Sample Size Calculations

For diagnostic accuracy & impact studies

Samuel G. Schumacher - McGill University
samuel.schumacher@mail.mcgill.ca

Epidemiology 1

Sample size calculations in randomised trials: mandatory and mystical

Lancet 2005; 365: 1348–53

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA
(K F Schulz PhD, D A Grimes MD)

Correspondence
Dr Kenneth
KSch

Kenneth F Schulz, David A Grimes

Investigators sho
describ

unassa
calculate sampl
details in their published
ogists describe the approaches in books and artic
Protocol committees and ethics review boards require adherence. CONSORT reporting guidelines clearly specify the reporting of sample size calculations.[1,2] Almost everyone agrees.

An important impetus to this unanimity burst on the medical world more than a quarter of a century ago. group of researchers, led by Tom Chalmers, publish

citati ally, that troubled him.[4] He regarded it damaging paper that he had ever d hy? We will describe his concerns later, so

ponents of sample size calculations

lculating sample sizes for trials with dichotomous outcomes (eg, sick *vs* well) requires four components: type I error (α), power, event rate in the control group, and a treatment effect of interest (or analogously an event rate in the treatment group). These basic components persist through calculations with other types of outcomes, except other assumptions can be necessary. For example, with quantitative outcomes and a typical statistical test, investigators might assume a difference between means and a variance for the means.

In clinical research, hypothesis testing risks two fundamental errors (panel 1). First, researchers can conclude that two treatments differ when, in fact, they do not. This type I error (α) measures the probability of making this false-positive conclusion. Conventionally, α is most frequently set at 0·05, meaning that investigators desire a less than 5% chance of making a false-positive conclusion. Second, researchers can conclude that two treatments do not differ when, in fact, they do—ie, a false-negative conclusion. This type II error (β) measures the probability of this false-negative conclusion. Conventionally, investigators set β at 0·20, meaning that they desire less than a 20% chance of making a false-negative conclusion.

Power derives from β error. Mathematically, it is the complement of β (1–β) and represents the probability of avoiding a false-negative conclusion. For example, for

sample size calculations = “predicting the future”

Outline

Part I

1. Intro on sample size & power
2. Sample size for accuracy
3. Literature examples for accuracy

Part II

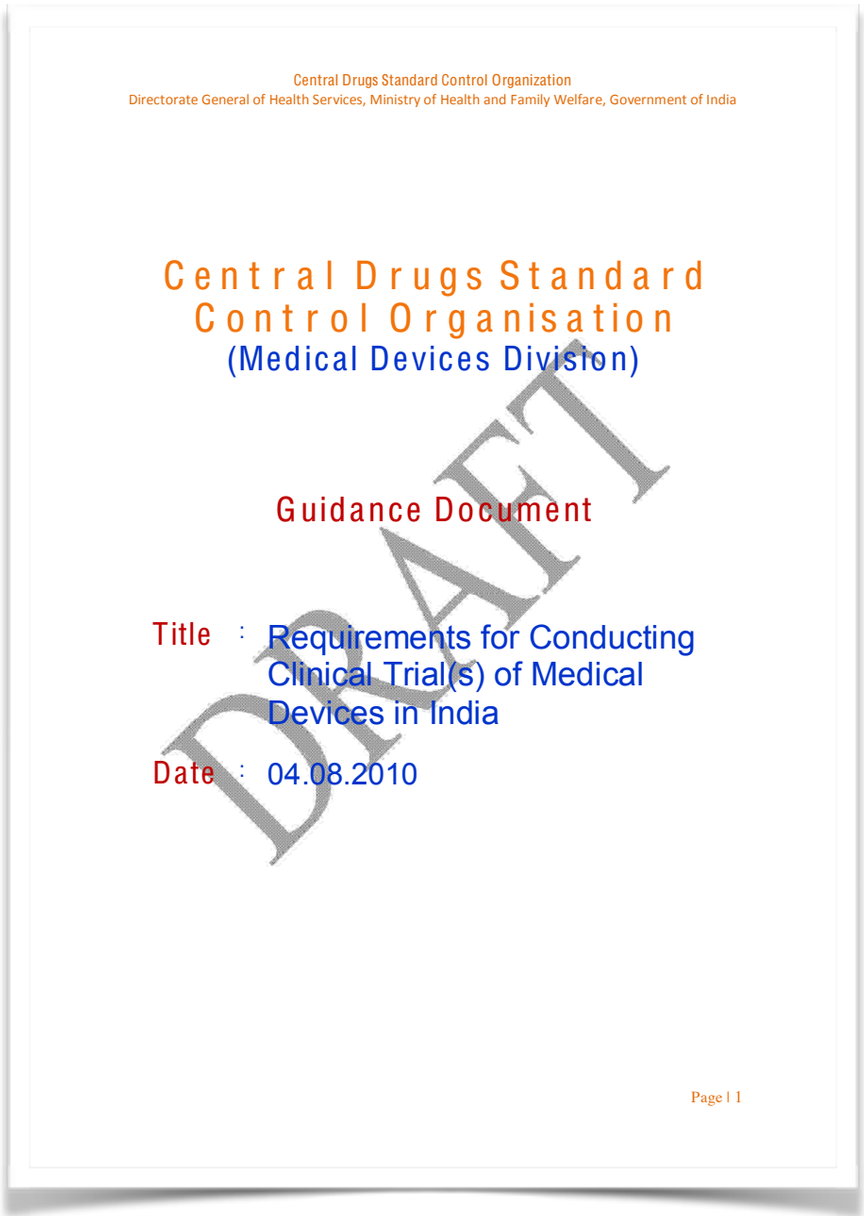
1. Additional considerations
2. Sample size for impact
3. Literature examples for impact

PART I

Intro on power & sample size

Why bother?

What happens if I don't do it?



Central Drugs Standard Control Organization
Directorate General of Health Services, Ministry of Health and Family Welfare, Government of India

Central Drugs Standard Control Organisation
(Medical Devices Division)

DRAFT

Guidance Document

Title : Requirements for Conducting Clinical Trial(s) of Medical Devices in India

Date : 04.08.2010

Page | 1

http://cdsco.nic.in/

o. Data Analysis:

Provide details of the statistical approach to be followed including sample size, how the sample size was determined, including as mptions made in making this determination, efficacy endpoints (primary as well as secondary) and safety endpoints.

DRAFT

Why bother?

What happens if I don't do it?

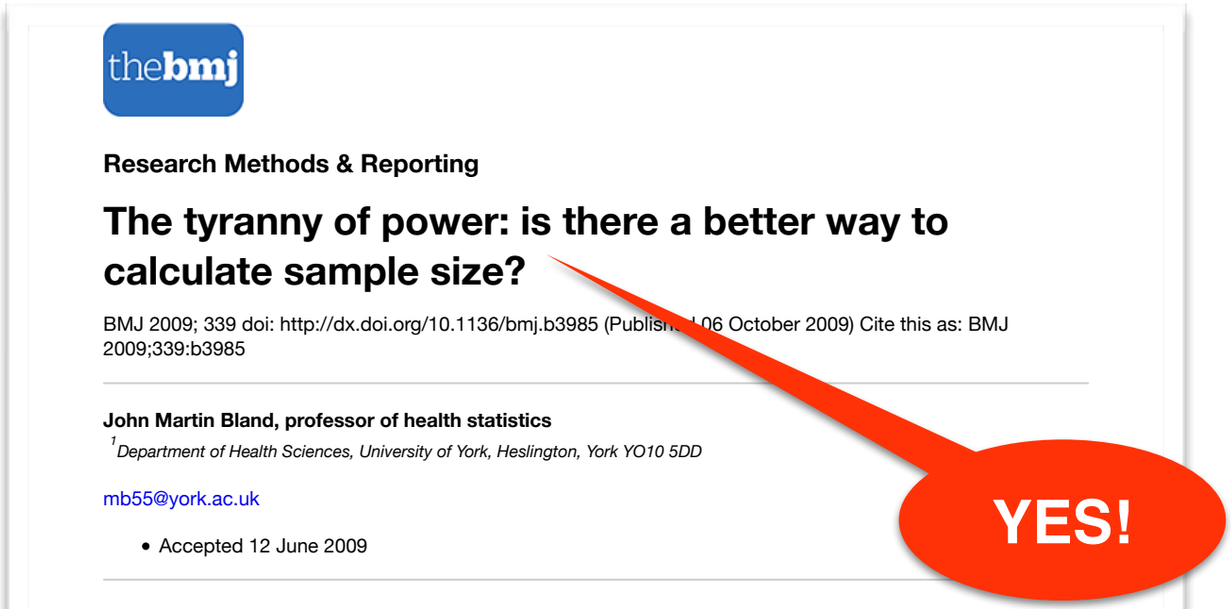
1. study **too small**: imprecision —> ethics

2. study **too large**: "too precise" —> ethics, $$$

3. **study planning**: design, logistics, feasibility etc.

Power vs Estimation

Two approaches to sample size calculations



Power vs Estimation

Two approaches to sample size calculations

1. **Power-based** ~ significance testing
   - useful if question is: “Is there a statistically significant difference between X and Y?”
   - usually NOT helpful for accuracy

2. **Estimation-based** ~ confidence interval
   - useful if question is: “How precisely can we estimate X or the difference between X and Y?”
   - method of choice for accuracy
   - generally preferable method
   - easier to understand

To understand sample size…

… you need to understand concepts of random variation & precision

- 1 sample: S-C+
- new test “T” is positive
  - ➡ YAY, my test is **100% sensitive!!! …right?**
  - ➡ **NO!!!** need to account for random variation
  - ➡ My **estimate** for Sensitivity is

    100% (95%CI **2.5% — 100%**)

To understand sample size…

… you need to understand concepts of random variation & precision

- 1 sampl… 100%)
- 2 sampl… 100%)
- 3 sampl… 100%)
- 4 samples S-C+T+: 100% (95%CI 40% — 100%)
- 5 sam… 0%)
- 10 sa… 00%)
- 50 sa… 00%)
- 100 samples S-C+T+: 100% (95%CI 96% — 100%)

precision ~ sample size

How many patients/samples do I need
to estimate accuracy with **acceptable precision**?

