Highlights from Week #2

- Estimation of average causal effects requires strong assumptions
  - Exchangeability, positivity, consistency
- Causal DAGs can be used to identify approaches to study design and analysis to eliminate systematic biases
- Confounding is a bias of common causes
  - Causal DAGs allow us to integrate expert knowledge into an analysis to identify true confounders
- Selection bias is a bias of common effects
  - Causal DAGS allow us to identify the structure of selection bias and methods to “adjust” for bias

Learning Objectives

- **Classification of epidemiologic studies**
- **Design features of cohort studies**
  - Cohort study as an approximation of an RCT design
  - Closed vs. Open cohorts (Fixed vs. Dynamic, reviewed in Epi 203)
  - Classifying person-time
  - Lag time
- **Field methods, i.e. nuts and bolts of conducting observational studies**
- **Introduction to instrumentation and data collection methods**

3

UCSF

How we are going to proceed today…

- Start with what you've learned in Epi 203
- Add and expand by linking with causal theories
- Explain common origins of RCT and cohort studies
- Give examples of field methods of data collection for cohort studies
- Showcase mistakes in data collection methods

4

UCSF

Epidemiological Methods (building upon Epi 203)

- **Classifications by:**
  - Unit of analysis (discussed in Epi 203)
  - Approach to data collection
  - Goal
  - Timing and directionality
  - ***New:*** **by outcome (N. Pearce 2012)**

5

UCSF

Classification by unit of analysis

- **What is a unit?**
  - Observations for which outcome and exposure are measured
- **Individual-level variables are properties of individuals**
- **ecological variables are properties of groups, organizations or places**

6

UCSF

Classification by approach to data collection

- **Experimental**
  - RCTs, field trials, community intervention and cluster randomized trials
- **Quasi-experimental**
  - natural disaster studies
- **Non-experimental or observational**
  - cohort, case-control, *ecological*

7

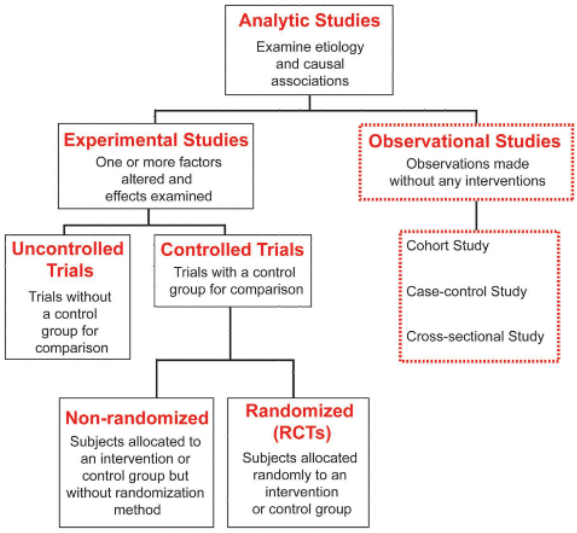
UCSF

Classification by goal

- **Descriptive**
  - ecological correlational studies, case reports, case series, cross-sectional surveys
- **Analytic**
  - observational studies and intervention studies (RCTs)

8

UCSF



9

UCSF

Classification by timing and directionality

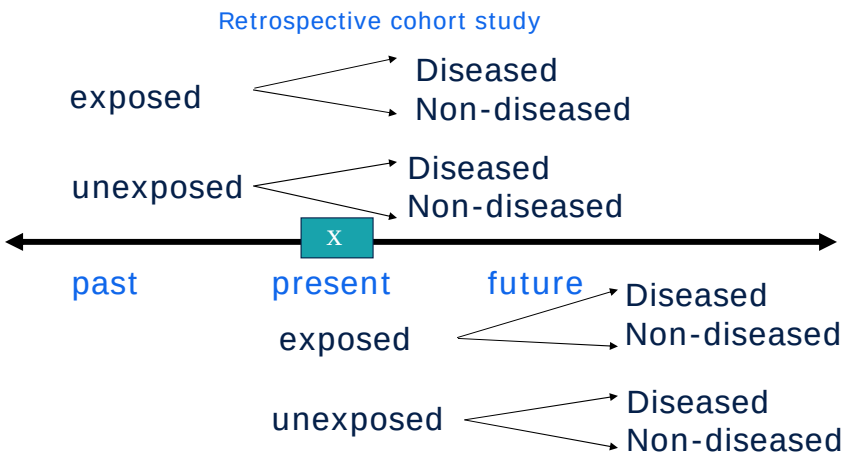
- **Directionality: "Which did you observe first, the exposure or the disease?"**
  - forward (RCT, cohort)
  - backwards (case-control)
- **Timing: "Has the information being studied already occurred before the study actually began?"**
  - retrospective and prospective cohort studies

10

UCSF

UCSF

Classification by timing and directionality



Classification by outcome

Table 1 Four basic study types

	Sampling on outcome	
Study outcome	No	Yes
Incidence	Incidence studies	Incidence case–control studies
Prevalence	Prevalence studies	Prevalence case–control studies

- Incidence studies identify causes of diseases
- Prevalence studies are more suitable for some diseases (e.g. asthma and diabetes), where incidence may be difficult to measure without very intensive follow-up
- **Caveat**: disease prevalence may differ between two groups because of differences in age-specific disease incidence, disease duration or other population parameters

Pearce 2012

Cohort Studies

13

UCSF

Q: At which of the following steps is epidemiologist's role critical to the success of the study?

A) Study design/data collection

B) Data analysis/manuscript writing

C) Implementation/advocacy

14

UCSF

UCSF

Q: At which of the following steps is epidemiologist's role critical to the success of the study?

