Epidemiological Study Designs

Dr. M.M. Mweu,

MBChB Level V Epidemiology,

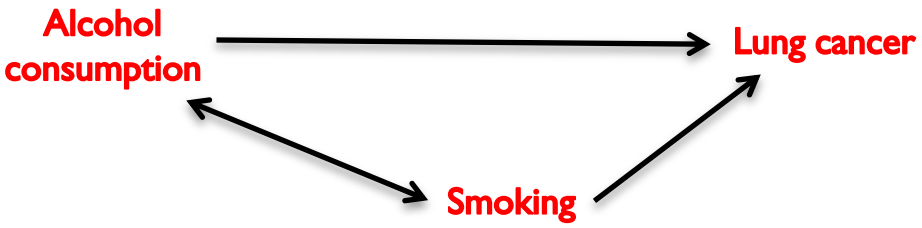
11 October, 2019

Introduction

- In epidemiological studies, we study:
  - The OUTCOME of interest
  - The EXPOSURE (S) or (Risk factors/Determinants) of interest, and
  - OTHER EXPOSURES that may influence the outcome (potential confounders)
- Primary exposure of interest is the one included in the hypothesis *e.g.* in the hypothesis – aflatoxin B_1 causes hepatocellular carcinoma – aflatoxin is the primary exposure of interest
- There might be more than one exposure e.g. in a study examining hypothesis that alcohol is a cause of lung cancer independent of smoking – both smoking and alcohol consumption should be measured

Introduction

- Smoking in this case is considered a “confounder”
- To be a confounder a factor must:
  - Be associated with the exposure of interest
  - Be independently associated with the outcome
  - Must not lie in the causal pathway between exposure & outcome
- *E.g.* if an association between alcohol & lung cancer were observed then the association may be partly/wholly due to the fact that people with high alcohol consumption are more likely to be smokers



Study types

Observational studies

- They collect info on events over which we have no control over – simply involved as observers
- Can be descriptive (outcome/exposure described without reference to the other) or analytic (exposure-outcome association is considered)

Study types

Intervention/Experimental studies

- The investigator deliberately allocates the exposure to individuals or communities (not harmful exposures) e.g. new therapeutic drug, vaccine etc
- The preferred form of intervention study is the *randomised controlled trial* in which the intervention (exposure) is randomly assigned at either group level (e.g. in community trials) or at individual level (clinical/field trial)
- All intervention studies are *analytical* since they study effect of exposures – outcomes are compared betwn exposure groups

Observational studies

a) Cross-sectional studies/surveys

- Collect info from each subject at *one point in time*
- Can be descriptive (describe frequency of outcome/exposure without reference to each other) or analytic (outcome measured in those with & without exposure of interest)
- Prevalence is the outcome of interest (proportion of pop with outcome/exposure of interest at a point in time e.g. prevalence of smoking)
- *Cases* are people with outcome of interest – those without outcome are *non-cases* (not controls)
- Are often conducted in a *sample* of the pop and the prevalence of exposure/outcome in sample is extrapolated to the rest of the pop
- Repeated cross-sectional studies (each with a different sample) over a period of time are useful for monitoring d'se trends or monitor interventions to see if they have impact on prevalence of d'se

Observational studies

a) Cross-sectional studies/surveys

Types of cross-sectional studies:

- **Descriptive** – info about outcome or exposure (but not both) is collected from individuals e.g. prevalence of cough in the pop
- **Analytical** – info about both outcome and exposure is measured from individuals at the *same time* (simultaneously) e.g. a study measuring current cough and risk factors such as indoor air pollution & current nutritional status (in order to test an association)
  - Prevalence of d'se is compared betwn the exposed group & the unexposed group by dividing prevalence of d'se in exposed by prevalence of d'se in unexposed – prevalence ratio (also called relative risk) e.g. prevalence of smoking in a population in 2005 was 33.1% (men) and 3.8% (women). *In this case smoking is outcome and gender is exposure:*

$$\textbf{Prevalence ratio} = p\left(\frac{D+}{F+}\right)/p\left(\frac{D+}{F-}\right) = \frac{\textit{Prevalence of smoking in men}}{\textit{Prevalence of smoking in women}} = \frac{33.1\%}{3.8\%} = 8.7$$

