Universidad del Cauca

ERASMUS UNIVERSITEIT ROTTERDAM

GENERAL AIMS OF DATA ANALYSIS

Jose Andres Calvache MD MSc PhD

Department of Anesthesiology, Universidad del Cauca, Colombia

Department of Anesthesiology, Easmus University Medical Centre, Rotterdam, The Netherlands

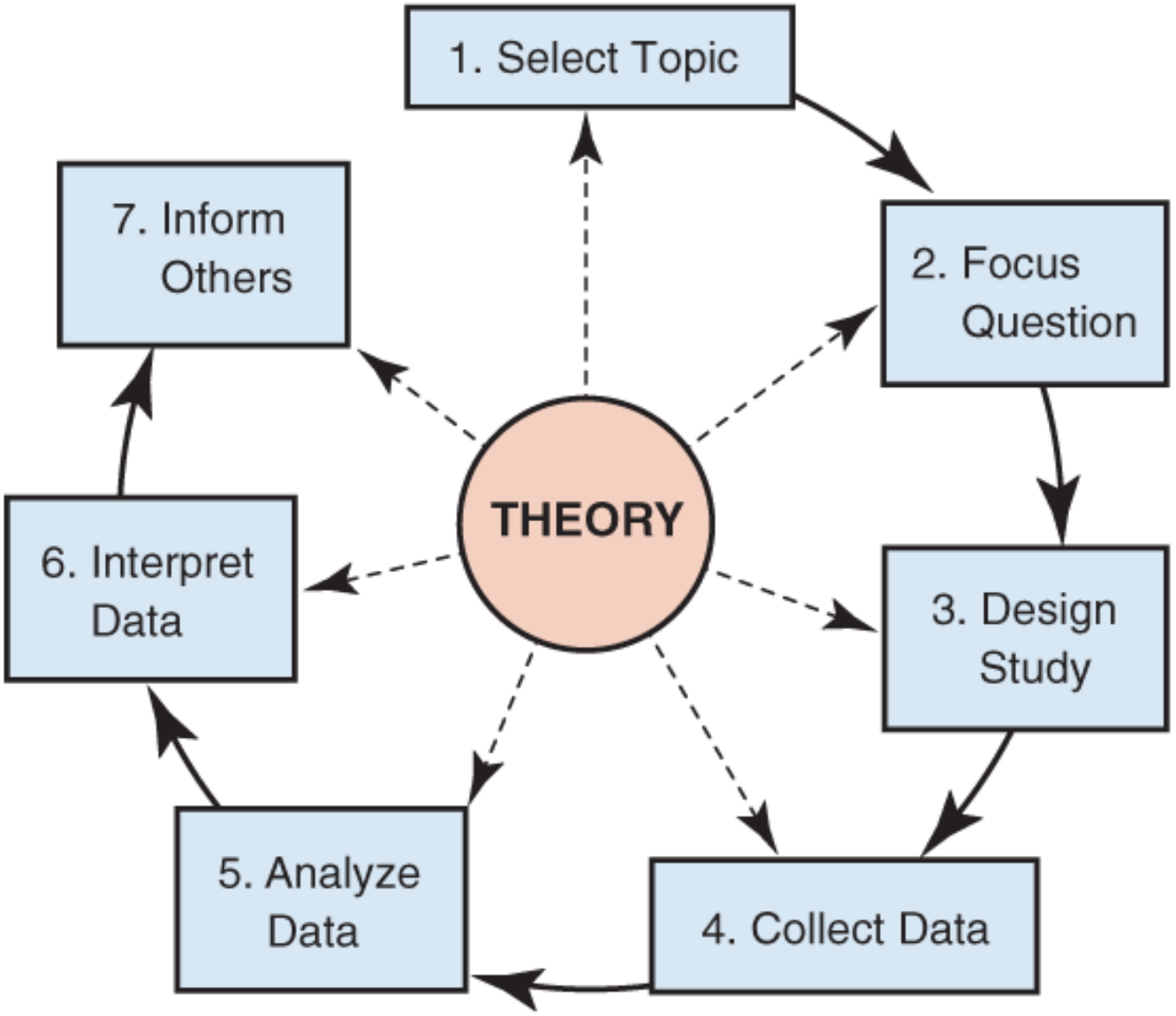


Figure 1.1 Steps in the Research Process

- Most data science activity can be divided into **three scientific tasks**, each with different methods **and philosophies**

