Cancer Research Methods Workshop

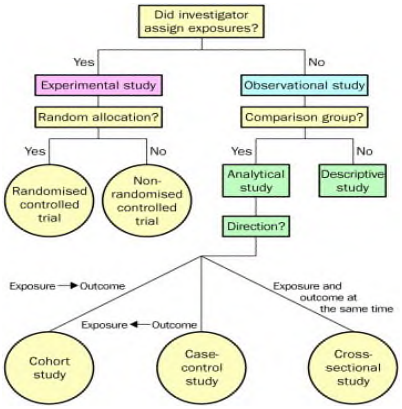
Mayur M. Desai, PhD, MPH

Mwanza, Tanzania

September 18-19, 2012

Yale SCHOOL OF PUBLIC HEALTH

Study Designs



Source: Lancet 2002;359:57–61.

Experimental study

- Randomized controlled trial: "the gold standard"
- Randomized clinical trial
  - unit of randomization is the individual
- Randomized community trial (also called randomized cluster or group randomized trial)
  - unit of randomization is at the aggregate level (e.g., communities, schools, hospitals, etc.)
- Non-randomized controlled trial (also called quasi-experimental study or natural experiment)

Randomized clinical trial
BASIC DESIGN

- Study subjects are randomized to be in either intervention or comparison group
- Intervention group: receives the new therapeutic or preventive intervention
- Comparison group: receives placebo, standard treatment, or no treatment, as appropriate
- The 2 groups are followed over time and compared with respect to the incidence of the outcome of interest (recovery, disease onset, death, etc.)

Randomized clinical trial
BASIC LOGIC

- If the sample size is large enough, randomization makes it likely (but doesn't guarantee) that the 2 groups will be similar with respect to potential confounders (measured and unmeasured) and that the only difference between the groups will be the intervention being studied

Randomization

- Protects against selection bias
- Helps defend against confounding
- Randomization procedure must be truly random and tamper-proof (allocation concealment)

Blinding

- Refers to keeping participants, investigators, and assessors unaware of group assignment, so that they are not influenced by that knowledge
- Of course, blinding is not always possible
- Use of a placebo does not guarantee blinding

Limitations of RCTs

- Cannot be used to study the effects of potentially harmful exposures
- Expensive
- Physicians are often reluctant to enroll their patients
- Patients are often reluctant to participate
- Eligibility criteria are often restrictive
- Trial efficacy doesn't always translate into "real-world" effectiveness

Randomized community (cluster) trial

(group randomized trial)

- Unit of randomization is at the aggregate level
- Intervention is delivered at the aggregate level
- Unit of analysis is at the individual level

Example of randomized community trial

© International Epidemiological Association 2003 Printed in Great Britain

International Journal of Epidemiology 2003;32:157–160
DOI 10.1093/ije/dyg019

BRIEF REPORT

A new tool for malaria prevention?: Results of a trial of permethrin-impregnated bedsheets (*shukas*) in an area of unstable transmission

Kate Macintyre,[1] Stephen Sosler,[1] Frederick Letipila,[2] Mohamed Lochigan,[2] Susan Hassig,[1] Sabah A Omar[3] and John Githure[4]

Malaria example

(Macintyre et al. Int J Epidemiol 2003;32:157-160.)

Objective:

- To evaluate whether treating personal clothing/sheets ("shukas") is protective against malaria infection

A *shuka* is a 6' by 4' sheet (smaller for children) of varying material that doubles as a wrap during the day and a sheet at night.

Malaria example

(Macintyre et al. Int J Epidemiol 2003;32:157-160.)

Methods:

- Setting: nomadic village of Ngilai, Samburu District, northern Kenya
- Sampling frame: 84 hamlets (manyattas) within a 2-hour walk of the Ngilai health clinic
- Study design: 36 manyattas were randomly selected and then randomized to be in either the experimental group or the control group (18 each)

Methods (cont.):

- Consent: approval from community elders; informed consent from all adults and parental consent for all children in the study
- Sample: N=234 and N=238 in the experimental and control manyattas, respectively
- Intervention: "Appropriate doses of sulphadoxine-pyrimethamine (Fansidar) were given to all members of the study at baseline. *Shukas* were dipped at baseline and an attempt was made to impregnate all *shukas* owned by the experimental group."