A) Study design/data collection

B) Data analysis/manuscript writing

C) Implementation/advocacy

15

UCSF

Highlights from Week #1

- **Sufficient-component cause model**
  - Useful for examining multiple risk factors ('web of causation'), but omits discussion of origins of causes, focuses on proximal causes, and ignores induction period
  - Does not address indirect effects
  - Causes of disease in individuals or groups of individuals, not in populations
  - Does not consider factors that control distribution of risk factors
  - Ignores dynamic non-linear relations
- **Counterfactual model**
  - Can be applied at individual or population level
  - Specifies what would happen to individuals or populations under alternative patterns of exposure
  - Forces researchers to think about operational definitions of cases and controls, sampling schemes, and other important design questions
  - One of the two conditions in the definitions of the effect measures must be contrary to fact – exposures or treatment vs. a reference condition

16

UCSF

UCSF

Main Epidemiologic Study Designs for Testing Hypotheses

- Randomized controlled trial = **Investigator assigns exposure and follows participants overtime to ascertain the outcome**
- Cohort study = **Analogous to RCT, but investigator does not assign exposure**
- Case-control study = **Analogous to cohort study, but with more efficient sampling**

17

UCSF

Principles of experimental studies

1. Randomization of treatment so groups are comparable on known and unknown confounders.
2. Use placebo in order to reduce bias.
3. Blinding to avoid bias in outcome ascertainment.

18

UCSF

UCSF

Principles of experimental studies applied to observational cohort studies

1. Randomization of treatment so groups are comparable on known and unknown confounders.

- RCTs are used to identify and quantify average causal effects because randomized assignment of treatments leads to **exchangeability**.

➢Can't randomize in an observational study so select a comparison group as alike as possible to the exposed group (try to achieve **exchangeability**)

19

UCSF

Principles of experimental studies applied to observational cohort studies

2. Use placebo in order to reduce bias.

➢Can't use placebo in observational studies so you must make the groups as comparable as possible.

20

UCSF

UCSF

Principles of experimental studies applied to observational cohort studies

3. Blinding to avoid bias in outcome ascertainment.

- In a cohort study, it is crucial to have high follow-up rates and comparable ascertainment of outcomes in the exposed and comparison groups.
- You can blind the investigators conducting follow up and confirming the outcomes.

When does observational study emulate an RCT design?

Using the Counterfactual Model of Causal Inference

- 1.The values of treatments under comparison correspond to well-defined interventions (stable-unit treatment, well-defined interventions, **stability**, consistency)
- 2.The conditional probability of receiving every value of treatment, though not decided by the investigators, depends only on the measured covariates (**exchangeability**)
- 3.The conditional probability of receiving every value of treatment is greater than zero, i.e., positive (**positivity**)

Various Definitions of Stability

- **Well-defined interventions (H&R2015)**
  - E.g., In a study of heart transplants, had the protocol not specified these details, it is possible that each doctor had conducted the heart transplant **in a different way**, perhaps using her preferred surgical technique or immunosuppressive therapy, i.e. different versions of the treatment "heart transplant" might have been applied to each patient in the study
  - presence of multiple versions of treatment is problematic when the causal effect varies across versions, i.e., when the versions of treatment are relevant for the outcome
- **Cole and Frangakis 2009**
  - There are a number of ways in which treatment X could have been assigned, but all those ways would have resulted in the same observed outcome
  - Assume that all k routes allowed by the (implicit or explicit) exposure assignment mechanism are equivalent

UCSF

Various Definitions of Stability, *cont.*

- **Hoffler 2005, stable-unit-treatment assingment assumption:**
  - After treatment assignment and the observation of the outcome, a sample of n individuals contains (at most) one realization of the outcome for each individual i where the outcome corresponds either to treatment level t or c.

UCSF

UCSF

Exchangeability

In observational studies, this is very hard to achieve because the distribution of outcome predictors generally vary between exposed and unexposed.

- Example: Heart transplants (E) are not assigned randomly but go to the sickest patients (condition L).
- Questions:
  - Does exchangeability hold here?
  - Can we verify if exchangeability holds by looking at the data?

Positivity

We must ensure that there is a positive probability of being assigned to each of the E levels.

- Question:
  - Can we verify if positivity holds by looking at the data?

Issues in design of cohort studies

1. Cohort specification

Closed cohort:

- **Fixed population**
  - After start of follow-up, no one can be added
  - Closely resemble RCT as follow-up starts at randomization
- **Membership-defining event**
  - Cohort of persons born in 2012 (calendar time)
  - Cohort formed at entry into UCSF medical school (an event)
- **Can directly calculate:**
  - Risk ratio
  - Incidence rate ratio
  - Odds ratio

27

UCSF

Issues in design of cohort studies

1. Cohort specification

- **Incidence proportion** (**IP,** also called cumulative incidence, average risk) **cannot be directly measured, but is more intuitive and easier to communicate**
- **Limitations of IP analyses:**
  - Loss to follow-up
  - Competing risks
  - For certain outcomes (e.g. death), IP tends toward 1 over time
  - Exposures may change over time
- **Solution**: **Calculate incidence rates (IR), assume that IR ~ IP**

28

UCSF

UCSF

Issues in design of cohort studies

1. Cohort specification

Open Cohorts (Dynamic Cohorts)

- No fixed roster – person-time accrued from a potentially changing roster of individuals
  - Residents of California, Members of Kaiser Permanente
- Individuals can contribute varying amounts of person-time
- Individuals can leave and re-enter the cohort
  - Connecticut Tumor Registry
- Individuals can contribute person-time to multiple exposure categories over time
  - Occupationally exposed workers

29

UCSF

Issues in design of cohort studies

1. Cohort specification

Open Cohorts (Dynamic Cohorts)

