible subjects would refuse participation in a study that [redacted]

   Therefore, since [redacted] *is essential to our study design, any enrolling program must have a large patient base.*

Case studies

ABSTRACT

The proposed project is a randomized, double-blind, placebo-controlled phase II trial of oral prolonged release melatonin for the treatment of delirium in older people with cancer. This study proposes to test the feasibility of conducting a future phase III randomized controlled trial by evaluating recruitment and retention rates and providing preliminary data for proposed efficacy and toxicity endpoints to inform the appropriate design of a phase III trial...Outcomes include feasibility (percentage of eligible participants screened who progress to be randomized, percentage who complete the study intervention, screened participants who meet the eligibility criteria, and reasons for non-eligibility, and ineligible participants screened) Delirium status and severity (Delirium Rating Scale – revised 98), Richmond Agitation Sedation Scale – Palliative, in-hospital medical complications and other toxicity (National Cancer Institute Common Terminology Criteria). The primary endpoint is defined as 60% or more participants completing study intervention in phase II pilot trial until delirium resolution or the following time-points lack of response at 10 days, withdrawal due to toxicity or death.

ABSTRACT

...We propose a single-blind, prospective, randomized controlled two-arm pilot trial to determine feasibility of music intervention in the ICU, estimate the efficacy of music intervention to reduce delirium incidence and severity, and study the effect of music on biomarkers of inflammation and delirium...

Statistical Analysis

With a sample size of 30 participants, we will be able to estimate recruitment rate of 60% to within a 95% confidence interval of ±17.53%, and an adherence rate of 80% to within a 95% confidence interval of ±14.3%.

Recommendations

Do perform pilot study where one will assist specifically with the **design** of a valid and more definitive investigation.

- Feasibility
- Measurements / Assessment
- Population / Participant Availability

Do distinguish between pilot and “pilot”

- **Do not** over-promise in re statistical power, but **do** take into account that funders may stipulate work product (e.g. publication)
- **Do** provide assessments of statistical power and uncertainty under **reasonable** constraints

Do not use pilot to determine **effect size** unless the design is specifically tailored to determine MCID

Part 2

Measurement development studies and the choice of a reference standard

Screening

The process of identifying individuals who may be at higher risk of a disease or condition among large populations of healthy people.

https://www.gov.uk/guidance/evidence-and-recommendations-nhs-population-screening#evidence-review-proces

Case finding

Early detection of symptomatic problems before they would normally be identified

Williamson J. Screening, surveillance and case-finding. In: Arie T (ed). Health care Qf the elderly. London: Croom Helm, 1981.

Diagnostic testing

The whole point of a diagnostic test is to use it to make a diagnosis.

Alman DG. Statistics Notes: Diagnostic Tests 2: predictive values. British Medical Journal 1994;309:102

Leave psychiatric diagnosis to physicians

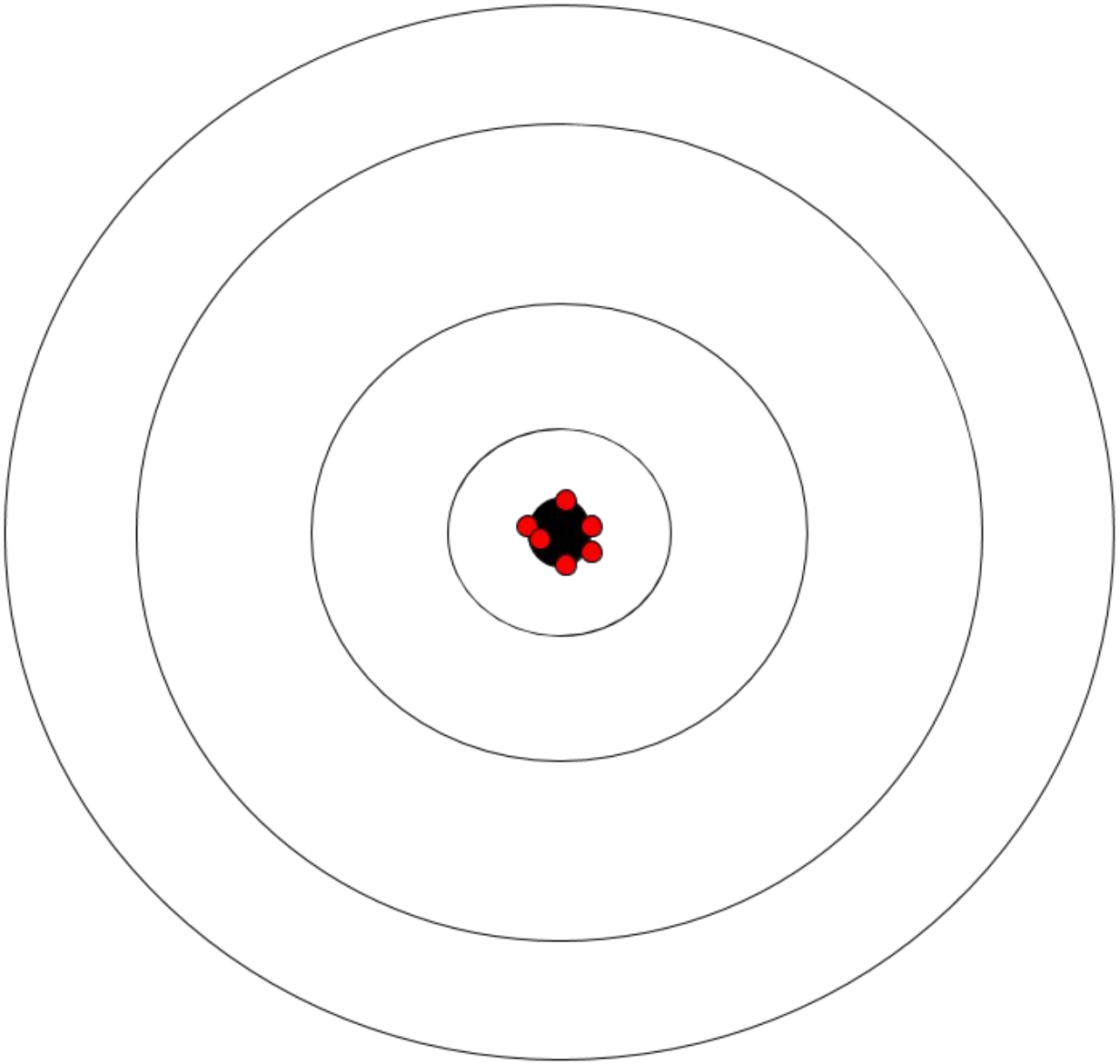
Find a euphemism, e.g.