Methods (cont.):

- Follow-up: active case detection to measure parasitaemia took place at 2-week intervals over 4 months
- Blinding: lab technicians were blinded
- Outcome analysis: efficacy was assessed by comparing the incidence of malaria in the experimental and control groups (incident case defined as individual's first episode of parasitaemia)

Table 1. Social and demographic characteristics of the study population

Population characteristics	Control (n = 238)	Control (n = 238)	Experimental (n = 234)	Experimental (n = 234)
No. households:	64		60	
	n	%	n	%
Sex				
Female	141	59.2	131	56.0
Male	97	40.8	103	44.0
Age: mean (sd)	20.2 (18.2)		20.6 (17.8)	
	n	%	n	%
≤5	47	19.7	50	21.4
6-10	44	18.5	41	17.5
11-19	58	24.4	48	20.5
20-29	28	11.8	31	13.2
30-39	24	10.1	23	9.8
40+	37	15.5	41	17.5
Shuka ownership: mean	1.9		1.7	

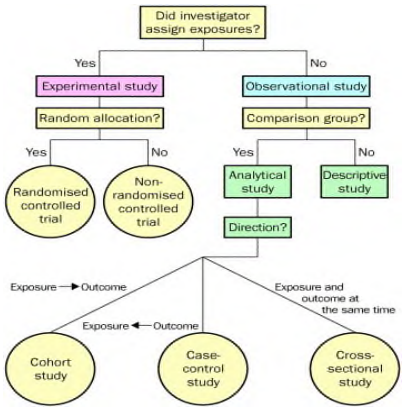
Table 2. Incidence density rates and ratios of control and experimental groups, by age

Age group	Control group No. cases	Control group IDR[a]	Experimental group No. cases	Experimental group IDR[a]	Incidence rate ratio IRR[b]	Incidence rate ratio (95% CI)[c]	Incidence rate ratio *P*
≤5 (n = 97)	2	5.35	4	10.03	1.874	(0.311-11.299)	0.49
>5 (n = 375)	11	7.49	2	1.41	0.187	(0.046-0.770)	0.02

a Incidence density rate per 10 000 person days at risk.
b Incidence rate ratio.
c Standard errors adjusted for clustering.

Conclusion:

- Permethrin-impregnated bedsheets may provide an alternative to impregnated bednets in communities where bednets are unaffordable or incompatible with the population's lifestyle



Source: Lancet 2002;359:57–61.

Observational study

Descriptive study

- Describes the distribution of a disease or other health-related state, event, or procedure with respect to person, place, and time
- Often relies on routinely-collected data sources
- Major uses include trend analysis, health care planning, and hypothesis generation

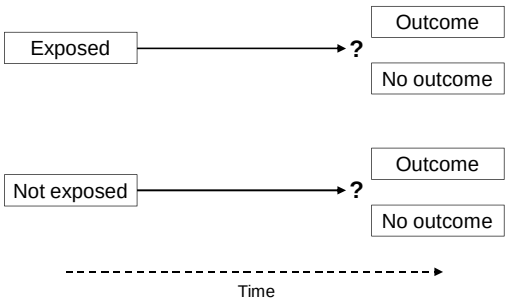
Observational study (cont.)

Analytical study

- Hypothesis-driven: test a specific hypothesis about a relationship between an exposure and an outcome
- Objective-driven: determine the prevalence and correlates of an outcome (descriptive and analytical studies are not necessarily mutually exclusive)

- Cohort
- Case-control
- Cross-sectional
- Ecological

Cohort study

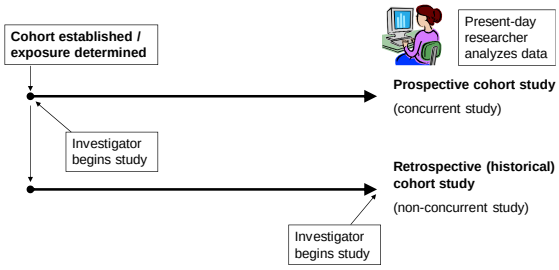


Cohort study

(also called prospective, follow-up, or longitudinal study)