- **Can directly calculate:**
  - IR
- **Can calculate IP from IR assuming:**
  - The IR is constant over the follow-up periods (j)
  - No competing risks or loss to follow-up related to disease risk
  - Number of events in each time interval is small relative to number of people at risk in that time interval
- **Incidence Proportion =** $1 - e^{-\sum IR_j \Delta t_j}$

30

UCSF

UCSF

Example

Incidence of a first episode of major depressive disorder (MDD) in young adults in their 30's

- 10 years of data from a large health insurance company that covers treatment for mental illness
- Assumptions: all cases of MDD result in a physician visit, correct diagnosis, and a correctly-coded insurance claim, so all of the numbers are accurate; the situation has been like this for many years, so that the incidence and population are stable
- Observed **incidence rate is 0.005 per year, or 5 per 1,000 per year**

31

UCSF

Q1: What proportion of 10,000 30-year-olds will experience a first episode of MDD by the time they reach age 40 years? **incidence rate=5 per 1,000 per year**

A) IP=5 per 10,000

B) IP=0.5%

C) IP=5%

32

UCSF

UCSF

Answer

- 10 years of data from a large health insurance company that covers treatment for mental illness
- Assumptions: all cases of MDD result in a physician visit, correct diagnosis, and a correctly-coded insurance claim, so all of the numbers are accurate; the situation has been like this for many years, so that the incidence and population are stable
- Observed incidence rate is 0.005 per year, or 5 per 1,000 per year
- Question 1: What proportion of 10,000 30-year-olds will experience a first episode of MDD by the time they reach age 40 years?
- Answer: C
  - The cumulative incidence increases each year as the cases continue to accumulate, but the denominator for cumulative incidence – the initial population at risk – remains fixed. Each year there are fewer people remaining at risk
  - Calculation: IP= 1-EXP(-0.005*10) =0.049 or **4.9%**
  - Calculation: IR= 50 per 10,000 per year *10 years =0.05 or **5.0%**
  - The annual incidence rate in the above example is so small that we don't see the impact of the new cases on the size of the population at risk each successive year

33

UCSF

Q2: What if incidence rate were larger – 5 per 100, instead of 5 per 1,000?

A) IP=50 per 1,000

B) IP=50%

C) IP=40%

34

UCSF

UCSF

Answer

- 10 years of data from a large health insurance company that covers treatment for mental illness
- Assumptions: all cases of MDD result in a physician visit, correct diagnosis, and a correctly-coded insurance claim, so all of the numbers are accurate; the situation has been like this for many years, so that the incidence and population are stable

- Question 2: What if incidence rate were larger – 5 per 100, instead of 5 per 1,000?
- Answer: C
  - Calculation: IP= 1-EXP(-0.05*10) =0.39 or 39% ~ 40%
  - Calculation: IR= 500 per 10,000 per year *10 years =0.5 or **50%**
  - Note that we increased the annual incidence rate by a factor of 10 (from 5 per 1,000 to 5 per 100), but the cumulative incidence increased only by a factor of 8 (from 4.9% to 39%) – because the population at risk was being depleted.

35

UCSF

Issues in design of cohort studies

2. Selection of exposed population

If **exposure is common**, you may want to use a general cohort that will facilitate accurate and complete ascertainment of data

Examples: Nurses Health Study, Health Professionals Health Study, well-defined Communities (Araihazar arsenic study)

36

UCSF

UCSF

Examples of general cohort studies

Framingham Heart Study
A Project of the National Heart, Lung, and Blood Institute and Boston University

- **Exposures: smoking, hypertension, family history**
- **Outcomes: heart disease, stroke, gout, etc.**

37

UCSF

Issues in design of cohort studies

2. Selection of exposed population

Examples of exposed populations:

- Groups undergoing particular medical treatment
- Groups with unusual dietary or life style factors
- Professional groups (nurses, doctors)
- Students or alumni of colleges
- Geographically defined areas (e.g. Framingham)
- Occupational groups

38

UCSF

Issues in design of cohort studies

2. Selection of exposed population

For **rare exposures**, you need to assemble special cohorts (occupational groups, groups with unusual diets etc.)

Examples of exposed populations:

- Groups undergoing particular medical treatment
- Groups with unusual dietary or life style factors
- Professional groups (nurses, doctors)
- Students or alumni of colleges
- Geographically defined areas (e.g. Framingham)
- Occupational groups

Examples of special cohort studies:

- Rubber workers in Akron, Ohio
- Exposure: industrial solvent
- Outcomes: cancer

39

UCSF

Issues in design of cohort studies

2. Selection of exposed population

- **Enrichment for primary exposure:**
  - To improve cost-effectiveness
  - To improve statistical power
  - Special cases: matched cohort studies, web-based recruitment oversamples high-SES population

40

UCSF

Issues in design of cohort studies

2. Selection of exposed population

- Enrichment presents analytical challenges not commonly discussed in epidemiology
- Enrichment entails oversampling 1 or more levels of the primary exposure and is sometimes referred to as "oversampling" or "exposure-dependent sampling"
- Could be by design on unintentional
  - E.g., cohort study based on a registry of prenatal medication use alongside an unexposed general population
- Matched cohort studies are effectively exposure-enriched

UCSF

- primary exposure X (fetal growth restriction),
- secondary exposure Z (maternal smoking)
- Y (preterm birth)

$$Z \longrightarrow X \longrightarrow Y$$

Q1: Prevalence of X?
- 100/(100+900)=10%

Q2: Prevalence X/Z?
- 80/(80+420)=16% (X^+/Z^+) and 20/(20+480)=4% (X^+/Z^-)

Q3: RR (Z, X)?
- 16%/4%=4.0

Q4: Risk of Y?
- 130/1,000=13%

Q5: Risk of Y/X?
- 40/100=40% (Y^+/X^+) and 90/900=10% (Y^+/X^-)