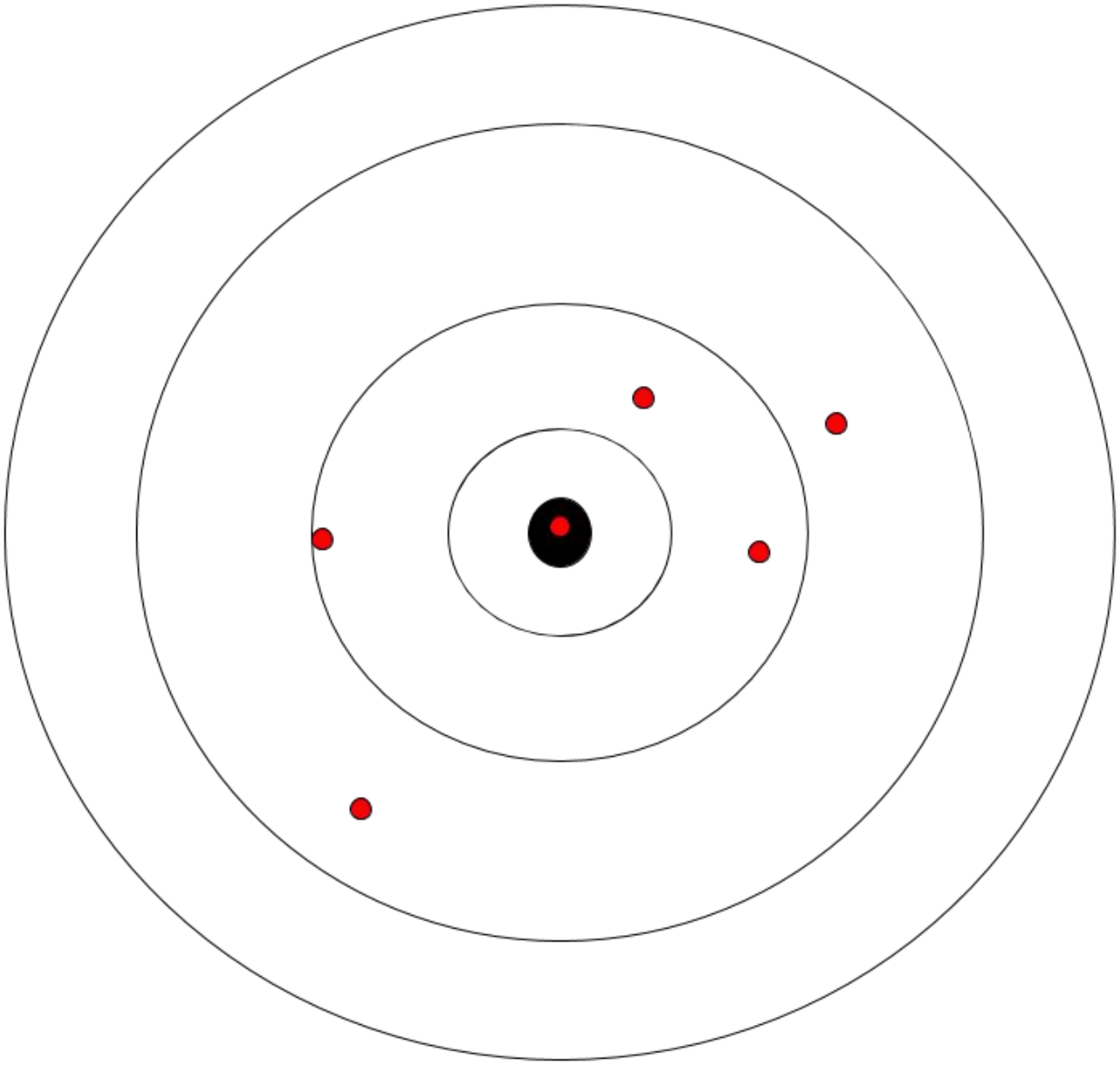
- Research diagnosis of delirium
- Research designation of delirium
- Probable delirium
- Likely delirium case

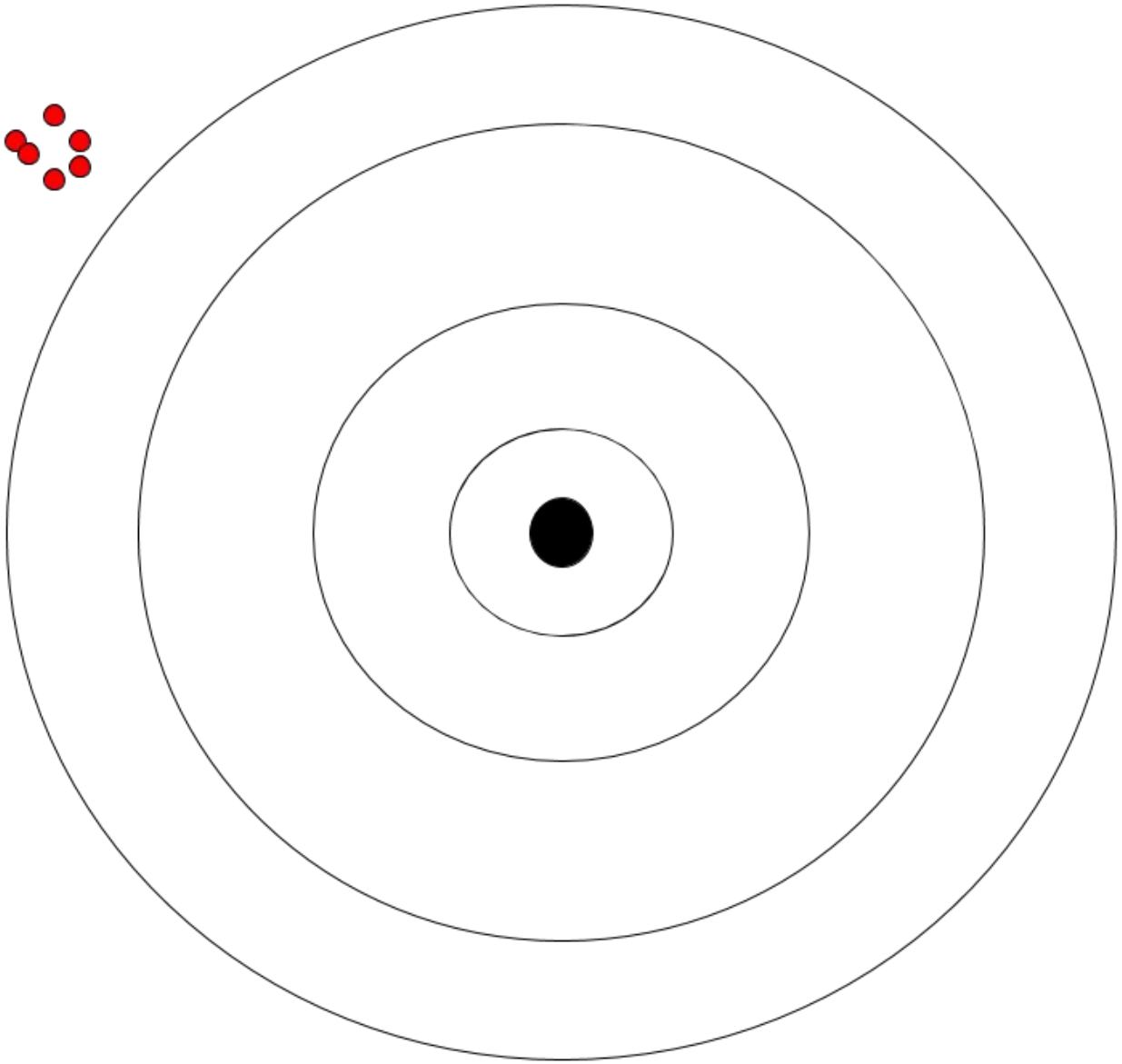
But this does not preclude reproducible procedures (consider structured evaluation)

