Alzheimer's Clinical Trial Simulation Concepts

Michael Donohue

Alzheimer's Therapeutic Research Institute
Department of Neurology
University of Southern California

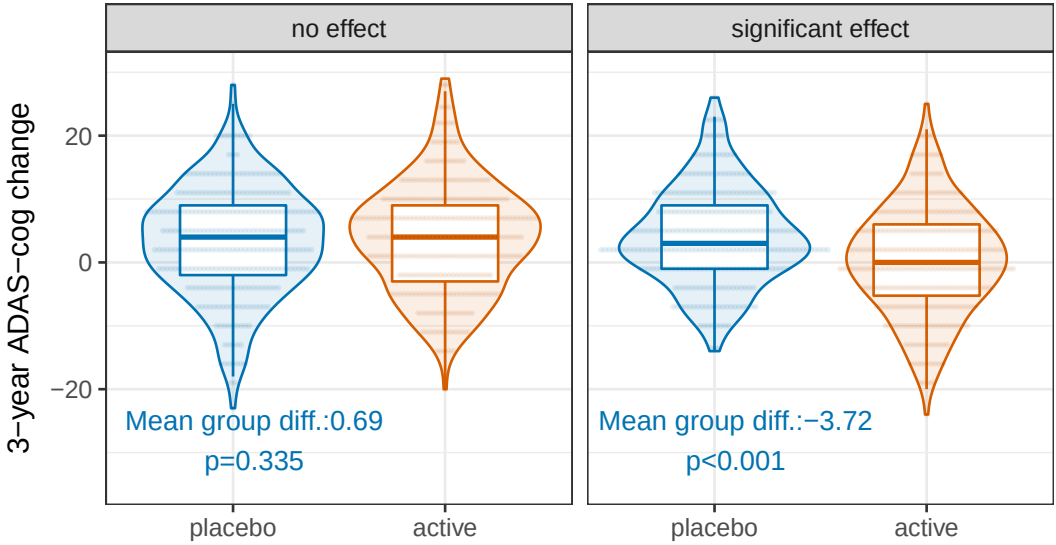
AAIC Educational Workshop: Contemporary Issues in Clinical Trials Methods
July 20, 2018

Learning objectives

- "Analytic" power/sample size calculations
  - two-sample t-test
  - Mixed Model of Repeated Measures (MMRM)
  - Time-to-event (TTE)
  - Missing data
- Simulated power/sample size
  - Why simulate?
  - What goes in the oven?
  - What comes out?

All of the code for this session is available from github.com/atrihub/AAIC2018ClinicalTrialMethods

Testing group diff. in 3-year ADAS-Cog change (simulated MCI data)



Hypothesis testing and power calculation review

- *Power*: probability of "success," i.e. concluding a treatment effect when one actually exists (typically 80-90%)
- *Type I error* (α): probability of concluding a treatment effect when **none** actually exists (typically 5%)

Power/sample size calculations solve for one of the following:

1. Assumed treatment effect (δ) ("minimum clinically meaningful difference")
2. Standard deviation of treatment effect (σ)
3. Type I error (α)
4. Power
5. Sample size

START SIMPLE!

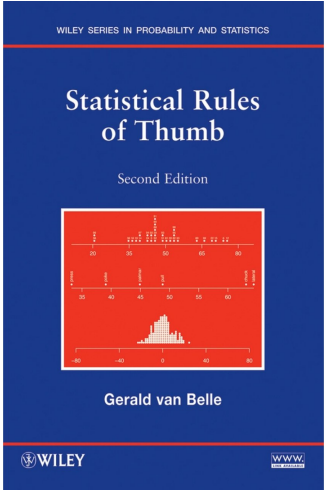
For two-sample t-test, required n per group is:

$$n = 2\left(\frac{z_{1-\alpha/2} + z_{\text{power}}}{\delta/\sigma}\right)^2$$

For $\alpha = 5\%$ and power$= 80\%$, simplifies to:

$$\begin{aligned} n &= 2\left(\frac{1.96 + 0.84}{\delta/\sigma}\right)^2 \\ &\approx \frac{16}{(\delta/\sigma)^2} \end{aligned}$$

van Belle's *Statistical Rules of Thumb*



Back to MCI example

How many subjects are required to detect a δ=2 point difference in ADAS-Cog change assuming σ=8.5, two-sided α=5%, and power=80%?