- Basic design: outcome-free, at-risk persons are categorized on the basis of exposure, followed-up for a certain period of time, and compared with respect to the incidence of the outcome
- Temporal sequence between exposure and outcome more clearly established
- Generally, not suited for the study of outcomes that are rare or take a long time to develop

Types of cohort studies



Prospective cohort study

- "Gold standard" in observational epidemiology
- Assemble the cohort and collect baseline data at the start of the study
- Calendar time and follow-up time are concurrent
- Advantages: baseline assessments and follow-up procedures are planned and implemented for the purposes of the study; no differential recall of exposure
- Disadvantages: time, money, loss to follow-up

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Obstructive Sleep Apnea as a Risk Factor for Stroke and Death

H. Klar Yaggi, M.D., M.P.H., John Concato, M.D., M.P.H., Walter N. Kernan, M.D., Judith H. Lichtman, Ph.D., M.P.H., Lawrence M. Brass, M.D., and Vahid Mohsenin, M.D.

(NEJM 2005;353:2034-41.)

Obstructive sleep apnea example

Objective

- To test the hypothesis that patients with obstructive sleep apnea have an increased risk of stroke or death that is independent of other cerebrovascular risk factors

Obstructive sleep apnea example

Methods

- Sample: patients (aged 50+ years) referred to the Yale Center for Sleep Medicine, specifically for the evaluation of sleep-disordered breathing
  - exclusion criteria included history of stroke
- Baseline data collection:
  - at least 2 hours of attended sleep monitoring (overnight polysomnography)
  - patient interview
  - height and weight measurements

Obstructive sleep apnea example

Methods (cont.)

- Exposure: apnea-hypopnea index of 5+ vs. <5 (number of events per hour of sleep)
- Outcome: composite outcome of stroke or death
  - follow-up patient or proxy questionnaire
  - reported strokes (and TIAs) validated by reviewing medical records
  - vital records from the Connecticut Department of Public Health and the Social Security Administration used to determine or confirm death

Obstructive sleep apnea example

Results

- Sample: between 1/1997 and 12/2000, 3,635 consecutive patients were referred to the sleep center, of whom **1,022** were eligible for the study
  - n=697 (68%) were categorized as having OSA
- Follow-up: between 6/2002 and 12/2003 (median follow-up ~3.4 years)
  - data on stroke events and death were obtained for n=842 (82%)
  - incident stroke or death occurred in n=88 (9%)

Incidence rates (stroke/death events per 100 person-yrs):
- OSA group: 3.48
- Unexp group: 1.60

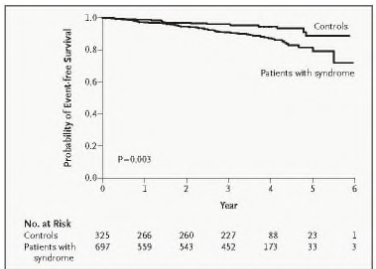
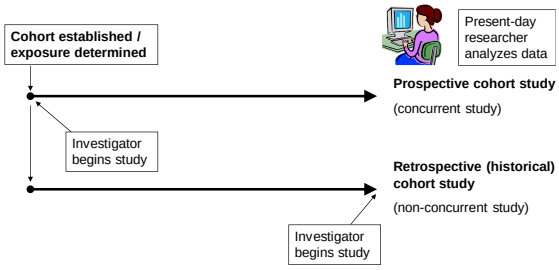


Figure 1. Kaplan–Meier Estimates of the Probability of Event-free Survival among Patients with the Obstructive Sleep Apnea Syndrome and Controls.

Even after controlling for covariates, the negative effect of obstructive sleep apnea remained significant.