How is it actually done?

Making use of existing formulas

Table 1. Estimating a population proportion with specified absolute precision

$$n = z^2_{1-\alpha/2} P(1-P)/d^2$$

Table 2. Estimating a population proportion with specified relative precision

$$n = z^2_{1-\alpha/2}(1-P)/\varepsilon^2 P$$

Table 4. Estimating the difference between two population proportions with specified absolute precision

$$n = z^2_{1-\alpha/2}[P_1(1-P_1)+P_2(1-P_2)]/d^2$$

or

$$n = z^2_{1-\alpha/2} V/d^2$$

where

$$V = P_1(1-P_1)+P_2(1-P_2)$$

Table 5. Hypothesis tests for two population proportions

For a one-sided test

$$n = \{z_{1-\alpha}\sqrt{[2\bar{P}(1-\bar{P})]} + z_{1-\beta}\sqrt{[P_1(1-P_1)+P_2(1-P_2)]}\}^2/(P_1-P_2)^2$$

where

$$\bar{P} = (P_1+P_2)/2.$$

For a two-sided test

$$n = \{z_{1-\alpha/2}\sqrt{[2\bar{P}(1-\bar{P})]} + z_{1-\beta}\sqrt{[P_1(1-P_1)+P_2(1-P_2)]}\}^2/(P_1-P_2)^2.$$

For a one-sided test for small proportions

$$n = (z_{1-\alpha}+z_{1-\beta})^2/[0.0006(\arcsin\sqrt{P_2}-\arcsin\sqrt{P_1})^2].$$

For a two-sided test for small proportions

$$n = (z_{1-\alpha/2}+z_{1-\beta})^2/[0.0006(\arcsin\sqrt{P_2}-\arcsin\sqrt{P_1})^2].$$

Table 6. Estimating an odds ratio with specified relative precision

$$n = z^2_{1-\alpha/2}\{1/[P_1^*(1-P_1^*)]+1/[P_2^*(1-P_2^*)]\}/[\log_e(1-\varepsilon)]^2$$

Table 8. Estimating a relative risk with specified relative precision

$$n = z^2_{1-\alpha/2}[(1-P_1)/P_1+(1-P_2)/P_2]/[\log_e(1-\varepsilon)]^2$$

Table 3. Hypothesis tests for a population proportion

For a one-sided test

$$n = \{z_{1-\alpha}\sqrt{[P_0(1-P_0)]}+z_{1-\beta}\sqrt{[P_a(1-P_a)]}\}^2/(P_0-P_a)^2.$$

For a two-sided test

$$n = \{z_{1-\alpha/2}\sqrt{[P_0(1-P_0)]}+z_{1-\beta}\sqrt{[P_a(1-P_a)]}\}^2/(P_0-P_a)^2.$$

Table 12. Estimating an incidence rate with specified relative precision

$$n = (z_{1-\alpha/2}/\varepsilon)^2$$

SOURCE: Lwanga SK, LEMESHOW S. Sample size determination in health studies. WHO Library. 1991

How is it actually done?

Making use of existing formulas

1. Sample size tables

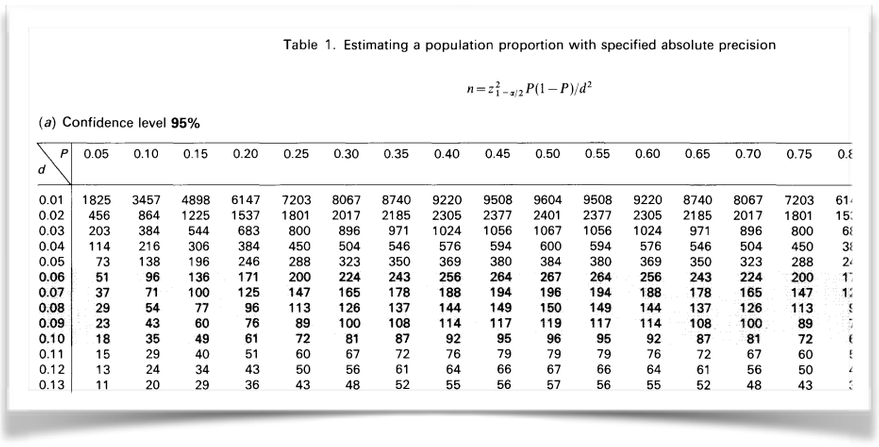


Table 1. Estimating a population proportion with specified absolute precision

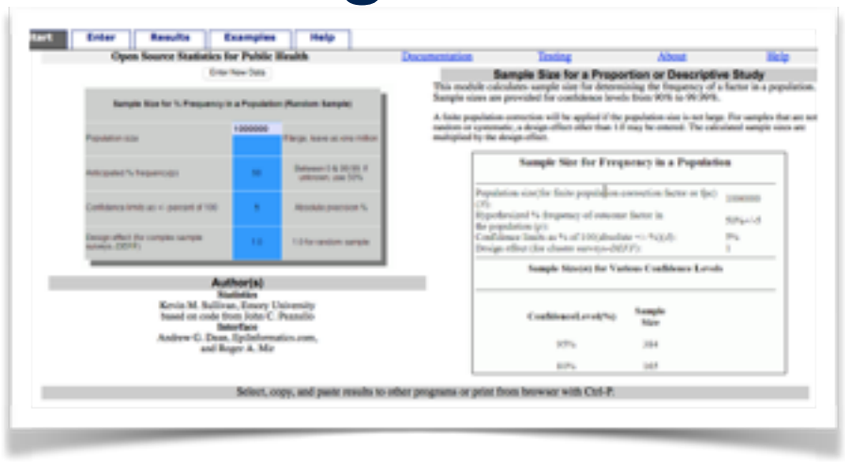
$$n = z^2_{1-\alpha/2} P(1-P)/d^2$$

(*a*) Confidence level **95%**

d \ *P*	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.8
0.01	1825	3457	4898	6147	7203	8067	8740	9220	9508	9604	9508	9220	8740	8067	7203	61
0.02	456	864	1225	1537	1801	2017	2185	2305	2377	2401	2377	2305	2185	2017	1801	15
0.03	203	384	544	683	800	896	971	1024	1056	1067	1056	1024	971	896	800	6
0.04	114	216	306	384	450	504	546	576	594	600	594	576	546	504	450	3
0.05	73	138	196	246	288	323	350	369	380	384	380	369	350	323	288	2
0.06	**51**	**96**	**136**	**171**	**200**	**224**	**243**	**256**	**264**	**267**	**264**	**256**	**243**	**224**	**200**	1
0.07	**37**	**71**	100	**125**	**147**	**165**	**178**	**188**	**194**	**196**	**194**	**188**	**178**	**165**	**147**	1
0.08	**29**	**54**	**77**	**96**	**113**	**126**	**137**	**144**	**149**	**150**	**149**	**144**	**137**	**126**	**113**	9
0.09	**23**	**43**	**60**	**76**	**89**	**100**	**108**	**114**	**117**	**119**	**117**	**114**	**108**	**100**	**89**	7
0.10	**18**	**35**	**49**	**61**	**72**	**81**	**87**	**92**	**95**	**96**	**95**	**92**	**87**	**81**	**72**	6
0.11	15	29	40	51	60	67	72	76	79	79	79	76	72	67	60	5
0.12	13	24	34	43	50	56	61	64	66	67	66	64	61	56	50	4
0.13	11	20	29	36	43	48	52	55	56	57	56	55	52	48	43	3

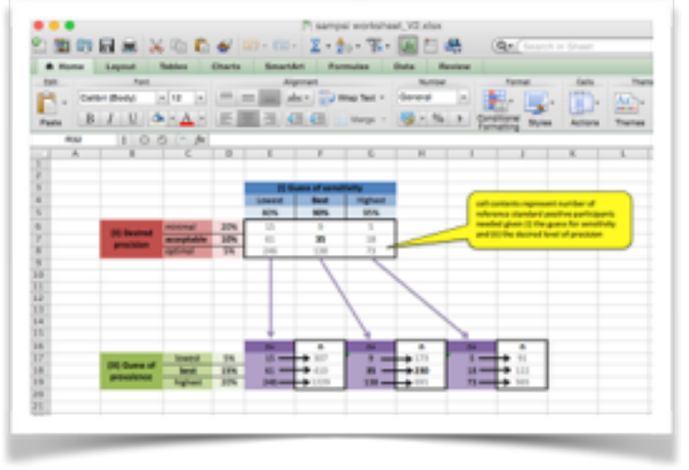
+ easy to use
+ quick impression
- limited parameter values
- limited formulas
- limited complexity

2. Packages / websites



+ easy to use
+ quick
- many not free
- slow if varying multiple parameters
- often power-based

3. Own program



+ flexible
+ free
+ dynamic
+ graphics
- effort in writing

Sample size for accuracy

Epidemiology 1

Sample size calculations in randomised trials: mandatory and mystical

Lancet 2005; 365: 1348–53

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA
(K F Schulz PhD, D A Grimes MD)

Correspondence
Dr Kenneth
KSch

Kenneth F Schulz, David A Grimes

Investigators sho
describ

unassa
calculate sampl
details in their published rep
ogists describe the approaches in books and artic
Protocol committees and ethics review boards require adherence. CONSORT reporting guidelines clearly specify the reporting of sample size calculations.[1,2] Almost everyone agrees.

