Karolinska Institutet

4.1 Incomplete/two-stage data

Example: ectopic pregnancy study

Case-control study of the association between ectopic pregnancy and sexually transmitted diseases

Total sample size = 979 (264 cases,715 controls)

Variables collected from beginning of study:

gonnorhoea, contraceptive use, sex partners.

From one year after study started, serum samples collected for **chlamydia antibody** test in *all* cases and in a 50% subsample of controls (*not a simple random sample!*)

As a result, *only 327 out of the 979 patients had measurements for chlamydia antibody*

Sherman et al. pregnancy. Sex Transm Dis. 1990;17(3):115-21

Validity of complete case analysis

- Valid if data are Missing Completely At Random (MCAR)
- Valid if missingness depends *only* on covariates (remember that regression model is model of Y *conditional on X*)
- Valid if missingness depends *only* on Y (e.g. case-control studies!)
- Invalid if missingness depends on Y *and* X, as then the relationship β between Y and X may be biased

Even when analysis is valid, loss of precision

Terminology

Can think of:

Having gathered outcome and *some* covariates (gonnorhoea, contraceptive use, sex partners) on *all* subjects at “**first stage**”

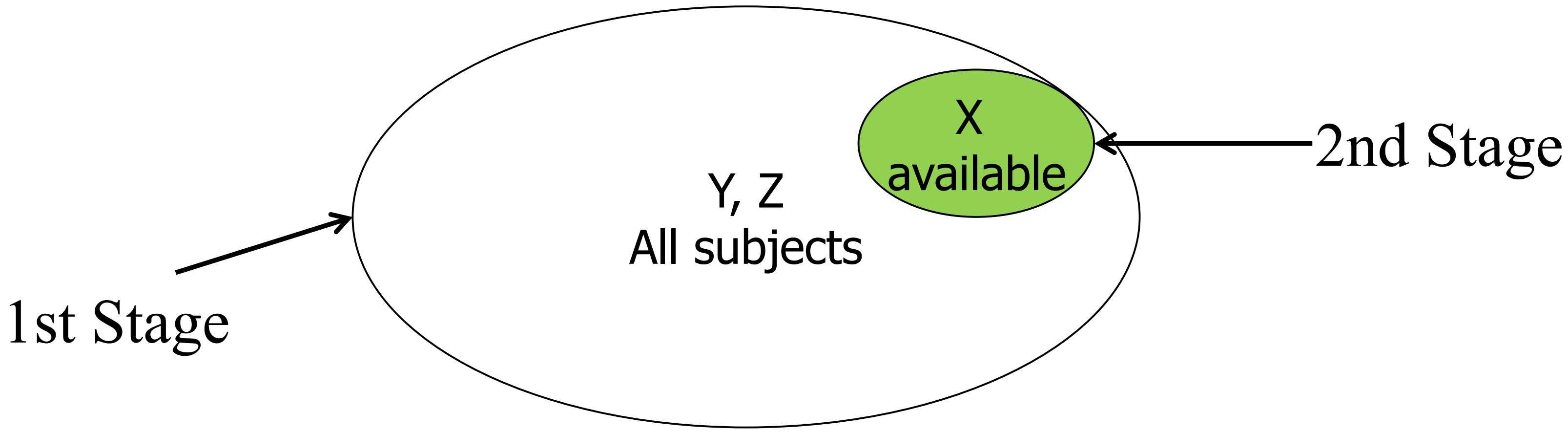
Collecting *special* covariate data (chlamydia antibody) on subsample of “**second stage**” subjects

true where *missing by design!*

Can we use *all* data in analysis?

