

APPROACH TO BIG DATA : THE GREY AREAS

Dr.Rajamohanam

School of Health Policy and Planning Studies-KUHS

APPROACH TO DATA ANALYSIS : BASICS

- Practice of epidemiology has three core elements: Design, conduct and analysis
- Epidemiological study is a measurement exercise and we get data from measurement . We use this data for estimation and inference which is called data driven conclusions.
- Overall goal of estimation is to get an accurate estimate based on data. ie. valid and precise estimate.
- Goal of comparison based on data is valid comparison for getting meaningful conclusions.
- Another important concept in epidemiological study is generalizability or how the results are applicable to the target population. This is taking the conclusions beyond the data set
- Think of all uses of Epidemiology and how approach of analysis will be different in different settings.

USING SECONDARY DATA SETS

- This is totally different concept from Big-data analysis.
- There is immense scope but limitations also.
- Understanding somebody else's data base structure is really difficult.
- Used by mostly armchair researchers.

WHAT IS BIG DATA?

“Big data” refers to any voluminous amount of structured, semi-structured, and unstructured data that has the potential to be mined for insights and information. (Ref: The National Academies Press, National Academy of Sciences, Washington, DC- 2016)

SPECIFIC ISSUES RELATED TO BIG DATA

- Access
- Interoperability
- Analysis
- Quality
- Validation
- Storage
- Privacy
- Security
- Liability

QUESTIONS FOR DEBATE: BASICS OF DATA ANALYSIS

- Why Epidemiological efficiency is not same as economic efficiency in relation to data analysis?
- During data analysis is there too much methodological complexity for gaining internal validity? Is simplification possible?
- How far data base incompleteness can be tolerated?

QUESTIONS FOR DEBATE: GOOD EPIDEMIOLOGICAL PRACTICE IN RELATION TO BIG DATA

- How far Personal attributes of epidemiologist matters in big data analysis: integrity, honesty, fairness(transparency)
- What are the specific ethical issues: Dilemmas in data sharing (Sharing data in responsible manner)
- How to be careful at each stage: Planning and protocol development(ill-designed), Signing MOU, getting commissioned(poorly launched), Conduct or team sub team development(badly implemented), analysis(inappropriately analyzed), Research communication(selectively reported) :Lead to irreproducible or inconsistent and invalid results.
- What is the basic debate?(Accountability of researcher to society or to the specific stakeholder(Individual, Funder, Employer or Govt. or study subjects). Is there a genuine conflict?

QUESTIONS FOR DEBATE: GOOD EPIDEMIOLOGICAL PRACTICE IN RELATION TO BIG DATA

- What are the questionable research practices?(in relation to data mining, data dredging, imputations)
- How far ignorance can be considered as a reason for labeling one as questionable research practice?
- What you think are the specific components of scientific rigor in Data analysis?
- To what extent contextual factors or urgency to get quick and reasonable results for policy advocacy justify the analysis?(This may be OK, in market research or climate research where volume of data is high, but for health sciences, can this be OK)

SPECIFIC APPLICATIONS

- Predicting disease outbreak(Surveillance)
- Digital disease detection (Forecasting)
- Risk mapping and clustering detection(Descriptive epidemiology)
- As a tool in artificial intelligence applications
- Molecular epidemiology
- Pathogen Phylo-dynamics
- Pharmaco-vigilance in one-health concept (Anti microbial resistance)
- Population observatory as an example of big data source

THANK YOU