Hernán et al 2019. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks, CHANCE, 32:1, 42-49, DOI: 10.1080/09332480.2019.1579578

Credits to @PWGTennant

Description (& visualisation)

- Focussed on **summarising**, **describing**, &/or **visualising** features of interest
- Data driven - involves simple calculations & unsupervised learning

Questions

- What happened?
- Who was affected?
- What was occurrence of **Y** in people with **X**?

Example

- Occurrence and spread of COVID-19

Credits to @PWGTennant

Prediction (AKA classification and regression)

- Focussed on **pattern recognition** and **forecasting**
- Data driven – involves statistical modelling and supervised learning

Questions

- What **will** happen?
- Who **will** be affected?
- Are people with **X** are more likely to have **Y**?

Examples

- Screening for COVID-19 with symptoms or CT scan
- Predicting prognosis or severity of infection

Credits to @PWGTennant

Causal inference (AKA counterfactual prediction)

- Focussed on understanding
- **NOT data driven** – involves fusion of external knowledge with statistical modelling and supervised learning

Questions

- What will happen **if**…?
- **Why** were they affected?
- If we **changed X**, how would it **change Y**?

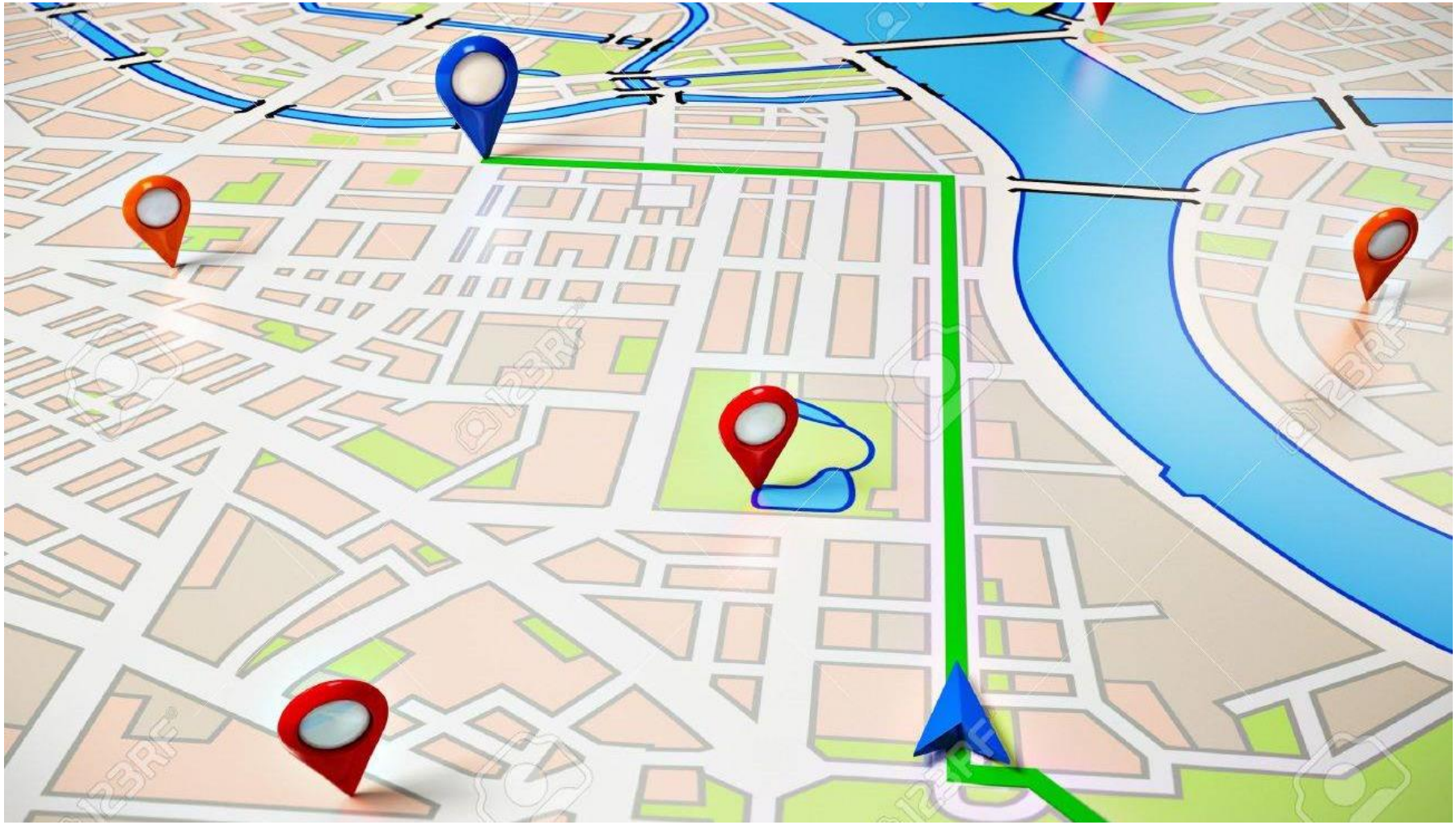
Examples

- Effect of opening/closing schools on infection spread
- Risks / benefits of ventilation

Credits to @PWGTennant

The research objectives

- Be focused to basic elements of the problem and research question
- Be measurable, achievable, observable
- Be clear and precise
- Follow a logical order (if they are more than one
- Infinitive verbs (english)



Appropriate objectives are the route map
of statistical analysis

across the US. The purposes of this study were to assess factors associated with death and to examine interhospital variation in treatment and outcomes in patients with COVID-19.

