

Two-stage studies and other special sampling designs

Jean-François Boivin

February 3, 2003

h:\lkowalski\JF\course_epid\two_stage_sampling.ppt

Background

1980s-1990s: Progress in use of administrative drug databases

- Spitzer WO, et al. The use of β_2 -agonists and the risk of death and near death from asthma. NEJM 1992; 326:501-506.
- Ernst P, et al. Risk of fatal and near-fatal asthma in relation to inhaled corticosteroid use. JAMA 1992; 268:3462-3464.
- Suissa S, et al. A cohort analysis of excess mortality in asthma and the use of inhaled β_2 -agonists. American Journal of Respiratory and Critical Care Medicine 1994; 149:604-610.

- Garbe E, et al. Inhaled and nasal glucocorticoids and the risks of ocular hypertension or open-angle glaucoma. *JAMA* 1997; 277:722-727.
- Hemmelgarn B, et al. Benzodiazepine use and the risk of motor vehicle crash in the elderly. *JAMA* 1997; 278:27-31.
- Garbe E, et al. Risk of ocular hypertension or open-angle glaucoma in elderly patients on oral glucocorticoids. *Lancet* 1997; 350:979-982.

- Blais L, et al. Inhaled corticosteroids and the prevention of readmission to hospital for asthma. American Journal of Respiratory and Critical Care Medicine 1998; 158:126-132.
- Blais L, et al. First treatment with inhaled corticosteroids and the prevention of admissions to hospital for asthma. Thorax 1998; 53:1025-1029.
- Collet JP, et al. Colorectal cancer prevention by non-steroidal anti-inflammatory drugs: effects of dosage and timing. British Journal of Cancer 1999; 81:62-68.

- Sharpe CR, et al. Nested case-cohort study of the effects of non-steroidal anti-inflammatory drugs on breast cancer risk and stage. *British Journal of Cancer* 2000; 83:112-120.
- Csizmadi I, et al. Use of postmenopausal estrogen replacement therapy from 1981 to 1997. *Canadian Medical Association Journal* 2002; 166:187-188.
- Sharpe CR, et al. The effects of tricyclic antidepressants on breast cancer risk. *British Journal of Cancer* 2002; 86:92-97.
- Moride Y, et al. Suboptimal duration of antidepressant treatments in the older ambulatory population of Quebec: Association with selected physician characteristics. *Journal of the American Geriatrics Society* 2002; 50:1365-1371.

Advantages

- Large
- Population-based
- Valid prescription data
- Long-time periods

Disadvantages

- Missing data on certain outcomes
Example: Glaucoma
- Temporal sequence not always clear
Example: Glucocorticoids → cataracts
OR
Cataract surgery → glucocorticoids
- Lack of data on confounders
 - reason for prescription
 - risk factors for outcome

= Today's topic

Exposure to NSAIDs and risk of breast cancer

Epidemiologic data

- Poor exposure data
 - Dose
 - Duration
 - Self-reports
- Small numbers
- Short follow-up
- Inadequate control of confounding

Exposure to NSAIDs and risk of breast cancer

- Cases: Saskatchewan cancer registry
n=6000 since 1981
- Controls: Random selection among Saskatchewan drug plan
matched on age and calendar time
n=24,000
- Drug exposure: Up to 15 years of computerized information
for most drugs

Missing:

- OTC drugs (aspirin, ibuprofen)
- Other confounding factors:
 - Age at first menarche
 - Age at menopause
 - Number of pregnancies
 - Family history of breast cancer
 - Obesity

Entire population (= truth)