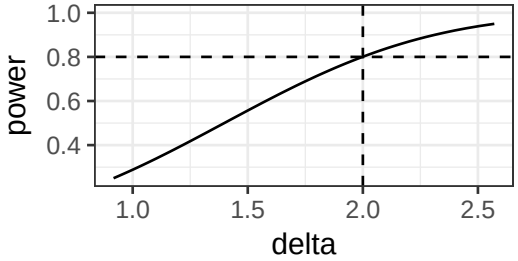
ANSWER: $n \approx 16/(2/8.5)^2 = 289$ subjects per group

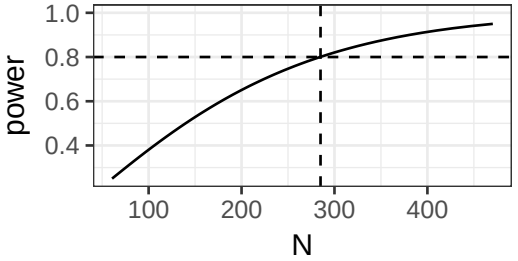
. . . the more exact answer

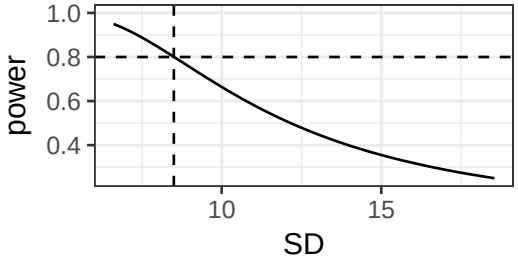
How many subjects are required to detect a δ=2 point difference in ADAS-Cog change assuming σ=8.5, two-sided α=5%, and power=80%?

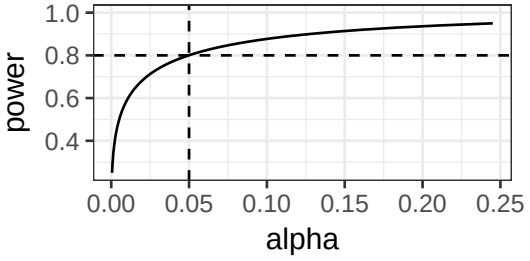
```
power.t.test(delta=2, sd=8.5, power=0.80, sig.level=0.05)


     Two-sample t test power calculation

              n = 285
          delta = 2
             sd = 8.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Power curves for each parameter









A basic simulation for power

```
for(i in 1:100000){
  placebo <- rnorm(n=285, mean=0, sd=8.5)
  active <- rnorm(n=285, mean=-2, sd=8.5)
  pvals[i] <- t.test(placebo, active)$p.value
}
sum(pvals<0.05)/100000
```

```
Power: 0.801
```

Confirms that n =285 per group provides 80% power to detect a δ=2 point difference in ADAS-Cog change assuming σ=8.5 and two-sided α=5%

A basic simulation for Type I error

```
for(i in 1:100000){
  placebo <- rnorm(n=285, mean=0, sd=8.5)
  active <- rnorm(n=285, mean=0, sd=8.5)
  pvals[i] <- t.test(placebo, active)$p.value
}
sum(pvals<0.05)/100000
```

```
Type I error: 0.0503
```

Confirms that our Type I error is 5%

Why simulate? Continuous vs integer values

- If simulations and calculation agree, why bother with simulations?
- Let's say we are worried that t-test calculation assumes continuous data, but ADAS is integer valued. Simulations show we shouldn't be too worried...