An important impetus to this unanimity burst on the medical world more than a quarter of a century ago. group of researchers, led by Tom Chalmers, publish

citati ally, that troubled him.[4] He regarded it damaging paper that he had ever d hy? We will describe his concerns later, so

ponents of sample size calculations

lculating sample sizes for trials with dichotomous outcomes (eg, sick *vs* well) requires four components: type I error (α), power, event rate in the control group, and a treatment effect of interest (or analogously an event rate in the treatment group). These basic components persist through calculations with other types of outcomes, except other assumptions can be necessary. For example, with quantitative outcomes and a typical statistical test, investigators might assume a difference between means and a variance for the means.

In clinical research, hypothesis testing risks two fundamental errors (panel 1). First, researchers can conclude that two treatments differ when, in fact, they do not. This type I error (α) measures the probability of making this false-positive conclusion. Conventionally, α is most frequently set at 0·05, meaning that investigators desire a less than 5% chance of making a false-positive conclusion. Second, researchers can conclude that two treatments do not differ when, in fact, they do—ie, a false-negative conclusion. This type II error (β) measures the probability of this false-negative conclusion. Conventionally, investigators set β at 0·20, meaning that they desire less than a 20% chance of making a false-negative conclusion.

Power derives from β error. Mathematically, it is the complement of β (1–β) and represents the probability of avoiding a false-negative conclusion. For example, for

sample size calculations = “predicting the future”

Predicting the future

Necessary Quantities for sample size calculation

1. pick target outcome (usually sensitivity)
2. pick desired precision for estimate of target quantity (e.g. ±10%)
3. make informed guesses about
   - 3.1 expected prevalence
   - 3.2 expected sensitivity

Example calculation

By Hand

1. target quantity: sensitivity
2. desired precision: ±10%
3. informed guesses about
   - 3.1 expected sensitivity: 90%
   - 3.2 expected prevalence: 15%

$$n(diseased) \geq \frac{(1.96)^2 sensitivity(1 - sensitivity)}{precision^2}$$

$$n(diseased) \geq \frac{(1.96)^2 0.9(1 - 0.9)}{0.1^2}$$

$$n(diseased) \geq \frac{0.35}{0.01}$$

$$n(diseased) \geq \mathbf{34.6}$$

$$n(total) = \frac{n(diseased)}{prevalence}$$

$$n(total) = \frac{35}{0.15} \approx \mathbf{233}$$

Example calculation

Sample Size Table from WHO book

Table 1. Estimating a population proportion with specified absolute precision

$$n = z^2_{1-\alpha/2} P(1-P)/d^2$$

(*a*) Confidence level **95%**

sens/spec

d \ *P*	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.01	1825	3457	4898	6147	7203	8067	8740	9220	9508	9604	9508	9220	8740	8067	7203	6147	4898	3457	1825
0.02	456	864	1225	1537	1801	2017	2185	2305	2377	2401	2377	2305	2185	2017	1801	1537	1225	864	456
0.03	203	384	544	683	800	896	971	1024	1056	1067	1056	1024	971	896	800	683	544	384	203
0.04	114	216	306	384	450	504	546	576	594	600	594	576	546	504	450	384	306	216	114
0.05	73	138	196	246	288	323	350	369	380	384	380	369	350	323	288	246	196	138	73
0.06	51	96	136	171	200	224	243	256	264	267	264	256	243	224	200	171	136	96	51
0.07	37	71	100	125	147	165	178	188	194	196	194	188	178	165	147	125	100	71	37
0.08	29	54	77	96	113	126	137	144	149	150	149	144	137	126	113	96	77	54	29
0.09	23	43	60	76	89	100	108	114	117	119	117	114	108	100	89	76	60	43	23
0.10	18	35	49	61	72	81	87	92	95	96	95	92	87	81	72	61	49	35	18
0.11	15	29	40	51	60	67	72	76	79	79	79	76	72	67	60	51	40	29	15
0.12	13	24	34	43	50	56	61	64	66	67	66	64	61	56	50	43	34	24	13
0.13	11	20	29	36	43	48	52	55	56	57	56	55	52	48	43	36	29	20	11
0.14	9	18	25	31	37	41	45	47	49	49	49	47	45	41	37	31	25	18	9
0.15	8	15	22	27	32	36	39	41	42	43	42	41	39	36	32	27	22	15	8
0.20	5	9	12	15	18	20	22	23	24	24	24	23	22	20	18	15	12	9	5
0.25	*	6	8	10	12	13	14	15	15	15	15	15	14	13	12	10	8	6	*

*Sample size less than 5.

precision

SOURCE: Lwanga SK, LEMESHOW S. Sample size determination in health studies. WHO Library. 1991

Example calculation

Sample Size Table from DEEP

Table 2 | **Relationship between sample size and 95% confidence interval**

Number of infected (non-infected) subjects required*	Estimated test sensitivity (or specificity)‡					
	50%	60%	70%	80%	90%	95%
50	13.9%	13.6%	12.7%	11.1%	8.3%	–
100	9.8%	9.6%	9.0%	7.8%	5.9%	4.3%
150	8.0%	7.8%	7.3%	6.4%	4.8%	3.5%
200	6.9%	6.8%	6.4%	5.5%	4.2%	3.0%
500	4.4%	4.3%	4.0%	3.5%	2.6%	1.9%
1,000	3.1%	3.0%	2.8%	2.5%	1.9%	1.4%

*As defined by the reference standard test. ‡95% confidence interval around the estimated sensitivity (+/– value in table).

sens/spec

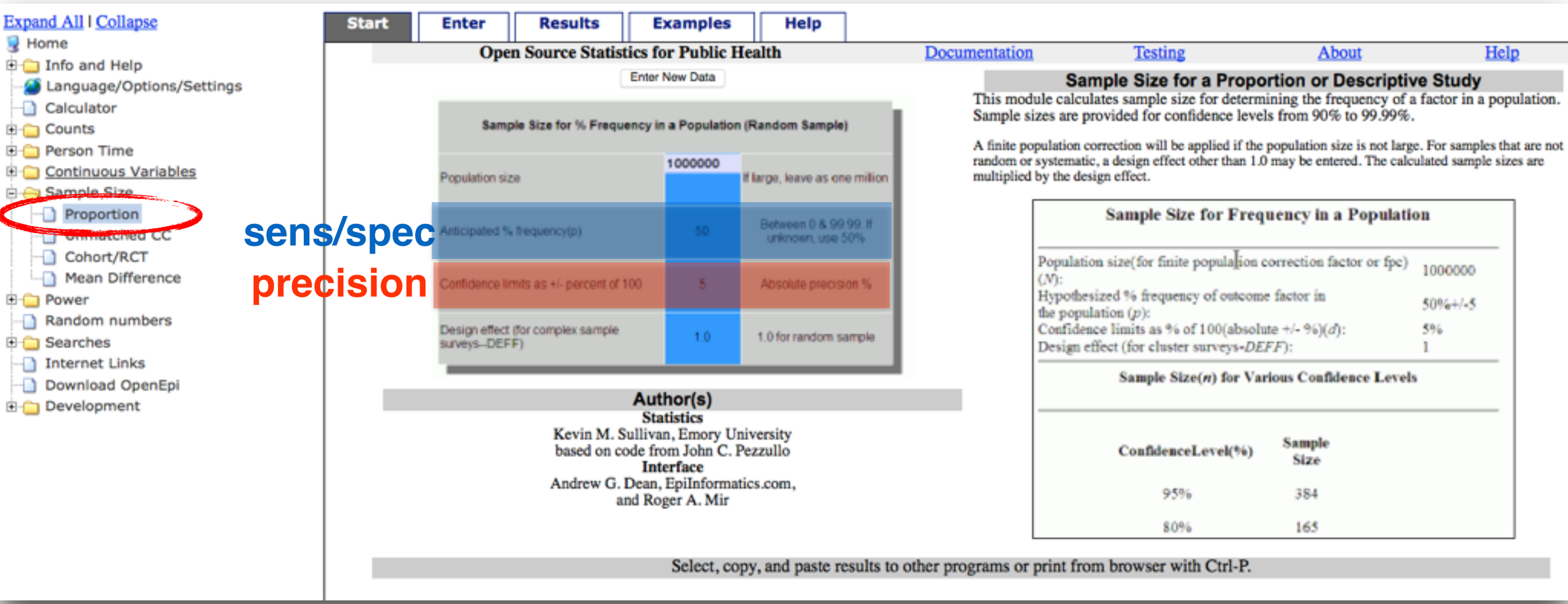
n+

precision

SOURCE: TDR. Evaluation of diagnostic tests for infectious diseases: general principles. Nat Rev Microbiol. 2006 Dec 1;4(12 Suppl):S20–32.