Obese women

		cancer	no cancer	
NSAIDs	+	2 000	10 000	OR=0.5
	-	40	100	
		2 040	10 100	

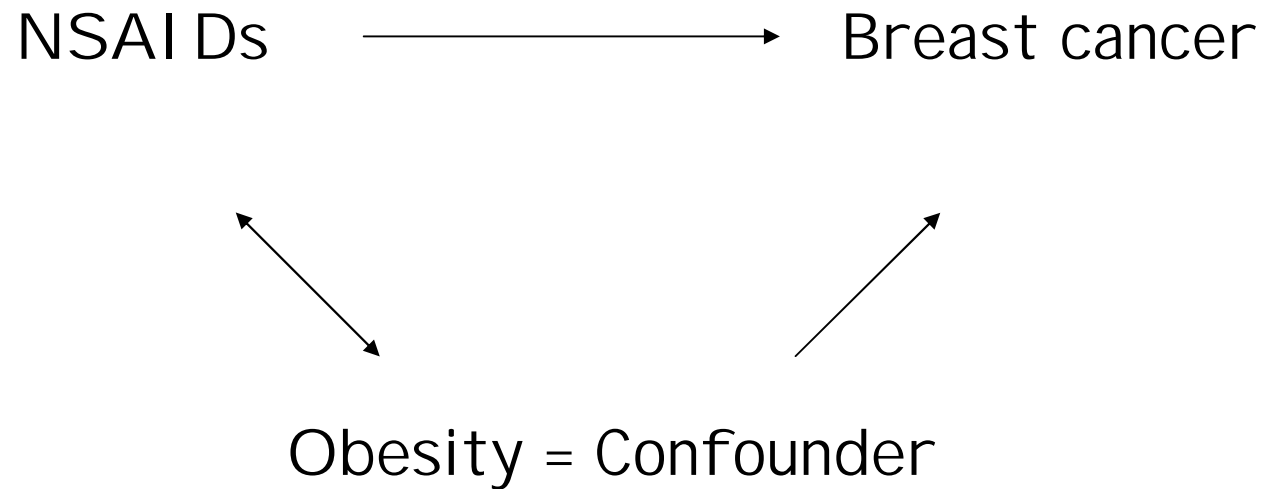
Not obese

		cancer	no cancer	
NSAIDs	+	200	10 000	OR=0.5
	-	400	10 000	
		600	20 000	

All women

		cancer	no cancer	
NSAIDs	+	2 200	20 000	OR=2.5
	-	440	10 100	
		2 640	30 100	32 740

Confounder: definition



Saskatchewan databases

Obese women

		cancer	no cancer
NSAIDs	+		
	-		

Not obese

		cancer	no cancer
NSAIDs	+		
	-		

data not
available in
computerized
databases

All women

		cancer	no cancer
NSAIDs	+	2 200	20 000
	-	440	10 100
		2 640	30 100

available in
computerized
databases

What to do about missing confounder data?

Design options

Option #1

Do not conduct research on that topic

Choose another research question

Option #2

Cohort or case-control study without data on confounder

Obese women

		cancer	no cancer
NSAIDs	+	?	?
	-	?	?

Not obese

		cancer	no cancer
NSAIDs	+	?	?
	-	?	?

All women

		cancer	no cancer
NSAIDs	+	2 200	20 000
	-	440	10 100

32 740

Advantages

- Cheaper
- May be scientifically reasonable for certain questions

Example: Tricyclic antidepressants and breast cancer risk (Sharpe CR, et al. British Journal of Cancer 2002)

Disadvantages

- Possibility of confounding

Sharpe et al: table

Option #3

Collect covariate data on a sample of the entire source population

- two-stage samples
- three-stage samples
- partial questionnaire
- case series only

Two-stage sample

Sampling approaches:

- simple random
- case-control
- cohort
- balanced

Stage-two random sample

Obese women

		cancer	no cancer
NSAIDs	+	60	300
	-	1	3

Not obese

		cancer	no cancer
NSAIDs	+	6	300
	-	12	300

All women

		cancer	no cancer	
NSAIDs	+	66	600	
	-	13	303	
				32 740 (I)
				x 3% (II)
				=982

Stage-two random sample

Provides valid estimates at
reduced cost

BUT: Can be very inefficient

Stage-two case-control sample

Obese women

		cancer	no cancer
NSAIDs	+	400	167
	-	8	2

Not obese

		cancer	no cancer
NSAIDs	+	40	167
	-	80	167

All women

		cancer	no cancer	
NSAIDs	+	440	334	
	-	88	169	
		2 640 (I) x 20%	30 100 (I) x 1.6%	
		<u> </u> =528 (II)	<u> </u> =503 (II)	1 031

Stage-two case-control sample

Provides valid estimates at
reduced cost

More efficient than random sample
when disease is rare

*Note: Same general reasoning applicable to
a stage-two cohort sample*

Stage-two balanced design

Obese women

		cancer	no cancer
NSAIDs	+	227	125
	-	23	2

Not obese

		cancer	no cancer
NSAIDs	+	23	125
	-	227	248

All women

		cancer	no cancer
NSAIDs	+	250	250
	-	250	250

32 740 (I)

Stage-two balanced design

Obese women

		cancer	no cancer
NSAIDs	+		
	-		

Not obese

		cancer	no cancer
NSAIDs	+		
	-		

All women

		cancer	no cancer
NSAIDs	+	250/2 200	250/20 000
	-	250/440	250/10 000

White JE. A two-stage design for the study of the relationship between a rare exposure and a rare disease. *AJE* 1982; 115:119-128

Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. *AJE* 1988; 128:1198-1206.