```
for(i in 1:100000){
  placebo <- round(rnorm(n=285, mean=0, sd=8.5), digits=0)
  active <- round(rnorm(n=285, mean=-2, sd=8.5), digits=0)
  pvals[i] <- t.test(placebo, active)$p.value
}
sum(pvals<0.05)/100000
```

```
Power: 0.8
```

Simulated Type I error with integer valued ADAS-Cog

```
for(i in 1:100000){
  placebo <- round(rnorm(n=285, mean=0, sd=8.5), digits=0)
  active <- round(rnorm(n=285, mean=0, sd=8.5), digits=0)
  pvals[i] <- t.test(placebo, active)$p.value
}
sum(pvals<0.05)/100000
```

```
Type I error: 0.0504
```

Why simulate? To dichotomize or not to dichotomize...

Let's say we are interested in **dichotomizing** ADAS-Cog change to consider the proportion of subjects who experience a **2 point improvement** or better in ADAS-Cog. Is this a good idea?

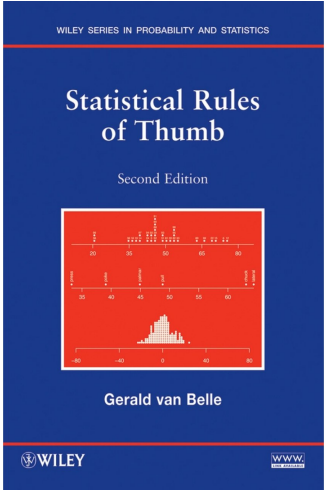
Power for continuous vs dichotomous outcome

```
cutPoint <- -2
for(i in 1:10000){
  dd <- data.frame(
      ADAS.ch = round(rnorm(n=285, mean=0, sd=8.5), digits=0),
      group = 'placebo') %>%
    bind_rows(data.frame(
      ADAS.ch = round(rnorm(n=285, mean=-2, sd=8.5), digits=0),
      group = 'active'))
  pvals[i] <- with(dd, t.test(ADAS.ch~group))$p.value
  pvalsBinary[i] <- with(dd, chisq.test(ADAS.ch<=cutPoint, group))$p.value
}
sum(pvalsBinary<0.05)/10000
sum(pvals<0.05)/10000
```

```
Power for binary outcome: 0.564
Power for continuous outcome: 0.801
```

DO NOT DICHOTOMIZE UNLESS ABSOLUTELY NECESSARY!

- van Belle's *Statistical Rules of Thumb* rule 4.11
- Testing for group differences requires $\pi/2 = 1.57$ times more subjects when you dichotomize compared to two-sample t-test (e.g. $n = 100$ vs $n = 64$; or $n = 447$ vs $n = 285$)

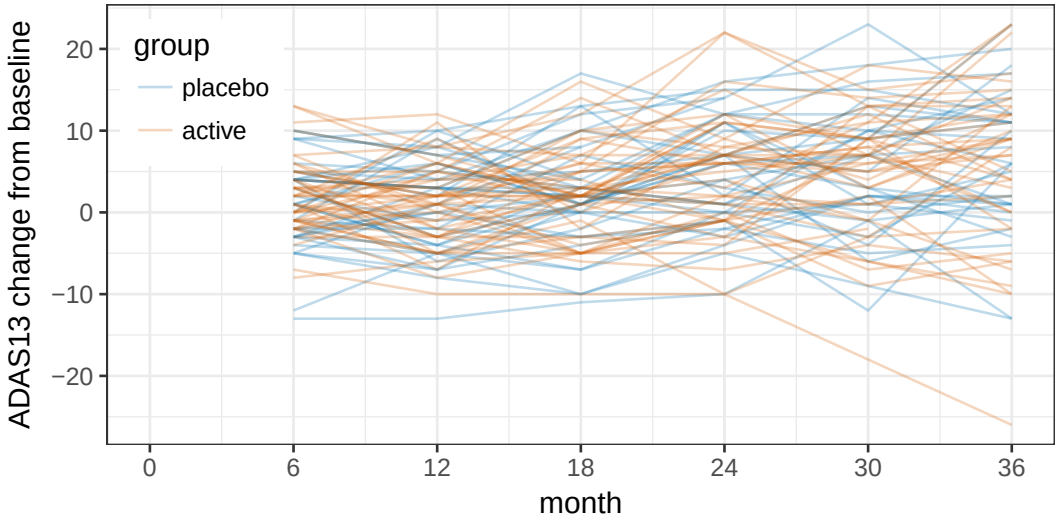


With $n = 1.57 \times 285 = 447$ per group:

