An investigation of the Kappa statistic under clustered data and group sequential testing, with an application to surgical rating

Mary M. Ryan

Advisor: Dr. Daniel L. Gillen

University of California, Irvine

SPOT GRADE Trial

- Researchers working on local hemostatic agent to stop bleeding on "low grade" wounds
- FDA required researchers to first develop scale to classify bleeds
  - Wanted surgeons to have better knowledge of what type of wounds appropriate to use agent on
  - Concerned surgeons would use agent on bleeds not be designed to stop

SPOT GRADE Trial

- Researchers working on local hemostatic agent to stop bleeding on "low grade" wounds
- FDA required researchers to first develop scale to classify bleeds
  - Wanted surgeons to have better knowledge of what type of wounds appropriate to use agent on
  - Concerned surgeons would use agent on bleeds not be designed to stop
- SPOT GRADE scale developed to standardize severity of blood loss[5]
  - Surface bleed severity scale (SBSS)
  - 6 categories, 0-5
  - Scores defined by flux/flow rate of blood from wound

SPOT GRADE Trial

- 14 surgeons watched video simulations in a randomized sequence and classified bleeding severity by SPOT GRADE category
  - 36 training videos
  - 36 testing videos
    - Each video viewed 3 times
- Kappa statistic used to assess inter- and intra-rater reliability

Goals

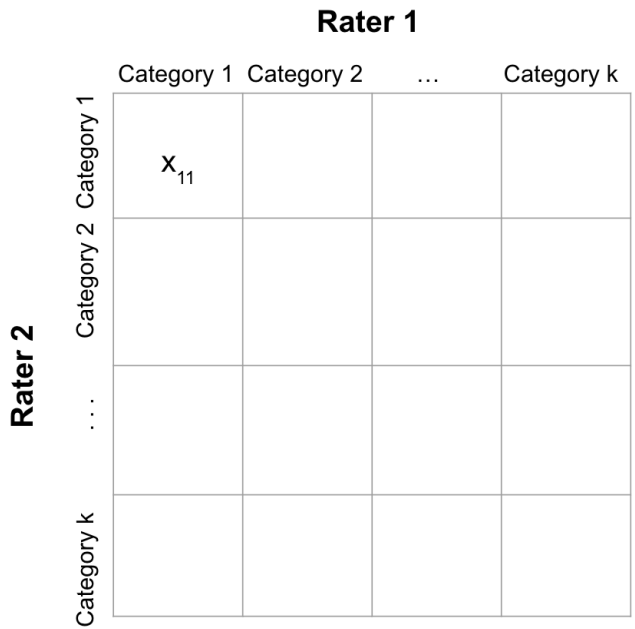
- Rating same video multiple times induces clustering that biases variance estimate
  - Operating characteristics of Kappa's asymptotic variance not yet explored under setting of heterogeneity within categories
  - Want to adapt Kappa statistic for clustered data and heterogeneity within categories by correcting variance estimate

Goals

- Rating same video multiple times induces clustering that biases variance estimate
  - Operating characteristics of Kappa's asymptotic variance not yet explored under setting of heterogeneity within categories
  - Want to adapt Kappa statistic for clustered data and heterogeneity within categories by correcting variance estimate
- Surgeon compensation costs quickly add up
  - Two sets of surgeons flown to training and test site for trial
  - Conducting study using group sequential design could conserve resources/increase study efficiency

Cohen's Kappa[1]

- Kappa statistic assesses likelihood-above-chance of two raters agreeing
- $\kappa = \frac{p_o - p_e}{1 - p_e} \in (-1, 1)$
  - $p_o = \sum_{i=1}^{k} p_{ii}$
  - $p_e = \sum_{i=1}^{k} p_{i.} p_{.i}$
- $\kappa = 0$ implies rater agreement on par with chance
- $\kappa \to 1$ implies raters agree more
- $\kappa \to -1$ implies raters disagree more