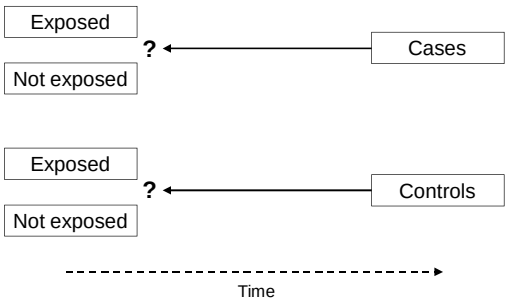
Types of cohort studies



Retrospective (historical) cohort study

- Study is initiated after exposure and outcome have occurred
- Use existing records (or self-reports) to:
  - assemble a cohort based on past exposure
  - determine the outcome experience of the cohort from a defined point in the past up to the present time
- Advantages: relatively low cost and takes less time than prospective cohort study
- Disadvantage: reliance on records collected for purposes other than conducting an epidemiologic study

Case-control study



Case-control study

Basic design

- Groups are defined by outcome (cases and controls)
- Look back in time to ascertain each person's exposure status
- Compare the odds or level of exposure in the groups

Major advantages of case-control study

- Generally, more efficient in terms of time, money, and effort than cohort studies
- Able to test hypotheses about multiple exposures and the outcome
- Well-suited for the study of rare outcomes

Major disadvantages of case-control study

- Not well-suited for rare exposures
- Challenge to identify appropriate control group
- Greater possibility of selection bias and information (recall) bias
- Cannot calculate incidence rates, risk, relative risk (odds ratio estimates relative risk)

Selection of cases

- Clear and explicit case definition
- Case selection must be independent of exposure status
- Hospital/clinic-based cases: easier and less expensive to identify and recruit, but more prone to selection bias
- Population-based cases: better representation of all cases in the population

Selection of cases
incident vs. prevalent

- Cases are usually considered "incident" at the time they are diagnosed
- Use of incident cases has several advantages:
  - since the outcome is newly diagnosed, better able to distinguish risk factors from prognostic factors and consequences of the outcome
  - recall of past exposures may be better due to decreased elapsed time between exposure and investigation
  - cases are likely to be more homogenous, since diagnostic criteria and procedures may change over time
  - better able to study outcomes with high fatality rate

Selection of cases
incident vs. prevalent

- Disadvantages of using incident cases:
  - time and cost involved in recruiting cases prospectively
  - potential influence of new diagnosis on recall of past exposures
- Generally, preferable to use incident cases – however, if suitable, advantage to using prevalent cases is that a larger number of cases are available at the start of the study

Selection of controls

- Free of the outcome of interest (not necessarily "healthy")
- Meant to indicate the frequency or level of exposure that would be expected among cases if there were no association between exposure and outcome
- Representative of the population at risk for the outcome (ideally, a random sample of the source population)
- Selected independent of the exposure of interest
- Would have been picked up as cases had they developed the outcome

Major sources for selecting controls

- Population controls
- Hospital/clinic controls
- Neighborhood controls
- Household/family/sibling controls

Risk factors for pulmonary tuberculosis: a clinic-based case control study in The Gambia

Philip C Hill*[1], Dolly Jackson-Sillah[1], Simon A Donkor[1], Jacob Otu[1], Richard A Adegbola[1] and Christian Lienhardt[2]

Address: [1]Tuberculosis Division, Medical Research Council Laboratories, Banjul, The Gambia and [2]Programme Tuberculose -UMR 145, Institut de Recherche pour le Développement, Montpellier, France

Email: Philip C Hill* - phill@mrc.gm; Dolly Jackson-Sillah - dolly.jackson-sillah@lshtm.ac.uk; Simon A Donkor - sdonkor@mrc.gm; Jacob Otu - jotu@mrc.gm; Richard A Adegbola - radegbola@mrc.gm; Christian Lienhardt - lienhardt@ird.sn

* Corresponding author

Published: 19 June 2006

BMC Public Health 2006, **6**:156 doi:10.1186/1471-2458-6-156

Received: 01 February 2006
Accepted: 19 June 2006

TB case-control example

Objective

- To identify risk factors for pulmonary tuberculosis (TB) in those attending a general outpatient clinic in The Gambia

TB case-control example

Methods

- **Cases**
  - consecutive, newly-diagnosed TB patients aged ≥15 years were recruited from the OPD (June 2002 to June 2004)
- **Controls**
  - for each case, two controls were selected
  - controls were matched on age (within 10 years) and sex
  - the first 'first-time' OPD attendee fitting the age and sex criteria was approached on a clinic day; if he/she refused to participate, the next eligible clinic attendee was approached
  - the control was seen by the study doctor to address their medical complaint and to have a chest x-ray to exclude TB

TB case-control example

Methods (cont.)