Q6: RR (X, Y)?
- 40%/10%=4.0

Conclusion: strong causal effects Z→X, X→Y

American Journal of Epidemiology

DOI: 10

Practice of Epidemiology

A Cautionary Note About Estimating Effects of Secondary Exposures in Cohort Studies

K. A. Ahrens, S. R. Cole, D. Westreich, R. W. Platt, and E. F. Schisterman*

	Random Sample		
	$Y=0$	$Y=1$	Total
$X=1$			
$Z=1$	48	32	80
$Z=0$	12	8	20
Subtotal	60	40	100
$X=0$			
$Z=1$	378	42	420
$Z=0$	432	48	480
Subtotal	810	90	900
Total	870	130	

Q7: RR (Z, Y)?
- (32+42)/(80+420)=14.8% (Y^+/Z^+), (8+48)/(20+480)=11.2% (Y^+/Z^-)
- 14.8% / 11.2% = 1.33 – modest effect

Q8: RR (Z/X, Y)?
- 1.00 – RR for Z-Y association, adjusted for X

This is a biased estimate!

UCSF

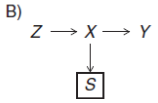
UCSF

- primary exposure X (fetal growth restriction),
- secondary exposure Z (maternal smoking)
- Y (preterm birth)

A)

$Z \longrightarrow X \longrightarrow Y$

B)



Q1: Prevalence of X? - 50%

Q2: Prevalence X/Z? - 63% (X^+/Z^+) and 27% (X^+/Z^-)

Q3: RR (Z, X)? - 63%/27%=2.3

Q4: Risk of Y? - 250/1,000=25%

Q5: Risk of Y/X? - 40% (Y^+/X^+) and 50/500=10% (Y^+/X^-)

Q6: RR (X, Y)?

- 40%/10%=4.0

Conclusion: strong causal effects Z→X, X→Y

Q7: RR (Z, Y)? – 28.9%, (Y^+/Z^+), 18.3% (Y^+/Z^-), 28.9% / 18.3% = 1.6 – modest effect – **This is a biased estimate!**

American Journal of Epidemiology

DOI: 10

Practice of Epidemiology

A Cautionary Note About Estimating Effects of Secondary Exposures in Cohort Studies

K. A. Ahrens, S. R. Cole, D. Westreich, R. W. Platt, and E. F. Schisterman*

	Random Sample A			Primary Exposure-Enriched Sample B		
	$Y=0$	$Y=1$	Total	$Y=0$	$Y=1$	Total
$X=1$						
$Z=1$	48	32	80	240	160	400
$Z=0$	12	8	20	60	40	100
Subtotal	60	40	100	300	200	500
$X=0$						
$Z=1$	378	42	420	210	23[b]	233
$Z=0$	432	48	480	240	27[c]	267
Subtotal	810	90	900	450	50	500
Total	870	130		750	250	

Q8: RR (Z/X, Y)?

- 1.00 – RR for Z-Y association, adjusted for X

This is a biased estimate!

43 Ahrens *AJE 2015*

UCSF

- primary exposure X (fetal growth restriction),
- secondary exposure Z (maternal smoking)
- Y (preterm birth)

American Journal of Epidemiology
© The Author 2015. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
DOI: 10

Practice of Epidemiology

A Cautionary Note About Estimating Effects of Secondary Exposures in Cohort Studies

K. A. Ahrens, S. R. Cole, D. Westreich, R. W. Platt, and E. F. Schisterman*

	Mean RR for the *Z*-*Y* Relationship	95% CI[b]	Mean SE[c]	Coverage, %[d]
Random sample				
Unadjusted (true effect) **True effect**	1.33	0.96, 1.84	0.167	95
X-adjusted	1.00	0.72, 1.40	0.169	61
Primary exposure-enriched sample				
Unadjusted **Biased estimates**	1.60	1.24, 2.05	0.128	70
X-adjusted	1.00	0.79, 1.28	0.123	39
Weighted[e]	1.33	0.90, 1.96	0.198	95

Abbreviations: CI, confidence interval; RR, risk ratio; SE, standard error.
[a] *X*, primary exposure (fetal growth restriction); *Y*, outcome (preterm birth); *Z*, secondary exposure (maternal smoking).
[b] 95% CIs were calculated using mean log RR ± (1.96 × mean log SE) and then exponentiated for tabular display.
[c] SE of the log RR.
[d] Coverage was based on sampling of 50,000 simulations, each of size 1,000.
[e] Robust variance estimate.

Q: What happened?

44 Ahrens *AJE 2015*

UCSF

UCSF

Q: What happened? What to do?

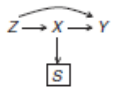


Figure 2. Causal diagram for the example cohort study representing the primary exposure-enriched sample with the addition of a direct effect of the secondary exposure on the outcome. *S*, selection into the study; *X*, primary exposure (fetal growth restriction); *Y*, outcome (preterm birth); *Z*, secondary exposure (maternal smoking).

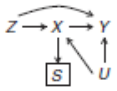


Figure 3. Causal diagram for the example cohort study representing the primary exposure-enriched sample with the additions of a direct effect of the secondary exposure on the outcome and a confounder of the primary exposure and the outcome. *S*, selection into the study; *X*, primary exposure (fetal growth restriction); *Y*, outcome (preterm birth); *Z*, secondary exposure (maternal smoking).

Answer:

We need to be careful with enriching cohort studies for exposure as it may have unintended consequences!

To correct this bias due to enrichment, use inverse probability (IP) weights to the 50% exposure enriched sample to create a reweighted sample with expected frequencies of maternal smoking, fetal growth restriction, and preterm birth identical to the random sample.