Cohen's Kappa

- Fleiss et al.[2] asserted that, by CLT:

$$\sqrt{n}(\kappa - \kappa_0) \dot{\sim} N\left(0, \sigma_\kappa^2\right),$$

where κ_0 is the true κ value, and σ_κ^2 is a function of p_e, p_o, and n

Cohen's Kappa

- Fleiss et al.[2] asserted that, by CLT:

$$\sqrt{n}(\kappa - \kappa_0) \dot{\sim} N\left(0, \sigma_\kappa^2\right),$$

  where κ_0 is the true κ value, and σ_κ^2 is a function of p_e, p_o, and n
- Since $\kappa \in (-1, 1)$, Normal approximation from Fleiss et al. likely to perform poorly in small samples
- Propose transformation of κ to map onto $\mathbb{R}$:

$$f(\kappa) = ln\left(\frac{1+\kappa}{1-\kappa}\right) \equiv \varphi$$

Data Generation

- If each rater classifies same item multiple times, clustering induced between observations
  - Here, sampling units are surgeons, not individual videos
  - Clustering not accounted for in variance estimate $\Rightarrow \hat{\sigma}^2_\kappa$ is biased
- Operating characteristics of Kappa's asymptotic variance unknown when heterogeneity within categories is present
- Want to explore variance bias in fixed sample setting

Data Generation

- If each rater classifies same item multiple times, clustering induced between observations
  - Here, sampling units are surgeons, not individual videos
  - Clustering not accounted for in variance estimate $\Rightarrow \hat{\sigma}^2_\kappa$ is biased
- Operating characteristics of Kappa's asymptotic variance unknown when heterogeneity within categories is present
- Want to explore variance bias in fixed sample setting
- Need simulated data that reflect:
  - Some SBSS categories are inherently easier (0, 5) or more difficult (2, 3) to correctly place than others
  - Some videos within an SBSS category may be easier/more difficult to correctly place than others

Data Generation

- Let π_{kmj} be the probability video j classified as category m when actually category k

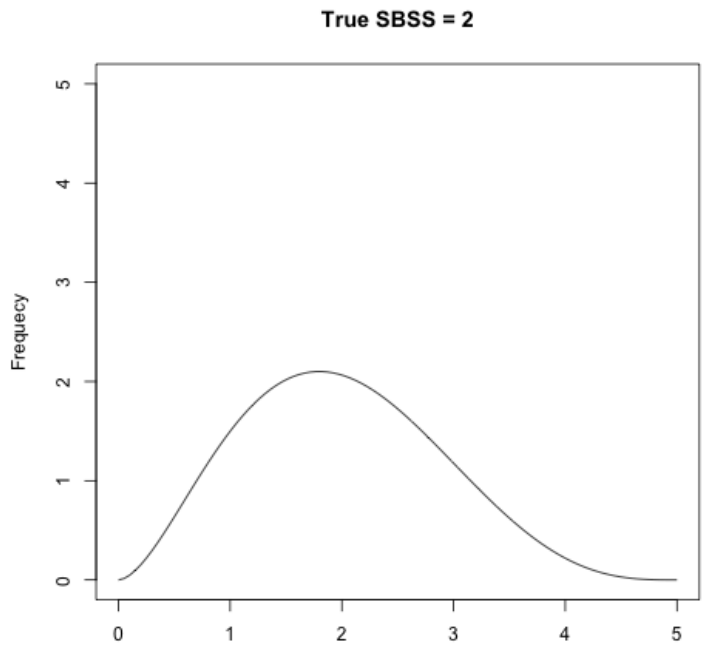
Data Generation

- Let π_{kmj} be the probability video j classified as category m when actually category k

- $\pi_{kmj} = \int_{m-0.5}^{m+0.5} \frac{\frac{1}{5}u^{\alpha_{kj}-1}(1-\frac{1}{5}u)^{\beta_{kj}-1}\Gamma(\alpha_{kj}+\beta_{kj})}{5\Gamma(\alpha_{kj})\Gamma(\beta_{kj})} du$
- $\frac{\alpha_{kj}}{\alpha_{kj}+\beta_{kj}} \times 5 = k$
- $log(\beta_{kj}) \overset{indep.}{\sim} N(\mu_k, \sigma_k^2)$
- $\alpha_{kj} = \frac{\beta_{kj} k}{5-k}$