Example calculation

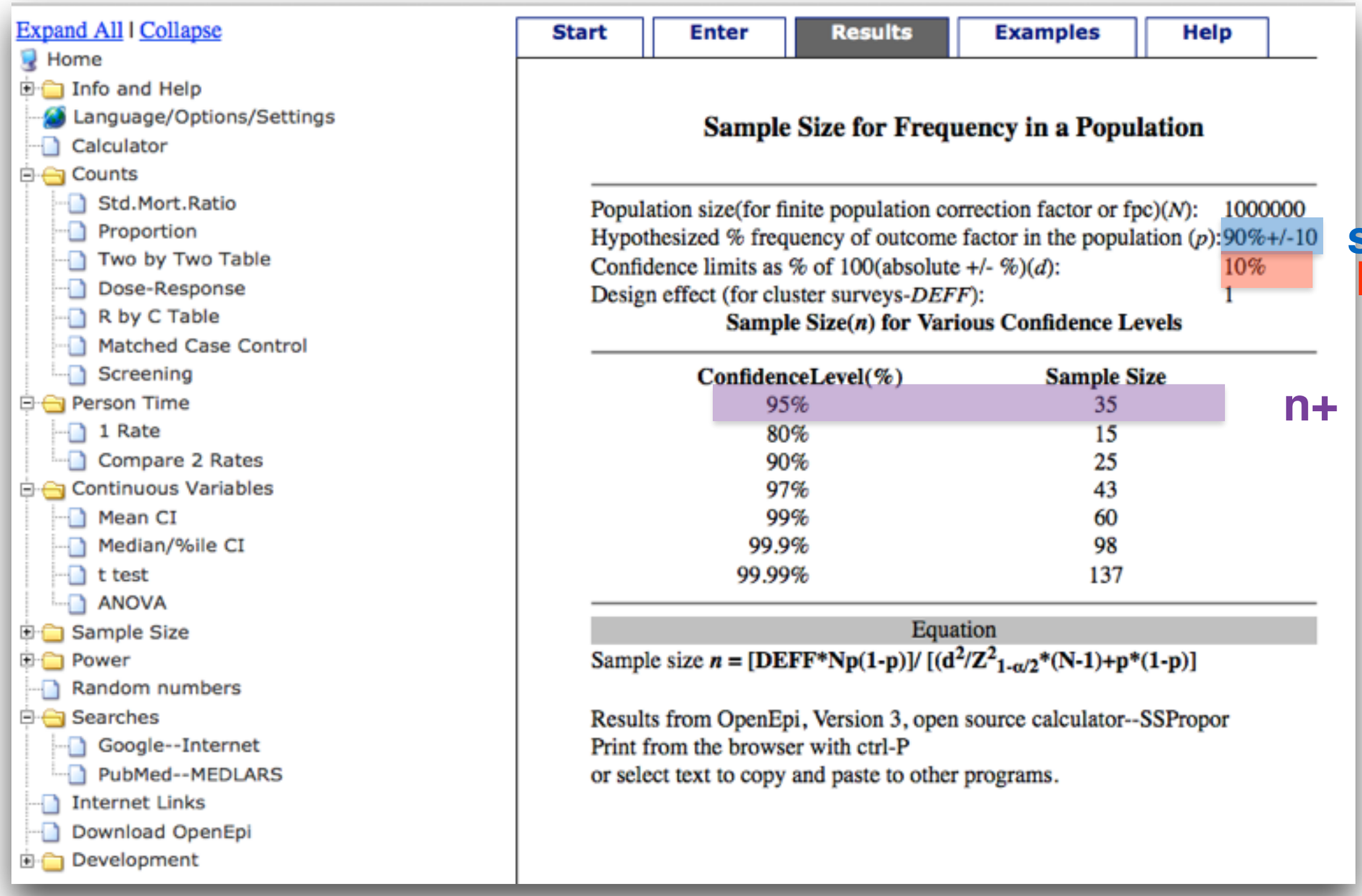
OpenEpi



http://www.**openepi.com**/SampleSize/SSPropor.htm

Example calculation

OpenEpi

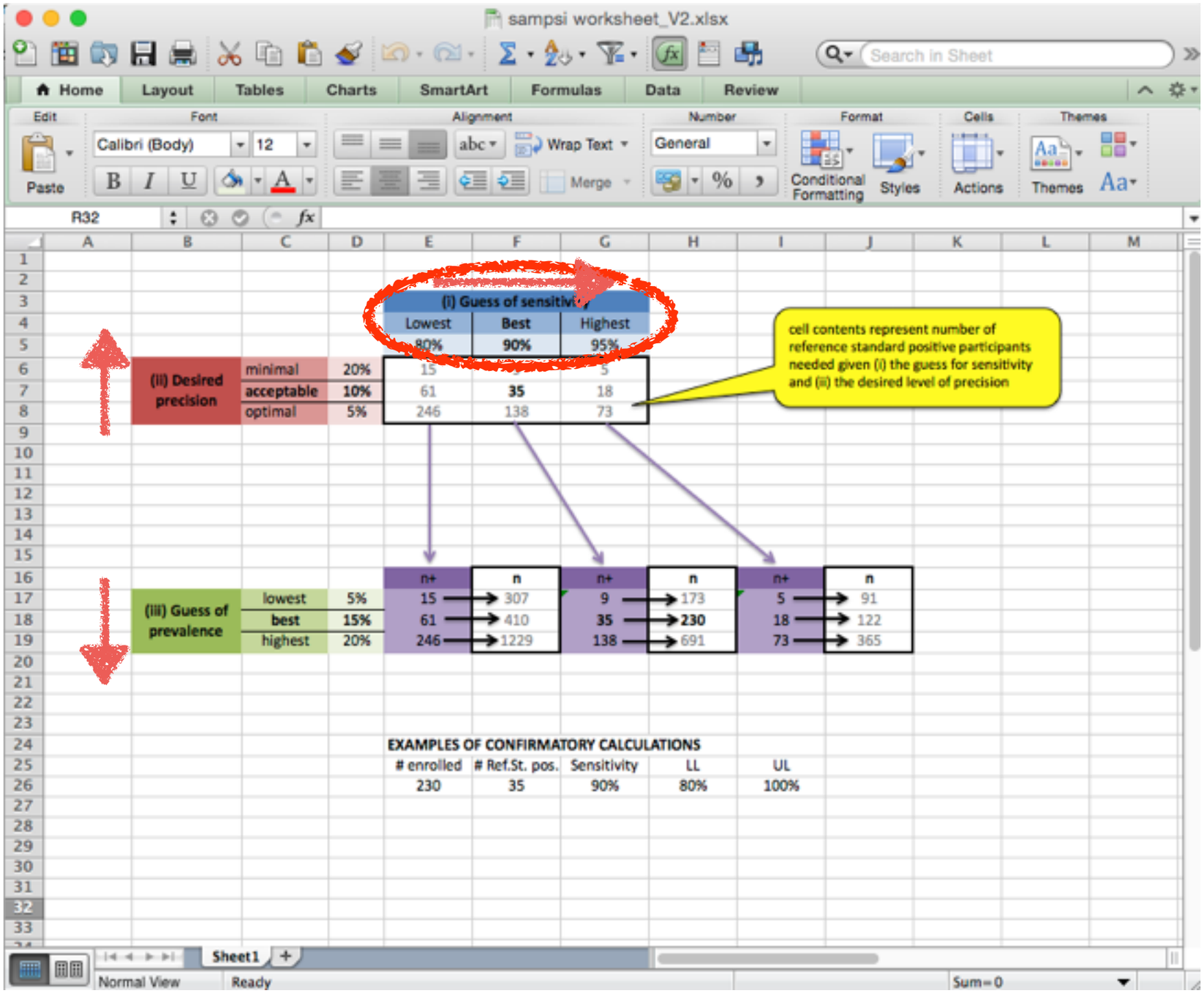


sens/spec
precision

n+

Example calculation

Xcel Sheet



Literature Examples for accuracy

Xpert NEJM

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 SEPTEMBER 9, 2010 VOL. 363 NO. 11

Rapid Molecular Detection of Tuberculosis and Rifampin Resistance

Catharina C. Boehme, M.D., Pamela Nabeta, M.D., Doris Hillemann, Ph.D., Mark P. Nicol, Ph.D., Shubhada Shenai, Ph.D., Fiorella Krapp, M.D., Jenny Allen, B.Tech., Rasim Tahirli, M.D., Robert Blakemore, B.S., Roxana Rustomjee, M.D., Ph.D., Ana Milovic, M.S., Martin Jones, Ph.D., Sean M. O'Brien, Ph.D., David H. Persing, M.D., Ph.D., Sabine Ruesch-Gerdes, M.D., Eduardo Gotuzzo, M.D., Camilla Rodrigues, M.D., David Alland, M.D., and Mark D. Perkins, M.D.

sens/spec precision n+ prevalence n

Group	Sensitivity or specificity target	Confidence interval (CI)	Required minimum group size	Average prevalence among TB suspects	Required TB suspects
s+, c+	95%	± 3%	203	18%	1128
s-, c+	90%	± 4.5%	171	11%	1554
Non-TB	98%	± 1%	335	46%	728
Rif- resistant	90%	± 7%	71	5%	1420
Rif-sensitive	95%	± 3%	203	22%	922

All values were calculated based on the estimation procedure. Based on these calculations, a total of **1550** patients with suspicion of PTB will be enrolled and tested according to the workflow described in Figure 1 and 2. Each site shall therefore aim to enroll **310** patients. Confidence intervals for *per specimen analysis* will be narrower than for *per patient analysis*. Calculations were based on the following assumptions:

1. Performance targets:
The primary objective of this study is to assess whether the product specifications for the Xpert™ MTB have been met. The critical performance targets are as follows:

Sensitivity in smear- and culture-positive patients:
Minimum >90%, optional >95%; since the feasibility study results showed a sensitivity >95%, the optional target was used for sample size calculation.

Sensitivity in smear-negative, culture-positive patients:
Minimum >60%, optional >90%; since we aim for a culture replacement test, the optional target has been used for sample size calculation.