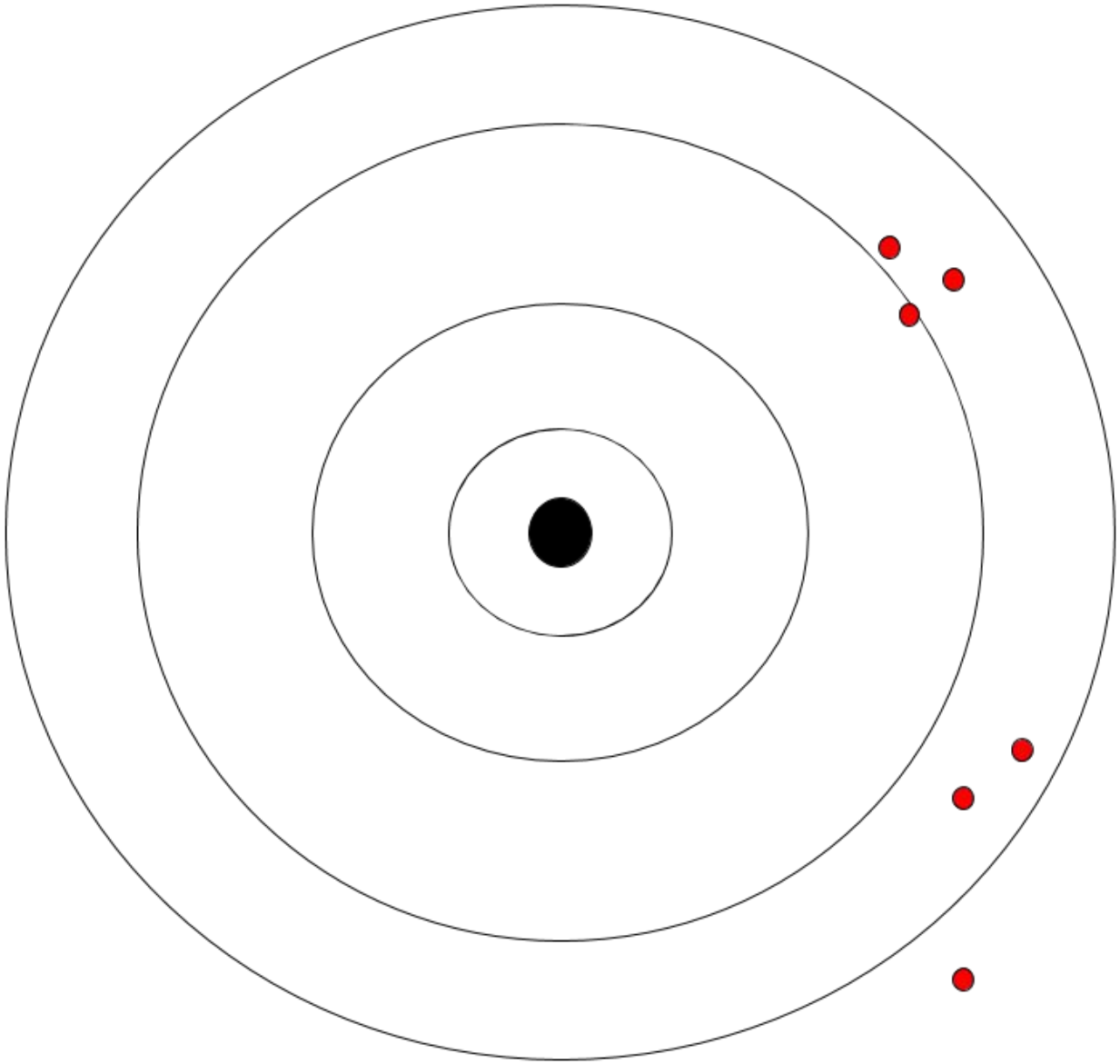
Evaluation

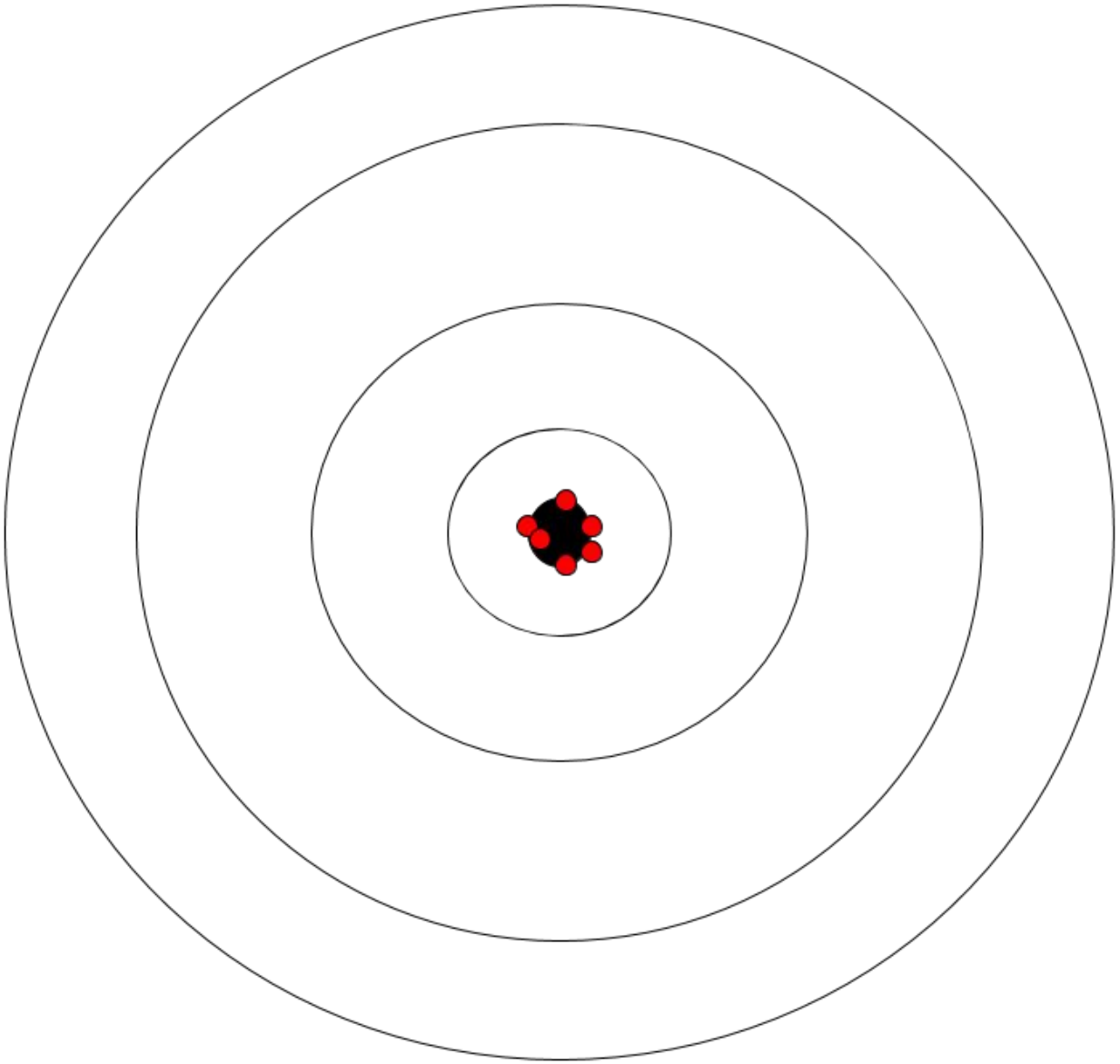
Validity (accuracy, sensitivity, specificity)
Reliability (consistency, repeatability)
Efficiency (costs, [also sometimes validity])
Generalizability (sample appropriateness)