45 Ahrens *AJE 2015*

UCSF

Issues in design of cohort studies

*3. Selection of **un**exposed population*

Counterfactual ideal: The ideal comparison group consists of exactly the same individuals in the exposed group had they not been exposed. Since it is impossible for the same person to be exposed and unexposed simultaneously, epidemiologists much select different sets of people who are as similar as possible.

46

UCSF

UCSF

Issues in design of cohort studies

3. Selection of unexposed population

Practice: You want the comparison (unexposed) group to be as similar as possible to the exposed group with respect to all other factors except the exposure. If the exposure has no effect on disease occurrence, then the rate of disease in the exposed and comparison groups will be the same.

47

UCSF

Issues in design of cohort studies

3. Selection of unexposed population

Three possible sources of comparison group

1. Internal comparison: unexposed members of same cohort

- E.g.: Framingham study

48

UCSF

UCSF

Issues in design of cohort studies

3. Selection of unexposed population

Three possible sources of comparison group

2. Comparison cohort: a cohort who is not exposed from another similar population

- E.g.: patients from the specialty practice receiving similar treatment regimens
- Asbestos textile vs. cotton textile workers

UCSF

Issues in design of cohort studies

3. Selection of unexposed population

Three possible sources of comparison group

3. General population data: Use pre-existing data from the general population as the basis for comparison. General population is commonly used in occupational studies. Usually find *healthy worker effect*

- E.g.: A study of asbestos and lung cancer with U.S. male population as the comparison group

UCSF

UCSF

Issues in design of cohort studies

3. Selection of unexposed population

Which of the three comparison groups is best?

UCSF

Issues in design of cohort studies

4. Sources of exposure information

* Pre-existing records - inexpensive, data recorded before disease occurrence but level of detail may be inadequate. Also, records may be missing, usually don't contain information on confounders

UCSF

UCSF

Issues in design of cohort studies

4. Sources of exposure information

- Questionnaires, interviews: good for information not routinely recorded but have potential for recall bias
- Direct physical exams, tests, environmental monitoring may be needed to ascertain certain exposures.

53

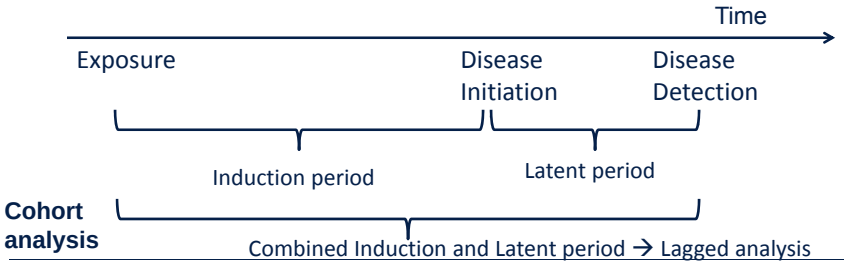
UCSF

Full cohort analysis

Issues in design of cohort studies

5. Classifying Person-Time

- Each unit of person-time contributed by an individual has its own exposure classification
- Must consider the etiologically relevant exposure
- Exposure may change over time (one-time vs. chronic)
- We want to exclude latent period from Person-Time, ***but can't***



54

UCSF

UCSF

Issues in design of cohort studies

5. Classifying Person-Time

- Time at which exposure occurs ≠ time at risk of exposure effects
- Only the time at risk for exposure effects should be counted in the denominator of the incidence rate for that level of exposure
- If the induction time is not known, can estimate empirically by calculating the incidence rates for differing categories of time since exposure

55

UCSF

Issues in design of cohort studies

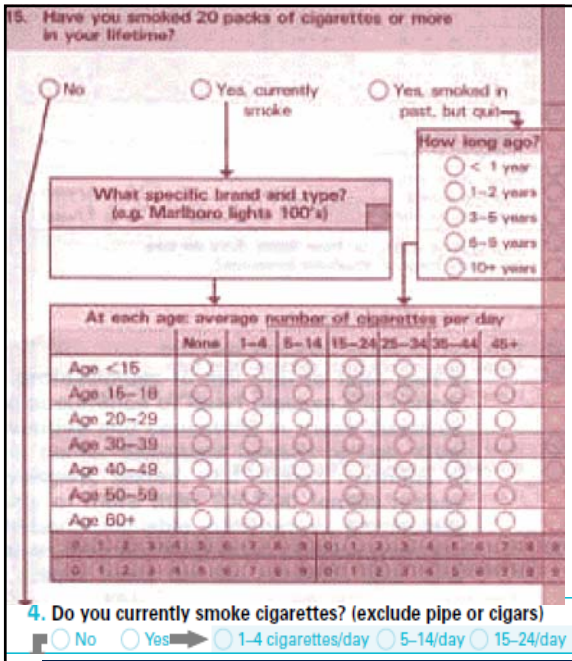
5. Classifying Person-Time

- **Exposures may change over time**
  - Measure exposure constantly and classify each unit of person-time
    - A given individual can contribute person-time to one or more exposure category in the same study
  - Assume one measure of exposure history is the only aspect of exposure associated with current risk
    - Current, average, cumulative, etc.

56

UCSF

UCSF



BASELINE QUESTIONNAIRE

Have you smoked 20 packs of cigarettes or more in your lifetime?

What specific brand and type?

If quit, how long ago?

At each age from <15 to 60+, what was the average number of cigarettes you smoked per day?

FOLLOW-UP QUESTIONNAIRES

Do you currently smoke cigarettes?

If so, how many per day?