Xpert Lancet

Articles

Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study

Catharina C Boehme, Mark P Nicol, Pamela Nabeta, Joy S Michael, Eduardo Gotuzzo, Rasim Tahirli, Ma Tarcela Gler, Robert Blakemore, William Worodria, Christen Gray, Laurence Huang, Tatiana Caceres, Rafail Mehdiyev, Lawrence Raymond, Andrew Whitelaw, Kalaiselvan Sagadevan, Heather Alexander, Heidi Albert, Frank Cobelens, Helen Cox, David Alland, Mark D Perkins

The sample size for this study is largely a reflection of WHO's requirement to see a new test evaluated in a geographically and otherwise representative number of routine settings. The sample size calculation was largely driven by the need to assess Xpert in a variety of different settings representative of the real-world situation in high burden countries. It will be important to determine how dependant Xpert performance is on population factors (notably HIV prevalence), daily workload, user skills and laboratory infrastructure aspects. We will have to assess whether the rate of DNA contamination goes up over time, to follow the evolution of sensitivity and specificity and the impact of laboratory technician's fatigue, and to monitor the robustness of the assay in the field (error rate, instrument robustness). All these endpoints require longer study duration and a variety of sites.

prevalence

Xpert sensitivity and specificity estimates were the primary endpoints for the minimum sample size calculation during phase 1 (validation phase). Based on discussions with the participating centers, we assumed a minimum average daily enrolment number per site of 5 patients and an average TB prevalence of 15%.

The following minimum performance criteria were set as go/no go decision criteria for moving to the next project phase

a) Xpert™ MTB/Rif sensitivity > 80% of culture positive cases (this should be the lower limit of the confidence interval)
b) Xpert™ MTB/Rif specificity > 90% of culture negative cases (this should be the lower limit of the confidence interval)

n **n+** **sens/spec** **precision**

For each site, enrolling 380 patients would provide 57 confirmed TB cases. A sensitivity of 90% or above would produce a lower confident interval limit of at least 80%. Of the 323 remaining, 5-10% would be of indeterminate diagnosis. However, only 55 TB negative patients are sufficient to have a lower confidence interval limit of at least 90% with given the expected specificity of 98%.

This would leave us with an overall sample size need of 2660 TB suspects for phase 1 and an enrolment duration for this phase of 3 months.

The overall sample size for the study (all phases) was also determined by Rifampicin resistance endpoints. While the expected sensitivity is 95%, the lower limit of the confidence interval should be at least 92% requiring at least 315 Rif-resistant cases among culture positives. With an MDR prevalence rate of 5% across all sites, this would require a total enrolment of 6300 TB suspects.

LAMP

n

In total, 900 patients with suspicion of PTB will be enrolled. Each patient will provide two sputum samples.

In order to prove superiority of the specificity in the modified test compared to the original evaluation results, we derived a minimum number of patients necessary for enrolment. We anticipate that the modified test will have a specificity of approximately 98% compared to the previous 94.7% demonstrated (97% being the minimum acceptable target). Proving superiority with 80% power and 95% confidence will required 504 Non-TB cases in that case. We expect an overall prevalence of TB across the sites to be less than approximately 30% which should comfortably provide the minimum number of patients to establish statistically significant comparison.

prevalence

n+

The roughly 250-300 culture positive anticipated in this study should be able to demonstrate whether the sensitivity for culture positives remains roughly similar. For S-C+ patients, an approximately 100 enrolees will provide a lower confidence limit of >40% if the point estimate is greater than 50%. For S+C+ patients, an approximately 150-200 enrolees will provide a lower confidence limit of >93% if the point estimate is similar to 97%.

precision

sens/spec

LPA

Sample size calculations were based on the number of sputum samples necessary to statistically determine non-inferiority between each new test and the reference test MTBDRplusV1 given the stated non-inferiority margins for RIF. INH will be evaluated as well. It is anticipated that sensitivity and specificity of each of tests under investigation will be at or very near the same point estimate as the reference test. A difference of 3% and 2% were deemed to be not clinically relevant for sensitivity and specificity, respectively. The calculations are for 80% power and 95% confidence.

	RIF	
	Sens	Spec
Anticipated performance	98%	99%
Non-Inferiority Margin	3%	2%
Lower limit for 95% confidence interval	95%	97%
Sample size	270	307

sens/spec

precision

n

prevalence

n+

prevalence

The study will include 900 sputum samples of which approximately 70% or 630 will be C+. Half of the positives will be RIF and/or INH resistant and half will not. This should be sufficient to demonstrate non-inferiority for RIF within the above constraints.

Take Home Points

Accuracy Studies

- Sample size important for planning & ethics
- Sample size calculations ≈ predicting the future
- Estimation-based better than power-based
- Use simple tools available for calculation
- Account for losses, clustering etc.
- Beware of large sample approximations

PART II

Additional considerations

Problems with small numbers…

Large sample approximation

			(i) Guess of sensitivity			
			Lowest	**Best**	Highest	
			80%	**90%**	95%	
(ii) Desired precision	minimal	20%	15	9	5	n+
	acceptable	**10%**	61	**35**	18	
	optimal	5%	246	138	73	

sample size formulas rely on large sample approximations

Problems with small numbers…

Large sample approximation

- Sensitivity: 95%, precision ±10% —> n=18
  - study results: 18*95% = 17.1 ≈ 17
  - study analysis: 17/18 = 94.4%

- multiple formulas to calculate confidence intervals
  - sampsi. calc.: should be from ~85% to 100%
  - Wald (large sample approx.): from 84% to 100%
  - Fisher (exact): from 73% to 100%

sample size calculation should match analytic approach

Starting with the sample size…

…calculating attainable precision

- Sensitivity: 95%, n=100 —> precision ?
  - study results: 100*95% = 95
  - study analysis: 95/100 = 95%

- 95% CI ???
  - Wald: from 91% to 99% (precision ±4%)
  - Fisher: from 89% to 98% (precision ±3-6%)

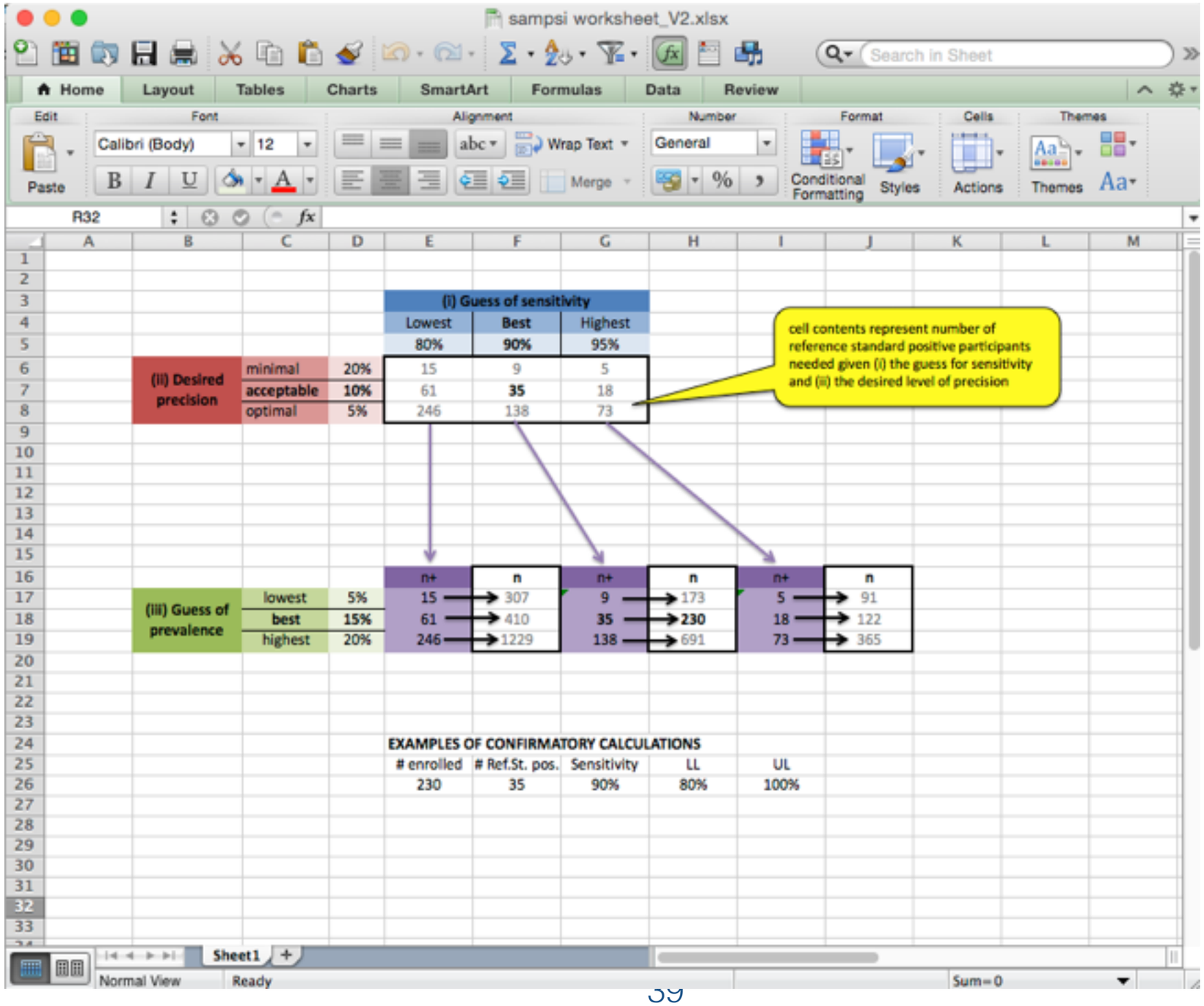
Starting with the sample size…

…calculating attainable precision

- n = 17: 95%CI from 73% to 100%
- n = 35: 95%CI from 81% to 99%
- n = 40: 95%CI from 83% to 99%
- n = 50: 95%CI from 86% to 100%