- Data collection
  - study participants answered a structured questionnaire administered in their own language by a trained health worker
  - information was collected on a wide range of potential host- and environmental-related risk factors for TB, focusing on those that might be amenable to an intervention

Matched on age group and sex

	Cases (n = 100)	Controls (n = 200)
Age, years		
Mean (std dev)	33.6 (14.4)	33.5 (13.2)
Median (range)	30 (15-75)	30 (15-75)
Sex, n (%)		
Male	65 (65)	130 (65)
Female	35 (35)	70 (35)
HIV positive, n(%)	6[a] (6.1)	2[b] (3.3)

[a]n = 98, [b]n = 60

Table 2: Host-related factors for tuberculosis: comparison of TB cases and clinic controls.

Variable	Controls	Cases	Total No. triplets	OR (95% CI)	P-value	Adj *OR (95% CI)	Adj* p value
Ethnic group, n(%)			100				
Mandinka	105 (52.5)	41 (41)					
Wolof	25 (12.5)	8 (8)					
Fula	28 (14)	11 (11)					
Jola	21 (10.5)	22 (22)					
Other	31 (15.5)	18 (18)					
Marital status, n(%)			100				
Married	118 (59)	48 (48)					
Single	74 (37)	43 (43)					
Widowed/divorced	8 (4)	9 (9)					
Smoker in past 6 months			100				
No	167 (83.5)	72 (72)					
Yes	33 (16.5)	28 (28)					
Alcohol							
Never	189 (94.5)	90 (90)					
Current/past	11 (5.5)	10 (10)					
BCG scar			100				
No/unsure	143 (71.5)	73 (73)					
Yes	57 (28.5)	27 (27)					
History of asthma			100				
No	198 (98)	99 (99)					
Yes	2 (1)	1 (1)					
History of diabetes			100				
No	200 (100)	97 (97)					
Yes	0 (0)	3 (3)					
Years of schooling			100				
None	42 (21)	25 (25)					
1-4	24 (12)	10 (10)					
5-8	40 (20)	27 (27)					
>8	94 (47)	38 (38)					
Occupation			100				
Unemployed	16 (8)	10 (10)					
Unskilled worker	29 (14.5)	22 (22)					
Skilled worker	96 (48)	41 (41)					
Professional	22 (11)	3 (3)					
Other	27 (18.5)	24 (24)					

*Adjusted through multivariable analysis

Table 3: Environmental factors for tuberculosis: comparison of TB cases and clinic controls.

Variable	Controls	Cases	Total No. triplets	OR (95% CI)	P-value	Adj *OR (95% CI)	Adj* p value
Electric or gas cooker			100				
Absent	159 (79.5)	86 (86%)					
Present	41 (20.5)	14 (14%)					
Ceiling			100				
Yes	155 (77.5)	63 (63)					
No	45 (22.5)	37 (37)					
Walls			100				
Cement	136 (68)	56 (56)					
Mud	4 (2)	11 (11)					
Mudbrick/other	60 (30)	33 (33)					
Floor			100				
Earth	27 (13.5)	18 (18)					
Concrete/tiles	170 (85)	71 (71)					
Number of windows			99				
0-1	60 (30.3)	29 (29.3)					
2-4	116 (58.5)	51 (51.5)					
>4	24 (12.1)	18 (18.2)					
Occupation head of house			74				
Unemployed	18 (10.2)	10 (12.5)					
Unskilled worker	37 (21.0)	10 (12.5)					
Skilled worker	18 (10.2)	3 (3.8)					
Professional	86 (48.8)	40 (50)					
Other	17 (9.7)	17 (21.3)					
Crowding index			91				
1	44 (23.2)	10 (11.8)					
2	31 (16.4)	11 (12.9)					
3	189 (60.3)	64 (75.3)					
Household member had TB			100				
No/unsure	178 (89)	55 (55)					
Yes	22 (11)	45 (45)					

*Adjusted through multivariable analysis

Table 4: Combined multivariable assessment of host and Environmental factors for tuberculosis: compa controls. (n = 274)