Data Generation

- Let π_{kmj} be the probability video j classified as category m when actually category k
- $\pi_{kmj} = \int_{m-0.5}^{m+0.5} \frac{\frac{1}{5}u^{\alpha_{kj}-1}(1-\frac{1}{5}u)^{\beta_{kj}-1}\Gamma(\alpha_{kj}+\beta_{kj})}{5\Gamma(\alpha_{kj})\Gamma(\beta_{kj})} du$
- $\frac{\alpha_{kj}}{\alpha_{kj}+\beta_{kj}} \times 5 = k$
- $log(\beta_{kj}) \overset{indep.}{\sim} N(\mu_k, \sigma_k^2)$
- $\alpha_{kj} = \frac{\beta_{kj} k}{5-k}$
- μ_k controls probability of correct classification
- σ_k^2 is increased or decreased to create random video effects for each unique video



$\mu_2 = 2.7,\ \sigma_2^2 = 1$

Variance Bias: Clustered Data

Clustered Kappa Setting

- 1,000 simulations
- n=14 surgeons per simulation
- Three Kappa values: 0.4, 0.6, 0.8
- Four video heterogeneity settings:

Heterogeneity Level	SBSS Category					
	0	1	2	3	4	5
None	0	0	0	0	0	0
Low	0.25	0.5	1	1	0.5	0.25
Medium	0.5	1	2	2	1	0.5
High	1	2	3	3	2	1

- Six videos per SBSS category, each rated three times per surgeon

Variance Bias: Clustered Data

- Variance ratio $= \dfrac{\text{Analytic variance}}{\text{Empirical variance}}$

Video Heterogeneity	$\kappa = 0.4$		$\kappa = 0.6$		$\kappa = 0.8$	
	Variance Ratio	Coverage	Variance Ratio	Coverage	Variance Ratio	Coverage
None	1.127	0.963	1.125	0.960	1.061	0.952

Variance Bias: Clustered Data

- Variance ratio $= \dfrac{\text{Analytic variance}}{\text{Empirical variance}}$

Video Heterogeneity	$\kappa = 0.4$		$\kappa = 0.6$		$\kappa = 0.8$	
	Variance Ratio	Coverage	Variance Ratio	Coverage	Variance Ratio	Coverage
None	1.127	0.963	1.125	0.960	1.061	0.952
Low	1.143	0.963	1.202	0.969	1.221	0.970

Variance Bias: Clustered Data

- Variance ratio $= \frac{\text{Analytic variance}}{\text{Empirical variance}}$

Video Heterogeneity	$\kappa = 0.4$		$\kappa = 0.6$		$\kappa = 0.8$	
	Variance Ratio	Coverage	Variance Ratio	Coverage	Variance Ratio	Coverage
None	1.127	0.963	1.125	0.960	1.061	0.952
Low	1.143	0.963	1.202	0.969	1.221	0.970
Medium	1.306	0.974	1.392	0.979	1.682	0.988
High	1.672	0.989	1.736	0.991	2.181	0.997

- Increases of video heterogeneity, combined with data clustering, inflates analytic variance

Variance Bias: Clustered Data

- Variance ratio $= \frac{\text{Analytic variance}}{\text{Empirical variance}}$

Video Heterogeneity	$\kappa = 0.4$		$\kappa = 0.6$		$\kappa = 0.8$	
	Variance Ratio	Coverage	Variance Ratio	Coverage	Variance Ratio	Coverage
None	1.127	0.963	1.125	0.960	1.061	0.952
Low	1.143	0.963	1.202	0.969	1.221	0.970
Medium	1.306	0.974	1.392	0.979	1.682	0.988
High	1.672	0.989	1.736	0.991	2.181	0.997

- Increases of video heterogeneity, combined with data clustering, inflates analytic variance
- May bootstrap new variance estimate to correct this

Variance Bias: Bootstrap

- Sampling units are surgeons, not videos
- Each bootstrap iteration will sample n surgeons

Algorithm 1: Bootstrap algorithm for Kappa statistic.

for *b in B* **do**
    Randomly choose n surgeons, with replacement;
    Add together all sampled surgeon contingency tables;
    Find statistic, κ_b;
end
Calculate $\hat{\sigma}^2_B = var(\vec{\kappa})$

Variance Bias: Bootstrap

- Sampling units are surgeons, not videos
- Each bootstrap iteration will sample n surgeons

Algorithm 1: Bootstrap algorithm for Kappa statistic.