Men are 8 times more likely to be smokers than women in the population – reveals a strong association betwn gender and tobacco use

Observational studies

a) Cross-sectional studies

- Advantages:
  - Cheap and easy to conduct
  - Take only a short time

  } Used for planning purposes

- Disadvantages:
  - Reverse causality – since info on exposure & outcome collected simultaneously hence suitable for hypothesis generation
- Suitable for exposures that are time-invariant e.g. genetic factors such as gender, blood groups
- Favourable for estimating prevalence of common diseases of long duration
- Random sampling to ensure representativeness is key

Observational studies

a) Cross-sectional studies

- Non-response can be a problem – responders invariably have different characteristics from non-responders – introduce *selection bias* (convenience sample) in cross-sectional studies e.g. in a study of anaemia in 1000 women in which there was 75% response (750 responders & 250 non-responders), the prevalence of anaemia was $\frac{75}{750} = \mathbf{10.0}\%$. If all non-responders were anaemic, the total prevalence would be $\frac{75+250}{1000} = \mathbf{32.5}\%$, if all non-anaemic, prevalence would be $\frac{75}{1000} = \mathbf{7.5}\%$
- If response rate is low < 80% non-responders should be followed up with reminders – prodding is key!

Observational studies

b) Case control studies

- Starting point is definition of a group of people with a particular disease or condition (**cases**)
- Suitable **controls** are then selected – without disease and *representing the population from which cases originated* (often random sample of the healthy pop)
- Cases are hence individuals in pop with d'se of interest & controls are a *representative* sample of individuals without d'se from same pop (*base* pop)
- Frequency of exposure amongst cases ($p\frac{F^+}{D^+}$) & controls ($p\frac{F^+}{D^-}$) are compared *e.g.* in a study of deaths from respiratory infection, cases could be children who died from pneumonia controls might be healthy children of same age
- If $p\frac{F^+}{D^+} > p\frac{F^+}{D^-}$ then exposure is a likely risk factor for d'se – otherwise its protective

Observational studies

b) Case control studies

- Many different exposures can be studied
- Useful for studying *rare* outcomes and diseases of *long latency*
- Cases should have a specific case definition e.g. based on histopathological results (cancers) or clinical patterns (measles)
- Can be *population-based* (all cases arising in a population in a defined period) or *hospital-based* (all cases fulfilling case definition and are attending one or more specific hospitals)
  - Hospital-based cases may not be representative of all cases fulfilling case definition – hospital attenders tend to differ from non-hospital attenders

Observational studies

b) Case control studies

- Incident cases are more preferable than prevalent cases:
  - Associations identified in prevalent cases may not only be due to factors related to *developing* the d'se but also *survival* with the d'se
  - Prevalent cases may also have changed exposure status because of d'se – *reverse causality*
- Controls should fulfil criteria defining cases apart from d'se itself

  e.g. if cases are females aged 14-44 yrs with rheumatoid arthritis then controls are also females aged 14-44 yrs *without* rheumatoid arthritis
- Source of controls depends on source of cases:
  - If cases are a *population-based* random sample of all incident cases, controls should be random sample of persons without disease from *same pop*
  - If cases are *hospital-based* then controls could be patients in the hospital having *other diseases* (since base pop that cases arose is unclear). However, controls should not have d'ses that are related to exposure of interest. E.g. if lung cancer is d'se of interest, controls should not be selected amongst patients with chronic bronchitis – as smoking status is likely to be *overrepresented* among controls
- Case control studies can be prospective or retrospective