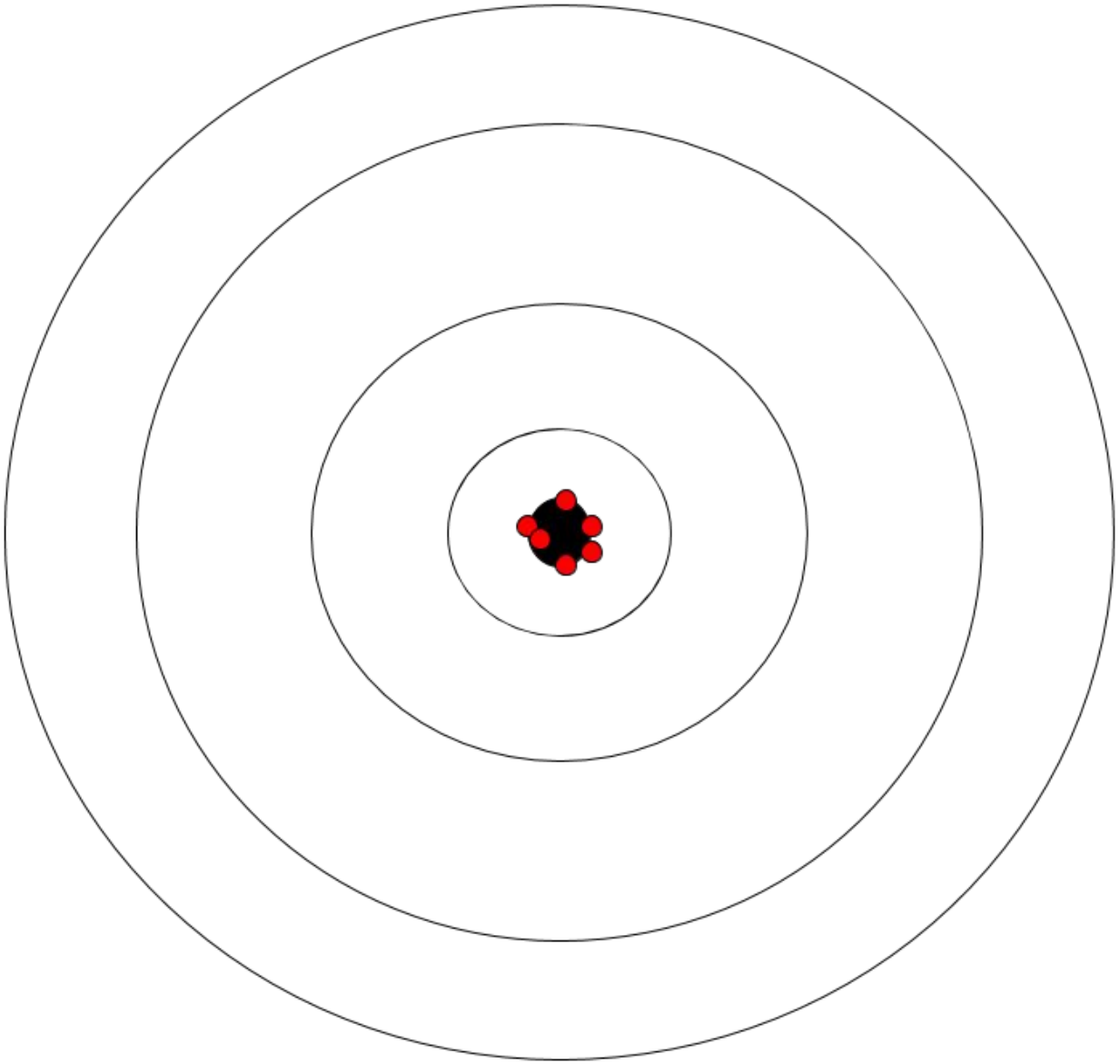


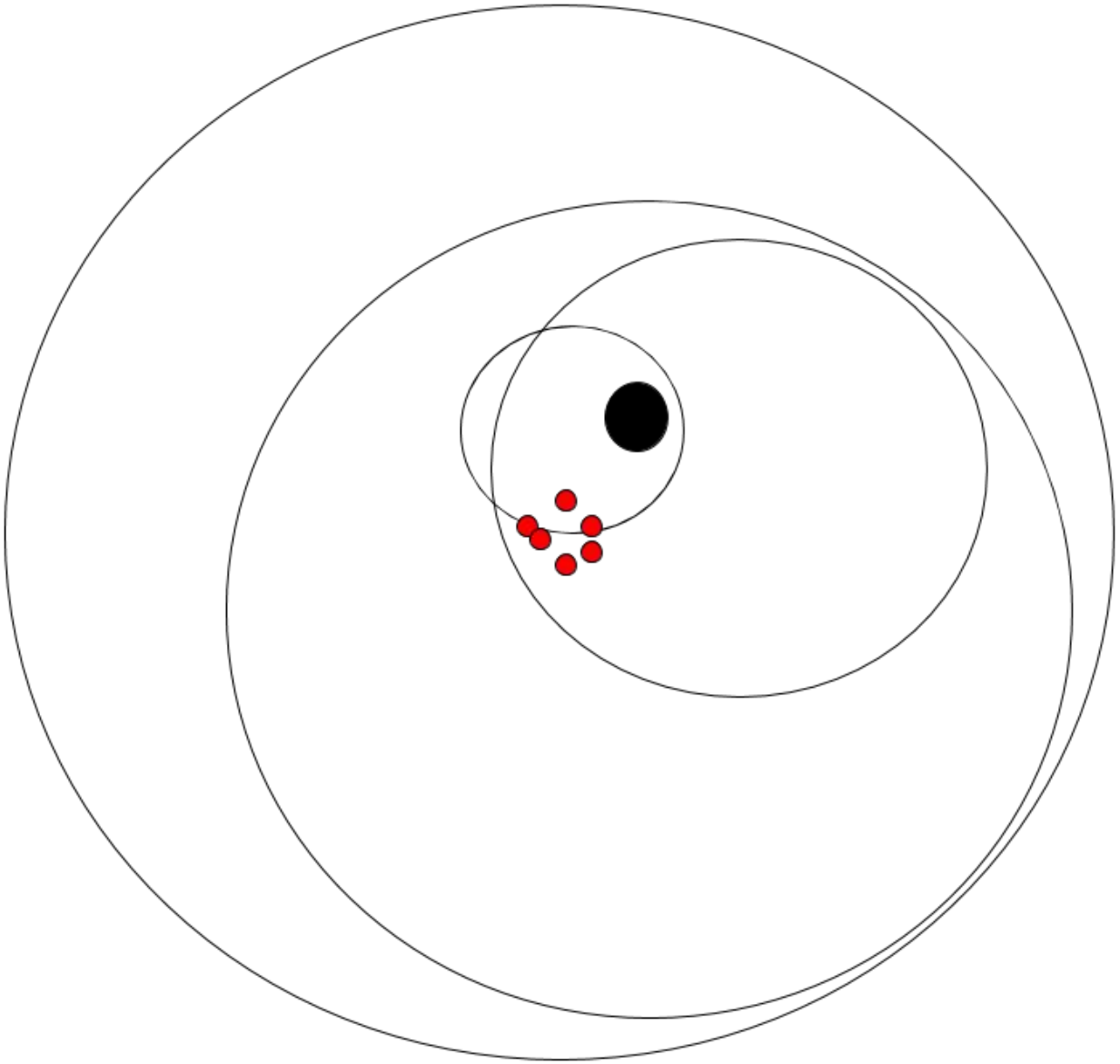






Reference standard





Strong design

Publication guidelines

STARD (EQUATOR Network)

Quality assessment tools

QUADAS-2

STARD

Standards for the reporting of diagnostic accuracy studies, a publication checklist (a la CONSORT).