UCSF

Issues in design of cohort studies

5. Classifying Person-Time

- Different exposure metric = different assumptions

Smoking

- Current v. Not
  - Assume former = never
- Never v. Former v. Current
  - Assume intensity & duration are not important
- Pack-years
  - Assume ½ pack for 10 y = 1 pack for 5 y = 5 packs for 1 y

UCSF

UCSF

Issues in design of cohort studies

5. Classifying Person-Time

- Pack-years is a composite of duration and intensity of smoking
- If the biologic effects of these differ, a study examining pack-years would not detect these differences, and could result in a misleading dose-response pattern
- Ideally, present results for pack-years and individual components of the primary exposure metric (e.g. smoking intensity and duration)

59

UCSF

Issues in design of cohort studies

5. Classifying Person-Time

Immortal Person-Time

- **Entry criteria in a cohort may depend on survival**
  - Study of pre-term birth and mortality in adulthood –
    - Individuals had to survive at least 1 year.
  - Study of occupationally exposed workers –
    - Workers had to be employed for at least 1 year.
- **First year of life = immortal person-time**
- **During this time, no one is at risk to become an event in the study**
- **Exclude immortal person-time from the denominator of incidence rates**

60

UCSF

UCSF

Issues in design of cohort studies

5. Classifying Person-Time

Lag Period (Induction Time)

- **Lag exposure to account for induction time between exposure and disease initiation**
  - i.e. using Sufficient-component-cause model: time that it takes for the causal mechanism to be completed by the action of the complementary component causes
- When lag period is not clear, need to hypothesize various induction times and reanalyze the data under each separate hypothesis. Also, can estimate the most appropriate time in the study by various statistical methods.

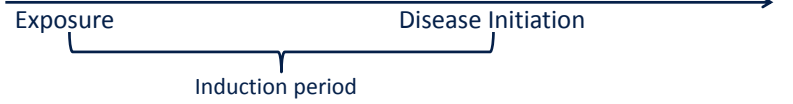
61

UCSF

Issues in design of cohort studies

5. Classifying Person-Time

- **Q: How do you classify person-time contributed by exposed subjects**
  a) before the minimum induction time has elapsed?
  b) after the maximum induction time has passed?
- **Example**:
  - Exposure = Rotavirus vaccine
  - Outcome = Intussusception
  - Assume induction period ranges from 1-7 days
  - Assume latent period in minimal and equal to zero days



62

UCSF

UCSF

Issues in design of cohort studies

5. Classifying Person-Time

Answer:

a) Consider the time unexposed

- BUT if the assumed induction time is incorrect, may misclassify exposed time as unexposed → underestimate exposure effects

b) Omit person-time of exposed subjects when they are not at risk for the exposure effects

- Must be a large number of unexposed cases

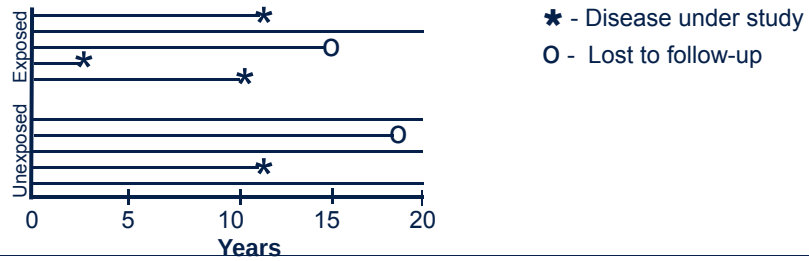
Also could model exposure effects as depending on time since exposure

63

UCSF

Exercise

- A prospective cohort study following 10 subjects for 20 years. All subjects were enrolled into the study at the same time at the beginning of the study. Assume no immortal time. The study investigates singular exposure and exposure status is determined at the time of enrollment into the study.

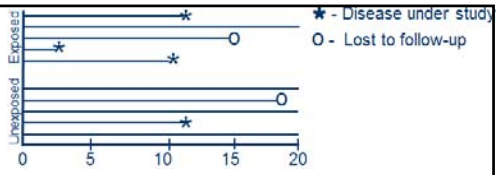


64

UCSF

UCSF

Q1: Calculate incidence rates - assuming no immortal time, no lag period



A) IR_{exp}=5.1 cases per 100 PY
IR_{unexp}=1.1 cases per 100 PY

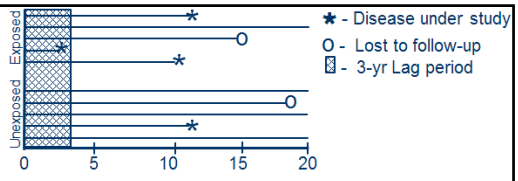
B) IR_{exp}=1.1 cases per 100 PY
IR_{unexp}=5.1 cases per 100 PY

C) IR_{exp}=3.1 cases per 100 PY
IR_{unexp}=1.1 cases per 100 PY

65

UCSF

Q2: Calculate incidence rates - assuming no immortal time and a 3-yr lag period



A) IR_{exp}=5.1 cases per 100 PY
IR_{unexp}=1.1 cases per 100 PY

B) IR_{exp}=4.4 cases per 100 PY
IR_{unexp}=1.9 cases per 100 PY

C) IR_{exp}=4.4 cases per 100 PY
IR_{unexp}=1.1 cases per 100 PY

66

UCSF

UCSF

Answers

- **Question 1:** Calculate incidence rates in exposed and unexposed.
- **Answer: A**

  IR_{exp}=3/59 PY=5.1 cases per 100 PY

  IR_{unexp}=1/89 PY=1.1 cases per 100 PY

- **Question 2:** Calculate incidence rates in exposed and unexposed with and without 3-yr induction period assuming a 3-yr lag (induction) period.
- **Answer : C or B**

  Taking into account 3-yr lag time:

  IR_{exp}=2/45 PY=4.4 cases per 100 PY

  a) *IR_{unexp}=1/89 PY=1.1 cases per 100 PY*
  b) *IR_{unexp}=(1+1)/(89+14) PY=2/103 PY=1.9 cases per 100 PY*

67

UCSF

Q3: If exposure does cause disease, what would you expect to happen with the rate in exposed people after taking into account an appropriate induction period?