Two-Stage Sampling for Etiologic Studies Sample Size and Power

Douglas Schaubel,^{1,2} James Hanley,¹ Jean-Paul Collet,^{1,3} Jean-François Bolvin,^{1,3} Colin Sharpe,^{1,3}
Howard I. Morrison,² and Yang Mao²

Preexisting computerized databases are potentially valuable sources of epidemiologic data. Since such databases are infrequently created specifically for etiologic research, data may be available for the exposure of interest and, through record linkage, for the endpoint of interest, but lacking for potential confounders. Because of the size of these databases, two-stage sampling is an efficient alternative to surveying the entire study population for confounder data. At stage 1, information on exposure and disease status is obtained for the entire study population. Confounder data are collected for probability-selected subsamples at stage 2. Logistic regression is performed on the stage 2 sample, with the parameter estimates and variances appropriately corrected to account for the stage 1 data. In this paper, the authors present methods for determining the required stage 2 sample size in the case of categorical exposure and confounding variables. Sample size tables, power curves, and a computer program have been produced to accommodate a binary exposure and a single binary confounder. With the increasing availability of preexisting yet incomplete databases, the potential for use of two-stage sampling will greatly increase in the future. This investigation provides a basis for estimating the number of participants to sample for the collection of confounder data at the second stage. *Am J Epidemiol* 1997;146:450-8.

biometry; case-control studies; confounding factors (epidemiology); epidemiologic methods; regression analysis; sample size; two-stage sampling

Many computerized data sources are potentially useful for epidemiologic research. Examples include physician claim files, hospital separation records, prepaid insurance plan databases, and occupational records. Studies using these databases can often be conducted more quickly and at a lower cost than those involving primary data collection. Unfortunately, such databases were seldom established with a view toward etiologic research. Although data on the exposure of interest may be available, with data on the endpoint of interest obtainable through record linkage, data on extraneous variables which could potentially confound the exposure-disease association are typically unavailable. The cost of surveying the entire study population may be prohibitive. A cost-effective alternative is to

collect confounder data on a subset of the original study population—an approach which has been termed “two-stage sampling” (1-4), since information pertaining to the crude and covariable-adjusted exposure effects is obtained in two separate phases of the investigation.

For efficiency, stage 2 sample selection is typically based jointly on exposure and disease status. Although methods for analyzing two-stage data exist (1, 3, 4), issues of sample size estimation have not been explicitly addressed. The objective of this article is to provide a method for deciding on the number of subjects to be selected at the second stage of a two-stage study.

EXAMPLES

Consider a hypothetical occupational cohort study of railway workers. Information on diesel exhaust exposure for cohort members could be estimated on the basis of employment histories. The cancer incidence experience of the cohort could be determined through linkage to vital statistics and cancer registry databases. However, the validity of any association would be compromised by the lack of information on smoking, since exposure might vary directly with smoking prevalence. If smoking data were not collected a priori, two-stage sampling could enhance the

Received for publication July 23, 1996, and in final form February 4, 1997.

Abbreviation: OR, odds ratio.
¹ Department of Epidemiology and Biostatistics, Faculty of Medicine, McGill University, Montréal, Québec, Canada.

² Laboratory Centre for Disease Control, Health Canada, Ottawa, Ontario, Canada.

³ Centre for Clinical Epidemiology and Community Studies, Sir Mortimer B. Davis-Jewish General Hospital, Montréal, Québec, Canada.

Reprint requests to Douglas Schaubel, Room 1368, LCDC Building 6601C1, Health Canada, Tunney's Pasture, Ottawa, Ontario K1A 0L2, Canada.

