Optimal two-stage sampling for mean estimation in multilevel populations when cluster size is informative

Francesco Innocenti, Math Candel, Frans Tan, & Gerard van Breukelen

Department of Methodology and Statistics
Maastricht University

44th Annual Conference of the ISCB
Joint Conference with the Italian Region of the IBS
August 27-31, 2023

Maastricht University

Outline

1. Motivating example
2. Definitions of population means
3. Sampling from a multilevel population
4. Unbiased estimation
5. Optimal and maximin design
6. Application
7. Guidelines and future research

Motivating example

Average alcohol consumption among adolescents:

- Adolescents clustered in schools
- Schools vary in # of enrolled students
- Adolescents' alcohol consumption can be related to school size (McNeely et al. [2002]; Resnick et al. [1997]; Thompson et al. [2006])

General framework:

- Two-level population
- Cluster size variation
- Informative cluster size (Nevalainen et al. [2014]; Seaman et al. [2014])

General framework

The outcome variable Y_{ij} is **quantitative**

$$y_{ij} = \beta_0 + u_j + \epsilon_{ij} \tag{1}$$

$$\epsilon_{ij} \sim N\left(0, \sigma_\epsilon^2\right),\ u_j \perp \epsilon_{ij}.$$

Informative cluster size $\left(N_j\right)$:

$$u_j = \gamma\left(N_j - \theta_N\right) + \upsilon_j, \tag{2}$$

$$\upsilon_j \sim N\left(0, \sigma_\upsilon^2\right),\ \upsilon_j \perp N_j$$

γ = Informativeness parameter
θ_N = Population mean of cluster size

Definitions of population means

1. The average of all individual outcomes:

   $\mu =$ *Expected outcome for an individual randomly sampled from the population ignoring cluster membership*

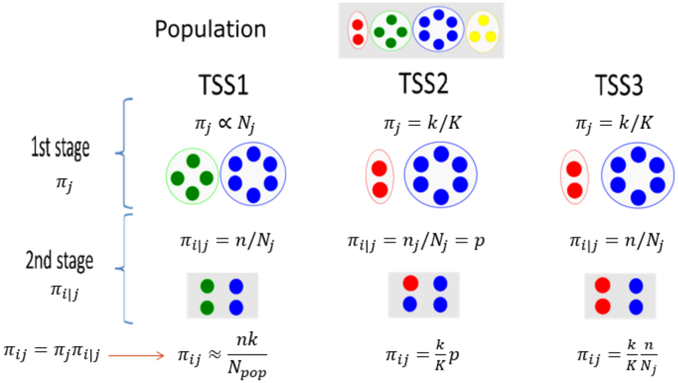
2. The average of all cluster-specific means:

   $\beta_0 =$ *Expected outcome for the average individual from the average cluster*

$$\mu = \beta_0 + \gamma \theta_N \tau_N^2 \tag{3}$$

$\gamma =$ Informativeness parameter
$\theta_N =$ Population mean of cluster size
$\tau_N =$ Population coefficient of variation of cluster size

Sampling from a multilevel population



- K = # of clusters in the population, k = # of clusters in the sample, N_j = cluster j size in the population, n or n_j = # of individuals sampled per cluster, N_{pop} = population size
- TSS1 requires prior knowledge of the whole cluster size distribution
- Sampling fraction is assumed to be negligible at each sampling stage

Unbiased estimation of μ

- Informative cluster size
  - SRS and TSS1: $\hat{\mu} = \sum_{j=1}^{k} \frac{\bar{y}_j}{k}$, $\bar{y}_j$ = cluster j mean, $k = \#$ of clusters
  - TSS2 and TSS3: $\hat{\mu} = \frac{\sum_{j=1}^{k} N_j \bar{y}_j}{\sum_{j=1}^{k} N_j} \Leftarrow$ <u>only asymptotically unbiased!</u>
  - For a given total sample size:

$$V_{SRS}(\hat{\mu}) \leq V_{TSS1}(\hat{\mu}) \leq V_{TSS2}(\hat{\mu}) \leq V_{TSS3}(\hat{\mu})$$

  TSS1 is the most efficient TSS for many cluster size distributions

- Non-informative cluster size (i.e. $\mu = \beta_0$)
  - SRS, TSS1, and TSS3: $\hat{\mu} = \sum_{j=1}^{k} \frac{\bar{y}_j}{k}$
  - TSS2: $\hat{\mu} = \frac{\sum_{j=1}^{k} V(\bar{y}_j)^{-1} \bar{y}_j}{\sum_{j=1}^{k} V \bar{y}_j)^{-1}}$, $V(\bar{y}_j)$ variance of cluster j mean
  - For a given total sample size:

$$V_{SRS}(\hat{\mu}) \leq V_{TSS1}(\hat{\mu}) = V_{TSS3}(\hat{\mu}) \leq V_{TSS2}(\hat{\mu})$$

Maastricht University

Optimal design (1/3)

- To maximize power and precision, is it better to sample more clusters or more individuals per cluster?

Optimal design (OD) = # of clusters (k) and # of individuals per cluster (n) that minimize $V(\hat{\mu})$ subject to $C = k\left(c_2 + c_1 n\right)$

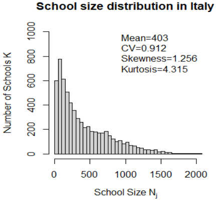
C = budget for sampling and measuring

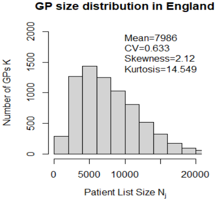
c_2 = (average) cost for sampling a cluster

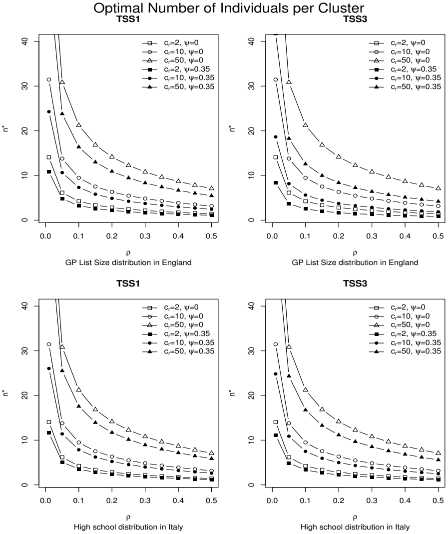
c_1 = (average) cost for sampling an individual from a sampled cluster

- OD maximizes power and precision for a fixed budget C, or minimizes the budget C for the required power or precision level