Cohen JF et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6:e012799. doi:10.1136/bmjopen-2016-012799

STARD: Reference standard

Clinical reference standard.
The best available method for establishing the presence or absence of the target condition.
A gold standard would be an error-free reference standard.

STARD: Reference standard

item 10b. Reference standard described in sufficient detail to allow replication.

item 11. Rationale for choosing the reference standard (if alternatives exist).

QUADAS-2

QUODAS-2 is a tool for grading quality of diagnostic accuracy studies, as, for instance, in conducting systematic reviews of diagnostic accuracy.

Whiting PF, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine. 2011;155:529-536

QUADAS-2 on the reference standard

Describe the reference standard and how it was conducted and interpreted

QUADAS-2 on the reference standard

Is the reference standard likely to correctly classify the target condition?

QUADAS-2 on the reference standard

Were the reference standard results interpreted without knowledge of the results of the index test?

QUADAS-2 on the reference standard

Could the reference standard, its conduct or its interpretation have introduced bias?

Reference standards in delirium

Neufeld, K. J., Nelliot, A., Inouye, S. K., Ely, E. W., Bienvenu, O. J., Lee, H. B., & Needham, D. M. (2014). Delirium diagnosis methodology used in research: a survey-based study. The American Journal of Geriatric Psychiatry, 22(12), 1513-1521.

Reference standards in delirium

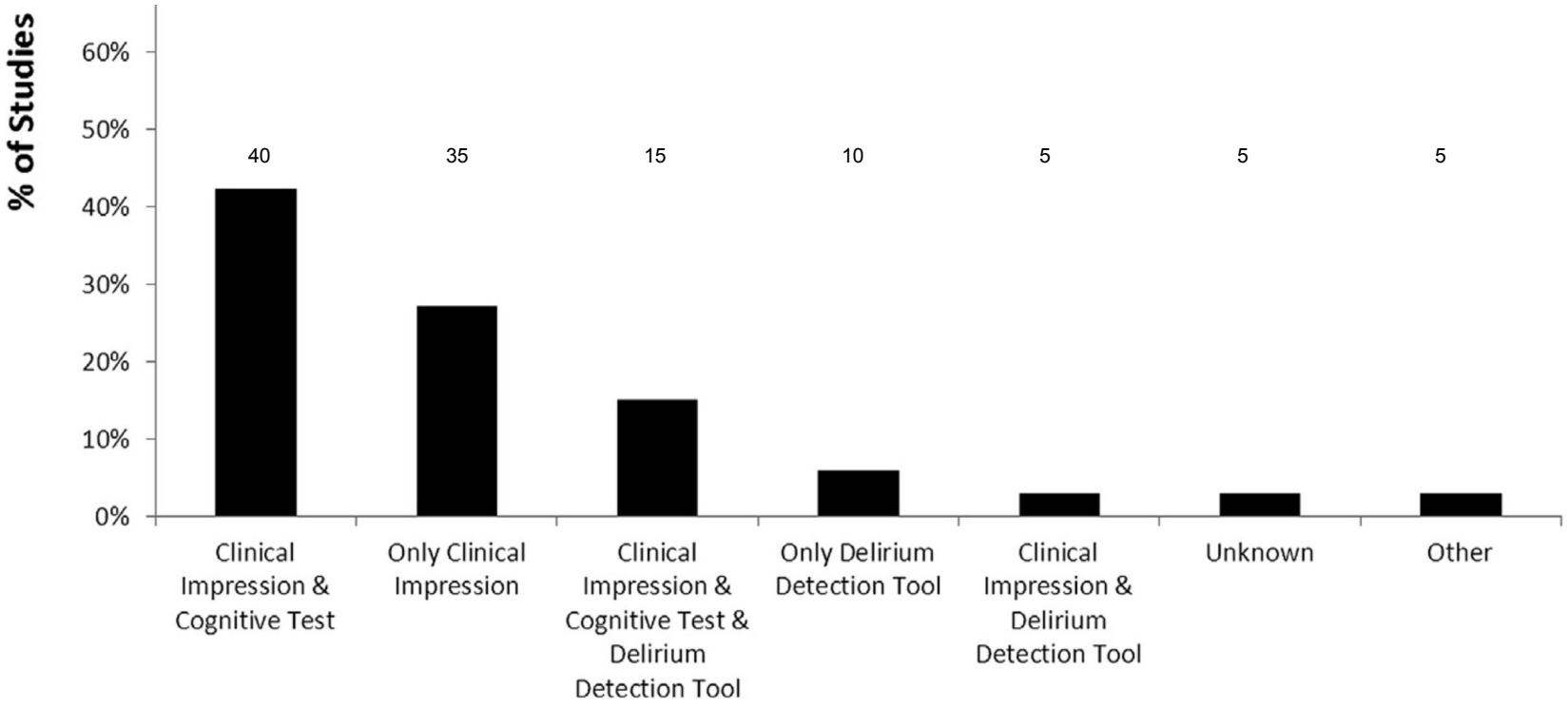
"One particular type of study that must, by design, regularly use an independent reference rater evaluation to serve as the reference standard is the development and evaluation of delirium detection tools."

Neufeld, K. J., Nelliot, A., Inouye, S. K., Ely, E. W., Bienvenu, O. J., Lee, H. B., & Needham, D. M. (2014). Delirium diagnosis methodology used in research: a survey-based study. *The American Journal of Geriatric Psychiatry, 22*(12), 1513-1521.

Reference standards in delirium

"Details of these reference rater methods are scant in most research publications"

Neufeld, K. J., Nelliot, A., Inouye, S. K., Ely, E. W., Bienvenu, O. J., Lee, H. B., & Needham, D. M. (2014). Delirium diagnosis methodology used in research: a survey-based study. *The American Journal of Geriatric Psychiatry, 22*(12), 1513-1521.



About 90% used “Clinical impression”, but in about a third of those that was it.

Neufeld, K. J., Nelliot, A., Inouye, S. K., Ely, E. W., Bienvenu, O. J., Lee, H. B., & Needham, D. M. (2014). Delirium diagnosis methodology used in research: a survey-based study. *The American Journal of Geriatric Psychiatry, 22*(12), 1513-1521.

NB: Those are Rich’s guesses at the proportions in each category. Neufeld et al did not label their bars. Take Gerald Van Belle’s (SROT 2009) advice and avoid bar charts.

Reference standards should be reproducible

e.g.,
structured exam*
semi-structured exam*

* exam: history, collateral sources, examination of patient, mental testing, review of laboratory

US-UK Study

Videotapes of diagnostic interviews with eight patients, three American and five English, were **shown to large audiences** of trained psychiatrists i**n the eastern United States and in different parts of the British Isles**...**there were major disagreements** ... the American concept of schizophrenia is much broader ... These serious differences in the usage of diagnostic terms have important implications for transatlantic communication, and indeed for international communication in general.

