

Sources of Data for Epidemiological Studies:

Introduction:

Development in technology has led to considerable increase in the number of individual based data sources, registers, data bases and information systems that may be of value in epidemiology study.

• Sources of data for epidemiological studies:

• Directly obtaining information from surveys, laboratory measure, clinical assessments from study participant are examples of primary data collection.

• The national vital statistic system is the oldest system of data collected according to uniform standards and procedures.

• The registration of vital events, births, marriage divorces and fetal deaths is authorized by the legal authority of the state, city or territory.

• The national center for the health statistics (Nchs) is another major source of us population based data.

• The NCHS data are collected through personal

interview or experimentation or by using system containing data collected from vital and medical records.

• Some NCHS data systems and surveys are ongoing annual system while other are conducted periodically.

• Also, private organisations and firms may collect and analyze health related information.

• Data from some health maintenance organizations are very useful for PE studies.

Example

1. Kaiser permanent in northern and southern California each include more than 1 million patients.

2. The organ Kaiser plan include between 300,000 to 4,00,000 patients.

Each of these programs has centralized computerized hospital discharge diagnosis information which can be used to identify cases.

3. Saskatchewan health has billing data for the

1 million residents.

4. Finally, an alternative source of computerized billing data is Medicaid data.

Medicaid data can be particularly useful when very large sample size are needed or when outpatient diagnoses are needed.

Computerized collection of medical billing data:

Advantages:

Size

cost

complete

Can be population based

Can include outpatient drugs and disease

No recall or interviewer bias

disadvantages:

- Cannot be used to performed RCT

- Incomplete information on potential confounders

- Variable data elements
- Outcome assessment usually cannot be quantitative but simply present or absent.
- Uncertain data validity, especially for diagnosis

Computerized collection of medical billing data :

program	advantages	disadvantages
Medicaid (eg : compass R)	Size Ease of use Outpatient dxs	Skewed population No inpatient drugs Eligibility changes
Saskatchewan health	Least skewed population Prolonged history	Size Limited formulary Little outpatient data Dxs affect reimburse
Kaiser permanent	Less skewed populations	Limited drug data computerized No outpatient dxs Limited formulary
Group health	Less skewed population Easiest access to medical records	Smallest No outpatient dxs Limited formulary

- Computerized collection of medical billing data: uniquely useful situation
- The study of uncommon exposure and/or uncommon outcomes
- When recall bias is a problem

• When interviewer bias is a problem

• Denominator needed to calculate incidence rate

• Short term drug effect, especially when they require specific drugs or surgical therapy.