- If we have:
- outcome variable and *some* (categorical) covariates on *all* subjects at "**first stage**"

special covariate data on subsample of "**second stage**" subjects



If second stage subjects *randomly selected within strata* defined by first stage data (outcome and covariates), we can do "weighted analysis", using sampling weights in different strata

Weighted logistic regression

Weighted likelihood of the complete data, where weights are the reciprocal of the "validation fraction"

The idea (for binary confounder Z)

First Stage

	Y=1	Y=0
Z=1	N_{11}	N_{01}
Z=0	N_{10}	N_{00}

Second Stage (X available)

Y=1	Y=0
n_{11}	n_{01}
n_{10}	n_{00}

individuals with X available are "upweighted" to represent the total number of individuals in that stratum, i.e. weight $=\frac{N_{ZY}}{n_{ZY}}$

”Statistical” explanation:

Y= outcome

X= exposure, available only for the second-stage sample

Z= other covariate(s), available for all (at the first stage)

If we had X,Y for all individuals:

Logistic regression finds β to maximise $\prod P_{\beta}(Y=1 \mid X)$

or equivalently the sum $\sum \ln\{\text{odds}(Y=1 \mid X)\}$

- The weighted regression uses the available individuals with X to “estimate/fill in” the contribution of those without X **who are in the same stratum**

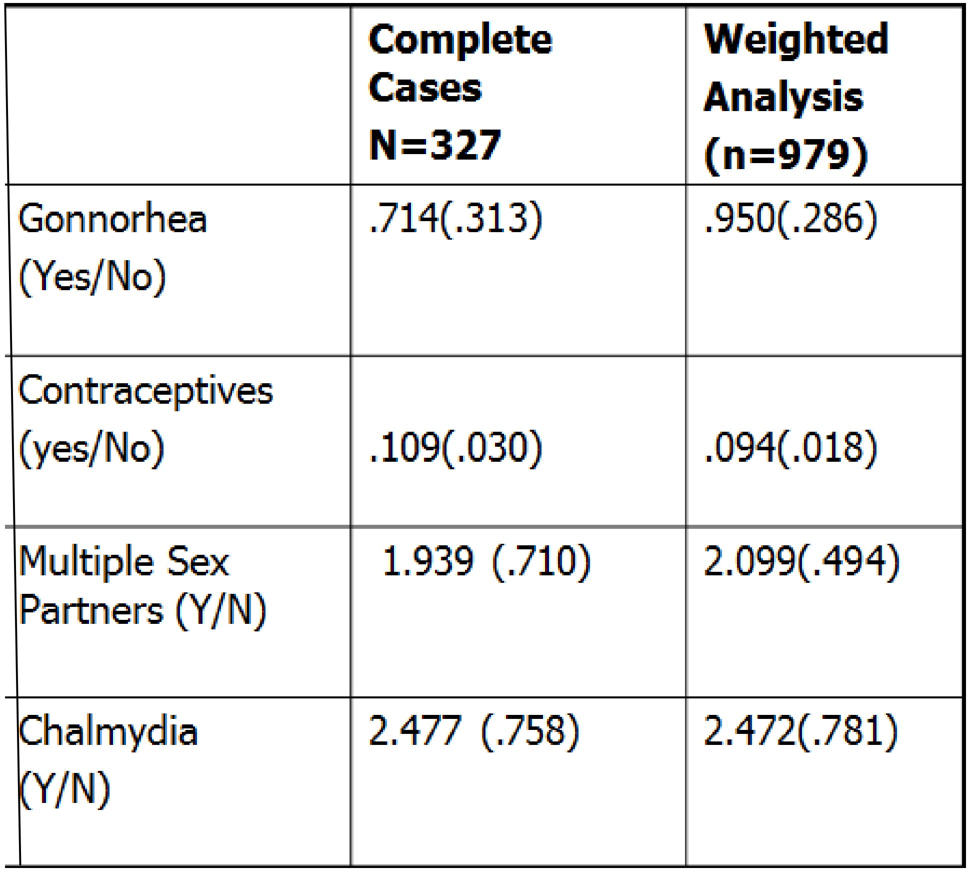
Weighted logistic regression

Simple to run once we have the weights

Statistical packages allow a weighting option in their regression models

We need an adjustment to the variance of the estimate as we do not have N ”real” observations (this more conservative variance is called a **robust variance**)