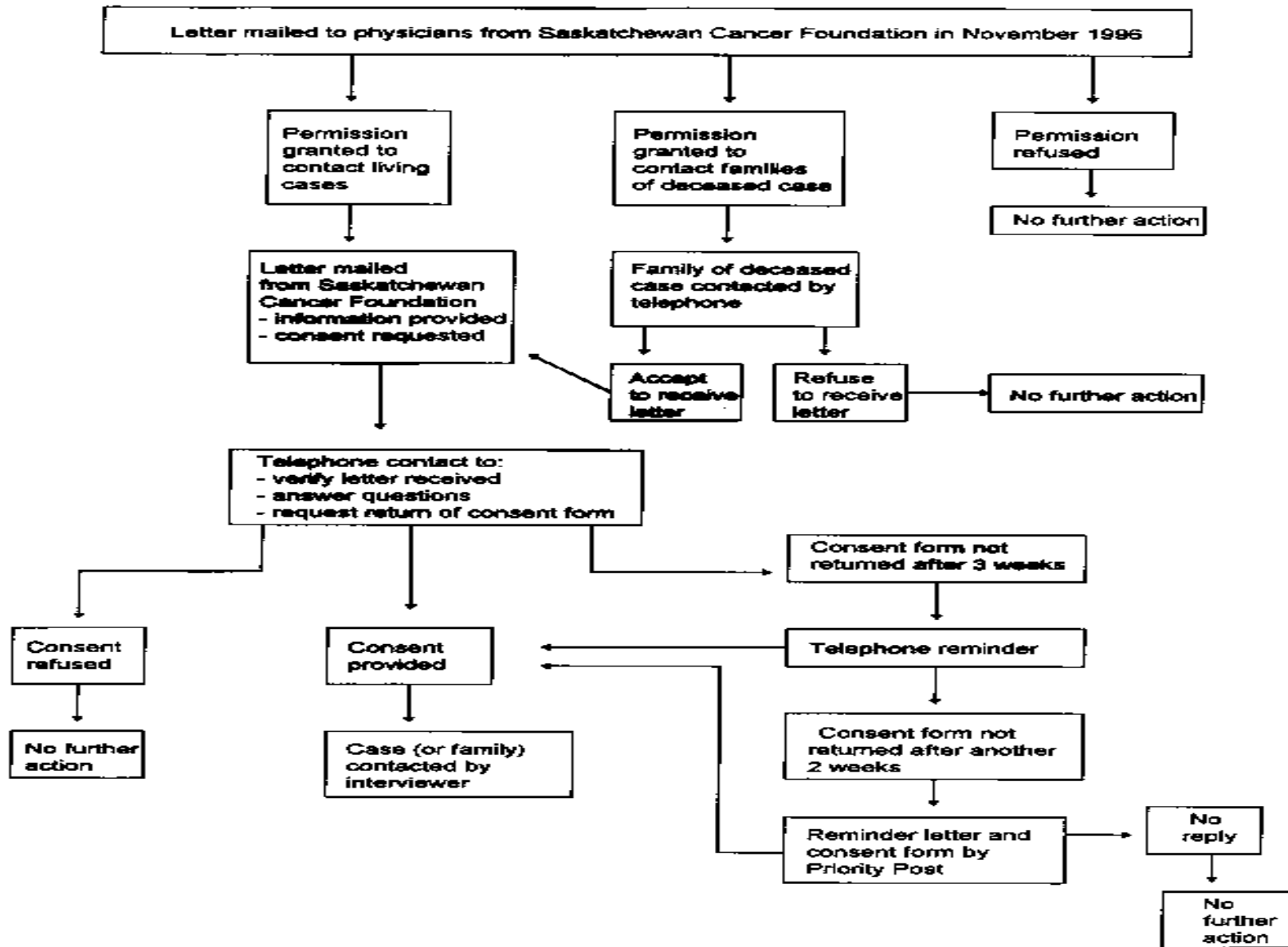
TABLE 1. Required stage 2 sample size* for two-stage case-control studies†

e^{\dagger}	P_C	θ	P_E ($N_1 = 1,000, N_0 = 2,000$)			P_E ($N_1 = 2,000, N_0 = 4,000$)		
			20%	30%	40%	20%	30%	40%
			2	10%	1.5	48	37	33
		3.0	136	99	84	85	72	64
		6.0	302	207	166	192	150	125
	30%	1.5	92	72	66	57	52	50
		3.0	245	195	179	156	143	136
		6.0	548	446	406	356	329	308
	50%	1.5	91	72	68	56	52	51
		3.0	220	184	179	139	134	135
		6.0	434	390	396	280	285	301
4	10%	1.5	192	148	135	119	107	102
		3.0	311	236	206	200	173	156
		6.0	528	387	320	352	285	242
	30%	1.5	308	248	232	194	180	175
		3.0	460	390	373	303	288	282
		6.0	762	680	666	516	504	500
	50%	1.5	260	211	201	162	154	152
		3.0	352	308	310	228	226	235
		6.0	507	464	522	337	358	393
6	10%	1.5	349	273	250	219	198	188
		3.0	490	381	339	323	282	257
		6.0	756	575	487	516	425	367
	30%	1.5	504	408	384	318	298	291
		3.0	639	560	548	428	414	414
		6.0	932	879	897	642	651	668
	50%	1.5	384	315	302	241	229	228
		3.0	443	397	408	291	293	308
		6.0	463	551	613	378	408	460