• Objective, laboratory driven diagnosis

• limited time

• limited budget

Computerized collection of medical billing data: problematic situation

• illnesses which do not reliably come to medical attention

• Inpatient drug exposures

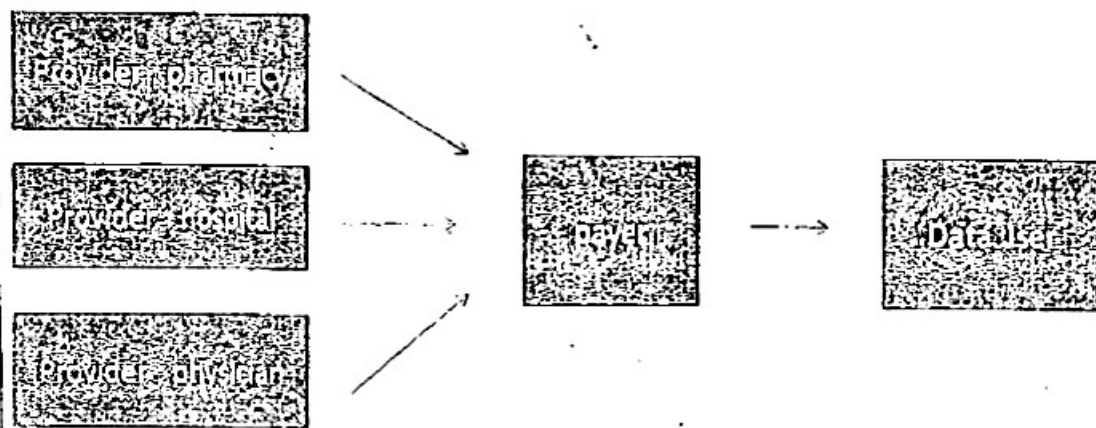
• Descriptive study

• Delayed drug effect

• Important unknown confounders

Claim databases

• Claim data arise from a person's use of the health care system.



When a patient goes to a pharmacy and gets a drug dispensed, the pharmacy bills the insurance carrier for the cost of that drug and has to identify which medication was dispensed, the milligram per tablet, No of tablets etc.

- If a patient goes to a hospital or to a physician for medical care, the providers of care bill the insurance carrier for the cost of the medical care, and has to justify the bill with the diagnosis.

- If there is a common patient identification number for both the pharmacy and the medical care claim, these elements could be linked and analyzed as a longitudinal medical record.

• Examples of automated databases:

- Group health cooperative
- Kaiser permanent medical care program
- HMO research network
- United health group
- Medicaid databases

Computerized databases:

• Strengths:

1. Computerized databases provide large sample size.
- This is especially important in the field of pharmaco-epidemiology, where achieving an adequate sample size is uniquely problematic
2. These databases are relatively inexpensive to use.
3. Complete data available for pharmacoepidemiological study
4. These databases can be population-based, they can

include outpatient drugs and diseases and there is no opportunity for recall and interviewer bias.

Weakness:

1. These database is the uncertain validity of diagnosis data (claim database) and outpatient data.

2. They lack information on some potential confounding variables e.g. in claim databases there are no data on smoking, alcohol, data of menopause etc.

3. Computerized databases also do not include information on medications obtained without a prescription or outside of the particular insurance carrier's prescription or outside of the particular insurance carrier's prescription plan. (Dietary supplement, NSAIDs etc)

4. Other disadvantages of claim based-data is the instability of the population due to job changes, employers changes of health plans, and changes

in coverage for specific employers and their family members.

5. Some results from studies that utilize these databases may not be generalized eg: on health care utilization.

6. Only those illness severe enough to overcome medical attention will be in databases.

Ad hoc data sources:

Introduction:

The term Ad Hoc refers to the billion of bytes of non-standard and continuously evolving data spread across all computer systems.

- Such data includes server logs, distributed system performance, telephone records, financial data.

- An Ad Hoc data source is any semistructured data source for which useful data analysis and transformation tools are not readily available.

- The data that constitutes a single, abstract source often comes from many different concrete, physical destinations distributed across the internet.

- Managing Ad Hoc data is a bane of the implementers of distributed systems. These systems may have hundreds or thousands of heterogeneous, distributed components; keeping these components running smoothly is a continuous maintenance task of significant complexity.

- Consequently, each component is a well-designed

Distributed system produces a continuous stream of log files that measure its performance and health.

- Example distributed Ad Hoc data sources



CoBlitz

File transfer system monitoring

At&T Arrakis

Website host monitoring

NCBI

Biotechnology info

Ad Hoc Case-control Studies

- Case-control studies can be performed as Ad Hoc studies as well. Cases can then be recruited from whatever source is appropriate for that disease, whether hospitals, outpatient practices, or some

other source Investigator in some geographic areas maintain ongoing relationships with a number of hospitals, to permit case finding for case-control studies.

- Case definitions must be clear. Cases should be "incident cases", that is, individuals who have recently developed the disease, so one can inquire about exposures that precede the onset of the disease. All individuals who meet the case definition should be enrolled, if possible, to decrease the risk of a selection bias. Finally, if all cases can be identified in a defined population, then the incidence of the disease in the population can be determined.