Estimates and SE from analysis of ectopic pregnancy data (naïve vs. weighted analysis)



	Complete Cases N=327	Weighted Analysis (n=979)
Gonnorhea (Yes/No)	.714(.313)	.950(.286)
Contraceptives (yes/No)	.109(.030)	.094(.018)
Multiple Sex Partners (Y/N)	1.939 (.710)	2.099(.494)
Chalmydia (Y/N)	2.477 (.758)	2.472(.781)

Complete Case analysis valid only if data is simple random sample

Not true in this study!

Small change in estimates in weighted analysis suggests only small bias.

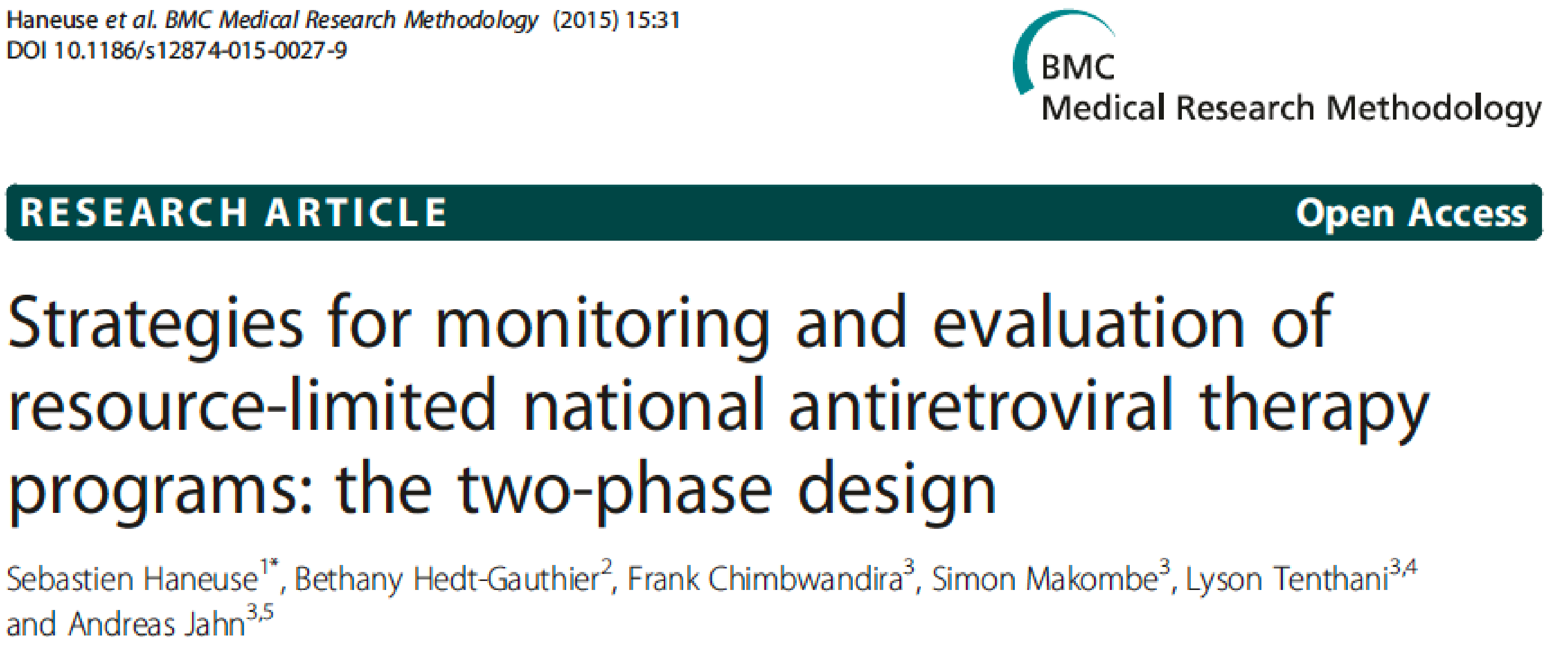
BUT *note downward bias in effect of gonnorhea (oversampling of gonn+ controls in study design!)*

Also note:Improved precision

Consequences for design

We can deliberately subsample (randomly) within strata defined by first-stage variables!