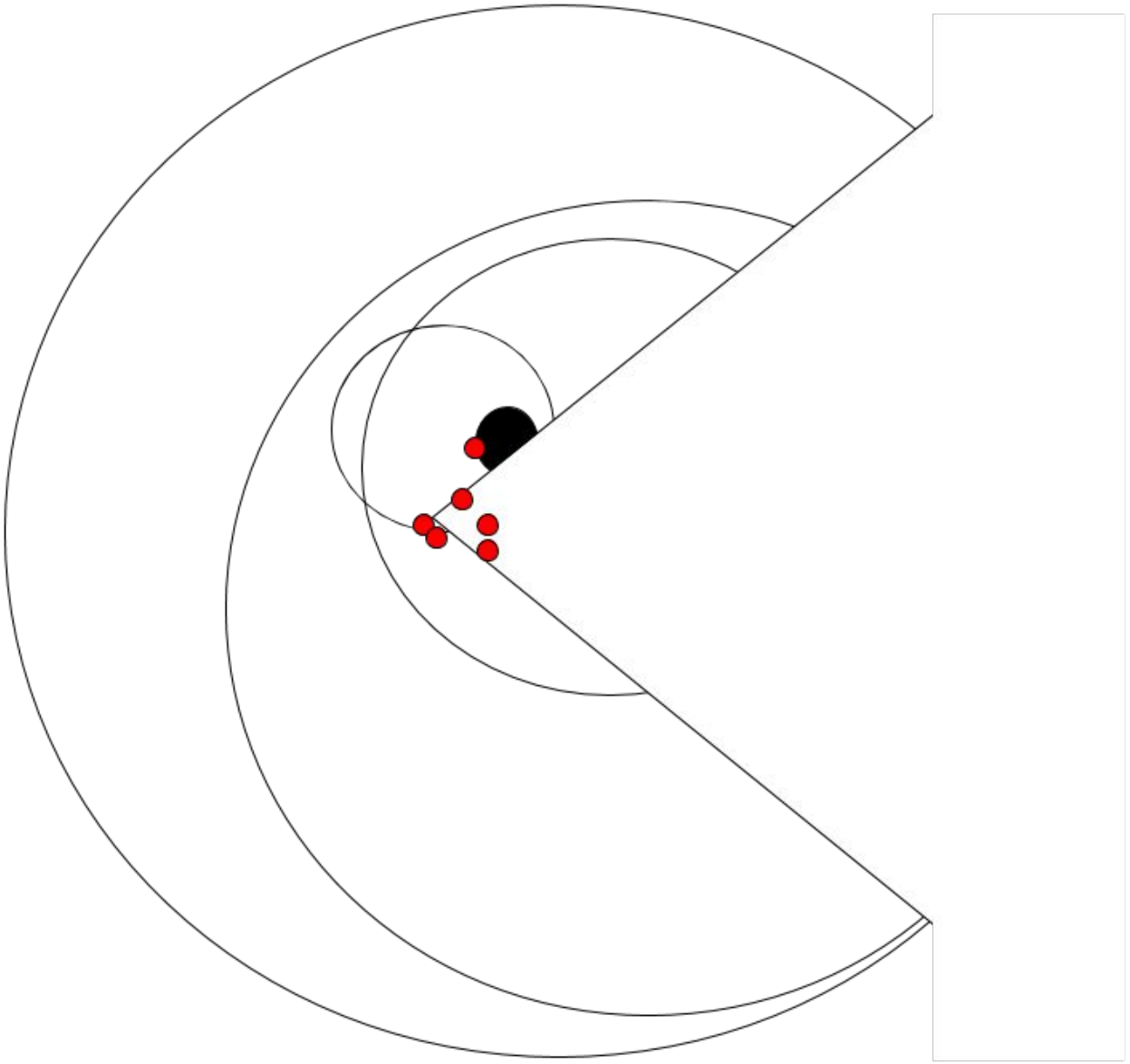
Kendell RE, et al. Diagnostic Criteria of American and British Psychiatrists. Arch Gen Psychiatry. 1971;25(2):123-130. doi:10.1001/archpsyc.1971.01750140027006

Reference standards should be “blind” to screening test

Not only how, and by whom, but when

Patients can get sick any day of the week, any time of day.

This applies to delirium too, perhaps of greater concern because the syndrome is by definition fluctuating and onsets acutely



Case studies

Concurrent validity was determined by comparing the INDEX TEST ratings with those of a psychiatrist. Concurrent validation was replicated in two clinically distinct samples. At two sites [redacted], each subject was evaluated independently by a geriatrician [redacted] and by a psychiatrist [redacted] within a maximum of 6 hours of each other. Only the geriatrician completed the INDEX TEST rating. The geriatrician and the psychiatrist were blinded to the results of each other's evaluation. The order in which patients were seen (whether by the geriatrician or the psychiatrist first) was varied. To minimize a learning effect, each investigator avoided correcting the subjects' responses.

The psychiatrist used standard psychiatric procedures to evaluate the patient and made final diagnoses in accordance with DSM[redacted] criteria. The evaluation included a detailed patient interview (complete psychiatric interview and mental status examination), family interview, medical record review, and nurse interview (for inpatients). The final diagnoses of the psychiatrist, as confirmed by follow-up medical record review, were the reference standard against which the INDEX TEST was assessed.

In a standardized interview, the geriatrician administered the Mini-Mental State Examination (20), checked immediate recall of a story (49), and made a INDEX TEST rating and a rating on the Visual Analog Scale for Confusion (20). In addition, either a family member or another observer (a nurse or physician) was interviewed briefly to assess whether an acute change in mental status had been noted. At site 2, an additional attention task, digit span forward (50, 51), was done. The geriatrician did not use information from any source other than the standardized interviews to score the INDEX TEST.

- RS is **described**, how conducted and interpreted (Q)
- RS is potentially **reproducible** (SD)
- RS is likely to **correctly classify** the target condition (Q)
- RS results are interpreted **without knowledge** of the results of the index test (Q)
- RS conduct and interpretation are unlikely to introduce **bias** (Q)
- RS choice is **justified** (SD)

[Is this a measurement development study?]

After developing the [redacted] score, the score is validated by an appeal to predictive validity (Aim 2)...This validation work would be executed within a retrospective cohort study of hospitalized patients using EHR data...The primary statistical question to be addressed in Aim 2 reflects the merits of [redacted] scoring as a valid predictor of delirium outcome. The primary outcome for this analysis will be thirty-day readmission with incident delirium...

- RS is **described**, how conducted and interpreted (Q)
- RS is potentially **reproducible** (SD)
- RS is likely to **correctly classify** the target condition (Q)
- RS results are interpreted **without knowledge** of the results of the index test (Q)
- RS conduct and interpretation are unlikely to introduce **bias** (Q)
- RS choice is **justified** (SD)

...The care of delirious patients cannot be improved unless delirium can be accurately diagnosed...

Study Design: We will conduct a pilot tool validation study according to established standards.33[<- STARD]

Reference Standard- [An ICU research nurse will conduct an assessment of delirium daily using DSM-5 criteria within two hours of the family assessments. All delirium assessments (i.e., nurse, family, research nurse) will be conducted independently and blinded to the other assessments. Assessors will be instructed not to discuss assessments with each other, though discussion of clinically relevant issues will not be precluded.] Consensus on the diagnosis of delirium will be reached between the ICU research nurse's reference standard assessment and a neuropsychiatrist [redacted] weekly, per DSM-5 criteria. To ensure standardized methodology is employed, the ICU research nurse and the neuropsychiatrist will independently conduct a minimum of 10 assessments during a pre-study training period, and will meet to establish a consensus on diagnosis (until kappa reaches ≥0.80) before recruitment proceeds.

- RS is **described**, how conducted and interpreted (Q)
- RS is potentially **reproducible** (SD)
- RS is likely to **correctly classify** the target condition (Q)
- RS results are interpreted **without knowledge** of the results of the index test (Q)
- RS conduct and interpretation are unlikely to introduce **bias** (Q)
- RS choice is **justified** (SD)

Recommendations (1)

Design your study in accordance with published quality reporting guidelines (STARD, QUADAS2).