Prevalence

A modifiable driver of sample size



Modifying prevalence…

…by modifying selection criteria

- You can increase prevalence
  - by selecting sites with high prevalence
  - by selecting participants with high prevalence

- BUT #1: beware of spectrum bias (do not enrol based on other test results)

- BUT #2: beware of spectrum effects, leading to reduced generalizability

- BUT #3: being more restrictive with selection criteria will reduce enrolment rate

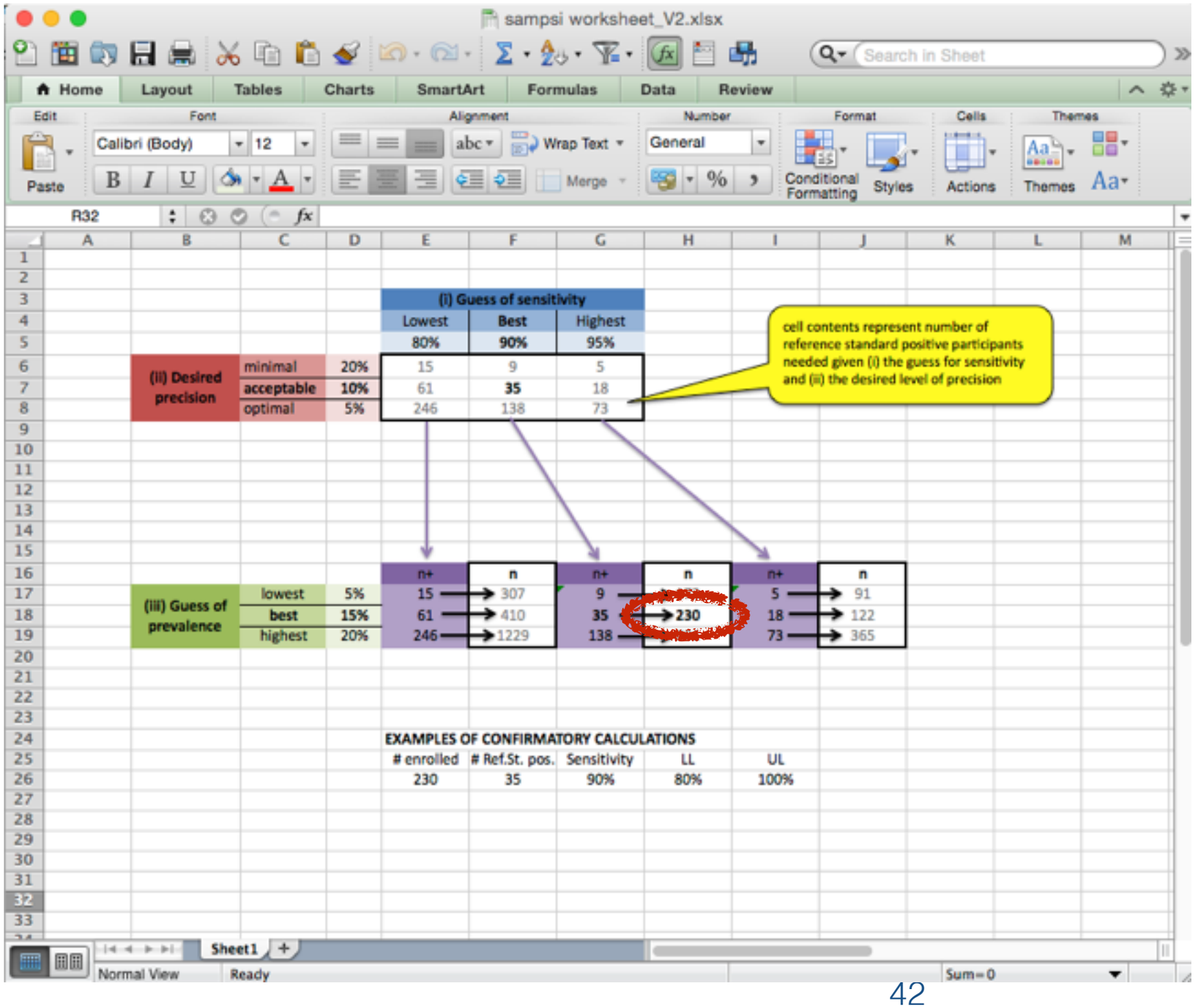
Accounting for losses

- **May occur due to**
  - insufficient # of samples or sputum volume
  - indeterminate results or cultures contaminated
  - missing culture- or speciation result
  - NTM
  - other missing data / incomplete CRF
  - drop-out
  - etc.

- **Account for this** by inflating sample size

Accounting for potential losses

…by inflating sample size



- you calculated you need n=230 TB suspects
- you anticipate 5-10% losses among those recruited

➡ **aim to recruit n=253**

Dealing with multiple outcomes or tests

- **Multiple outcomes**: focus on primary outcome
- **Multiple tests**: focus on main comparison

Given **chosen sample size** you can still calculate **achievable precision** for secondary outcomes or comparisons

- if you MUST have adequate precision for multiple outcomes or tests: pick larger sample size

Comparative Studies

Superiority, equivalence & non-inferiority

decreasing "n" needed

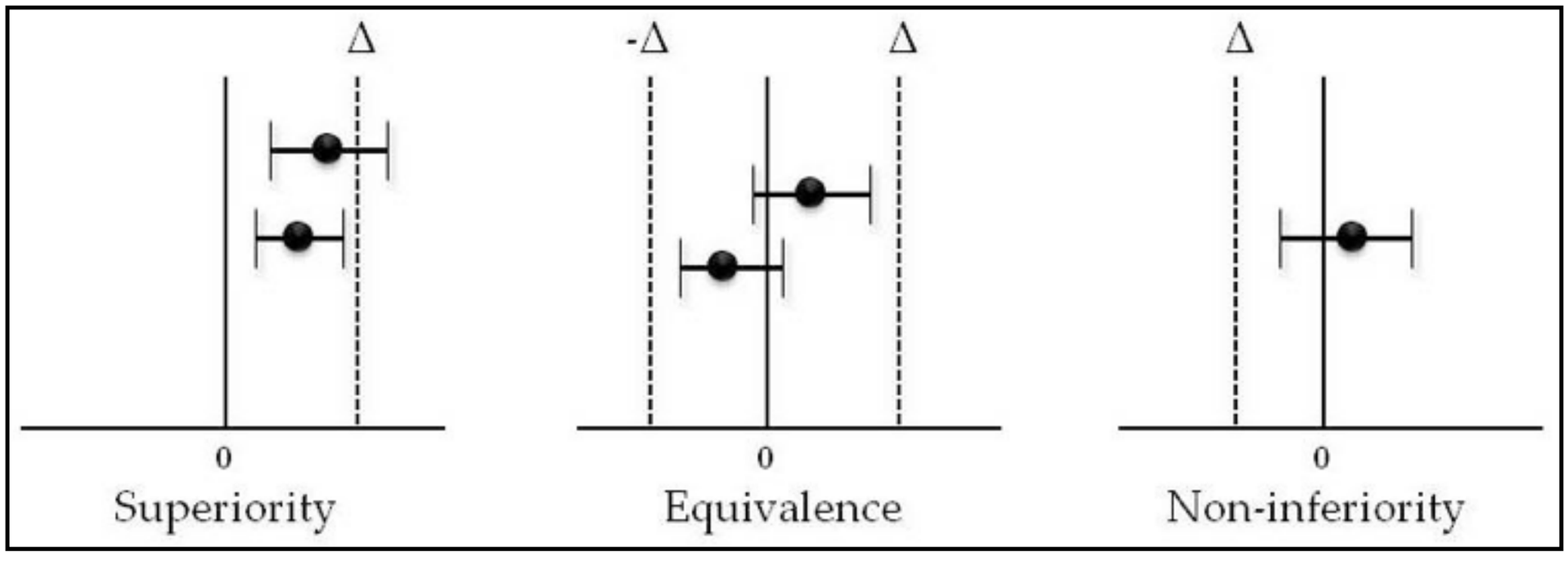
- **Superiority:** index test is <u>better</u> than established test
- **Equivalence:** index test is the same or <u>not unacceptably different</u>* from established test
- **Non-inferiority:** index test is <u>not unacceptably worse</u>* than established test

* Equivalence & non-inferiority studies require specification of a margin of equivalence/non-inferiority