Example of two-stage/two-phase design



Haneuse *et al. BMC Medical Research Methodology* (2015) 15:31
DOI 10.1186/s12874-015-0027-9

BMC
Medical Research Methodology

RESEARCH ARTICLE **Open Access**

Strategies for monitoring and evaluation of resource-limited national antiretroviral therapy programs: the two-phase design

Sebastien Haneuse[1*], Bethany Hedt-Gauthier[2], Frank Chimbwandira[3], Simon Makombe[3], Lyson Tenthani[3,4] and Andreas Jahn[3,5]

Example of two-stage/two-phase design (ctd)

Setting: Cross-sectional survey by the Malawian Ministry of Health of 82,887 patients registered at 189 ART clinics in 2005-2007

HIV positive = 16,141 HIV negative = 66,746

Objectives:

a) to identify risk factors for patient outcome

b) test for interaction between clinic and year

Example of two-stage/two-phase design (ctd)

HIV positive = 16,141 HIV negative = 66,746

Could do a simple case- control study

(random sample of positives and negatives)

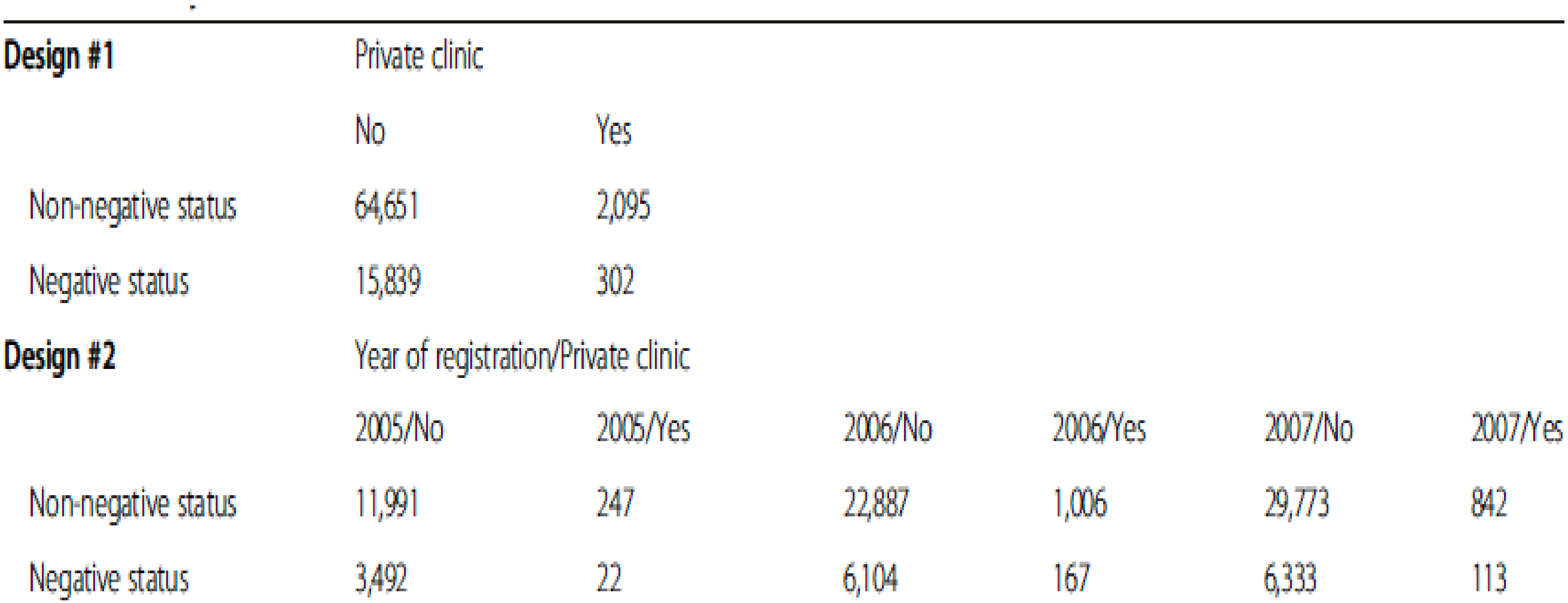
But this does not make use of many details collected and recorded quarterly for the clinic cohorts