A. IR_{exp} would increase.
B. IR_{exp} would decrease.
C. IR_{exp} would stay the same.

68

UCSF

UCSF

Answers

- **Question 3:** If exposure does cause disease, would you expect that the rate in exposed people would increase, decrease, or stay the same after taking into account an appropriate induction period?
- **Answer: A** - *Increase. The rate based on the inappropriate induction period is a blend of the rate related to exposure (using the appropriate induction period) and the rate unrelated to exposure. The resulting rate is between the high rate related to exposure (high because exposure causes disease) and the low rate not related to exposure. If the induction period is taken into account, the rate among exposed people should become higher.*

  Question 4: So, why did in our exercise the IR_{exp} decreased from 5.1 to 4.4 cases per 100 PY?

  Answer: The study is based on a very small number of cases!

69

UCSF

Recommended References

1. Background stratified Poisson regression analysis of cohort data. Richardson DB, Langholz B. Radiat Environ Biophys. 2012 Mar;51(1):15-22. 2.
2. Lagging exposure information in cumulative exposure-response analyses. Richardson DB, Cole SR, Chu H, Langholz B. Am J Epidemiol. 2011 Dec 15;174(12):1416-22.
3. Hierarchical latency models for dose-time-response associations. Richardson DB, MacLehose RF, Langholz B, Cole SR. Am J Epidemiol. 2011 Mar 15;173(6):695-702.

70

UCSF

UCSF

Issues in design of cohort studies

6. Sources of outcome information

- Death certificates
- Physician, hospital, health plan, pharmacy records
- Electronic medical records
- Questionnaires (could be verified by records) – "patient-reported outcomes"
- Medical exams

71

UCSF

Issues in design of cohort studies

6. Sources of outcome information

Goal is to obtain complete follow-up information on all subjects regardless of exposure status. You can use blinding (like an experimental study) to ensure that there is comparable ascertainment of the outcome in both groups.

72

UCSF

UCSF

Issues in design of cohort studies

6. Sources of outcome information

Timing of Outcome Events

- Events and the person-time being accumulated at the moment of the event are assigned to the same category of exposure
- Must clearly define the outcome event, including a protocol for determining the timing of the event
- Goal is to detect events at onset, however this may not be possible
  - Date of diagnosis for cancer
  - Hospitalization for end-stage renal disease

73

UCSF

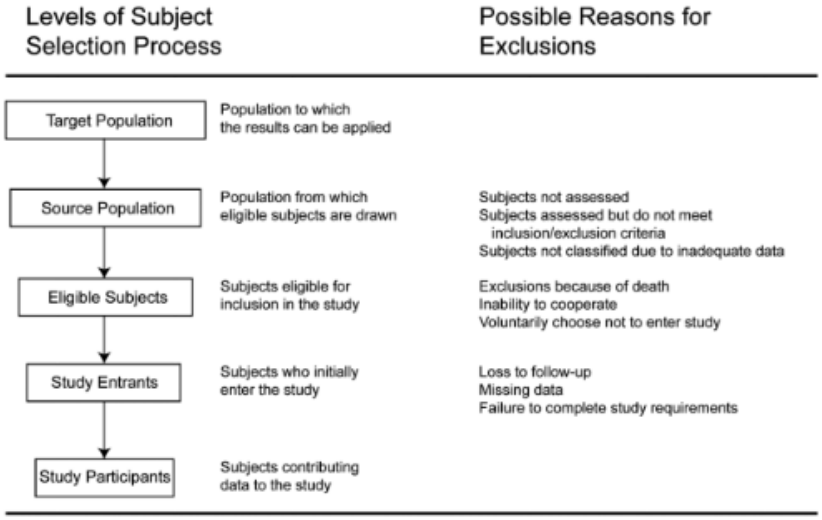
Issues in design of cohort studies

7. Approaches to follow-up

- In any cohort study, the ascertainment of outcome data involves tracing or following all subjects from exposure into the future
- Various **indices evaluating study quality**: tracing, recruitment, retention, compliance and loss to follow-up rates.

74

UCSF



Therefore, each cohort study should report **tracing rates, recruitment rates, retention rates, etc**. to give readers an idea of *generalizability* of study findings.

- Example: Determinants of young Adult Social well-being and Health (DASH) study
- culturally-appropriate retention strategies in an ethnically diverse cohort
- compared 45 different cohort retention strategies: 1) incentives, 2) reminder methods, repeat visits or repeat questionnaires, alternative modes of data collection or 3) other methods
- Incentives were associated with an increase in retention rates, which improved with greater incentive value. Whether cash was the most effective incentive was not clear from studies that compared cash and gifts of similar value. The average increase in retention rate was 12% for reminder letters, 5% for reminder calls and 12% for repeat questionnaires.

Issues in design of cohort studies

7. Approaches to follow-up

- Resources utilized to conduct follow-up: town lists, Polk directories, telephone books; birth, death, marriage records; driver's license lists, physician and hospital records; relatives, friends.
- New ways of keeping in touch with study participants

- This is a time consuming process but high losses to follow-up raise doubts about validity of study.