```
cutPoint <- -2
for(i in 1:10000){
  dd <- data.frame(
      ADAS.ch = round(rnorm(n=447, mean=0, sd=8.5), digits=0),
      group = 'placebo') %>%
    bind_rows(data.frame(
      ADAS.ch = round(rnorm(n=447, mean=-2, sd=8.5), digits=0),
      group = 'active'))
  pvalsBinary[i] <- with(dd, chisq.test(ADAS.ch<=cutPoint, group))$p.value
  pvals[i] <- with(dd, t.test(ADAS.ch~group))$p.value
}
sum(pvals<0.05)/10000
sum(pvalsBinary<0.05)/10000
```

```
Power for binary outcome: 0.782
Power for continuous outcome: 0.936
```

Mixed Model of Repeated Measures (MMRM)



Power/sample size calculations for MMRM

Familiar parameters:

- Mean of treatment effect (δ)
- Standard deviation of treatment effect (σ)
- Type I error (typically $\alpha = 5\%$)
- Power (typically 80 or 90%)
- Sample size

New parameters

- Visit schedule
- Attrition per visit
- Visit-to-visit correlations

Let's consider a hypothetical MCI trial...

- Two groups: placebo vs active (hypothetical)
- Alzheimer's Disease Assessment Scale (ADAS-Cog) assessed at 0, 6, 12, ..., 36 months
- Placebo group behaves like ADNI participants
- A treatment which slows ADAS-Cog progression by **2 points at 36 months**
- Assume:
  - Attrition per visit: 5, 10, 15, 20, 25, 30%
  - Visit-to-visit correlation $\rho = 0.6$
  - Residual standard deviation at last visit $\sigma = 10$

Start simple

Take simple t-test calculations and inflate sample size for attrition, i.e. divide by $(1 - 30\%)$ (n per group):

```
16/(2/10)^2 / (1-0.30)

[1] 571

power.t.test(delta=2, sd=10, power=0.80)$n / (1-0.30)

[1] 562
```

This doesn't account for information we get from follow-up prior to attrition.

Analytic sample size calculation for MMRM in R

```
followup <- c(6,12,18,24,30,36)
cc <- matrix(0.6, nrow=length(followup), ncol=length(followup))
diag(cc) <- 1
attrition_rate <- c(0, 15, 15, 20, 25, 30)/100
longpower::power.mmrm(Ra=cc, ra=1-attrition_rate, sigmaa=10, delta=2, power=0.80)
```

```
     Power for Mixed Model of Repeated Measures (Lu, Luo, & Chen, 2008)

             n1 = 485
             n2 = 485
     retention1 = 1.00, 0.85, 0.85, 0.80, 0.75, 0.70
     retention2 = 1.00, 0.85, 0.85, 0.80, 0.75, 0.70
          delta = 2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Let's simulate MMRM data from a hypothetical clinical trial...