Authors explore two stratified (two-stage) designs

(i) stratified by public/private clinic

(ii) public/private and year of registration

Example of two-stage/two-phase design (ctd)

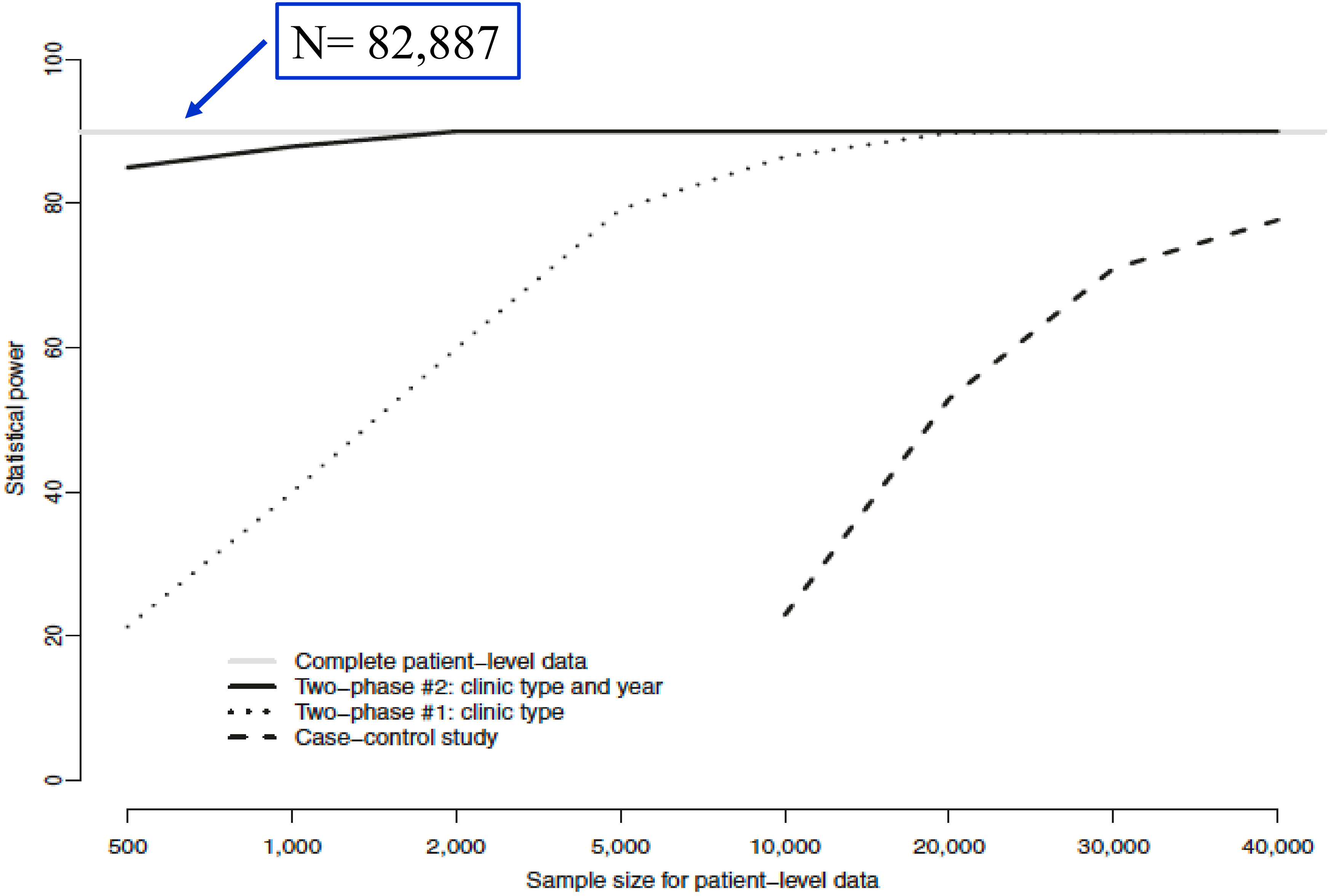


Design #1	Private clinic					
	No	Yes				
Non-negative status	64,651	2,095				
Negative status	15,839	302				