Highlight this in a proposal section labeled "***Rigor, Transparency, and Reproducibility***"

Recommendations (2)

Delirium is fluctuating by definition. Strategize and make explicit your plan to capture this with your reference standard.

Help advance the field by using a structured reference standard diagnostic assessment. Or at least semi-structured.

Consider supplement structured assessment with structured chart review and caregiver interview.

Part 3

A checklist for preparing a complete sample size justification

Sample size/power

- Each aim has a power/sample/minimum detectable effect size documented
- Match between model for power/sample size and planned analysis
- Estimates on which power/sample size are based are
  - Appropriate
  - Derive from adequately powered preliminary studies or otherwise well justified
- Clarity and transparency in power/sample size presentation

Bookmarks

Checklist

https://goo.gl/lGjfYY

Explanatory text

https://sites.google.com/site/ifarwf/home/samplesizeandpower

SAMPLE SIZE AND POWER ANALYSIS CHECKLIST

Click here to read an elaboration and explanation of the points below.

☐ **1. Effect size specification.** The *minimum detectable* effect size is specified, OR, the *hypothesized effect* size is specified
- ☐ Each hypothesis/aim has its own effect size specification

☐ **2. Effect size justification.** There is a scientific justification for the effect size to be detected. The effect to be detected must be convincingly...
- ☐ Feasible to observe
- ☐ Based on previously published or pilot data from adequately powered studies, an essential part of scientific rigor
- ☐ Clinical relevance/practical importance of effect to be detected is justified
- ☐ Each hypothesis/aim has its own effect size justification

☐ **3. Sample size specification.** The *number* of units to be studied (people, animals) is specified

☐ **4. Sample size justification.** The proposal contains sufficient justification that the number of units to be studied is both
- ☐ Feasible; the population and recruitment plan will support target enrollment
- ☐ Achievable; the research team has the capacity to achieve goals

☐ **5. Statistical power.** Statistical power is a function of the hypothesized effect size and the sample size
- ☐ Each hypothesis/aim needs its own statement of statistical power

☐ **6. Special design considerations.** The following design considerations are reflected in the power analysis and sample size justification
- ☐ Attrition or loss to follow-up
- ☐ Missing data from other causes
- ☐ Clustering or other non-independence of observations
- ☐ Repeated measurements of outcome variables
- ☐ Multiple comparisons in terms of multiple outcomes and/or pre-planned subgroup analyses, and whether or not to adjust significance levels
- ☐ Preliminary analysis, or early stopping rules and their effect on P-values
- ☐ Covariates or other confounders and adjustment accounted for

☐ **7. Strategies to deal with challenges.** Strategies are described for monitoring and dealing with shortfalls in achieving target sample

☐ **8. Presentation details**
- ☐ Jargon is avoided
- ☐ Assumptions are conservative and justified
- ☐ Simple methods that can be easily confirmed are used
- ☐ Methods match between the power or sample size calculation and the analysis to be performed

November 1, 2016
Rich Jones (rich_jones@brown.edu)
Quantitative Science Program, Department of Psychiatry and Human Behavior, Department of Neurology
Warren Alpert Medical School, Brown University

SAMPLE SIZE AND POWER ANALYSIS CHECKLIST

Click here to read an elaboration and explanation of the points below.

☐ **1. Effect size specification.** The *minimum detectable* effect size is specified, OR, the *hypothesized effect* size is specified

- ☐ Each hypothesis/aim has its own effect size specification

☐ **2. Effect size justification.** There is a scientific justification for the effect size to be detected. The effect to be detected must be convincingly...

- ☐ Feasible to observe
- ☐ Based on previously published or pilot data from adequately powered studies, an essential part of scientific rigor
- ☐ Clinical relevance/practical importance of effect to be detected is justified
- ☐ Each hypothesis/aim has its own effect size justification

☐ **3. Sample size specification.** The *number* of units to be studied (people, animals) is specified

☐ **4. Sample size justification.** The proposal contains sufficient justification that the number of units to be studied is both

- ☐ Feasible; the population and recruitment plan will support target enrollment
- ☐ Achievable; the research team has the capacity to achieve goals

☐ **5. Statistical power.** Statistical power is a function of the hypothesized effect size and the sample size

- ☐ Each hypothesis/aim needs its own statement of statistical power

☐ **6. Special design considerations.** The following design considerations are reflected in the power analysis and sample size justification

- ☐ Attrition or loss to follow-up
- ☐ Missing data from other causes
- ☐ Clustering or other non-independence of observations
- ☐ Repeated measurements of outcome variables
- ☐ Multiple comparisons in terms of multiple outcomes and/or pre-planned subgroup analyses, and whether or not to adjust significance levels
- ☐ Preliminary analysis, or early stopping rules and their effect on P-values
- ☐ Covariates or other confounders and adjustment accounted for

☐ **7. Strategies to deal with challenges.** Strategies are described for monitoring and dealing with shortfalls in achieving target sample

☐ **8. Presentation details**

- ☐ Jargon is avoided
- ☐ Assumptions are conservative and justified
- ☐ Simple methods that can be easily confirmed are used
- ☐ Methods match between the power or sample size calculation and the analysis to be performed

Example

Aim	To test the hypothesis that aspirin use reduces headache pain. We will evaluate this hypothesis in the context of a double-blind, placebo-controlled randomized controlled trial.
Analysis	We will use ANCOVA to test the hypothesis that self-reported headache pain 2 hours after drug administration is lower among those who received aspirin relative to those who received placebo. We will control for baseline level of headache pain in a regression framework.
Power/SS	We have determined that a clinically meaningful change in the headache impact score is 8 points or more (about 0.5 standard deviation units). The minimum sample size to detect an effect of that magnitude, under the conservative assumption that baseline pain and follow-up pain are uncorrelated, is 64 persons per group to achieve a type-II error rate of 20% and type-I error rate of 5% (Lehr, 1992).

Power/SS	We have determined that a clinically meaningful change in the headache impact score is 8 points or more (about 0.5 standard deviation units). The minimum sample size to detect an effect of that magnitude, under the conservative assumption that baseline pain and follow-up pain are uncorrelated, is 64 persons per group to achieve a type-II error rate of 20% and type-I error rate of 5% (Lehr, 1992).
Missing Data	We will analyze our data under an intent-to-treat framework, and use multiple imputation with 50 imputations to account for missing data. We will conduct extreme value sensitivity analyses for data that are missing as a check on our inferences. We expect 20% missing data, and are therefore inflating our sample size requirement to 80 per group to ensure an evaluable effective sample of 64 persons per group.

Missing Data	We will analyze our data under an intent-to-treat framework, and use multiple imputation with 50 imputations to account for missing data. We will conduct extreme value sensitivity analyses for data that are missing as a check on our inferences. We expect 20% missing data, and are therefore inflating our sample size requirement to 80 per group to ensure an evaluable effective sample of 64 persons per group.
Justification	Prior research by Soandso et al (2014) has demonstrated that the minimum clinically important difference on the headache impact scale is 8 points.
Feasibility of effect	Previously, Whozatnow et al (2014) demonstrated a 0.3SD difference on the headache impact scale between persons appearing an emergency room who had versus those who had not taken an aspirin within the past 2 hours. Therefore we believe our target effect size to detect of 0.5SD units is feasible to observe.