Parameters for mean structure:

```
Beta <- c(
     'ADAS13.bl'=-0.07, # ADAS13 change per unit baseline ADAS13
            'm6'= 0.90, # worsening at month 6 in pbo
           'm12'= 1.30, # worsening at month 12 in pbo
           'm18'= 2.90, # worsening at month 18 in pbo
           'm24'= 4.25, # worsening at month 24 in pbo
           'm30'= 5.50, # worsening at month 30 in pbo
           'm36'= 6.70, # worsening at month 36 in pbo
     'm6:active'=-0.05, # relative improvement at month 6 with treatment
    'm12:active'=-0.10, # relative improvement at month 12 with treatment
    'm18:active'=-0.50, # relative improvement at month 18 with treatment
    'm24:active'=-1.00, # relative improvement at month 24 with treatment
    'm30:active'=-1.50, # relative improvement at month 30 with treatment
    'm36:active'=-2.00) # relative improvement at month 36 with treatment
```

Let's simulate MMRM data from a hypothetical clinical trial...

Other parameters:

```
# other design parameters
followup <- c(6, 12, 18, 24, 30, 36)
n <- 485 # per group
attrition_rate <- c(0.00, 0.15, 0.15, 0.20, 0.25, 0.30)

# var-cov parameters:
# standard deviation scale parameter:
SD <- 5
# hetergeneous variance weights:
vv <- diag(c(1.00, 1.00, 1.25, 1.30, 1.50, 2.00))
# correlation matrix
cc <- matrix(0.60, nrow=length(followup), ncol=length(followup))
diag(cc) <- 1
```

Simulated MMRM power

```
for(i in 1:1000){
  trial <- arrange(design, id, month) %>%
    mutate(
      residual = as.numeric(t(
        rmvnorm(length(unique(design$id)), mean=rep(0,nrow(vv)), sigma=SD^2*vv%*%cc%*%vv))),
      ADAS13.ch = round(ADAS13.ch.noResidual + residual, digits = 0)) %>%
    filter(!missing)

  trial_fit <- gls(ADAS13.ch ~ -1+ADAS13.bl+
      (m6+m12+m18+m24+m30+m36)+
      (m6+m12+m18+m24+m30+m36):active,
    data=trial, correlation = corCompSymm(form = ~ visNo | id),
    weights = varIdent(form = ~ 1 | m))
  pvals[i] <- summary(trial_fit)$tTable['m36:active','p-value']
  txEffect[i] <- summary(trial_fit)$tTable['m36:active','Value']
}
sum(pvals<0.05)/1000
summary(txEffect)
```