Design #2	Year of registration/Private clinic					
	2005/No	2005/Yes	2006/No	2006/Yes	2007/No	2007/Yes
Non-negative status	11,991	247	22,887	1,006	29,773	842
Negative status	3,492	22	6,104	167	6,333	113

“Balanced” sample of 5000:

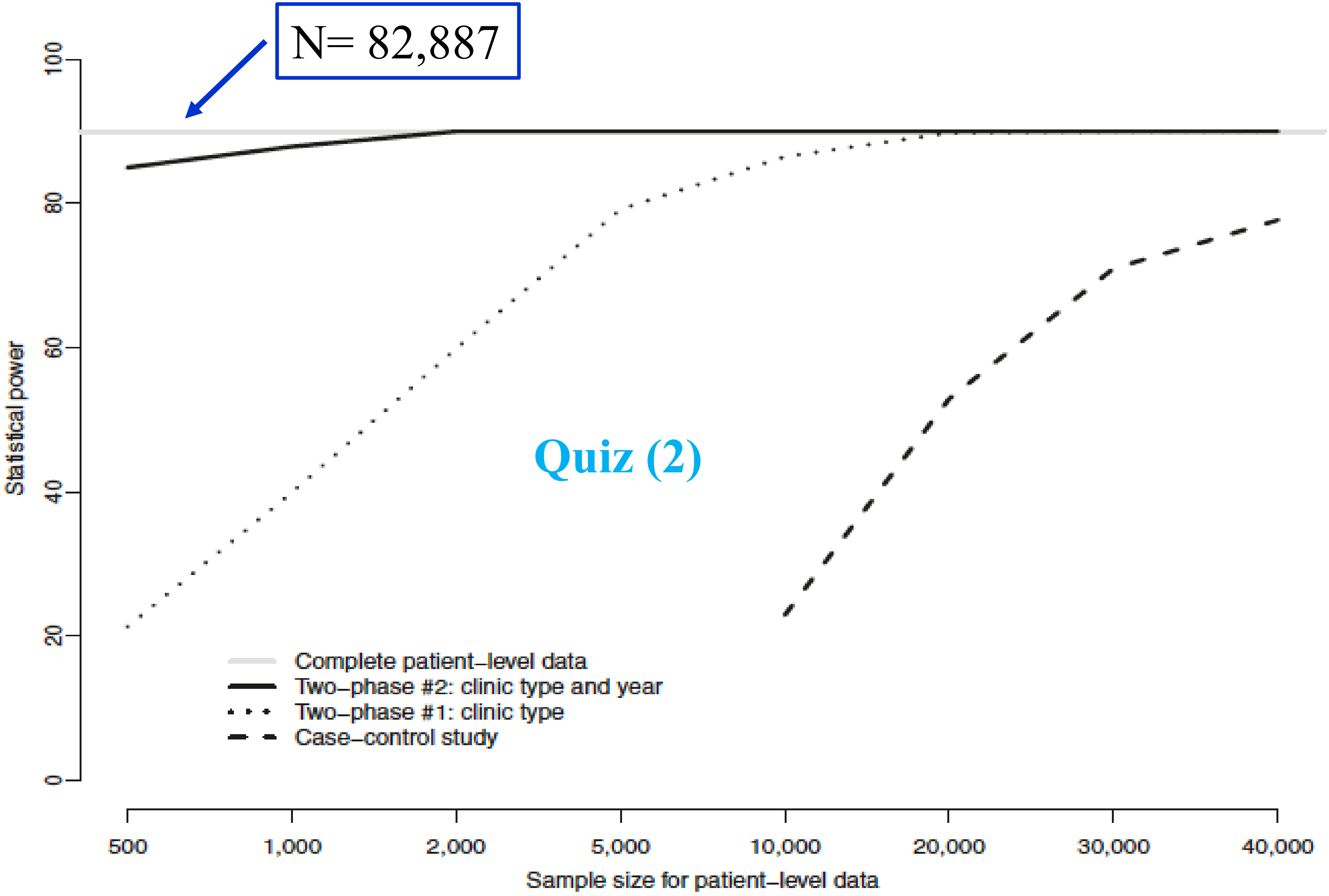
1,250 patients from each of the 4 strata in Design #1
416 patients from each of the 12 strata in Design #2.