- Optimal # of clusters: $k^* = \dfrac{C}{c_1(c_r+n^*)}$
- $c_r = \dfrac{c_2}{c_1}$ cluster-to-individual cost ratio
- $\rho = \dfrac{\sigma_v^2}{\sigma_v^2+\sigma_\epsilon^2}$ intraclass correlation coefficient (ICC)
- $\psi = \dfrac{corr(u,N)^2}{1-corr(u,N)^2}$ cluster size informativeness





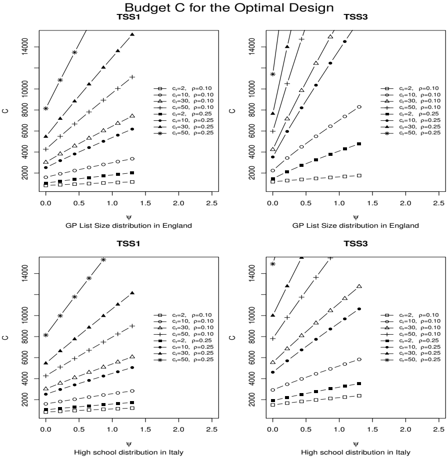


Optimal design (3/3)

- **Given the research budget** C, the OD is robust against misspecification of cluster size informativeness ψ

$$\frac{V(\hat{\mu})\text{ under OD for }\psi > 0}{V(\hat{\mu})\text{ under OD for }\psi = 0} \approx 1$$

- **Given the desired power level and effect size**, the required budget C for the OD is sensitive to misspecification of ψ
- Required C for TSS1 < Required C for TSS2 < Required C for TSS3



- c_r = cluster-to-individual cost ratio
- ψ = cluster size informativeness
- ρ = ICC

Maastricht University

Maximin design (1/2)

- **Local optimality problem:** OD depends on prior knowledge of ICC (ρ), cluster size informativeness (ψ), coefficient of variation (τ_N), skewness (ζ_N), and kurtosis (η_N) of the cluster size distribution

- A solution ⇒ **Maximin approach** (Van Breukelen and Candel [2018]):
  1. Define parameter space (e.g. $\rho \in [0, 0.10]$)
  2. Define design space (i.e. set of all candidate desgins (n, k))
  3. For each design (n, k), find those values of ρ, ψ, τ_N, ζ_N, and η_N that **minimize efficiency** $V(\hat{\mu})^{-1}$
  4. Choose n and k that **maximize** the **minimum efficiency** $V(\hat{\mu})^{-1}$ as determined in step 3

Maximin design (2/2)

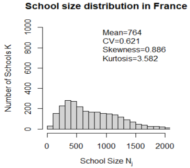
- Maximin design = OD for the **worst-case scenario** of the unknown ICC (ρ), cluster size informativeness (ψ), coefficient of variation (τ_N), skewness (ζ_N), and kurtosis (η_N) of the cluster size distribution
  - Maximin TSS1 = Optimal TSS1 for the largest plausible values of ρ and ψ
  - Maximin TSS2 and TSS3 = Optimal TSS2/TSS3 for the largest plausible values of ρ, ψ, η_N, ζ_N, and τ_N if largest $\tau_N \leq 1$
    - ✓ If largest $\tau_N > 1$, the worst-case values for τ_N and ζ_N are obtained via a numerical evaluation (R function)
    - ✓ It depends on some approximations used to derive $V(\hat{\mu})$, which are accurate (bias $\leq 5\%$) only if $k \geq 20$ clusters are sampled ($k \geq 100$ if η_N and ζ_N are extreme) ⇒ sample 10% more clusters
- Advantages:
  - Simple to implement
  - By maximizing the minimum efficiency over the parameter space, it is **robust against misspecification of the unknown parameters**

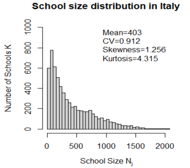
Maastricht University

Application (1/3)

- Sample size calculation for cross-population comparions with TSS1:
  1. Optimal sample sizes
  2. **Optimal budget split** between populations

- Example: Comparison of the average alcohol consumption among adolescents in France and Italy
  $\Rightarrow H_0 : \mu_F = \mu_I$ versus $H_1 : \mu_F \neq \mu_I$





- Inputs for **R code**:
  - Coefficient of variation and skewness of cluster size distribution per country
  - Sampling costs per country
  - Largest realistic ICC: $\rho = 0.10$
  - Largest realistic cluster size informativeness: $\psi = 0.35$
  - Range for the ratio of the outcome SDs: $\frac{\sigma_{y,F}}{\sigma_{y,I}} \in \left[\frac{1}{3}, 3\right]$
  - Effect size $d = 0.5$, power $= 90\%$, and $\alpha = 0.05$

Maastricht University

Budget to detect $d = 0.5$, power level 90%, Type I error rate $\alpha = 0.05$

c_2 = cost for sampling a cluster

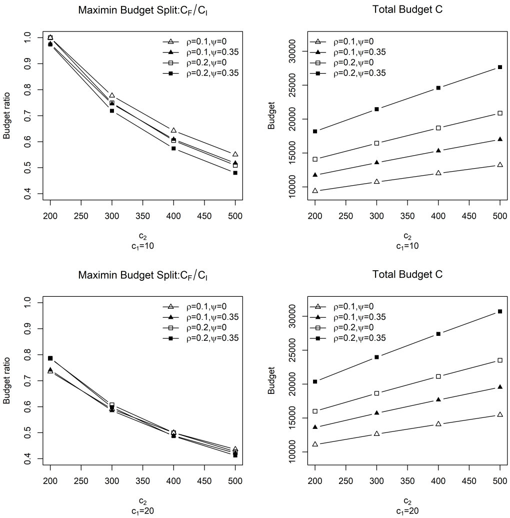
c_1 = cost for sampling an individual

ρ = ICC

ψ = cluster size informativeness

 : $c_1 = 10$ and $c_2 = 200$

 : $c_1 \in \{10, 20\}$ and $c_2 \in [200, 500]$



Application (3/3)