death requires a new approach. We therefore set out to deliver a secure analytics platform inside the data centre of major electronic health records vendors, running across the full live linked pseudonymised electronic health records of a very large population of NHS patients, to determine factors associated with COVID-19 related death in England (referred to as "death" in text that follows).

In this study, the epidemiological characteristics of patients with COVID-19 in Wuhan through March 8, 2020, were described, and the rate of confirmed cases and effective reproduction number in different periods according to key events and interventions were compared to evaluate the temporal associations of multiple public health interventions with control of the COVID-19 outbreak in Wuhan.

We aimed to systematically review and critically appraise currently available prediction models for covid-19, in particular diagnostic and prognostic models for the disease. This systematic review was

factors. For this reason, the aim of the study was to estimate the prevalence of depression and mistreatment and identify the factors associated in medical interns of Peruvian hospitals.

tratamiento. El objetivo de este trabajo es establecer los factores asociados con más de un intento de suicidio registrados durante 2016 en la población colombiana.

El objetivo de este estudio es explorar si hay relación y con qué fuerza entre los factores psicosociales y la HTA en la ciudad de Medellín.

and 76 in the other trial.[12,13] Accordingly, the primary objective of the International Nocturnal Oxygen (INOX) trial was to determine, in patients with COPD with nocturnal arterial oxygen desaturation who do not qualify for long-term oxygen therapy because of the absence of severe daytime hypoxemia, whether nocturnal oxygen provided for a period of 3 to 4 years affects mortality or the progression of COPD such that patients meet current specifications for long-term oxygen therapy. Secondary objectives were to examine whether nocturnal oxygen changes disease-specific quality of life or modifies exacerbation and hospitalization rates.