77

UCSF

Field methods (research methods, research administration methods)

- Use your research proposal (e.g., DCR protocol, NIH grant application) to prepare study documents:
  - Manual of Operations and Procedures (MOP): policies for conducting research tasks; *a priori* exposure and outcome definitions, recruitment strategies
    - If you do something without a pre-specified plan, this could introduce unintended consequences!
    - MOP and Clinical Intervention Study Protocol Templates: https://www.nia.nih.gov/research/dgcg/clinical-research-study-investigators-toolbox/startup
  - Administrative documents: policies for managing staff conducting research tasks, e.g., reporting of adverse events, quality control and re-training
  - IRB application, ethical guidelines for project staff

78

UCSF

UCSF

Special case: Using EMRs

- **Develop abstraction form and pre-test it**
  - Review a selection of typical and atypical charts
- **Write down policies for ambiguous situations, missing data**
  - Impute missing race/ethnicity by using U.S. Census Location and Surname Data
- **Prepare written training materials to train data abstractors**
  - Beware of possible biases at the beginning of the study and as study personnel changes
- **Keep written logs of decisions (e.g., how to interpret physician's narrative notes, specific symptoms and signs, etc.)**
- **Supplement with non-EMR sources (patient-reported data, admin/pharmacy/lab data)**

79

UCSF

Paper charts: To what extent do clinical notes reflect actual events?

Table 1. Agreement between actions which had actually been recorded and actions which could have been recorded by general practitioners in the clinical notes of simulated patients, by category and by priority.

	Mean content score (interquartile range) for the care of simulated patients				
	With tension headache (n = 28)	With acute diarrhoea (n = 27)	With shoulder pain (n = 24)	Having diabetes check up (n = 22)	Across all four cases (n = 18)
Category					
History	0.27 (0 to 0.42)	0.32 (0.14 to 0.50)	0.23 (0.04 to 0.33)	0.08 (0 to 0)	0.29 (0.18 to 0.40)
Physical examination	0.42 (0 to 1.0)	0.50 (0 to 1.0)	0.09 (0 to 0.20)	0.92 (0.92 to 1.0)	0.31 (0.20 to 0.42)
Laboratory examination[a]	0.33 (0 to 1.0)	1.0 (1.0 to 1.0)	–	0.75 (0.33 to 1.0)	0.64 (0.14 to 1.0)
Guidance and advice	0.18 (0 to 0.33)	0.23 (0 to 0.33)	0.30 (0.21 to 0.33)	0.04 (0 to 0)	0.22 (0.15 to 0.29)
Medication and therapy	0.72 (0.25 to 1.0)	0.88 (0 to 1.0)	0.90 (1.0 to 1.0)	0.07 (0 to 0)	0.68 (0.53 to 0.83)
Follow up	0.05 (0 to 0)	0	0.22 (0 to 1.0)	0.15 (0 to 0)	0.22 (0 to 0.47)
Total	0.30 (0.15 to 0.46)	0.36 (0.24 to 0.50)	0.25 (0.19 to 0.28)	0.26 (0.13 to 0.30)	0.32 (0.27 to 0.37)
Priority					
Essential action	0.28 (0.11 to 0.38)	0.36 (0.22 to 0.50)	0.22 (0.14 to 0.29)	0.25 (0.09 to 0.41)	0.30 (0.23 to 0.37)
Intermediate action	0.38 (0.15 to 0.67)	0.51 (0.33 to 0.75)	0.67 (0.50 to 1.0)	0.90 (0.75 to 1.0)	0.46 (0.37 to 0.55)

n = number of general practitioners. [a]Laboratory examination not used in shoulder pain category.

Of all actions undertaken, only 32% had been recorded

80 Rethans et al. *BJGP* 1994

UCSF

UCSF

EMRs: To what extent do clinical notes reflect actual events?

Table 2: Registration of lifestyle habits in the initial visit and the last year's notes (*n* = 209 medical records); values are presented as *n* (%)

Lifestyle habit	Initial visit	Last year's notes	Initial visit plus last year's notes
Smoking	164 (78)	81 (39)	189 (90)
Alcohol use	151 (72)	47 (22)	171 (82)
Physical activity	58 (28)	49 (23)	105 (50)
Eating habits	47 (22)	13 (6)	57 (27)

n, number of cases

Table 3: Unhealthy lifestyle habits according to the last year's records versus LSQ (*n* = 209 patients); values are presented as *n* (%)

Lifestyle habit	Unhealthy in last year's record notes	Unhealthy in LSQ	*P*[a], N = 209	X^2,[c] N = 209
Smoking	20 (31)	29	0.20	1.480
Alcohol use	7 (83)	42	< 0.001	26.725
Physical activity	27 (54)	59	< 0.001	14.069
Eating habits	3 (97)	147	< 0.001	212.629

P = significance value; [a] = significance values based on chi-square tests; [c] = degrees of freedom = 1
% = fraction underreported unhealthy habits in records compared to LSQ

Feedback, advice or referral was documented in 8% for smoking, 3% for alcohol use, 12% for physical activity, and 26% for eating habits.

81 Fouwels et al. *BMC* 2009

UCSF

New technologies

- **MediClass, an automated, rule-based classifier of the EMR that incorporates natural language processing**
- **Example**: Assess physician's advice in smoking cessation
- **Exposure**: Assess whether clinicians: (1) asked if the patient smoked; (2) advised them to stop; (3) assessed their readiness to quit; (4) assisted them in quitting by providing information or medications; and (5) arranged for appropriate follow-up care (i.e., the 5A's of smoking-cessation care)
- **Conclusion**: Chance-corrected agreement between the human raters and MediClass is very high (kappa=05-0.9)

82 Hazlehurst et al. 2005

UCSF

UCSF

Summary

- Cohort studies should be conducted in such a way that they approximate RCT study design
- Subject selection and exposure and outcome assessment play crucial role in getting valid results from cohort studies which would approximate the results from RCT
- Subject selection, recruitment, interviewing and retention have important effects on the study findings
- Field methods should be pre-specified ahead of time

83

UCSF