- Controls can then be recruited from either the site of medical care for the cases or from the community where the cases come from.

Community controls can be recruited from friends of the cases, from neighbors of the cases, from a reverse telephone directory, by using random digit dialing, or from some comprehensive listing of the target population.

- A list, organized medical care programs which can provide listings of eligible individuals, such as general practitioner's and other special situations.
- An example would be a study limited to the elderly, which could obtain listing of those eligible for Medicare in the local area.
- Ad Hoc case-control studies can be very useful for studying uncommon conditions, but the choice of the control group needs to be made carefully, preferably choosing population-based cases and controls.
- Ad Hoc cohort studies can be useful to study new drugs, but are expensive and large, and need to have a concurrent control group to avoid attributing all adverse events to the drug.
- Ad Hoc studies can be whatever size the researcher desires for the study at hand.
- Historically, Ad Hoc epidemiologic studies have been employed to assess potential vaccine adverse events. Newer projects based on automated

SOURCES OF DATA IN PHARMACOEPIDEMIOLOGICAL STUDIES

- An ideal databases would include records from inpatient and outpatient care, emergency care, all lab and radiological tests, pharmacological and non pharmacological medications.
- The covered population should be large enough to detect rare events and should be stable from birth to death; updated and have information on potential confounders variables such as smoking stature and alcohol consumption.
- Despite of availability of databases, their use as a research tool maybe a complex task. The implementation of studies through the database require a clear and comprehensive understanding of the strengths and weakness of PE methods, how data were collected, enrollment and coverage factors and other factors that might have affected data quality and validity also knowledge of social, political, cultural, historical settings is also crucial. Failure to understand weaknesses, limitations of databases can lead to selection of inappropriate study methods and hinder the interpretation of results.

SOURCES OF PE DATA

- Spontaneous AE reporting
- Global drug surveillance
- Case control surveillance
- Prescription event monitoring
- Automated data bases

SPONTANEOUS AE REPORTING

- They are the most common method used in PV to generate signals On new or rare AE not discovered during clinical trials.
- They are very useful in hypothesis generation with need to explore possible explanation for AE in question.
- Spontaneous report is a clinical observation that originates outside of a formal study

In order to ensure that safe and effective pharmaceuticals are available, FDA relies on:-

- Voluntary reporting
- Mandatory reporting

الزمن، الجاد

- Individual spontaneous report of ADRs, medication error and product quality problems are sent to FDA directly or indirectly combined with clinical data and data from medical and scientific literature; comprise the primary data source upon which post marketing surveillance depend.

→ After confirmation of a signal FDA can initiate various regulatory actions like :-

- Labelling change such as a boxed warning
- Restricted use of drug
- Naming or packaging changes
- Health care professional letter
- Withdrawal of medical product from market

GLOBAL DRUG SURVEILLANCE

- Global reporting of concerns about suspected ADRs is a vital alerting tool
 - The WHO collaborating centre for international drug monitoring centre in Uppsala.
 - Today - 73 countries full of official members
12 countries associated members
 - National centres should report at a minimum monthly frequency
- Recently there has been an international effort to :-

- ✓ Harmonize the terms
- ✓ Set criteria and definitions
- To harmonize the way data are stored and communicating internationally.

Main agencies involved in are WHO, CIOMS, ICH and EU

The medical dictionary for regulatory activities is been used more and more world wide.

Case - control surveillance

- Like prescription drugs, non prescription drugs can also have serious AEs and unintended benefits.

Hence in case control studies multiple case control studies are conducted simultaneously in order to monitor the effects of prescription and OTC medications.

- CCS depends on self reports of medication and dietary supplement use.
- It asks about 43 indications or medication categories.

STRENGTHS :-

- CCS has the capacity to monitor :
 - Prescription drugs
 - Potentially important confounders
 - OTC drugs
 - Dietary supplements
- It can have high statistical power because of large number of cases accrued.
- It has the capacity to discover unsuspected associations

Example - inverse association between aspirin and risk of colorectal cancer was documented in CCS.