c_2 = cost for sampling a cluster

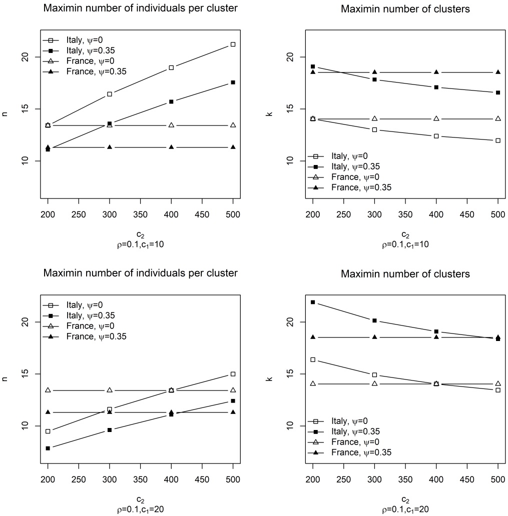
c_1 = cost for sampling an individual

ρ = ICC

ψ = cluster size informativeness

 $= \triangle$: $c_1 = 10$ and $c_2 = 200$

 $= \square$: $c_1 \in \{10, 20\}$ and $c_2 \in [200, 500]$



Application (3/3)

c_2 = cost for sampling a cluster

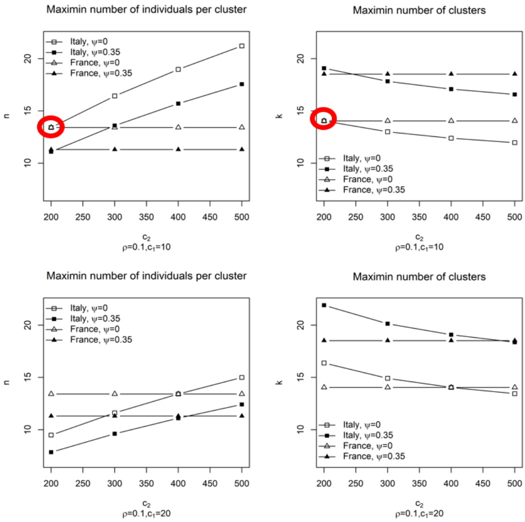
c_1 = cost for sampling an individual

ρ = ICC

ψ = cluster size informativeness

$= \triangle$: $c_1 = 10$ and $c_2 = 200$

$= \square$: $c_1 \in \{10, 20\}$ and $c_2 \in [200, 500]$



Maastricht University

Application (3/3)

c_2 = cost for sampling a cluster

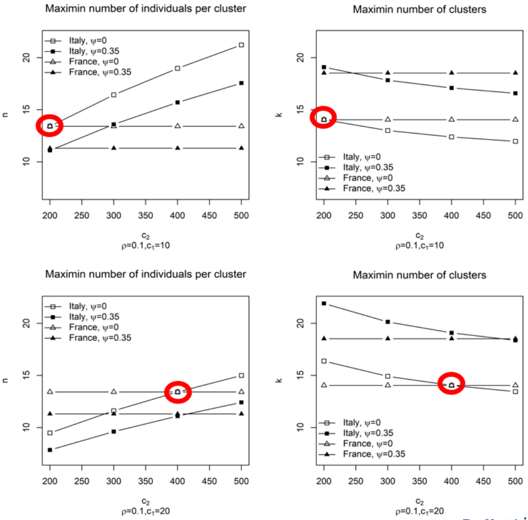
c_1 = cost for sampling an individual

ρ = ICC

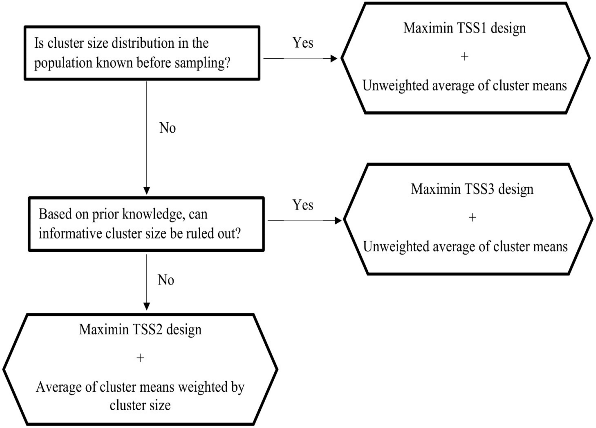
ψ = cluster size informativeness


$= \triangle$: $c_1 = 10$ and $c_2 = 200$


$= \square$: $c_1 \in \{10, 20\}$ and $c_2 \in [200, 500]$



Guidelines



Parameter	**Range of plausible values**
ICC (ρ)	$[0, 0.10]$ in health and medical research (Adams et al. [2004]; Eldridge et al. [2004]) $[0, 0.25]$ in educational research (Hedges and Hedberg [2007]; Shackleton et al. [2016])
Informativeness parameter (ψ)	$[0, 0.35]$ wich corresponds to a correlation of $[-0.51, +0.51]$
CV of cluster size $\left(\tau_N\right)$	$[0, 1]$
Skewness of cluster size $\left(\zeta_N\right)$	$[0.5, 2]$
Kurtosis of cluster size $\left(\eta_N\right)$	$[3, 15]$

Future research

- Binary outcome variables
- Three-level populations
- Extension to non-linear effect of cluster size
- Multipurpose surveys

Thank you for your attention!

francesco.innocenti@maastrichtuniversity.nl

Maastricht University

References

Adams, G., Gulliford, M. C., Ukoumunne, O. C., Eldridge, S., Chinn, S., and Campbell, M. J. (2004). Patterns of intracluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*, 57:785–794.

DGCASIS (2018). Studenti per anno di corso e fascia di etá: scuola statale. Technical report, Direzione generale per i contratti, gli acquisti e per i sistemi informativi e la statistica.

Eldridge, S. M., Ashby, D., Feder, G. S.and Rudnicka, A. R., and Ukoumunne, O. C. (2004). Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clinical Trials*, 1:80–90.

Hedges, L. V. and Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29:60–87.

Innocenti, F., Candel, M. J. J. M., Tan, F. E. S., and van Breukelen, G. J. P. (2019). Relative efficiencies of two-stage sampling schemes for mean estimation in multilevel populations when cluster size is informative. *Statistics in Medicine*, 38:1817–1834.

Innocenti, F., Candel, M. J. J. M., Tan, F. E. S., and van Breukelen, G. J. P. (2021). Optimal two-stage sampling for mean estimation in multilevel populations when cluster size is informative. *Statistical Methods in Medical Research*, 30:357–375.

McNeely, C., Nonnemaker, J., and Blum, R. (2002). Promoting school connectedness: Evidence from the national longitudinal study of adolescent health. *Journal of School Health*, 72:138–146.

MENJVA (2015). L'Éducation nationale publie les effectifs des établissements scolaires. Technical report, Ministère de l'Éducation nationale, de la Jeunesse et de la Vie associative.

Nevalainen, J., Datta, S., and Oja, H. (2014). Inference on the marginal distribution of clustered data with informative cluster size. *Statistical Papers*, 55:71–92.