Observational studies

b) Case control studies

- Several controls could be *matched* for each case
  - Matching refers to procedure whereby *one* or *more* controls are selected for each case on basis of *similarity for certain characteristics* other than the factor under investigation e.g. age & sex
  - Matching may be *individual* or *frequency* based
    - Individual based matching – cases are individually matched to one or more controls – matched analysis is necessary
    - Frequency based matching – distribution of the level of the matching variable e.g. age is same in cases and controls
  - Matching increases efficiency of a study
- No. of cases available in some studies may be limited but this may not apply to controls – necessary to therefore increase no. of controls per case in order to improve statistical power of the study (often 4 controls per case is considered the max)

Observational studies

b) Case control studies

- Data on exposure status may be gathered through:
  - Interviews – can be personal, postal or telephone
  - Medical or occupational records
  - Use of biological samples
- Minimisation of bias (information bias) when collecting exposure info is key:
  - *Recall bias* – recall of past exposures is differential betwn cases & controls – being a case increases likelihood of recall
  - *Observer bias* – investigator may gather info differently depending on whether one is case or control
- The measure of association in case-control studies is OR.
- *Odds of exposure* in *cases* $(\frac{a}{c})$ is compared to *odds of exposure* in *controls* $(\frac{b}{d})$

Observational studies

b) Case control studies

	Cases	Controls	
Exposed	a	b	a+b
Unexposed	c	d	c+d
	a+c	b+d	a+b+c+d

$$OR = \frac{a}{c} \div \frac{b}{d} = \frac{ad}{bc}$$

Which actually becomes odds ratio of d'se in those with and without d'se

Observational studies

b) Case control studies

- Advantages:
  - Relatively cheap & quick
  - Useful for rare diseases and diseases of long latency
  - Can study multiple risk factors
  - Can test hypotheses
- Disadvantages:
  - Prone to selection & information bias
  - Temporal sequence might be absent especially in retrospective studies
  - Unsuitable for studying rare exposures
  - Can not obtain estimate of incidence

Observational studies

c) Cohort studies

- Starting point is definition of a group of people (*initially free of d'se of interest*) by their *exposure* status e.g. group of smokers (**exposed**) & non-smokers (**unexposed**)
- Exposures may not be *all-or-nothing* – pop could be divided by degrees of exposure e.g. no. of packs of cigarettes smoked per year, breastfeeding by duration, dose of exposure, age at exposure etc
- Since selection of study groups is by exposure – allows the study of rare exposures e.g. asbestos which is rare in general pop in which case an industrial cohort would be selected (more people likely exposed hence high statistical power)
- In cohort studies one hopes to mimic intervention study where individuals are randomly allocated to exposure hence 2 groups similar with respect to all factors other than exposure itself – hence exposed & unexposed groups in cohort studies should be similar in *all respects except exposure status*

Observational studies

c) Cohort studies

- *E.g.* In industrial cohorts exposed group could be factory workers exposed to chemical whereas unexposed group could be workers in same factory working in different site (*internal comparison group*) or people from other factories near where factory is located (*external comparison group*)

- Comparison group should be *truly* unexposed – often a problem since people in comparison group may have previously moved from jobs where they had been already exposed or may disguise exposure
- Once classified they are *followed* up over time until they develop the outcome of interest (**prospective** study) e.g. lung cancer
- However, studies can also be **retrospective** (historical) [where d'se has already developed] – rely on records of past exposure in individuals
- Historical cohorts are ideal for d'ses of long latency betwn exposure & d'se
- Historical cohorts have disadvantage in that info on exposure & d'se may be incomplete and there may be limited/no info on confounders collected
- Assessment of outcome should be at regular intervals e.g. through physical exam or questionnaires

Observational studies

c) Cohort studies

- Risk of disease in exposed ($P\frac{D^+}{F^+}$) is compared to risk of disease in unexposed group ($P\frac{D^+}{F^-}$) – to get measure of association (Risk/rate/odds ratio)
- Cohort studies are generally suitable for diseases of *short duration* between exposure & outcome and when outcome is *common*
- Advantages:
  - Exposure is measured temporally before disease onset – no reverse causality
  - Rare exposures can be examined e.g. exposure to vinyl chloride
  - Multiple outcomes and multiple exposures can be studied
  - Useful for testing hypotheses