Variable	Adj[a]OR (95% CI)	Adj[a] p value
Ethnic group, n(%)		
Mandinka	1.0	
Wolof	2.76 (0.65-2.01)	0.17
Fula	0.78 (023-2.64)	0.70
Jola	3.83 (1.35-10.86)	0.012
Other	0.94 (0.32-2.74)	0.91
Smoker in past 6 months		
No	1.0	
Yes	1.88 (0.83-4.26)	0.13
Occupation		
Unemployed	2.70 (0.56-12.94)	0.22
Unskilled worker	1.0	
Skilled worker	0.65 (0.24-1.71)	0.38
Professional	0.22 (0.03-1.72)	0.15
Other	1.17 (0.38-3.61)	0.79
Ceiling		
Yes	1.0	
No	2.12 (0.91-4.93)	0.082
Crowding index		
1	1.0	
2	2.58 (0.69-9.68)	0.16
3	5.05 (1.75-14.62)	0.003
Household member had TB		
No/unsure		
Yes	10.17 (4.08-25.63)	<0.0001

[a]Adjusted

TB case-control example

Conclusion

- In The Gambia, household crowding and past household exposure to a known TB case are independent risk factors for TB disease
- Further research is needed to identify why risk of TB seems to differ according to ethnicity

Cross-sectional study

- Information on prevalence of “exposure” and/or “outcome” collected for a specified population at a particular point in time
- Describe current status of a population
- Can be descriptive (prevalence estimation) or analytical (relationship between “exposure” and “outcome”)
- Study may include entire specified population or sample of specified population

Some important considerations in cross-sectional studies

- Representativeness/generalizability of the sample
- Response rate
- Observed “cases” are not all of the cases produced by the study population (i.e., some may have died, recovered, moved away, refused to participate)

Major strengths of cross-sectional studies

- Relatively quick, economical, and easy to conduct
- Answers the question of what the population of interest looks like now

Major limitations of cross-sectional studies

- Inefficient when exposures or outcomes are rare
- Direction of association often cannot be determined (e.g., cross-sectional association between depression and obesity)
- Difficulty distinguishing risk factors from prognostic (survival) factors and consequences of the outcome

AIDS PATIENT CARE and STDs
Volume 18, Number 5, 2004
© Mary Ann Liebert, Inc.

Nonadherence to Antiretroviral Therapy in HIV-Positive Patients in Costa Rica

BRIAN DOUGLAS STOUT, M.P.H.,[1] MARIA PAZ LEON, M.D.,[2] and LINDA M. NICCOLAI, Ph.D.[1]

Nonadherence to HIV therapy in Costa Rica

(Stout et al., 2004)

Objective:

- To determine the prevalence and predictors of nonadherence among a sample of HIV-infected patients in Costa Rica

Methods:

- Setting: outpatient clinic of a large referral hospital in San Jose, Costa Rica (summer 2002)
- Sample: convenience sample, N=88 (87% participation rate), aged ≥18 years, taking ART for ≥6 months
- Data collection: interviewer-administered questionnaire and medical record abstraction

Table 1. Subject Characteristics (*n* = 88)

Characteristic	*No. (%)*
Gender	
Male	75 (85)
Female	13 (15)
Age	
18-29	13 (15)
30-59	70 (79)
60-79	5 (6)
Median	38.2 years
Education	
Primary	25 (28)
Secondary	13 (15)
Some college	14 (16)
College graduate	11 (13)
Postgraduate	22 (25)
Other	3 (3)
Lives alone	
No	79 (90)
Yes	9 (10)
Works for Pay	
No	60 (68)
Yes	28 (32)
Most recent CD4 T cell count	
0-199	4 (4)
200-499	43 (49)
500+	41 (47)
Most recent serum HIV-1 viral load	65 (74)
Undetectable	17 (20)
<10,000	3 (3)
10,000-49,000	3 (3)
50,000+	