Resnick, M. D., Bearman, P. S., Blum, R. W., and et al. (1997). Protecting adolescents from harm: Findings from the national longitudinal study on adolescent health. *Journal of the American Medical Association*, 278:823–832.

Salt, K. (2017). Patients registered at a gp practice. october 2017; special topic-practice list size comparison. Technical report, Health and Social Care Information Centre.

Seaman, S., Pavlou, M., and Copas, A. (2014). Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Statistics in Medicine*, 33:5371–5387.

Shackleton, N., Hale, D., Bonell, C., and Viner, R. M. (2016). Intraclass correlation values for adolescent health outcomes in secondary schools in 21 european countries. *SSM Population Health*, 2:217–225.

Thompson, D., Iachan, R., Overpeck, M., and et al. (2006). School connectedness in the health behavior in school-aged children study: The role of student, school, and school neighborhood connectedness. *Journal of School Health*, 76:379–386.

Van Breukelen, G. J. P. and Candel, M. J. J. M. (2018). Efficient design of cluster randomized trials with treatment-dependent costs and treatment-dependent unknown variances. *Statistics in Medicine*, 37:3027–3046.

Appendices

1. Sampling variances
   - Equations
   - Simulation study
2. Optimal design
   - Equations
   - Robustness against misspecification of ψ
3. Relative efficiency for a fixed total sample size
   - Equations
   - Figures
4. Relative efficiency for a fixed budget
   - Results
   - Equations
   - Figures
5. Sample size calculation for cross-population comparisons
6. Examples of real cluster size distributions
7. Model-based versus design-based inference

Maastricht University

Sampling variances

- SRS:

$$V(\hat{\mu}) = \frac{\sigma_y^2}{m}\left\{1 + \rho\psi\left[\tau_N\left(\zeta_N - \tau_N\right) + 1\right]\right\}$$

- TSS1:

$$V(\hat{\mu}) = \frac{\sigma_y^2}{nk}\left\{1 + \rho\left[(n-1) + n\psi\left(\tau_N\left(\zeta_N - \tau_N\right) + 1\right)\right]\right\}$$

- TSS2:

$$V(\hat{\mu}) \approx \frac{\sigma_y^2}{nk}\left\{1 + \rho\left[n\left(\left(\tau_N^2 + 1\right) + \psi\left(\tau_N^4 + \tau_N^2\left(\eta_N - 3\right) + 2\zeta_N\tau_N\left(1 - \tau_N^2\right) + 1\right)\right) - 1\right]\right\}$$

- TSS3:

$$V(\hat{\mu}) \approx \frac{\sigma_y^2}{nk}\left\{\tau_N^2 + 1 + \rho\left[\left(\tau_N^2 + 1\right)(n-1) + n\psi\left(\tau_N^4 + \tau_N^2\left(\eta_N - 3\right) + 2\zeta_N\tau_N\left(1 - \tau_N^2\right) + 1\right)\right]\right\}$$

- $V(\hat{\mu})$ for TSS2 and TSS3 are derived using the **delta method** (see subsection 2) and are based on a large k approximation (i.e. k such that $\frac{\tau_N^2}{k} \approx 0$, $\frac{k-1}{k} \approx 1$, and $\frac{k-3}{k-1} \approx 1$)

Simulation study

$V(\hat{\mu})$ for TSS2 and TSS3 are based on the delta method, so their accuracy were evaluated through a simulation study:

- Sampling $k = 20$ clusters guarantees nearly unbiased estimates of μ under TSS2 and TSS3
- Sampling $k = 20$ clusters guarantees fair accuracy (i.e. relative bias $\leq 5\%$) of $V(\hat{\mu})$ for TSS2 and TSS3 when $|corr(u, N)| \leq 0.75$, $\rho \leq 0.3$, and ζ_N and η_N are relatively close (say, ± 1.5) to those of the Normal distribution (i.e. $\zeta_N = 0$ and $\eta_N = 3$)
- For cluster size distributions with extreme skewness and kurtosis (e.g. $\zeta_N \geq 2$ and $\eta_N \geq 9$) at least $k = 100$ clusters must be sampled to achieve a reasonable accuracy (i.e. bias $\leq 6\%$) of $V(\hat{\mu})$, for $|corr(u, N)| \leq 0.5$ and $\rho \leq 0.3$
- These lower-bounds for k (i.e. 20 and 100) guarantee the corresponding accuracy level across different values for n (at least for $2 \leq n \leq 100$) (Shackleton et al. [2016]: in the ESPAD study $k \in [36, 531]$, median=123, and $\bar{n} \in [5.92, 119.62]$, median=20.74)

Optimal designs (1/2)

- SRS:

$$V\left(\hat{\mu}\right)^{*} = \frac{c_{srs}\sigma_y^2\left(1+\rho\psi\left[\tau_N\left(\zeta_N-\tau_N\right)+1\right]\right)}{C-c_0}$$

where c_{srs} is the average cost for sampling an individual directly from the population, and c_0 represents the extra-cost due to constructing the sampling frame for a SRS compared with the sampling frame for a TSS.

- TSS1: $n^{*} = \sqrt{c_r\left(\frac{1-\rho}{\rho}\right)\left(\frac{1}{1+\psi\left[\tau_N\left(\zeta_N-\tau_N\right)+1\right]}\right)}$

$$V\left(\hat{\mu}\right)^{*} = \frac{c_1\sigma_y^2\left(\sqrt{c_r\rho\left(1+\psi\left[\tau_N\left(\zeta_N-\tau_N\right)+1\right]\right)}+\sqrt{1-\rho}\right)^2}{C}$$

Maastricht University

Optimal designs (2/2)

- TSS2: $n^* = p^*\theta_N = \sqrt{c_r\left(\frac{1-\rho}{\rho}\right)\frac{1}{(\tau_N^2+1)+\psi\left[\tau_N^4+\tau_N^2(\eta_N-3)+2\zeta_N\tau_N(1-\tau_N^2)+1\right]}}$

$$V(\hat{\mu})^* = \frac{c_1\sigma_y^2\left(\sqrt{c_r\rho\left[\tau_N^2+1+\psi(\tau_N^4+\tau_N^2(\eta_N-3)+2\zeta_N\tau_N(1-\tau_N^2)+1)\right]}+\sqrt{1-\rho}\right)^2}{C}$$