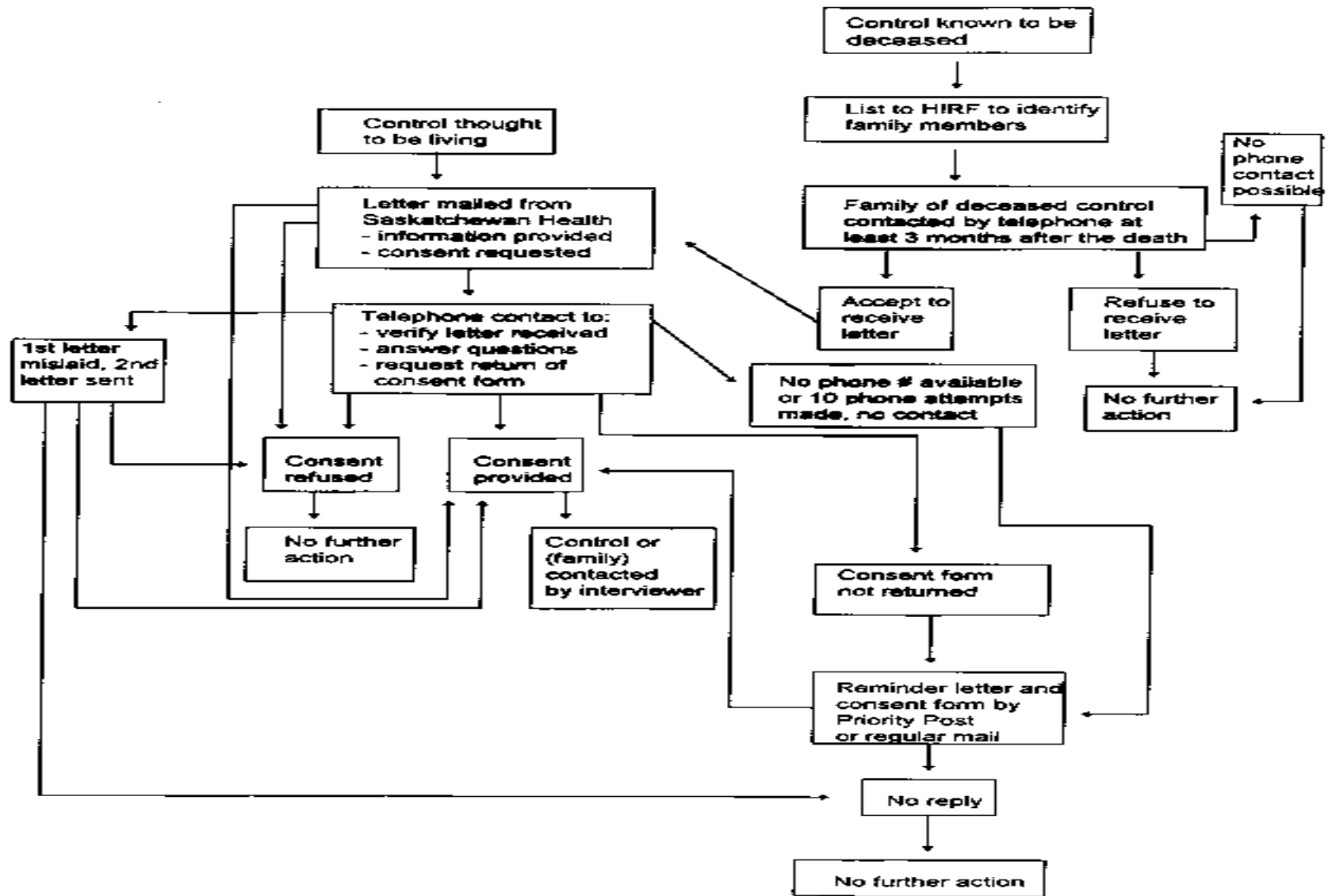
* Stage 2 sample size required to detect an exposure odds ratio (OR) of $e^{\theta} = 1.5$ with 90% power and type I error of $\alpha = 0.05$ (two-sided).

† A case-control study (N_1 cases and N_0 controls at stage 1) designed to evaluate the effect of exposure recorded on a binary scale, with adjustment for a single binary confounder with the following quantities anticipated: exposure prevalence = p_E , confounder prevalence = p_C , exposure OR = e^{θ} , confounder OR = e^{γ} , and (E,C) cross-product ratio = θ .