- Assessment of effects after long intervals or duration of use.

Example - AE of oestrogen supplements on risk of endometrial cancer persisted for 15-19 yrs after cessation of use.

- Allows for assessment of whether genetic polymorphism modify the effect of a medication or supplement on the risk of the illness.

LIMITATIONS

- Potential for selection bias and recall bias

PRESCRIPTION EVENT MONITORING

It is a study in which a cohort of users of a medicine is defined from prescriptions and followed-up for a defined period so as to identify all adverse events occurring in the early post treatment period.

It is one of the form of pharmacovigilance and is complementary to spontaneous reporting of suspected ADRs

It is a Method of both hypothesis generation and testing.

AUTOMATED DATABASES

There is a growing use of computerized databases containing medical care data called automated databases which are potential data sources for PE studies.

The different types are :-

- **CLAIMED DATABASES** - it arise from persons use of health care system and submission of claims to insurance companies for payment.
- **MEDICAL RECORD DATABASES** - it arise from increasing use of computerization in medical care.

ADVANTAGES

- Provide very large sample size, so common outcomes can be detected.
- Provides denominator needed to calculate incidence rate.
- Relatively inexpensive to use.
- No opportunity for recall or interviewer bias
- When time is limited
- When budget is limited

example of automated database :

- group health cooperative
- kaiser permanent medical care program
- HMO research network
- united health group
- medicaid database
- health services database in Saskatchewan
- automated pharmacy record linkage in Netherlands
- tayside medicine monitoring unit
- UK general practice research database

advantage :

- computerized database provide large sample size
- this is especially important in the field of PE where achieving an adequate sample size is problematic.
- these database are relatively inexpensive to use .
- complete data available for PE study
- these databases can be population based ,they can include outpatient drugs and disease and there is no opportunity for recall and interviewer bias .

disadvantage :

- these database is the uncertain validity of diagnosis data (claim data base and outpatient data)
- they lack information on some potential confounding variables ex: in claim databases there are so data on smoking, alcohol, data of menopause etc.
- computerized databases also do not include information on medication obtained without a prescription or outside of particular insurance carrier prescription or outside of the particular insurance carrier's prescription plan (dietary supplement, NSAIDS etc)
- other disadvantage of claim based data is the instability of the population, due to job changes, employers changes of health plans, and changes in coverage for specific employers and their family members

- some results from studies that utilize these database may not be generalized ex: on health care utilization
- only those illness severe enough to overcome to medical attention will be in data base

AD HOC DATA SOURCES

- the term ad hoc refers to the billion of bytes of non standard and continuously evolving data spread across all computer system.
- such data includes server logs, disturbed system performance, telephone records, financial data.
- an ad hoc data source is any semistructured data source for which useful data analysis and transformation tools are not really available.
- the data that constitutes a single, abstract source often comes from many different concrete, physical destinations distributed across the internet.
- Managing ad hoc data is a bane of the implementers of distributed systems. these may have hundreds or thousand of heterogenous, distributed components, keeping these components running smoothly is continouse maintenance task of significant complexity. Consequently, each component is a well designed distributed system produces a continuous stream of log files that measures its performance and health.
- Ad hoc analysis is a business intelligence process designed to answer a single, specific question. The product of ad hoc analysis is typically a statistical model, analytical report or other type of data summary.
- AD HOC means for the particular case at hand without consideration of wider applications. The purpose of this analysis is to fill in gaps left by the business static, regular reporting.
- Ad doc data analysis may be used to create a report that does not already exist or drill deeper into a static report to get details about accounts, transactions or records. The process may be also used to get more current data for the existing areas covered by a static report.

ADVANTAGES

- Cost
- speed

DISADVANTAGES

- Incomplete data
- Lack of quality control