- TSS3: $n^* = \sqrt{c_r\left(\frac{1-\rho}{\rho}\right)\frac{(\tau_N^2+1)}{(\tau_N^2+1)+\psi\left[\tau_N^4+\tau_N^2(\eta_N-3)+2\zeta_N\tau_N(1-\tau_N^2)+1\right]}}$

$$V(\hat{\mu})^* = \frac{c_1\sigma_y^2\left(\sqrt{c_r\rho\left[\tau_N^2+1+\psi(\tau_N^4+\tau_N^2(\eta_N-3)+2\zeta_N\tau_N(1-\tau_N^2)+1)\right]}+\sqrt{(1-\rho)(\tau_N^2+1)}\right)^2}{C}$$

- optimal number of clusters for any TSS: $k^* = \frac{C}{c_1(c_r+n^*)}$

Maastricht University

Robustness of the optimal design against misspecification of ψ

Assuming the general practice list size distribution in England, $\rho = 0.05$, $c_r = 10$, and $C/c_1 = 1000$

	TSS1		TSS2		TSS3	
	$\psi = 0$	$\psi = 1/3$	$\psi = 0$	$\psi = 1/3$	$\psi = 0$	$\psi = 1/3$
n^*	13.78	10.74	11.65	7.01	13.78	8.30
k^*	42.04	48.22	46.2	58.79	42.04	54.65
$Var(\hat{\mu})/\sigma_y^2$ if $\psi = 1/3$	0.00360	0.00354	0.00595	0.00559	0.00690	0.00647
$\frac{Var(\hat{\mu}\vert\psi = 1/3)}{Var(\hat{\mu}\vert\psi = 0)}$	0.983		0.939		0.938	
	$\psi = 0$	$\psi = 1$	$\psi = 0$	$\psi = 1$	$\psi = 0$	$\psi = 1$
n^*	13.78	8.04	11.65	4.65	13.78	5.50
k^*	42.04	55.44	46.2	68.27	42.04	64.52
$Var(\hat{\mu})/\sigma_y^2$ if $\psi = 1$	0.00514	0.00478	0.01129	0.00944	0.01276	0.01057
$\frac{Var(\hat{\mu}\vert\psi = 1)}{Var(\hat{\mu}\vert\psi = 0)}$	0.930		0.836		0.828	

Maastricht University

Relative efficiency for a fixed total sample size (1/4)

- $RE\,(TSS1\ vs\ SRS) = \dfrac{\left(1-corr(u_j,N_j)^2\right)+corr(u_j,N_j)^2\rho[\tau_N(\zeta_N-\tau_N)+1]}{\left(1-corr(u_j,N_j)^2\right)[1+(n-1)\rho]+corr(u_j,N_j)^2 n\rho[\tau_N(\zeta_N-\tau_N)+1]}$
- $RE\,(TSS2\ vs\ SRS) =$

$$\frac{\left(1-corr(u_j,N_j)^2\right)+corr(u_j,N_j)^2\rho[\tau_N(\zeta_N-\tau_N)+1]}{\left(1-corr(u_j,N_j)^2\right)\left[1+\left(\bar{n}\left(\frac{k\left(\tau_N^2+1\right)}{\tau_N^2+k}\right)-1\right)\rho\right]+corr(u_j,N_j)^2\bar{n}\rho\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N-\frac{k-3}{k-1}+\tau_N(\tau_N-2\zeta_N)\right)+2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N-\tau_N)+1\right]}$$

- $RE\,(TSS3\ vs\ SRS) =$

$$\frac{\left(1-corr(u_j,N_j)^2\right)+corr(u_j,N_j)^2\rho[\tau_N(\zeta_N-\tau_N)+1]}{\left(1-corr(u_j,N_j)^2\right)\left[\left(\frac{k\left(\tau_N^2+1\right)}{\tau_N^2+k}\right)(1+(n-1)\rho)\right]+corr(u_j,N_j)^2 n\rho\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N-\frac{k-3}{k-1}+\tau_N(\tau_N-2\zeta_N)\right)+2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N-\tau_N)+1\right]}$$

Maastricht University

Relative efficiency for a fixed total sample size (2/4)

- $RE\,(TSS2\ vs\ TSS1) =$

$$\frac{\left(1-corr(u_j,N_j)^2\right)[1+(n-1)\rho]+corr(u_j,N_j)^2 n\rho[(\tau_N(\zeta_N-\tau_N)+1)]}{\left(1-corr(u_j,N_j)^2\right)\left[1+\left(\bar{n}\left(\frac{k(\tau_N^2+1)}{\tau_N^2+k}\right)-1\right)\rho\right]+corr(u_j,N_j)^2\bar{n}\rho\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N-\frac{k-3}{k-1}+\tau_N(\tau_N-2\zeta_N)\right)+2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N-\tau_N)+1\right]}$$

- $RE\,(TSS3\ vs\ TSS1) =$

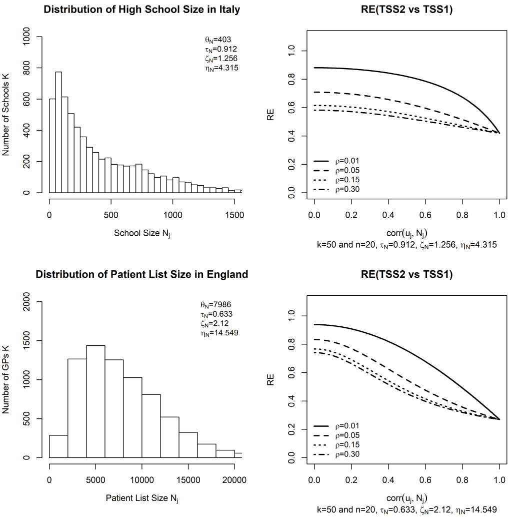
$$\frac{\left(1-corr(u_j,N_j)^2\right)[1+(n-1)\rho]+corr(u_j,N_j)^2 n\rho[\tau_N(\zeta_N-\tau_N)+1]}{\left(1-corr(u_j,N_j)^2\right)\left[\left(\frac{k(\tau_N^2+1)}{\tau_N^2+k}\right)(1+(n-1)\rho)\right]+corr(u_j,N_j)^2 n\rho\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N-\frac{k-3}{k-1}+\tau_N(\tau_N-2\zeta_N)\right)+2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N-\tau_N)+1\right]}$$