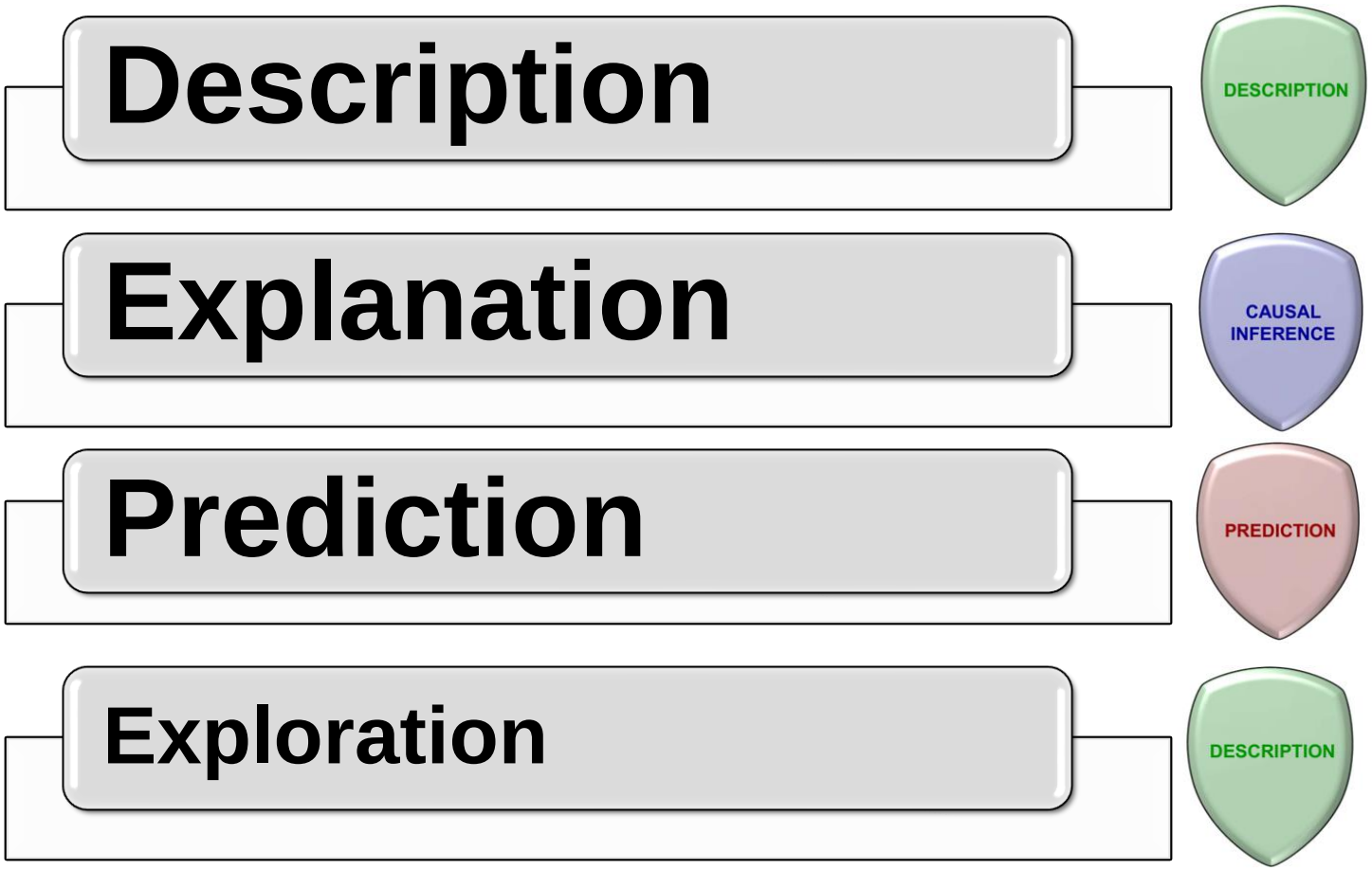
belimumab.[11] Those observations led us to conduct the current trial, Belimumab International Study in Lupus Nephritis (BLISS-LN), to evaluate the efficacy and safety of belimumab plus standard therapy (mycophenolate mofetil or cyclophosphamide–azathioprine) in patients with active lupus nephritis.