Comparative Studies

Superiority, equivalence & non-inferiority

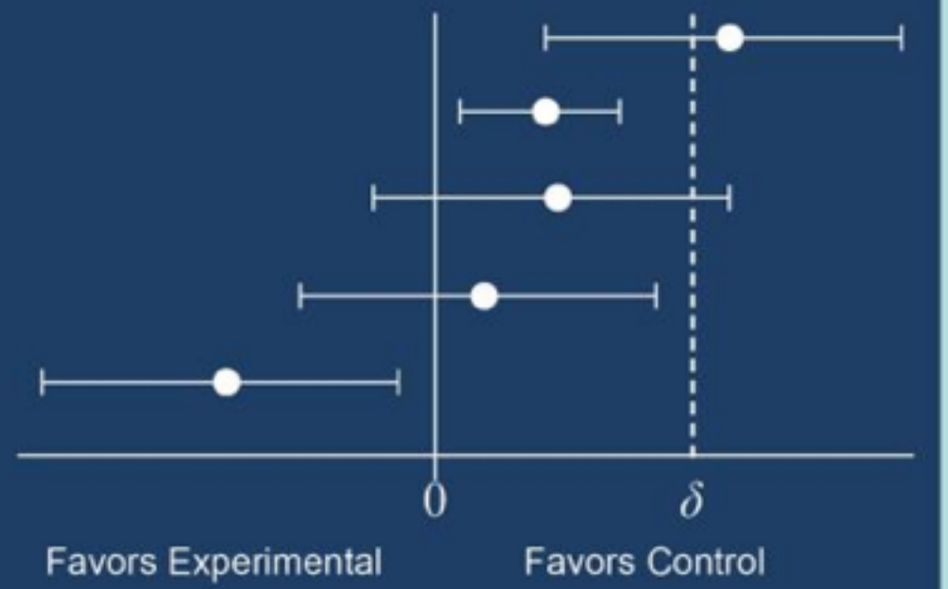
- **Superiority:** index test is better than established test
  - e.g. difference in sensitivity +30% (95%CI +20, +40)
- **Equivalence:** index test is as good as established test, given a pre-specified margin
  - e.g. difference in sensitivity +0% (95%CI -5%, +5) for a pre-specified margin of ±5%
- **Non-inferiority:** index test is either better than established test, or inferior but less than a pre-specified margin
  - e.g. difference in sensitivity -3% (95%CI -5%, -1%) for a pre-specified margin of -5%



Source: Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. Trials. 2011;12(1):106.

Chapman & Hall/CRC Biostatistics Series

Design and Analysis of Non-Inferiority Trials



Mark D. Rothmann
Brian L. Wiens
Ivan S. F. Chan

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Clustered Designs

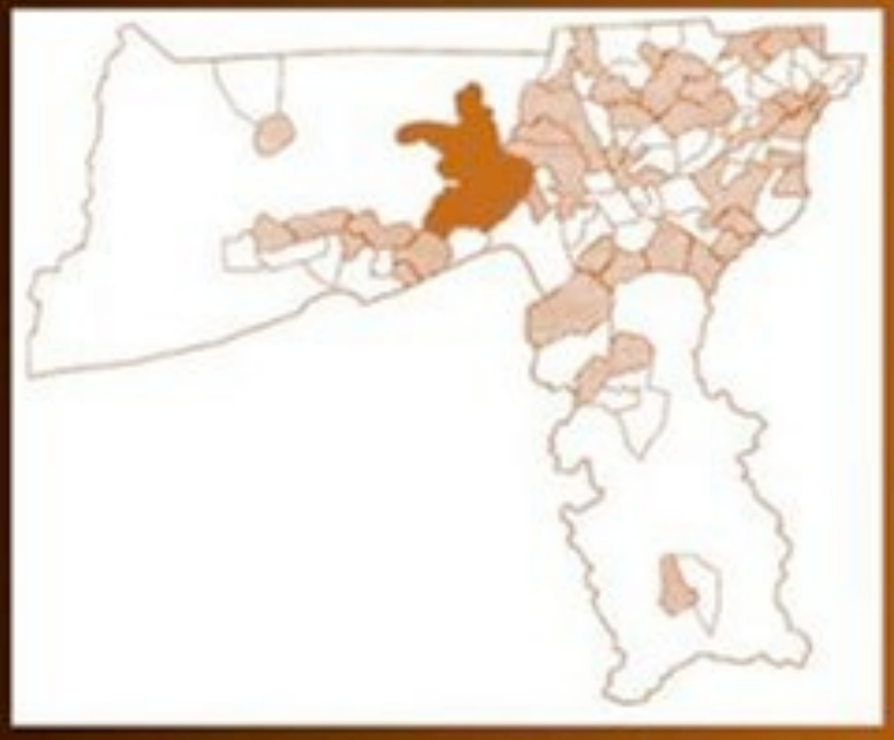
- **Clustered designs** require increases sample size
  - ➡ need to account for between-cluster variation

- **Design Effect** / variance inflation factor (Deff)
  - ratio of actual variance for a given sampling design to variance assuming the same sample size, but using simple random sampling without replacement
  - Deff = 1 + (m-1)*ICC [m: number of observations per cluster)

- **Intraclass correlation coefficient** (ICC)
  - describes how strongly units in the same group resemble each other, compared to resemblance across groups
  - ICC = between-cluster variance / total variance [total variance = between-cluster+within-cluster variance]

Chapman & Hall/CRC
Interdisciplinary Statistics Series

CLUSTER RANDOMISED TRIALS



Richard J. Hayes
Lawrence H. Moulton

CRC Press
A CHAPMAN & HALL BOOK

SANDRA ELDRIDGE AND SALLY KERRY

A Practical Guide to Cluster Randomised Trials in Health Services Research

WILEY
STATISTICS IN PRACTICE

Sample size for impact

Sample size for impact

How is it different from sample size for accuracy?

1. **Questions about impact are comparative**
   - e.g. we are interested in the difference/ratio of two proportions (rather than just one proportion)

   ➡ need different formulas

2. **Questions about impact are not always about proportions**
   - **Proportion** e.g. reduction in % LTFU, % unfavourable treatment outcome, % mortality
   - **Time to event** e.g. reduction in time to Dx/Rx, time to smear status conversion
   - **Continuous** e.g. reduction in morbidity score

   ➡ need different formulas for different variable types

Sample size for impact

How is it different from sample size for accuracy?

1. **Questions about impact are comparative**
   - e.g. we are interested in the difference/ratio of two proportions (rather than just one proportion)

   ➡ need different formulas

2. **Questions about impact are not always about proportions**
   - **Proportion** e.g. reduction in % LTFU, % unfavourable treatment outcome, % mortality
   - **Time to event** e.g. reduction in time to Dx/Rx, time to smear status conversion
   - **Continuous** e.g. reduction in morbidity score

   ➡ need different formulas for different variable types

Predicting the future

Necessary Quantities for sample size calculation

1. Expected baseline proportion

2. Expected proportion in intervention group (or or risk ratio or risk difference)

3. Desired precision around risk ratio / risk difference

Example calculation

Hand calculation

1. Expected baseline proportion: 10%
2. Expected proportion in intervention group: 5%
3. Desired precision around risk ratio: 20% (i.e. we expect RR=0.5 with 95%CI 0.4, 0.6)

$$n = (1.96)^2 \frac{\frac{1-P_1}{P_1} + \frac{1-P_2}{P_2}}{ln(1-precision)^2}$$

$$n = (1.96)^2 \frac{\frac{1-0.1}{0.1} + \frac{1-0.05}{0.05}}{ln(1-0.2)^2}$$

$$n \approx 3.84 \frac{28}{0.05}$$

$$\underline{n \approx 2{,}160}$$

Example calculation

Sample Size Table

Table 8 (*continued*)

(*b*) Confidence level **95%**, relative precision **20%**

precision

RR

P_2 \ RR	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00	4.25	4.50	4.75	5.00
0.01	15276	13733	12705	11970	11419	10990	10647	10367	10133	9935	9766	9619	9490	9377	9276	9186	9104
0.02	7561	6790	6275	5908	5633	5418	5247	5107	4990	4891	4806	4732	4668	4611	4561	4516	4475
0.03	4990	4475	4132	3887	3704	3561	3447	3353	3275	3209	3153	3104	3061	3023	2989	2959	2932
0.04	3704	3318	3061	2877	2739	2632	2546	2476	2418	2368	2326	2289	2257	2229	2204	2181	2161
0.05	2932	2624	2418	2271	2161	2075	2006	1950	1904	1864	1830	1801	1775	1752	1732	1714	1698
0.10	1389	1235	1132	1059	[illegible]	961	926	898	875	855	838	823	811	799	789	780	772
0.15	875	772	703	654	618	589	566	548	532	519	507	498	489	482	475	469	463
0.20	618	541	489	452	425	403	386	372	361	351	342	335	328	323	318	313	309
0.25	463	402	361	331	309	292	278	267	258	250	243	237	232				
0.30	361	309	275	250	232	218	206	197	189	182							
0.35	287	243	214	193	177	165	155	147									
0.40	232	193	168	149	136	125	116										
0.45	189	155	132	116	103												
0.50	155	124	103	89	78												
0.55	127	99	80	67													
0.60	103	78	61														
0.70	67	45															
0.80	39	20															
0.90	18																

Since $RR = P_1/P_2$, $RR \leqslant 1/P_2$.

For $RR < 1$, use the column value corresponding to $1/RR$ and the row value corresponding to P_1.