- $RE\,(TSS3\ vs\ TSS2) =$

$$\frac{\left(1-corr(u_j,N_j)^2\right)\left[1+\left(\bar{n}\left(\frac{k(\tau_N^2+1)}{\tau_N^2+k}\right)-1\right)\rho\right]+corr(u_j,N_j)^2\bar{n}\rho\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N-\frac{k-3}{k-1}+\tau_N(\tau_N-2\zeta_N)\right)+2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N-\tau_N)+1\right]}{\left(1-corr(u_j,N_j)^2\right)\left[\left(\frac{k(\tau_N^2+1)}{\tau_N^2+k}\right)(1+(n-1)\rho)\right]+corr(u_j,N_j)^2 n\rho\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N-\frac{k-3}{k-1}+\tau_N(\tau_N-2\zeta_N)\right)+2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N-\tau_N)+1\right]}$$

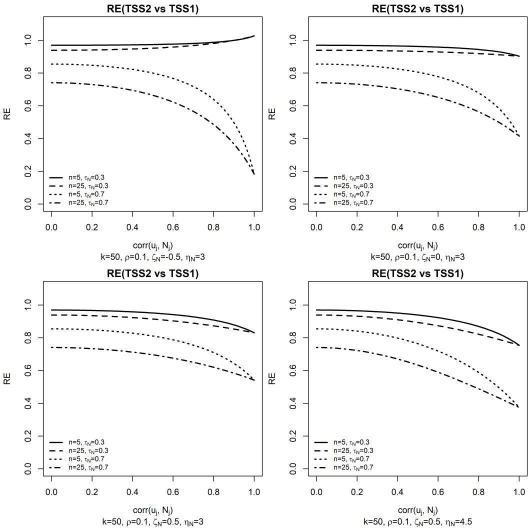
- TSS1 is more efficient than TSS2 and TSS3 if one of the following conditions is met: the cluster size distribution is positively skewed with $\tau_N \in [0, \zeta_N]$, or is symmetric with $\tau_N \in [0, 1]$ and $k \in \left[1,\ \frac{(2-\tau_N^2)+\sqrt{2-\tau_N^2}}{(1-\tau_N^2)}\right]$, or is Normal.

Relative efficiency for a fixed total sample size (3/4)



Maastricht University

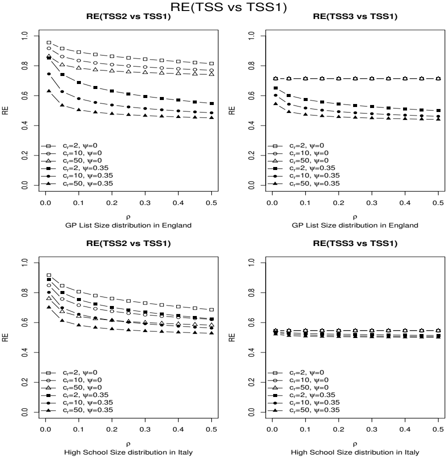
Relative efficiency for a fixed total sample size (4/4)



Maastricht University

Relative efficiency for a given budget

- Relative efficiency of OD_1 versus OD_2: $RE = \frac{V_{OD_2}(\hat{\mu})}{V_{OD_1}(\hat{\mu})}$
- Informative cluster size
  - RE depends on cluster size distribution
  - TSS1 is the most efficient TSS for many cluster size distributions
  - TSS3 is always the least efficient TSS
- Non-informative cluster size
  - TSS1 and TSS3 are equally efficient and outperform TSS2



- c_r = cluster-to-individual cost ratio
- ψ = cluster size informativeness
- ρ = ICC

Maastricht University

Relative efficiency for a fixed budget (1/6)

- TSS1 vs SRS:

$$\frac{1+\rho\psi[\tau_N(\zeta_N-\tau_N)+1]}{\left(\sqrt{c_r\rho(1+\psi[\tau_N(\zeta_N-\tau_N)+1])}+\sqrt{1-\rho}\right)^2}\times\left(\frac{c_{srs}}{c_1}\right)\times\left(\frac{C}{C-c_0}\right)$$

which is ≤ 1 if $\zeta_N \geq \tau_N - \frac{1}{\tau_N} - \frac{1}{\tau_N\psi}$ and $\left(\frac{c_{srs}}{c_1}\right) = \left(\frac{C}{C-c_0}\right) = 1$

- TSS2 vs SRS:

$$\frac{1+\rho\psi[\tau_N(\zeta_N-\tau_N)+1]}{\left(\sqrt{c_r\rho[\tau_N^2+1+\psi(\tau_N^4+\tau_N^2(\eta_N-3)+2\zeta_N\tau_N(1-\tau_N^2)+1)]}+\sqrt{1-\rho}\right)^2}\times\left(\frac{c_{srs}}{c_1}\right)\times\left(\frac{C}{C-c_0}\right)$$

which is ≤ 1 if $\zeta_N \leq \tau_N - \frac{1}{\tau_N}$ or $\zeta_N \geq \tau_N + \frac{1}{\tau_N c_r} - \frac{1}{\tau_N}$ or $N_j \sim N\left(\theta_N, \sigma_N^2\right)$, and $\left(\frac{c_{srs}}{c_1}\right) = \left(\frac{C}{C-c_0}\right) = 1$

- TSS3 vs SRS:

$$\frac{1+\rho\psi[\tau_N(\zeta_N-\tau_N)+1]}{\left(\sqrt{c_r\rho[\tau_N^2+1+\psi(\tau_N^4+\tau_N^2(\eta_N-3)+2\zeta_N\tau_N(1-\tau_N^2)+1)]}+\sqrt{(1-\rho)(\tau_N^2+1)}\right)^2}\times\left(\frac{c_{srs}}{c_1}\right)\times\left(\frac{C}{C-c_0}\right)$$

which is ≤ 1 if $\zeta_N \leq \tau_N - \frac{1}{\tau_N}$ or $\zeta_N \geq \tau_N + \frac{1}{\tau_N c_r} - \frac{1}{\tau_N}$ or $N_j \sim N\left(\theta_N, \sigma_N^2\right)$, and $\left(\frac{c_{srs}}{c_1}\right) = \left(\frac{C}{C-c_0}\right) = 1$

Relative efficiency for a fixed budget (2/6)

- TSS2 vs TSS1:

$$\frac{\left(\sqrt{c_r\rho[1+\psi(\tau_N(\zeta_N-\tau_N)+1)]}+\sqrt{1-\rho}\right)^2}{\left(\sqrt{c_r\rho[\tau_N^2+1+\psi(\tau_N^4+\tau_N^2(\eta_N-3)+2\zeta_N\tau_N(1-\tau_N^2)+1)]}+\sqrt{1-\rho}\right)^2}$$

which is ≤ 1 if $\tau_N - \frac{1}{\tau_N} - \frac{1}{\tau_N\psi} \leq \zeta_N \leq \tau_N - \frac{1}{\tau_N}$ or $\zeta_N \geq \tau_N$ or $N_j \sim N\left(\theta_N, \sigma_N^2\right)$