Example of gain in power to detect interaction between clinic and year of registration



N= 82,887

Example of gain in power to detect interaction between clinic and year of registration



N= 82,887

Quiz (2)

In practice…...

When “exposures” are expensive/difficult to measure e.g. diet, biochemical or genetic markers

- Measure on just a sub-sample of study subjects
- Sub-selection often ad-hoc

(as in **ectopic pregnancy study)**

How many? simple “sums” of time and cost

Which? Intuition regarding most “informative”

But sample may be recognised/handled as 2-stage

Example: Transmission of H.pylori*

- cross-sectional serological survey of 679 school children aged 10–12 years in 11 Stockholm schools
- Risk factors already identified: family from country with high *H. pylori* prevalence, socioeconomic factors
- New question: risk to children from infected family members?

To avoid testing all family members of all children, investigators tested families of ***all*** children from the four schools with highest *H. pylori* prevalence, only infected children from the other seven schools

* Kivi, Johansson, et al. *Stat Med*. 2005 Dec 30;24(24):4045-54.

Standard Analysis

...of those with complete data: if valid?

…of certain restricted subsets

(original paper analysed the 4 schools that were fully sampled)

Two-stage analysis*

Second-stage children assumed randomly sampled in strata defined by SES, immigrant background

(This is what investigators were targeting as most "informative" families by choosing schools with high prevalence)

Weighted logistic regression of all schools with:

SES and immigrant background as first stage variables

(i.e. known for every child)

family members infection status as second stage

(only known for some children)