Observational studies

c) Cohort studies

- Disadvantages:
  - Losses to follow-up are common especially for long follow-ups
  - High monetary cost
  - Time-consuming

Intervention/Experimental studies (Trials)

- Most important limitation of observational studies is that an observed association betwn an exposure & outcome may be due to differences betwn exposed & unexposed groups with respect to *other risk factors* for outcome - *confounders*
- Even after removal of effects of confounding either in design or analysis stage in observational studies, *residual confounding* may persist
- Also there may still be effects of *unknown confounders* – other risk factors for the outcome yet to be identified
- In intervention study the investigator determines which individuals are exposed to factor of interest & which are unexposed i.e. *allocates exposure*
- *Randomisation* is used to ensure that exposed & unexposed groups are similar with respect to *all other factors* (known & unknown)

Intervention studies (Trials)

- They play a central role in *evidence-based medicine*
- Objectives of intervention studies:
  - To test a specific causal hypothesis concerning a d'se e.g. In a Gambia Hepatitis Intervention study, the objective was to test the hypothesis whether hepatitis B virus infection causes liver cancer – some given hepatitis B vaccine and some placebo and followed over time where liver cancer rates are then compared
  - To measure effect of a particular intervention. E.g. effect of a drug/other treatment on established d'se (clinical trial) or a preventive intervention (field trial) e.g. vaccine/health promotion programme
- When conducting a trial it's important to consider the:
  - *Reference pop* – pop which the results of the study are intended to apply e.g. all patients with a specific d'se
  - Study pop – pop in which trial is actually conducted. Often more limited geographically than reference population. Need to consider how representative the study pop is of the reference pop (generalisability). Often a stable & co-operative pop (less migrative & likely to be compliant) is chosen to ensure high coverage and follow-up rates
- Simplest design is *two-arm trial* – subjects recruited from study pop are allocated either to intervention arm (receiving treatment) or to control arm (receiving either nothing, placebo or current/existing treatment) through randomisation

Intervention studies (Trials)

- To preserve full benefits of randomisation and avoid bias *concealment of the allocation* from subject or investigator is done – called *blinding*
- If treatment group to which subject is allocated is known in advance this may affect whether the individual is recruited leading to bias.
- Systematic allocation e.g. based on odd/even nos. is unsatisfactory as allocation concealment is difficult to ensure
- Collection of baseline data on recruited subjects e.g. age, sex, occupation, education – done to check whether comparability between treatment arms has been achieved. Adjustments for non-comparability may be done in the analysis stage

Intervention studies (Trials)

- Losses to follow-up (withdrawals) are important sources of bias since those lost from study may be different from those seen
- Careful choice of study pop & revisits when subjects are not at home for household visits – should ensure high follow-up rates
- Bias may result if subjects/evaluators are aware of treatment allocation of subjects since reporting/recording of outcomes may be influenced by knowledge of what intervention has been received – hence blinding is done

Intervention studies (Trials)

- Single, double or triple-blinding can be done involving:
  - The subject participating in the trial e.g. subjects receiving vaccine might be less careful about taking other preventive measures increasing their proneness to d'se
  - Recruiters and carers of the subjects e.g. clinician – might be tempted to provide alternative treatment for patients in control arm
  - Evaluators e.g. interviewers, lab personnel – may bias outcome measurements
- Giving placebo to control group is a means of blinding
- Rate/risk of d'se (outcome) is compared in the different treatment arms
- In field & clinical trials, vaccine/treatment efficacy is calculated as:

$$VE/TE = \frac{R(control) - R(intervention)}{R(control)}$$

Intervention studies (Trials)

- In interpreting the results of the trial key sources of bias should be considered - questions to be asked:
  - Was randomisation effective? Were treatment arms similar at baseline?
  - To what extent was blinding achieved? Is their likely to be bias in reporting/recording of outcomes?
  - What proportion of participants were successfully followed up?
  - Did participants comply with the intervention? How many participants defaulted/changed to a different treatment?