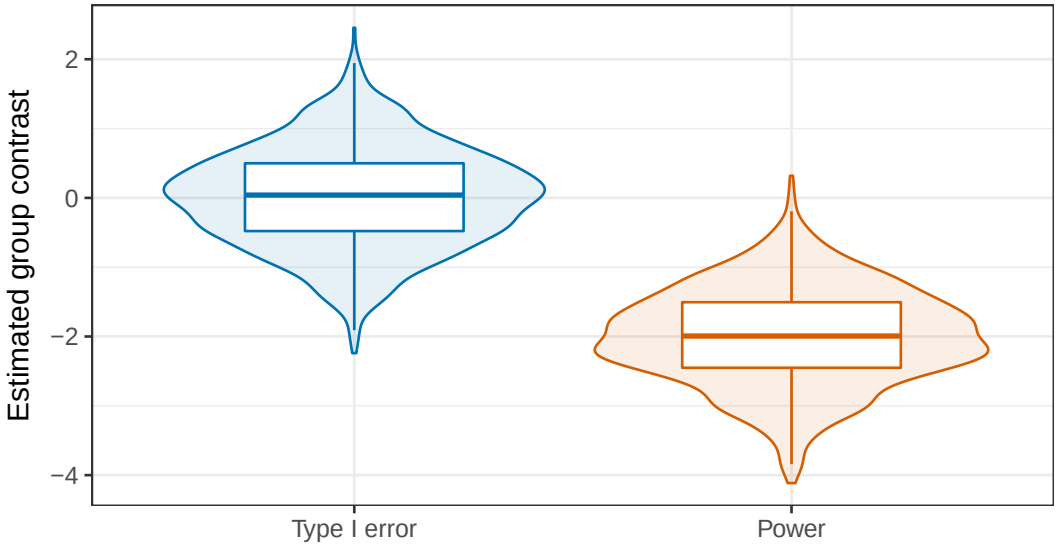
```
Power for MMRM: 0.802
```

Simulated MMRM Type I error

```
for(i in 1:1000){
  trial <- arrange(design, id, month) %>%
    mutate(
      residual = as.numeric(t(
        rmvnorm(length(unique(design$id)), mean=rep(0,nrow(vv)), sigma=SD^2*vv%*%cc%*%vv))),
      ADAS13.ch = round(ADAS13.ch.noResidual.noTxEffect + residual, digits = 0)) %>%
    filter(!missing)

  trial_fit <- gls(ADAS13.ch ~ -1+ADAS13.bl+
      (m6+m12+m18+m24+m30+m36)+
      (m6+m12+m18+m24+m30+m36):active,
    data=trial, correlation = corCompSymm(form = ~ visNo | id),
    weights = varIdent(form = ~ 1 | m))
  pvals[i] <- summary(trial_fit)$tTable['m36:active','p-value']
  txEffect[i] <- summary(trial_fit)$tTable['m36:active','Value']
}
sum(pvals<0.05)/1000
summary(txEffect)
```

```
Type I error for MMMR: 0.049
```

MMRM simulation summaries



Time-to-dementia

- An alternative to MMRM is an analysis of time to progression to dementia
- Outcome variable is distilled to
  - time of conversion, for those who convert
  - time of last follow-up, for those who do not convert
- No information about sub-conversion changes is used in analysis
- In the ADCS trial of Donepezil in MCI (Petersen, et al. 2005), conversion was defined by a clinical criteria (McKhann, et al. 1984) reviewed by a central committee
- The central committee reviewed data from the neuropsychological evaluation, but no specific point changes were required
- The lack of specific rules makes it difficult to simulate simultaneous conversion events and assessment scores

Sample size inflation for time-to-threshold vs "slope model"

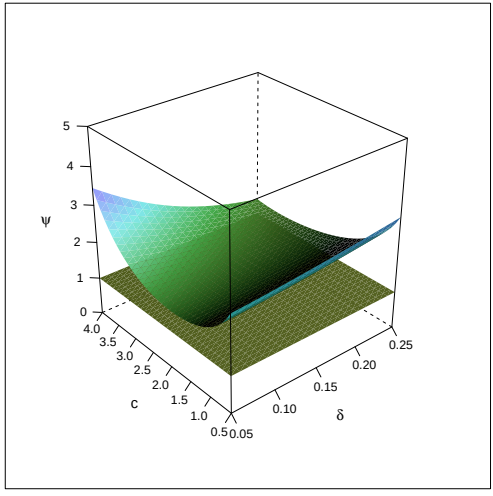
$$\psi = \frac{n_{\text{PH}}}{n_{\text{LM}}} = \frac{E_{\text{PH}}}{r n_{\text{LM}}} = \frac{(\theta_B - \theta_A)^2}{\xi r} \log\left(\frac{\log(P[T_A > t])}{\log(P[T_B > t])}\right)^2 \quad (1)$$

- n_{PH} and n_{LM} are sample size required for time-to-*threshold* and "slope model"
- $\theta_B - \theta_A$ is group difference in slopes
- T_A & T_B represent the time-to-threshold for an individual randomized in groups A & B
- Not a particularly simple "rule of thumb"

Donohue, Gamst, Thomas, Xu, Beckett, Petersen, Weiner, & Aisen. (2011). The relative efficiency of time-to-threshold and rate of change in longitudinal data. *Contemporary clinical trials*, 32(5), 685-693.