**Kivi, Johansson et al 2005*

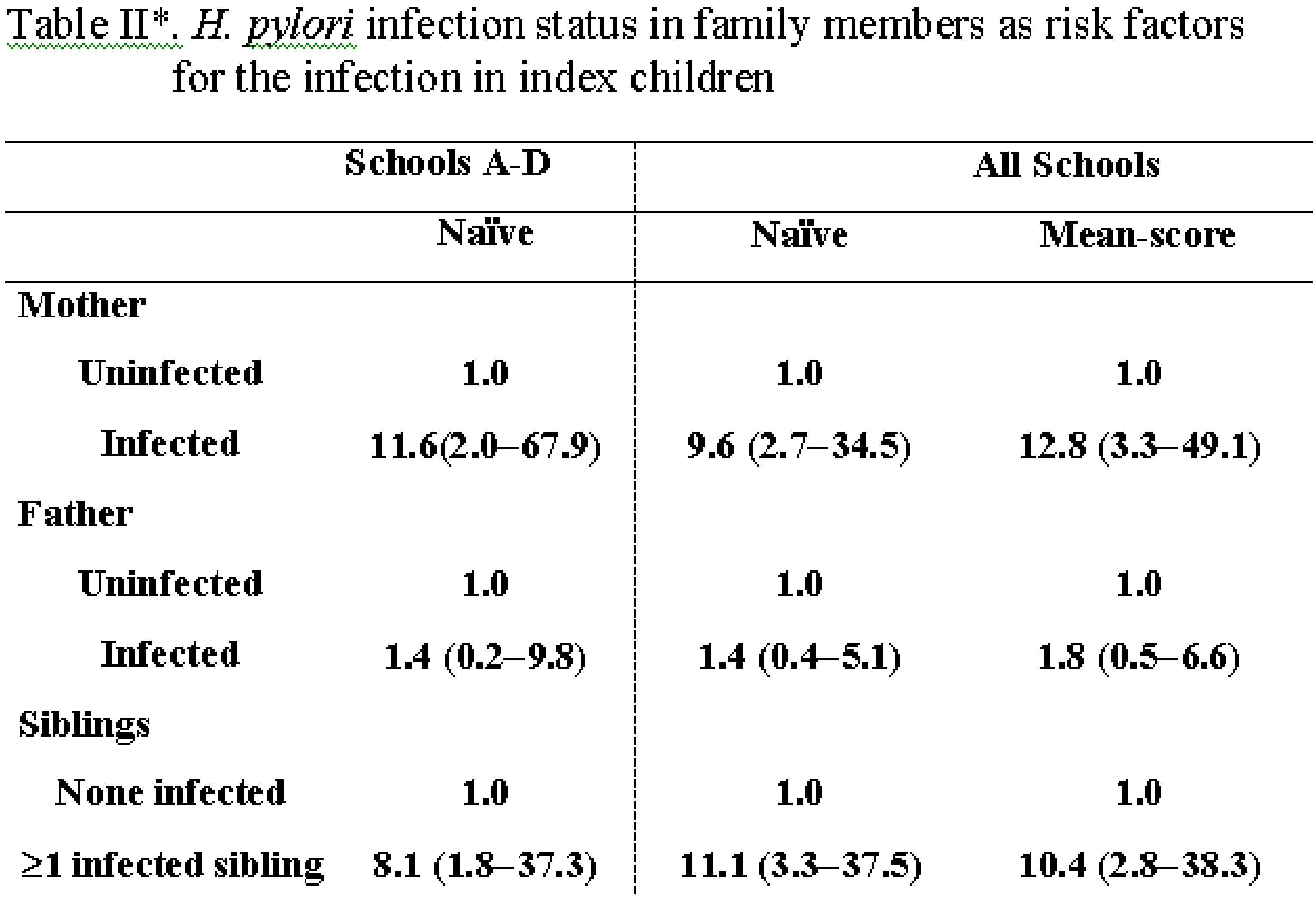


Table II*. *H. pylori* infection status in family members as risk factors for the infection in index children

	Schools A-D	All Schools	
	Naïve	Naïve	Mean-score
Mother			
Uninfected	1.0	1.0	1.0
Infected	11.6(2.0–67.9)	9.6 (2.7–34.5)	12.8 (3.3–49.1)
Father			
Uninfected	1.0	1.0	1.0
Infected	1.4 (0.2–9.8)	1.4 (0.4–5.1)	1.8 (0.5–6.6)
Siblings			
None infected	1.0	1.0	1.0
≥1 infected sibling	8.1 (1.8–37.3)	11.1 (3.3–37.5)	10.4 (2.8–38.3)

**Kivi, Johansson et al 2005*

Weighted analysis

Simple commands in **Stata** and **R**

Special command in Stata called “meanscor”

user specifies: logistic model, first stage variables defining strata sampled, second stage variable(s)

(the command gets the weights)

Survey package in R (by Thomas Lumley) on CRAN

wide range of designs, includes Cox model for two-phase sampling (more later today)

Alternatively, compute weights and run weighted model

Exercise 4.1: Analysis of 2-stage data.