Feasibility of effect	Previously, Whozatnow et al (2014) demonstrated a 0.3SD difference on the headache impact scale between persons appearing an emergency room who had versus those who had not taken an aspirin within the past 2 hours. Therefore we believe our target effect size to detect of 0.5SD units is feasible to observe.
Feasibility of sample	Dr. Seniorgal, co-investigator, recently completed a RCT of ibuprofen for headache pain and used similar sampling and recruitment methodologies as we propose in this study. In that study we were able to recruit 120 patients in a similar time frame. Therefore, we believe our target sample size is feasible to accrue.

Case studies

Aim	Aim 1. To describe the predictive risk factors for [redacted] in a critically ill cohort.
Analysis	For Aim #1b, we will model the risk factors for [redacted] compared to the other three groups using a multinomial regression model.
Power/SS	Analysis for Aim #1b will use multivariable linear regression to account *for potential confounders introduced by imbalances* in groups created by death and loss to follow-up. Hence sample size or allowable model complexity for this aim will be based on the general rule that a model must fit no more than m/10 parameters to allow for proper multivariable analysis and to be generalizable to future patients, where m is the effective sample size.

Aim	To determine the interaction of X and Y on outcome in a cohort of critically ill patients.
Analysis	Test the hypothesis that patients with X and Y will have the highest risk of outcome (vs. what?; and, two independent risk factors will produce group with highest risk of outcome without interaction)
Power/SS	Our study with N subjects will have greater than 80% power to detect a standardized effect size of 0.2 at the 3-month time point (Does the effect size reflect the main effect or interaction effect? Consider the correlation between X and Y; if X and Y are highly correlated, the power to detect the interaction effect may be limited)

Aim	To identify if X is a risk factor for mortality in a cohort of critically ill patients
Analysis	We will fit a Cox proportional hazards regression model with censoring...
Power/SS	Assuming a type I error rate of 5%, 30% of the cohort exposed to X, and 40% survival rate in the unexposed group, we estimate that N patients will have at least 80% power to detect a hazard ratio of 1.5. (Contrast these assumptions with "we expect 80% will survive at 90 days and 90% of survivors will be interviewed..." for the earlier aims)

Aim	Aim 2. To determine the interaction of [redacted] and delirium on cognitive and psychological sequelae of critical illness. We hypothesize that there is a graded effect of brain dysfunction on clinically relevant patient outcome measures.
Analysis	Aim #2 will test the hypothesis that patients with [redacted] and delirium will have the highest risk of long-term cognitive impairment and depression ... Due to the ordinal nature of all the outcomes, we will use multivariable proportional odds regression model to evaluate this hypothesis.
Power/SS	For Aim #2, using the effect size index approach (56). For power computation, our study with 288 subjects will have greater than 80% power to detect a standardized effect size of 0.2 at the 3-month time point.

Aim	[AIM 1] Characterize the trajectory of [BIOMARKERS IN PATIENTS] undergoing cardiac surgery requiring bypass.
Analysis	We will compare ... with delirium and without delirium ... using the Wilcoxon rank sum test for non-normally distributed continuous variables, Student t test for normally distributed continuous variables, or chi-square test for categorical variables. We will construct receiver operator characteristic (ROC) curves for the maximum value attained by each biomarker during the perioperative course to classify [patients] with delirium.
Power/SS	Sample size for this study was computed to provide a sufficient number of patients to provide the lower CI of 70%. Assuming a conservative prevalence of 25% for pediatric delirium, and point estimate for sensitivity of 90%, a sample size of 58 patients will be required for this pilot study.

Aim	[AIM 2] Determine the incidence of delirium [IN PATIENTS] undergoing cardiac surgery requiring bypass.
Analysis	We will compare ... with delirium and without delirium ... using the Wilcoxon rank sum test for non-normally distributed continuous variables, Student t test for normally distributed continuous variables, or chi-square test for categorical variables. We will construct receiver operator characteristic (ROC) curves for the maximum value attained by each biomarker during the perioperative course to classify [patients] with delirium.
Power/SS	Sample size for this study was computed to provide a sufficient number of patients to provide the lower CI of 70%. Assuming a conservative prevalence of 25% for pediatric delirium, and point estimate for sensitivity of 90%, a sample size of 58 patients will be required for this pilot study.

Aim	[AIM 3] Examine the association between changes in plasma biomarkers and delirium occurrence in [PATIENTS] undergoing cardiac surgery requiring cardiopulmonary bypass.
Analysis	We will compare ... with delirium and without delirium ... using the Wilcoxon rank sum test for non-normally distributed continuous variables, Student t test for normally distributed continuous variables, or chi-square test for categorical variables. We will construct receiver operator characteristic (ROC) curves for the maximum value attained by each biomarker during the perioperative course to classify [patients] with delirium.
Power/SS	Sample size for this study was computed to provide a sufficient number of patients to provide the lower CI of 70%. Assuming a conservative prevalence of 25% for pediatric delirium, and point estimate for sensitivity of 90%, a sample size of 58 patients will be required for this pilot study.

Aim	[AIM 4] Determine if combinations of biomarkers increase the sensitivity and specificity for delirium prediction.
Analysis	We will compare ... with delirium and without delirium ... using the Wilcoxon rank sum test for non-normally distributed continuous variables, Student t test for normally distributed continuous variables, or chi-square test for categorical variables. We will construct receiver operator characteristic (ROC) curves for the maximum value attained by each biomarker during the perioperative course to classify [patients] with delirium.
Power/SS	Sample size for this study was computed to provide a sufficient number of patients to provide the lower CI of 70%. Assuming a conservative prevalence of 25% for pediatric delirium, and point estimate for sensitivity of 90%, a sample size of 58 patients will be required for this pilot study.

Aim	[AIM 4] Determine if combinations of biomarkers increase the sensitivity and specificity for delirium prediction.
Analysis	We will compare ... with delirium and without delirium ... using the Wilcoxon rank sum test for non-normally distributed continuous variables, Student t test for normally distributed continuous variables, or chi-square test for categorical variables. We will construct receiver operator characteristic (ROC) curves for the maximum value attained by each biomarker during the perioperative course to classify [patients] with delirium.
Power/SS	Sample size for this study was computed to provide a sufficient number of patients to provide the lower CI of 70%. Assuming a conservative prevalence of 25% for pediatric delirium, and point estimate for sensitivity of 90%, a sample size of 58 patients will be required for this pilot study.

Aim	Compare commonly used [CONSTRUCT] and delirium scores in [PATIENTS] testing correlation of entire score and of various components to identify clinically significant overlap, and discrimination in diagnosis. ***Hypothesis*** – there will be significant correlation between [ABC] and [DEL] scores, however several individual domains will show higher relationship to clinical syndrome.
Analysis	To allow for the skewed distribution of scores, we will use the Spearman correlation coefficient to assess the association between [ABC] and [DEL], as well as any associations between the scores and continuous variables such as anticholinergic burden score or severity of illness. Evaluation of each component of both scores and its relationship to the … diagnosis … will be conducted by stepwise logistic regression analysis, building a model to best predict the outcome of [EITHER ONE OF TWO OUTCOMES].
Power/SS	Testing the first aim, using an alpha value of <0.05 to detect a clinically significant correlation between the two scores, estimating a value of at least 0.30, a sample size of at least 85 subjects would be required to achieve a power of 80%. Each domain of the scores will be tested using logistic regression, for the 10 different domains, a sample size of 100 will allow at least 10 subjects for each domain.