for b *in* B **do**
    Randomly choose n surgeons, with replacement;
    Add together all sampled surgeon contingency tables;
    Find statistic, κ_b;
end
Calculate $\hat{\sigma}^2_B = var(\vec{\kappa})$

- Use $\hat{\sigma}^2_B$ instead of analytic variance estimate

Variance Bias: Clustered Data

- Employing bootstrap (200 samples) attenuates variance ratio back toward 1:

Video Heterogeneity	$\kappa = 0.4$		$\kappa = 0.6$		$\kappa = 0.8$	
	Variance Ratio	Coverage	Variance Ratio	Coverage	Variance Ratio	Coverage
None	0.963	0.931	0.968	0.927	0.886	0.908
Low	0.895	0.913	0.906	0.910	0.971	0.927
Medium	0.965	0.918	1.002	0.927	1.052	0.936
High	1.007	0.918	0.947	0.929	0.872	0.905

- When n=50 surgeons:

Video Heterogeneity	$\kappa = 0.4$		$\kappa = 0.6$		$\kappa = 0.8$	
	Variance Ratio	Coverage	Variance Ratio	Coverage	Variance Ratio	Coverage
None	0.984	0.940	0.993	0.947	0.965	0.938
Low	1.009	0.942	0.983	0.942	0.962	0.937
Medium	0.971	0.940	1.004	0.942	0.983	0.939
High	0.973	0.940	0.979	0.937	0.994	0.941

- Bootstrap procedure corrects variance overestimation

Group Sequential Testing

- Study framework used to assess early signs of study futility or efficacy
- Potential to conserve resources (time, expenses, samples, etc.)
- Hypothesis tests performed at multiple points throughout data accrual (interim analyses) to determine if sufficient evidence to draw a conclusion early
- Statistics at each interim analysis compared against stopping boundaries that control for type I error rate
  - Pocock[4] (constant on Z-statistic scale)
  - O'Brien-Fleming[3] (constant on partial sum statistic scale)

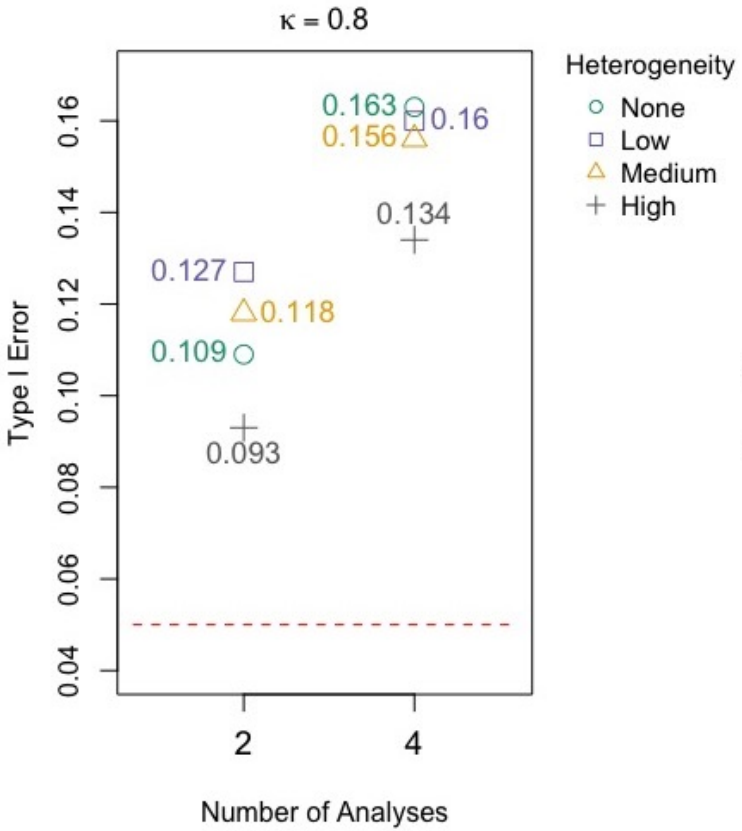
Simulation Study: Sequential & Clustered Kappa