P

53

Table 8

SOURCE: Lwanga SK, LEMESHOW S. Sample size determination in health studies. WHO Library. 1991

Literature Examples for impact

Xpert XTEND

Xpert MTB/RIF vs microscopy as the first line TB test in South Africa: mortality, yield, initial loss to follow up and proportion treated. The XTEND study

GJ Churchyard

On behalf of the XTEND team

(Xpert for TB - Evaluating a New Diagnostic)

Prof Churchyard has no financial relationships with commercial entities to disclose

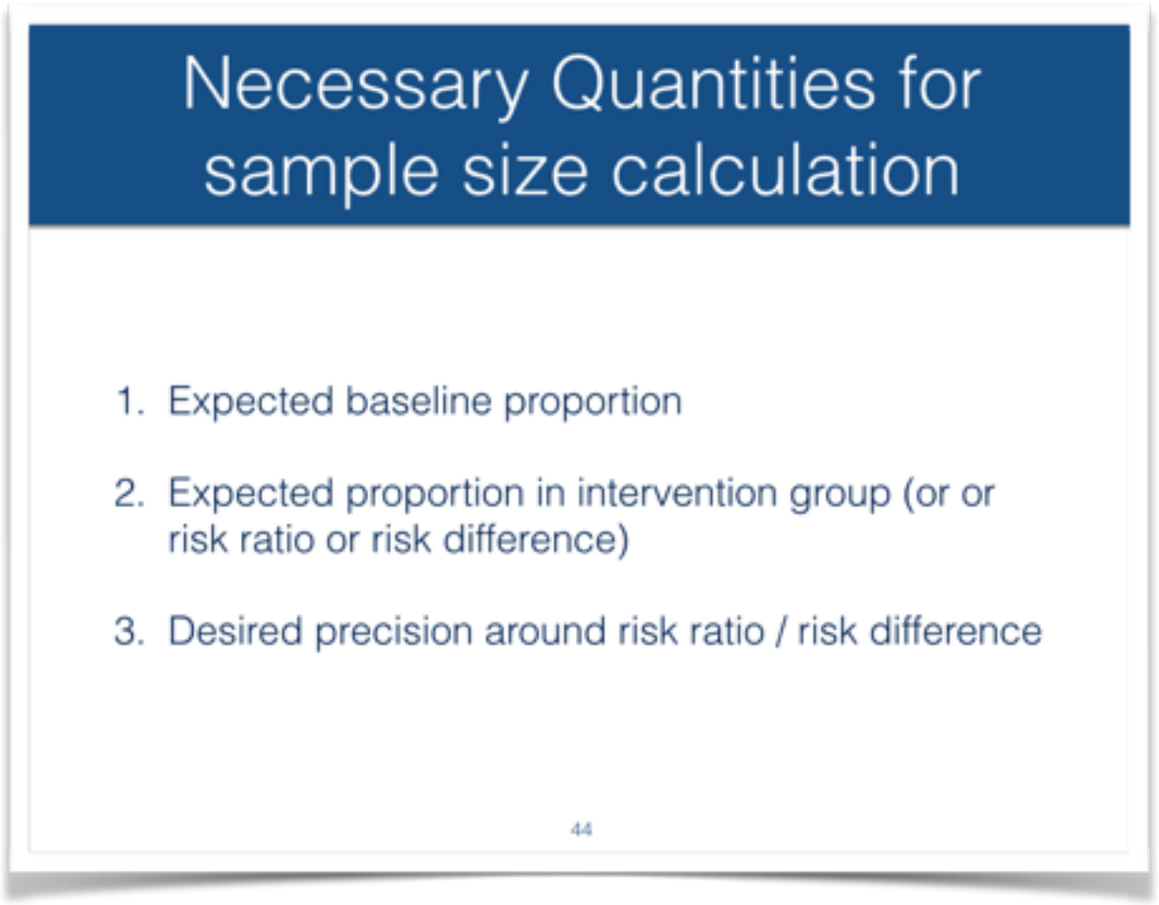
6.1.1. Primary outcome: early mortality among TB suspects

The sample size estimates are based on 6 month mortality among TB suspects. Data from Western Cape and Kwa-Zulu Natal in South Africa suggest a mortality of 2-11% at 2 months (Western Cape, M Nicol personal communication; Kwa-Zulu Natal, amongst TB suspects). In the Kwa-Zulu Natal study, of those with known HIV status, the HIV prevalence was 84%.

Assuming a six month mortality among TB suspects of 5% in the "standard of care arm", 10 clusters per arm, 220 TB suspects per cluster and a coefficient of variation of 0.25, there would be approximately 90% power to detect a 50% reduction of mortality in the Xpert MTB/RIF arm. If the coefficient of variation is 0.3 we would have 87% power to detect a 50% reduction (Table 61).

Table 61. The power for various scenarios where the total number of clusters per arm is fixed at 10.

% mortality at 6 months in SoC arm)	% reduction	# TB suspects per cluster	Power (k=0.25)	Power (k=0.30)
5%	50%	300	.95	.91
	50%	220	.91	.87
	50%	200	.90	.86
	50%	100	.70	.66
7.5%	50%	300	.98	.96
	50%	220	.97	.93
	50%	200	.96	.92
	50%	100	.83	.78
7.5%	40%	300	.87	.79
	40%	220	.83	.74
	40%	200	.80	.72
	40%	100	.61	.55
10%	50%	300	.99	.97
	50%	220	.98	.96
	50%	200	.98	.95
	50%	100	.90	.86
10%	40%	300	.91	.83
	40%	220	.87	.79
	40%	200	.86	.78
	40%	100	.70	.63



Xpert TBNEAT

Articles

Feasibility, accuracy, and clinical effect of point-of-care Xpert MTB/RIF testing for tuberculosis in primary-care settings in Africa: a multicentre, randomised, controlled trial

*Grant Theron, Lynn Zijenah, Duncan Chanda, Petra Clowes, Andrea Rachow, Maia Lesosky, Wilbert Bara, Stanley Mungofa, Madhukar Pai, Michael Hoelscher, David Dowdy, Alex Pym, Peter Mwaba, Peter Mason, Jonny Peter, Keertan Dheda, for the TB-NEAT team**

4. *A priori* sample size calculations for morbidity (TBscore and Karnofsky performance score)

TBscore: We projected the difference in TBscore between arms to be one (the minimally important clinical difference). We assumed, based on previous studies,[4-6] that the within group standard deviation would be two points in each arm. With an alpha value of 5% (two-sided) and a desired power of 80%, and assuming equal numbers in each arm, we required approximately 63 culture-positive patients in each arm. To account for deaths, loss to follow-up, withdrawals, and missing data, we inflated this by 30% (~82 culture-positive). We conservatively estimated the overall study TB prevalence to be 15%, meaning we aimed to recruit ~550 patients in each arm.

Karnofsky performance score: We based our calculations on data from a study in 147 patients with suspected TB from South Africa[7], where culture-positive patients with a baseline (IQR) KPS of 70 (60-80) improved to a score of 90 (80-90) after two months of anti-TB treatment. We calculated that, in order to detect a difference in KPS of 10 (i.e. the minimum change of one grade on the scale shown in Table S2) and assuming a standard deviation of 20, with power of 80% and an alpha of 0.05, we would need 63 culture-positive patients in each arm. Inflating this by 30% to account for early deaths, loss to follow-up, withdrawals, and missing data, meant we would need to recruit 82 culture-positive patients in each arm or, at an assumed conservative prevalence of 15%, 546 patients in each arm.

5. *Post hoc* sensitivity calculations for morbidity (TBscore and Karnofsky performance score)

All sites

At two months we had 87 and 108 culture-positive patients in the smear microscopy and MTB/RIF arms who were initiated on treatment and seen within two weeks of their desired follow-up date, with mean (SD) TBscores of 2.184 (2.060) and 2.065 (1.866), respectively, and mean KPSs of 85.29 (14.77) and 85.13 (11.64), respectively. At six months, there were 80 and 97 culture-positive patients in each arm who were initiated on treatment and seen within two weeks, with mean TBscores of 1.488 (1.661) and 1.691 (1.692), respectively, and mean KPSs of 97.28 (8.064) and 95.05 (13.080), respectively. Post-hoc sample size calculations show 80% power ($\alpha = 0.05$) to detect an effect size of 0.4151 and 0.4354 or a difference between the means of 0.8150 and 0.7301 for TBscore and 5.481 and 4.6030 for KPS at two and six months, respectively – less than the minimally important clinical difference of one sign or symptom at each time point or one grade on the KPS scale.

Xpert weekly cRCT

OPEN ACCESS Freely available online

PLOS | MEDICINE

Impact of Xpert MTB/RIF for TB Diagnosis in a Primary Care Clinic with High TB and HIV Prevalence in South Africa: A Pragmatic Randomised Trial

Helen S. Cox[1,2]*, Slindile Mbhele[3], Neisha Mohess[3], Andrew Whitelaw[3,4], Odelia Muller[2], Widaad Zemanay[3], Francesca Little[5], Virginia Azevedo[6], John Simpson[4], Catharina C. Boehme[7], Mark P Nicol[1]

1 Division of Medical Microbiology and Institute for Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town, South Africa, **2** Médecins Sans Frontières, Khayelitsha, South Africa, **3** Division of Medical Microbiology, University of Cape Town, Cape Town, South Africa, **4** National Health Laboratory Service, Johannesburg, South Africa, **5** Department of Statistical Science, University of Cape Town, Cape Town, South Africa, **6** Khayelitsha Health, City of Cape Town, Cape Town, South Africa, **7** Foundation for Innovative New Diagnostics, Geneva, Switzerland

Sample Size and Statistical Analysis

The sample size was calculated based on an expected higher yield of bacteriologically confirmed TB cases of 20% to 30% among individuals with presumptive TB in the Xpert arm compared to the routine arm (80% power and one-sided significance $p<0.05$). We assumed 40 patients would be seen per week and assumed a weak intra-cluster correlation coefficient of 0.05, resulting in a design effect of 3 and a required sample size of 882 per study arm.

Take Home Points

Impact Studies

- Need different formulas because
  - they ask comparative questions
  - they may look at outcomes other than proportions

- Estimation-based better than power-based
  - can be difficult for some outcomes (requiring distributional assumptions or non-parametric approaches)

- Use simple tools available for calculation

- Account for losses, clustering etc.