- TSS3 vs TSS1:

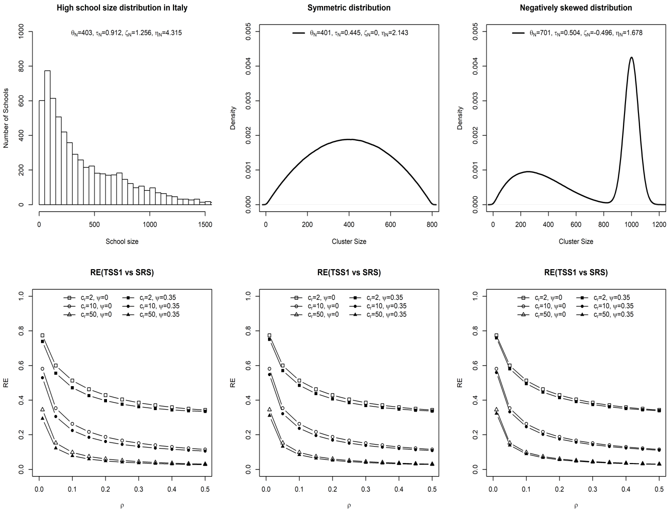
$$\frac{\left(\sqrt{c_r\rho[1+\psi(\tau_N(\zeta_N-\tau_N)+1)]}+\sqrt{1-\rho}\right)^2}{\left(\sqrt{c_r\rho[\tau_N^2+1+\psi(\tau_N^4+\tau_N^2(\eta_N-3)+2\zeta_N\tau_N(1-\tau_N^2)+1)]}+\sqrt{(1-\rho)(\tau_N^2+1)}\right)^2}$$

which is ≤ 1 if $\tau_N - \frac{1}{\tau_N} - \frac{1}{\tau_N\psi} \leq \zeta_N \leq \tau_N - \frac{1}{\tau_N}$ or $\zeta_N \geq \tau_N$ or $N_j \sim N\left(\theta_N, \sigma_N^2\right)$

- TSS3 vs TSS2:

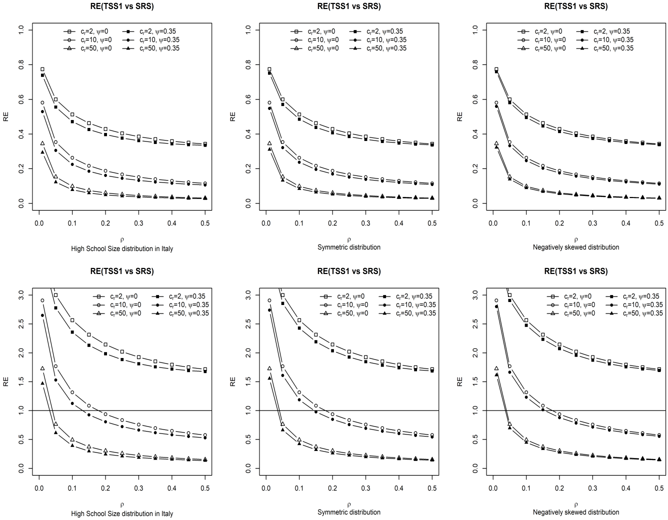
$$\frac{\left(\sqrt{c_r\rho[\tau_N^2+1+\psi(\tau_N^4+\tau_N^2(\eta_N-3)+2\zeta_N\tau_N(1-\tau_N^2)+1)]}+\sqrt{1-\rho}\right)^2}{\left(\sqrt{c_r\rho[\tau_N^2+1+\psi(\tau_N^4+\tau_N^2(\eta_N-3)+2\zeta_N\tau_N(1-\tau_N^2)+1)]}+\sqrt{(1-\rho)(\tau_N^2+1)}\right)^2} \leq 1$$

Relative efficiency for a fixed budget (3/6)



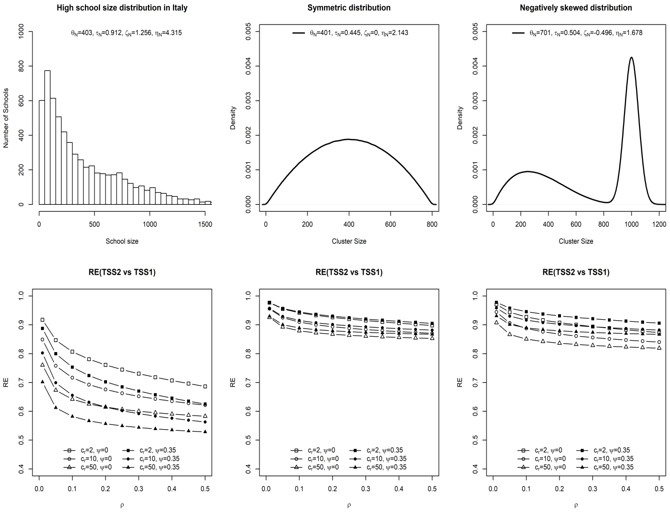
No extra costs for SRS

Relative efficiency for a fixed budget (4/6)



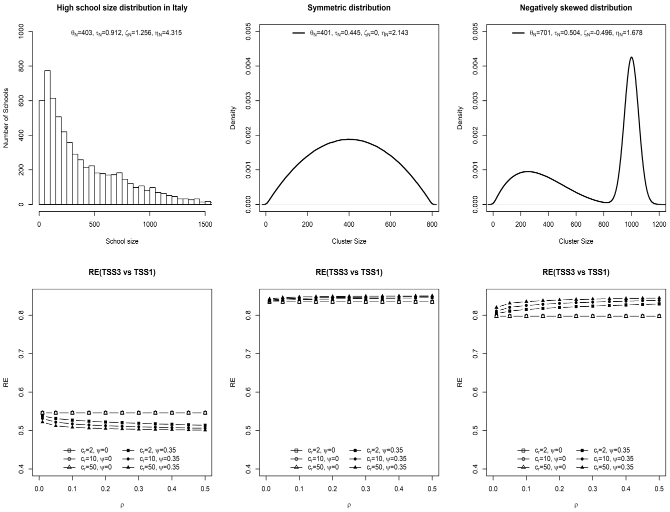
Extra costs for SRS: $c_{SRS} = 4c_1$ and $c_0 = 20\%$

Relative efficiency for a fixed budget (5/6)



Maastricht University

Relative efficiency for a fixed budget (6/6)



Sample size calculation for cross-population comparisons (1/3)