CONTACTING CASES OR FAMILIES OF DECEASED CASES



CONTACTING CONTROLS OR FAMILIES OF DECEASED CONTROLS



CONTROLS RECRUITMENT
(as of February 6, 1997)

FIRST MAIL OUT TO CONTROLS **n = 713**

Non deceased	n = 701	(98%)
No response	87	(12%)
Consented	272	(39%)
Refused	330	(47%)
Undeliverable	12	(2%)
Deceased	n = 12	(2%)
Consented	0	(0%)
Refused	2	(17%)
Undeliverable	10	(83%)

Total Consented **272** **(39%)**

Reason for refusing

Not interested	68
Too ill/depressed	34
Deceased	1
Communication problem	19
Questionnaire too long	6
Questions too personal	3
Too busy	24
Other reasons	28
No reason	147

CASES RECRUITMENT

(as of Feb. 6, 1997)

MAIL OUT TO PHYSICIANS:			n = 982 in December 1996
No response	16		(1.6%)
Gave consent	775		(78.9%)
Refused consent	189		(19.2%)
Undeliverable	3		(0.3%)

MAIL OUT TO PATIENTS:			n = 745
Non-deceased:			
	n = 690		(93%)
Consented	452		(66%) or (46%)
Refused	187		(27%)
Undeliverable	5		(0.7%)
Still to receive	46		(7%)
Deceased			
	n = 55		(7%)
Family consented	32		(40%)
Family refused	22		(58%)
Undeliverable	1		(0.5%)
Total Consented			n = 484 (65% or 49%)
Reason for refusing			
Not interested	25		
Too ill/depressed	8		
Communication problem	11		
Questionnaire too long	1		
Questions too personal	2		
No reason	126		

Other related complex sampling designs

- three-stage sampling
- partial questionnaire

Collection of covariate data
on cases only

Ray, Griffin 1989

Obese women

	cancer	no cancer
NSAIDs +	2 000	?
-	40	?

Not obese

	cancer	no cancer
NSAIDs +	200	?
-	400	?

All women

	cancer	no cancer
NSAIDs +	2 200	20 000
-	440	10 100
		32 740

Confounding totally accounts for exposure effect. A cohort study

Vitamin deficiency

Old

		Yes	No	
Depression	+	1 200	60	$\frac{1\ 200}{2\ 000} / \frac{60}{100} = 1.0$
	-	800	40	
		2 040	100	

Young

		Yes	No	
Depression	+	20	200	$\frac{20}{100} / \frac{200}{1\ 000} = 1.0$
	-	80	800	
		100	1 000	

All

		Yes	No	
Depression	+	1 220	260	$\frac{1\ 200}{2\ 100} / \frac{260}{1\ 100} = 2.5$
	-	880	840	
		2 100	1 100	

Conditions for confounding in the example of vitamin deficiency and depression

1. Age is associated with depression

In subjects without vitamin deficiency

		Old	Young
Depression	+	60	200
	-	40	200
		100	1 000

$$\frac{60}{100} / \frac{200}{1\ 000} = 3.0$$

In subjects with vitamin deficiency

		Old	Young
Depression	+	1 200	20
	-	800	80
		2 000	100

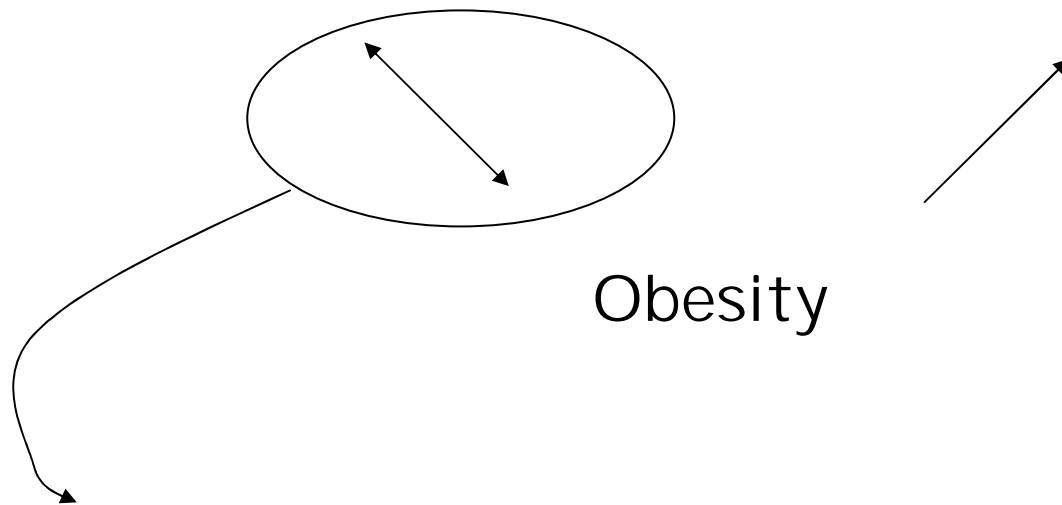
$$\frac{1\ 200}{2\ 000} / \frac{20}{100} = 3.0$$