Table 2. Predictors of Nonadherence in Past Three Days

Characteristic	*No. (%)*	*Adherent (%)*	*Unadjusted OR (CI)*	*Adjusted OR (CI)*
Transportation				
Difficult	7 (54)	16 (21)	4.3 (1.3-14.6)**	6.3 (1.5-26.9)**
Not difficult	6 (46)	59 (79)	1.0	1.0
Takes meds on empty stomach[a]				
Difficult	9 (69)	29 (39)	3.3 (0.9-11.6)*	6.7 (1.3-35.7)**
Not difficult	4 (31)	42 (61)	1.0	1.0
Maintaining a good attitude or mental state[b]				
Not helpful	2 (15)	1 (1)	13.3 (1.1-166.6)**	NS
Helpful	11 (85)	73 (99)	1.0	
Laying pills out				
Yes	7 (54)	57(77)	2.9 (0.9-9.7)*	9.9 (1.7-55.4)**
No	6 (46)	17 (23)	1.0	1.0
Pill burden				
High	1 (8)	25 (33)	6.0 (0.7-48.8)*	NS
Low	12 (92)	50 (67)	1.0	

*$p < 0.10$.
**$p < 0.05$.
P values are Fisher's exact two-sided *p* values.
[a](*n* = 84) 4 missing because of no restrictions on medication.
[b]n = 87.

Nonadherence to HIV therapy in Costa Rica

(Stout et al., 2004)

Some limitations noted by the authors:

- Possible over-reporting of adherence
  - however, interview not done by clinician
- Small convenience sample
  - however, sample size is larger than for many other studies, and only adherence study to date in Costa Rica

Ecological study

- The key characteristic is that the units of observation are populations, as opposed to individuals
- Correlation coefficient ("r") usually used to quantify association between independent and dependent variables

Unit of observation or analysis

- Usually, geographically-defined populations (e.g., cities, regions, countries)
  - Examine different populations at a given point in time

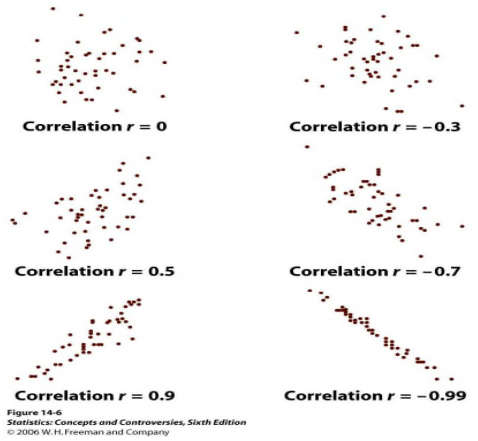
    or
  - Examine the same population at different points in time

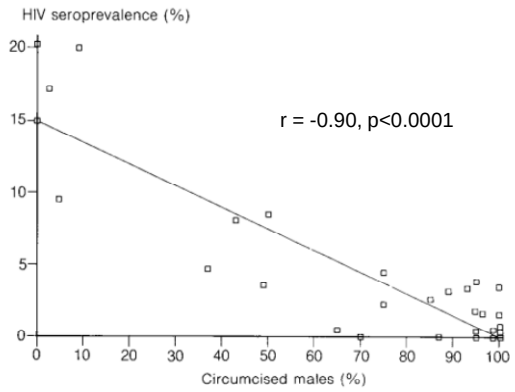
Examples

- Association between prevalence of gun ownership and rate of homicide among different countries in 2000
- Association between amount of cigarette tax and prevalence of smoking in the USA between 1995 and 2005

Correlation coefficient (*r*)

- The strength of the linear relationship between 2 continuous variables is measured with a correlation coefficient (mostly Pearson correlation)
- *r* has two components: sign (+ or -) and magnitude
- *r* ranges from -1 to +1:
  - +1 perfect positive correlation
  - 0 no correlation
  - -1 perfect negative correlation





Major strengths of ecological studies

- Relatively quick, cheap, and easy to conduct – largely rely on secondary data sources with ecological-level variables (e.g., Census data, vital statistics, disease registries, etc.)
- Generate hypotheses

Major limitations of ecological studies

- Associations among groups may not hold at the individual level – ecological fallacy
- Secondary data sources may be inaccurate or incomplete
- Summary measures of exposure or outcome may hide wide variability in the data, leading to imprecise measures of association
- It may be impossible to isolate the independent effects of specific variables on the outcome due to multicollinearity
- It may be difficult to identify and control for potential confounders at the group level
- It may be difficult to establish a correct temporal sequence between exposure and outcome
- There may be uncertain effects due to migration of subjects across ecological units