Clustered Kappa Setting

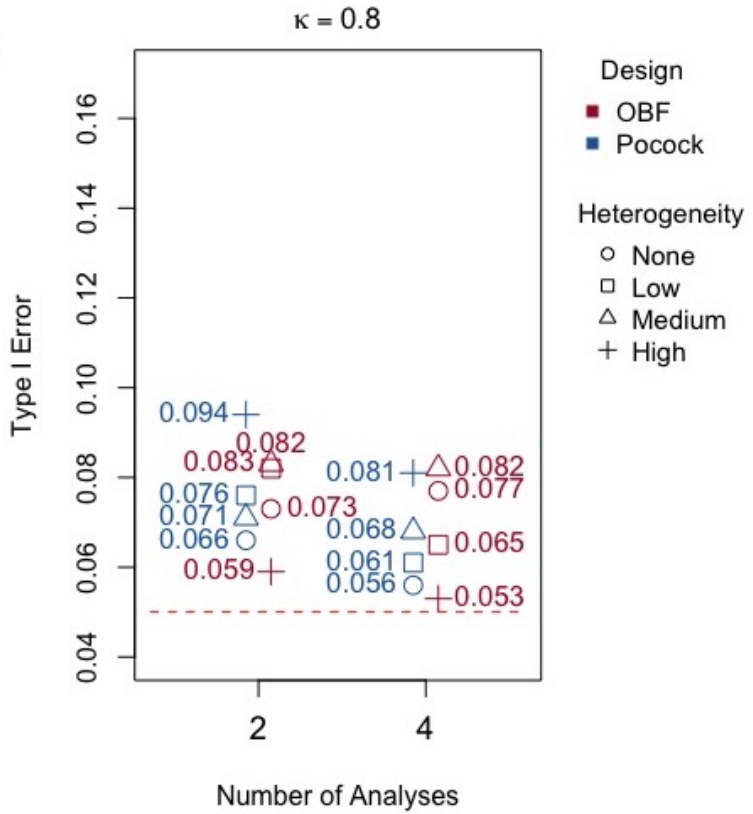
- 1,000 simulations, n=14 surgeons per simulation
- 200 bootstrap samples
- Three Kappa values: 0.4, 0.6, 0.8 (shown)
- Four video heterogeneity settings: none, low, medium, high
- Two maximal analysis amounts: 2,4
- 6 videos per SBSS category, each video rated three times per surgeon
- $H_0 : \kappa \leq \kappa_0$ vs. $H_1 : \kappa > \kappa_0$
- Comparing: Naive boundaries (all boundaries Z=1.645) vs. group sequential boundaries (O'Brien-Fleming, Pocock)