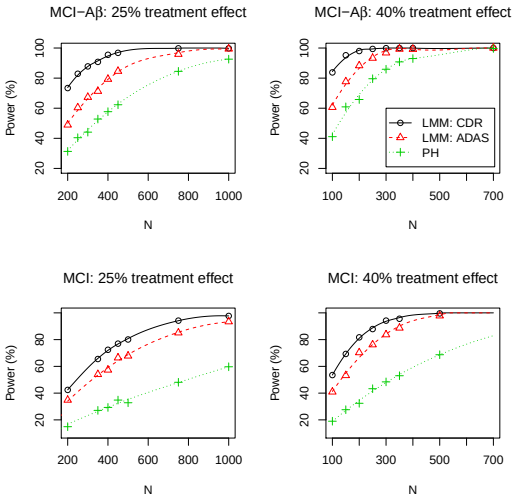
Sample size inflation for time-to-threshold

- $\psi =$ inflation factor
- $c =$ threshold
- $\theta_B = 0.1$
- $\theta_A = \theta_B + \delta$
- $\sigma = 0.5$
- We could not find useful scenarios where $\psi < 1$



Simulated time-to-dementia vs slope models

- Dementia "algorithm" learned from ADNI clinical diagnoses
- Based on ADAS-Cog, CDR-SB, and FAQ
- Simultaneously simulated continuous outcomes and dementia events
- Explored CSF defined Aβ+ MCI
- Very larger sample size inflation at 80% power



MMRM bias with informative missingness

- So far, our MMRM simulations have assumed data "missing completely at random" (MCAR)
- Let's assume all ADAS-Cog change scores ≥ 12 go missing
- How bad are the MMRM estimates under this diabolical "informative missingness"?

MMRM biased with informative missingness?

```
for(i in 1:1000){
  trial <- arrange(design, id, month) %>%
    mutate(
      residual = as.numeric(t(
        rmvnorm(length(unique(design$id)), mean=rep(0,nrow(vv)), sigma=SD^2*vv%*%cc%*%vv))),
      ADAS13.ch = round(ADAS13.ch.noResidual + residual, digits = 0)) %>%
    filter(ADAS13.ch<12)

  trial_fit <- gls(ADAS13.ch ~ -1+ADAS13.bl+
      (m6+m12+m18+m24+m30+m36)+
      (m6+m12+m18+m24+m30+m36):active,
    data=trial, correlation = corCompSymm(form = ~ visNo | id),
    weights = varIdent(form = ~ 1 | m))
  pvals[i] <- summary(trial_fit)$tTable['m36:active','p-value']
  txEffect[i] <- summary(trial_fit)$tTable['m36:active','Value']
}
sum(pvals<0.05)/1000
summary(txEffect)
```

```
Power: 0.727
```

MMRM biased with informative missingness?

Bias is about:

```
summary(txEffect) - -2

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.334   0.257   0.659   0.646   1.003   2.413
```

- So, on average, treatment looks 0.646 ADAS points *worse* compared to MAR.
- Censoring ADAS13 change ≥ 12 results in attrition at 36 months of about 32% for placebo vs 25% for active.

MMRM Type I error with informative missingness?

```
for(i in 1:1000){
  trial <- arrange(design, id, month) %>%
    mutate(
      residual = as.numeric(t(
        rmvnorm(length(unique(design$id)), mean=rep(0,nrow(vv)), sigma=SD^2*vv%*%cc%*%vv))),
      ADAS13.ch = round(ADAS13.ch.noResidual.noTxEffect + residual, digits = 0)) %>%
    filter(ADAS13.ch<12)

  trial_fit <- gls(ADAS13.ch ~ -1+ADAS13.bl+
      (m6+m12+m18+m24+m30+m36)+
      (m6+m12+m18+m24+m30+m36):active,
    data=trial, correlation = corCompSymm(form = ~ visNo | id),
    weights = varIdent(form = ~ 1 | m))
  pvals[i] <- summary(trial_fit)$tTable['m36:active','p-value']
  txEffect[i] <- summary(trial_fit)$tTable['m36:active','Value']
}
sum(pvals<0.05)/1000
summary(txEffect)
```

```
Type I error: 0.048
```

MMRM bias with *imbalanced* informative missingness

- What about tolerability?
- Let's assume all ADAS-Cog change scores ≥ 16 go missing **only in the active group**.
- How bad are the MMRM estimates now?