2. Age is associated with vitamin deficiency

		Old	Young
Vitamin deficiency	+	2 000	100
	-	100	1 000

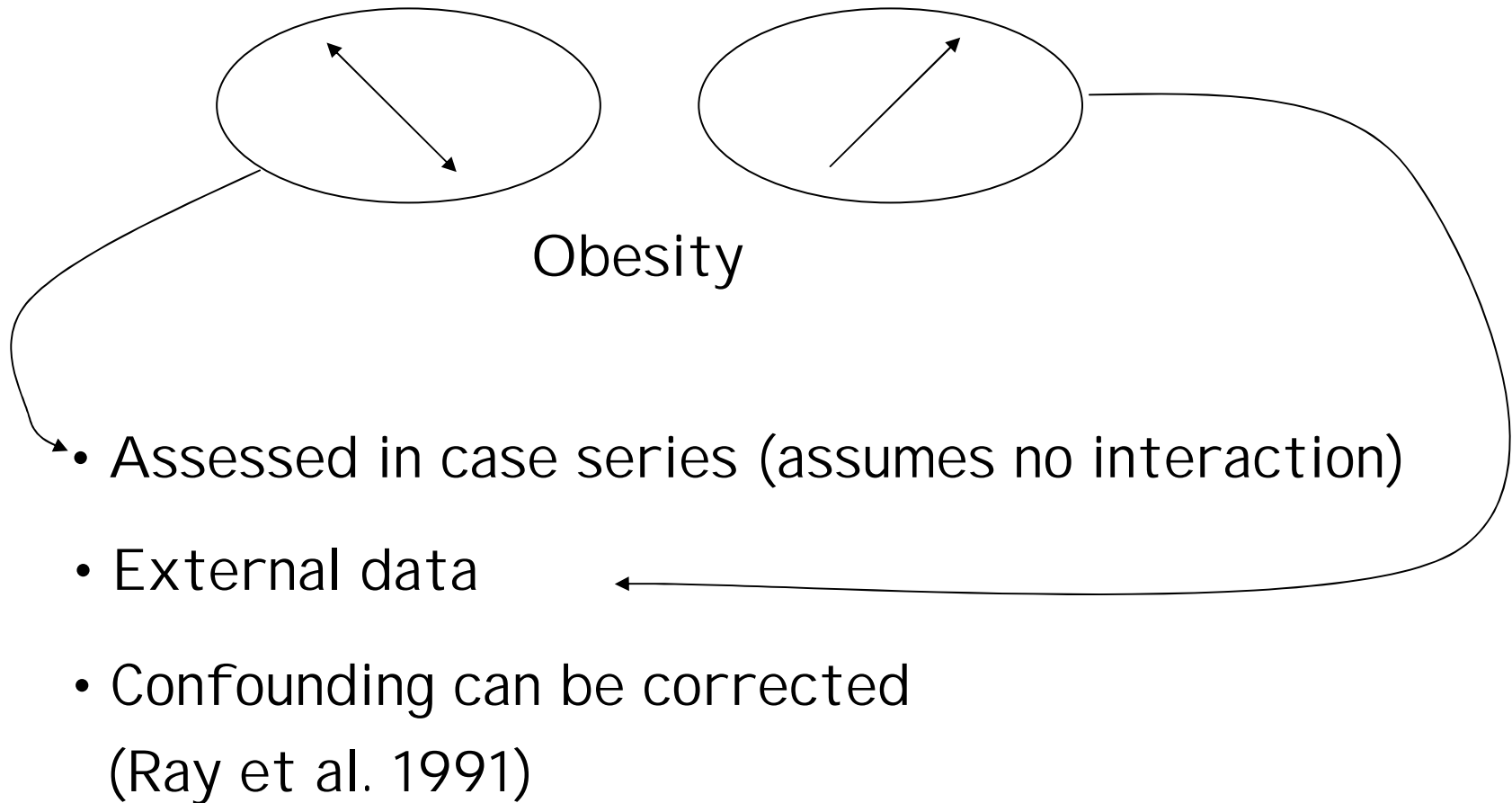
$$\text{Odds ratio} = \frac{2\ 000 \times 1\ 000}{100 \times 100} = 200$$