Example: $H_0 : \mu_F = \mu_I$ versus $H_1 : \mu_F \neq \mu_I$ where e.g.  and 

1. Specify sampling costs $\left(c_1, c_2\right)$ per population, largest realistic ρ and ψ values, smallest relevant standardized difference d, $\frac{\sigma_{y,F}}{\sigma_{y,I}} \in \left[\frac{1}{q}, q\right]$, power level and Type I error rate
2. Compute the maximum allowable $V\left(\hat{\mu}_F - \hat{\mu}_I\right)$ to guarantee the desired power
3. Compute the maximin n_F^{MD} and n_I^{MD}
4. Compute the maximin budget split $\frac{C_F}{C_I}$
5. Compute the total budget C by equating the maximum variance for the maximin design with $V\left(\hat{\mu}_F - \hat{\mu}_I\right)$ as computed in step 2
6. Compute the separate budget per population using C and $\frac{C_F}{C_I}$
7. Compute the maximin k_F^{MD} and k_I^{MD}

Sample size calculation for cross-population comparisons (2/3)

1. Specify $c_{1,F}$, $c_{1,I}$, $c_{2,F}$, $c_{2,I}$, $\rho\,(max)$, $\psi\,(max)$, $\min\left\{\mu_F - \mu_I\right\}$, $V_{max} \geq \sigma^2_{y,F} + \sigma^2_{y,I}$, $\frac{\sigma_{y,F}}{\sigma_{y,I}} \in \left[\frac{1}{q}, q\right]$, and $d = \sqrt{\frac{\mu_F - \mu_I}{V_{max}/2}}$
2. Compute $V\left(\hat{\mu}_F - \hat{\mu}_I\right) = \left(\frac{\mu_F - \mu_I}{z_{1-\frac{\alpha}{2}} + z_{1-\gamma}}\right)^2$
3. Compute the maximin n_F^{MD} and n_I^{MD} (see subsection 1)
4. Compute the maximin budget split $\frac{C_F}{C_I}$ using h and $\left[\frac{1}{q}, q\right]$ (see next slide)
5. Compute the total budget C by equating the maximum variance for the maximin design with $V\left(\hat{\mu}_F - \hat{\mu}_I\right)$ as computed in step 2
6. Using C from step 6 and $\frac{C_F}{C_I}$ from step 5, compute C_F and C_I
7. Compute the maximin k_F^{MD} and k_I^{MD} (see subsection 1)

Sample size calculation for cross-population comparisons (3/3)

$$h = \sqrt{\frac{g_F\left(\rho\left(max\right),\ \psi\left(max\right)\right)}{g_I\left(\rho\left(max\right),\ \psi\left(max\right)\right)}}$$

$$= \sqrt{\frac{c_{1,F}\left(\sqrt{c_{r,F}\rho_F\left(1+\psi_F\left[\tau_{N,F}\left(\zeta_{N,F}-\tau_{N,F}\right)+1\right]\right)}+\sqrt{1-\rho_F}\right)^2}{c_{1,I}\left(\sqrt{c_{r,I}\rho_I\left(1+\psi_I\left[\tau_{N,I}\left(\zeta_{N,I}-\tau_{N,I}\right)+1\right]\right)}+\sqrt{1-\rho_I}\right)^2}}$$

Relation of h to q	Maximin budget split	Maximum variance for MD
$\frac{1}{q} \leq h \leq q$	h^2	$\frac{g_I(\rho(max),\,\psi(max))V_{max}}{C} \times \left(1+h^2\right)$
$h > q$	hq	$\frac{g_I(\rho(max),\,\psi(max))V_{max}}{C} \times \frac{(hq+1)^2}{(q^2+1)}$
$h < \frac{1}{q}$	$\frac{h}{q}$	$\frac{g_I(\rho(max),\,\psi(max))V_{max}}{C} \times \frac{(h+q)^2}{(q^2+1)}$

Real cluster size distributions

Cluster Size distribution	θ_N	τ_N	ζ_N	η_N
GP List size distribution in England (Salt [2017])	7,986	0.633	2.12	14.549
High School size distribution in Italy (DGCASIS [2018])	403	0.912	1.256	4.315
High School size distribution in France (MENJVA [2015])	764	0.621	0.886	3.582
Lower Secondary School size distribution in Italy (DGCASIS [2018])	225	0.789	1.351	5.303
Lower Secondary School size distribution in France (MENJVA [2015])	493	0.387	0.63	5.47
Primary School size distribution in Italy (DGCASIS [2018])	171	0.761	1.451	5.740
Primary School size distribution in France (MENJVA [2015])	135	0.71	1.045	4.084

Model-based versus design-based inference (1/2)

Model-based approach:

- Y_{ij} is random
- Inference based on the stochastic model for Y_{ij}
- Advantage: It simplifies sample size planning and sampling schemes comparisons

Design-based approach:

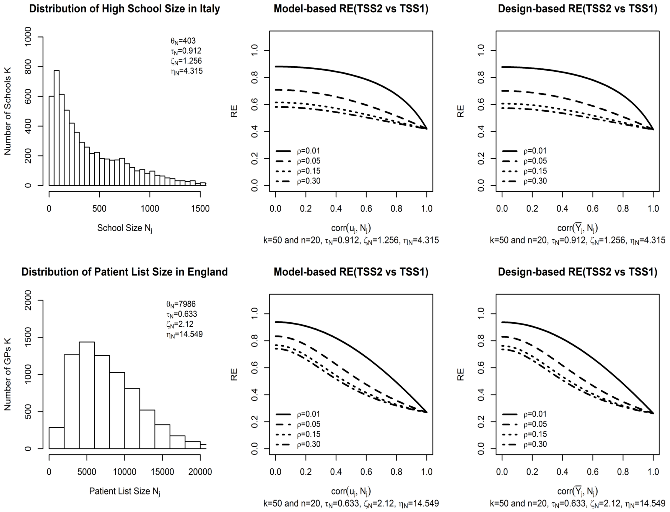
- Y_{ij} is fixed but unknown. The inclusion indicator I_{ij} is random (i.e. $I_{ij} = 1$ with π_{ij}, and $I_{ij} = 0$ otherwise)
- Inference based on the distribution of I_{ij} over repeated sampling with a given sampling design
- Advantage: Robustness

In the considered setting, the two approaches yield almost the same results (if the model assumptions are met) (Innocenti et al. [2019]):

- same estimators of the population mean
- approximately the same relative efficiencies (i.e. for k sufficiently large)

Maastricht University

Model-based versus design-based inference (2/2)



Maastricht University