MMRM bias with *imbalanced* informative missingness

```
for(i in 1:1000){
  trial <- arrange(design, id, month) %>%
    mutate(
      residual = as.numeric(t(
        rmvnorm(length(unique(design$id)), mean=rep(0,nrow(vv)), sigma=SD^2*vv%*%cc%*%vv))),
      ADAS13.ch = round(ADAS13.ch.noResidual + residual, digits = 0)) %>%
    filter(ADAS13.ch<18 | active==0)

  trial_fit <- gls(ADAS13.ch ~ -1+ADAS13.bl+
      (m6+m12+m18+m24+m30+m36)+
      (m6+m12+m18+m24+m30+m36):active,
    data=trial, correlation = corCompSymm(form = ~ visNo | id),
    weights = varIdent(form = ~ 1 | m))
  pvals[i] <- summary(trial_fit)$tTable['m36:active','p-value']
  txEffect[i] <- summary(trial_fit)$tTable['m36:active','Value']
}
sum(pvals<0.05)/1000
summary(txEffect)
```

```
Power: 0.998
```

MMRM bias with *imbalanced* informative missingness

Bias is about:

```
summary(txEffect) - -2

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.637  -1.472  -1.051  -1.061  -0.651   1.322
```

- So, on average, treatment looks -1.061 ADAS points *better* compared to MAR.
- Censoring ADAS13 change $\geq$ 18 only in active group results in attrition at 36 months of about 0% for placebo vs 10% for active in simulation for power (15% for active in simulation for Type I error).

MMRM Type I error with *imbalanced* informative missingness?

```
for(i in 1:1000){
  trial <- arrange(design, id, month) %>%
    mutate(
      residual = as.numeric(t(
        rmvnorm(length(unique(design$id)), mean=rep(0,nrow(vv)), sigma=SD^2*vv%*%cc%*%vv))),
      ADAS13.ch = round(ADAS13.ch.noResidual.noTxEffect + residual, digits = 0)) %>%
    filter(ADAS13.ch<18 | active==0)

  trial_fit <- gls(ADAS13.ch ~ -1+ADAS13.bl+
      (m6+m12+m18+m24+m30+m36)+
      (m6+m12+m18+m24+m30+m36):active,
    data=trial, correlation = corCompSymm(form = ~ visNo | id),
    weights = varIdent(form = ~ 1 | m))
  pvals[i] <- summary(trial_fit)$tTable['m36:active','p-value']
  txEffect[i] <- summary(trial_fit)$tTable['m36:active','Value']
}
sum(pvals<0.05)/1000
summary(txEffect)
```

```
Type I error: 0.616
```

Take home

- Start with simple calculations to approximate power/sample size
- Next, use fancier calculations when available
- Simulate when assumptions of the fancy calculations are not met
- Better to discover design problems in simulations, rather than during trial
- Do not dichotomize unless absolutely necessary!
- Be wary of MNAR!

Further reading

- Van Belle, G. (2011). *Statistical Tules of Thumb*. John Wiley & Sons.
- Mallinckrodt, C.H., et al. (2003). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry*, 53(8), 754-760.
- Pinheiro, J.C., & Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.
- Lu, K. (2010). On efficiency of constrained longitudinal data analysis versus longitudinal analysis of covariance. *Biometrics*, 66(3), 891-896.
- Donohue, M.C. and Aisen, P. S. (2012). Mixed model of repeated measures versus slope models in Alzheimer's disease clinical trials. *The Journal of Nutrition, Health & Aging*. 16(4), 360-364.
- McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 1984;34:939-44.
- Lu, Luo, & Chen. (2008). Sample size estimation for repeated measures analysis in randomized clinical trials with missing data. *The International Journal of Biostatistics*, 4(1).