NSAIDs → Breast cancer

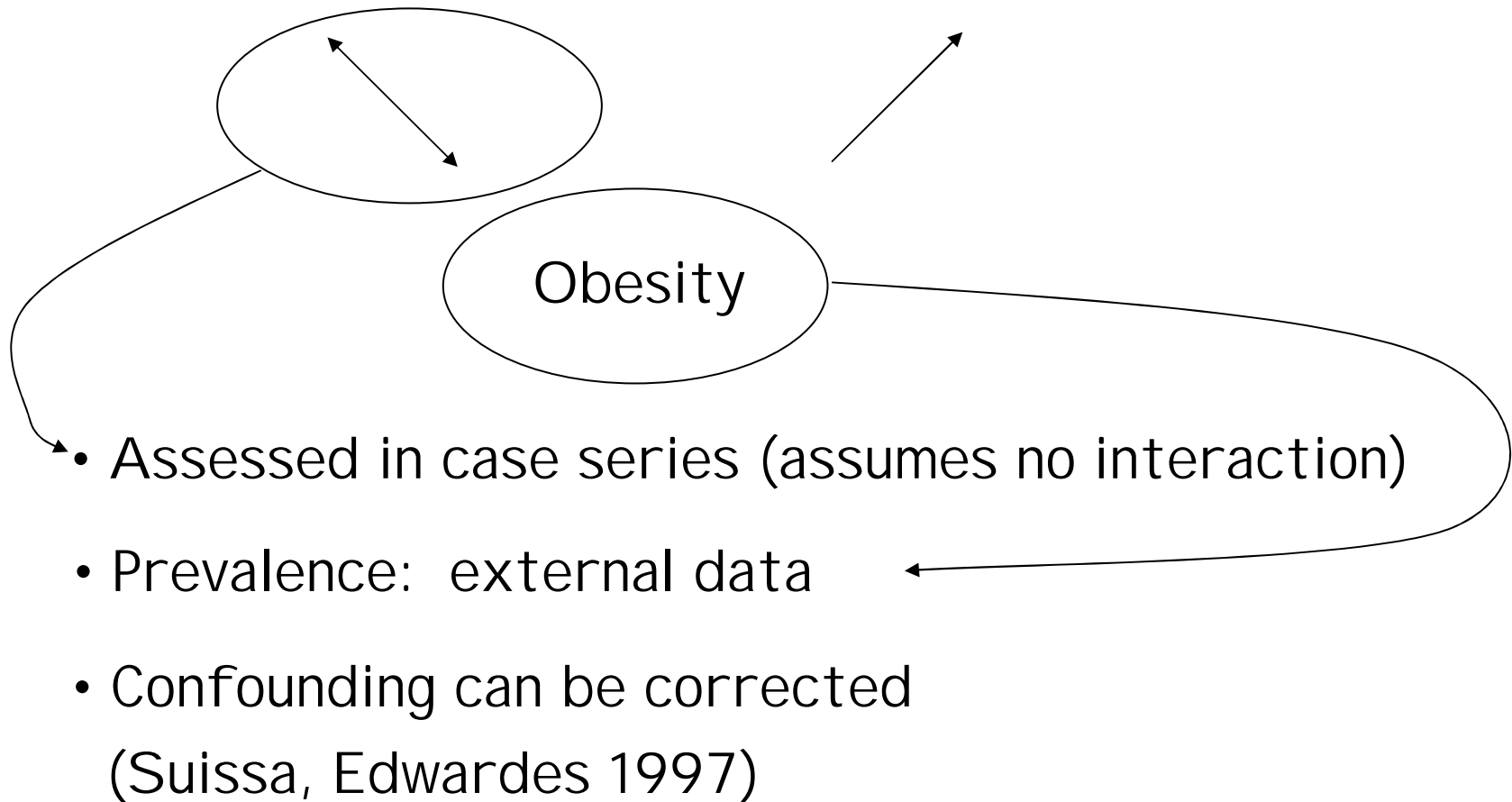


- Assessed in case series (assumes no interaction)
- Confounding can be diagnosed but not corrected
(Ray, Griffin 1989)

NSAIDs → Breast cancer



NSAIDs → Breast cancer



In summary

May be easily applicable in certain studies

Provides valid estimates at reduced cost

However:

- Assumption of absence of interaction
- External data must exist for correction of confounding

Example of interaction

Obese women

		cancer	no cancer	
NSAIDs	+			OR = 0.4
	-			

Not obese

		cancer	no cancer	
NSAIDs	+			OR = 1.0
	-			

All women

		cancer	no cancer	
NSAIDs	+			
	-			