Simulation Study: Sequential & Clustered Kappa

Naive Boundaries

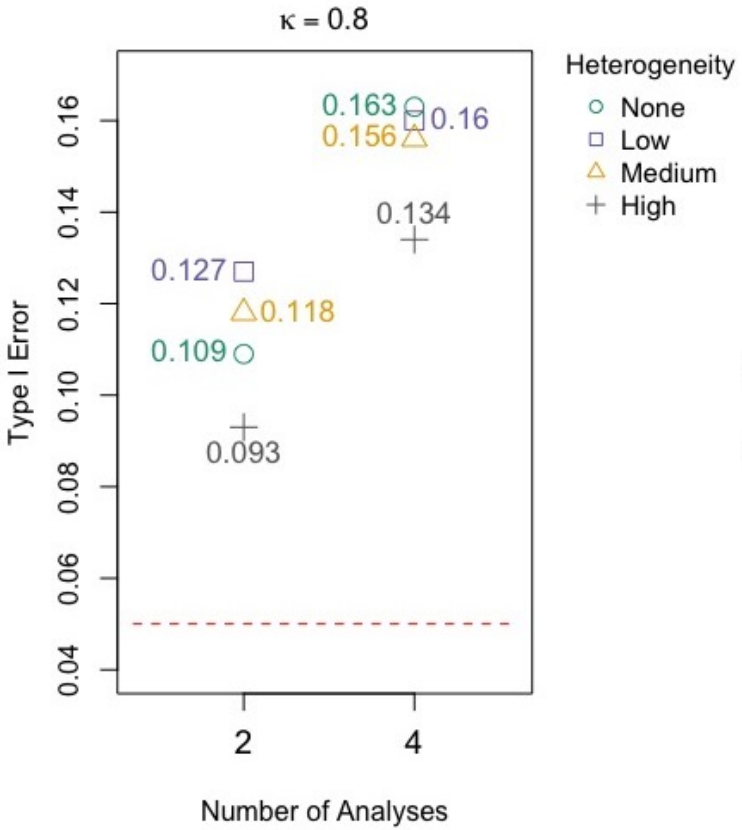


Group Sequential Boundaries

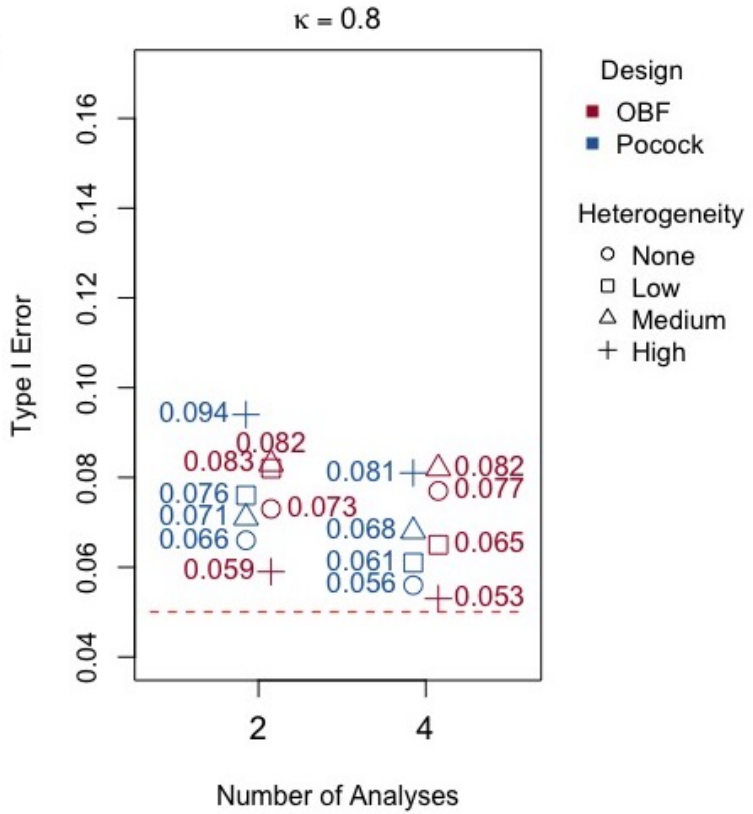


Simulation Study: Sequential & Clustered Kappa

Naive Boundaries



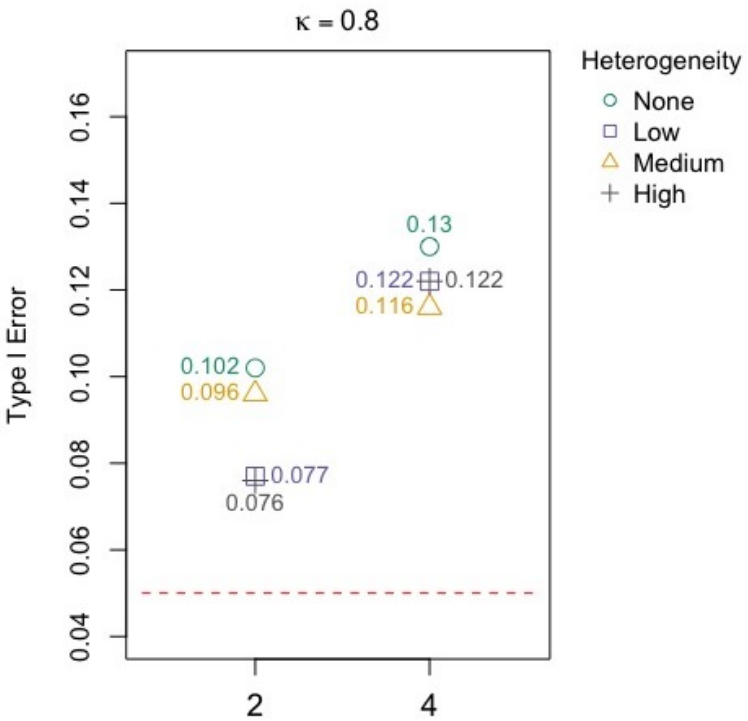
Group Sequential Boundaries



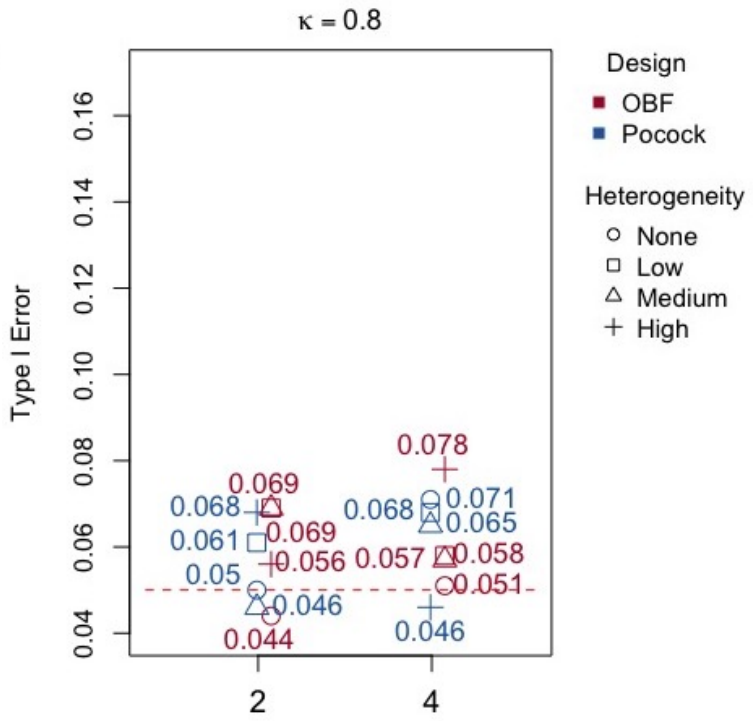
$$P(\kappa \in \Theta_1 | \kappa \in \Theta_0) = 1 - (1-\alpha)^J$$

Simulation Study: Sequential & Clustered Kappa

Naive Boundaries (n=50)



Group Sequential Boundaries (n=50)



Simulation Study: Sequential & Clustered Kappa

- Average analysis number (n) at which Kappa would cross its respective stopping boundary

Video Heterogeneity	$\kappa = 0.8$	
	Pocock	OBF
None	1.650 (5.76)	2.616 (9.16)
Low	1.583 (5.54)	2.600 (9.10)
Medium	1.656 (5.80)	2.568 (8.99)
High	1.618 (5.66)	2.545 (8.91)

- Under Pocock design, nine fewer surgeons needed than fixed sample
- Under the O'Brien-Fleming design, five fewer surgeons needed than fixed sample

Application to SPOT GRADE: Identification of Eligibility

- For development later clinical trial of local hemostatic device, important to be able to identify study-eligible bleeds (SBSS 1-3) from study-ineligible bleeds (SBSS 4-5)
- Testing hypothesis

$$H_0 : \kappa \leq 0.60 \quad \text{vs.} \quad H_1 : \kappa > 0.60$$

Application to SPOT GRADE: Identification of Eligibility

- For development later clinical trial of local hemostatic device, important to be able to identify study-eligible bleeds (SBSS 1-3) from study-ineligible bleeds (SBSS 4-5)
- Testing hypothesis

$$H_0 : \kappa \leq 0.60 \quad \text{vs.} \quad H_1 : \kappa > 0.60$$

κ	Z-statistic	Stopping Time (n)	Decision
0.816	2.402	1 (4)	Reject

- Original study's statistic: 0.833

References

[1] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, Apr. 1960.

[2] J. L. Fleiss, J. Cohen, and B. S. Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327, 1969.

[3] P. C. O'Brien and T. R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556, Sept. 1979.

[4] S. J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, Aug. 1977.

[5] W. D. Spotnitz, D. Zielske, V. Centis, R. Hoffman, D. L. Gillen, C. Wittmann, V. Guyot, D. M. Campos, P. Forest, A. Pearson, and P. C. McAfee. The SPOT GRADE: A New Method for Reproducibly Quantifying Surgical Wound Bleeding. *Spine*, 43(11):E664, June 2018.