Rivaroxaban is a potent, oral, highly selective direct inhibitor of factor Xa and is effective for primary and secondary thromboprophylaxis.[12,13] Therefore, we conducted the CASSINI trial to assess the efficacy and safety of rivaroxaban thromboprophylaxis in patients with a solid tumor or lymphoma who had a Khorana score of 2 or higher and were initiating a new systemic cancer regimen.

The greatest invention of the nineteenth century was the invention of the method of invention.

—A. N. Whitehead

Additional scope

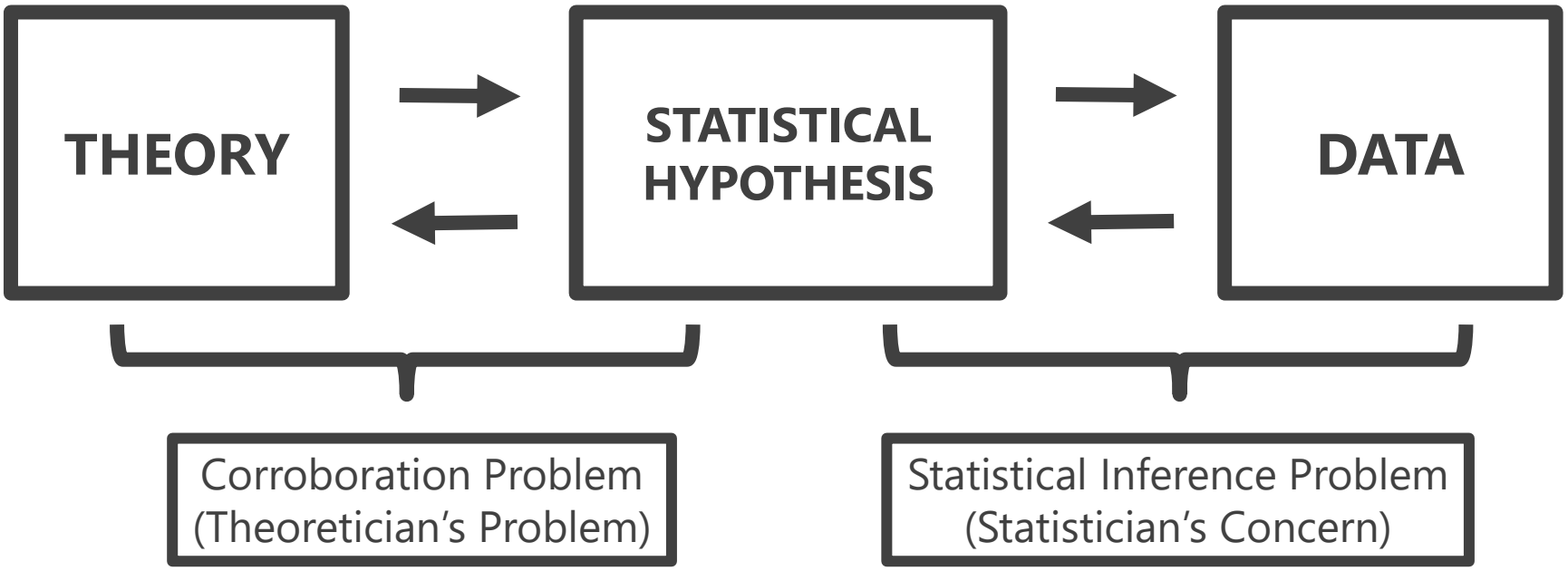


Descriptive modeling

- Uses statistical models to summarize data
- The focus is on measurement
- Less on constructs or theory.

Explanatory modeling (causal inference)

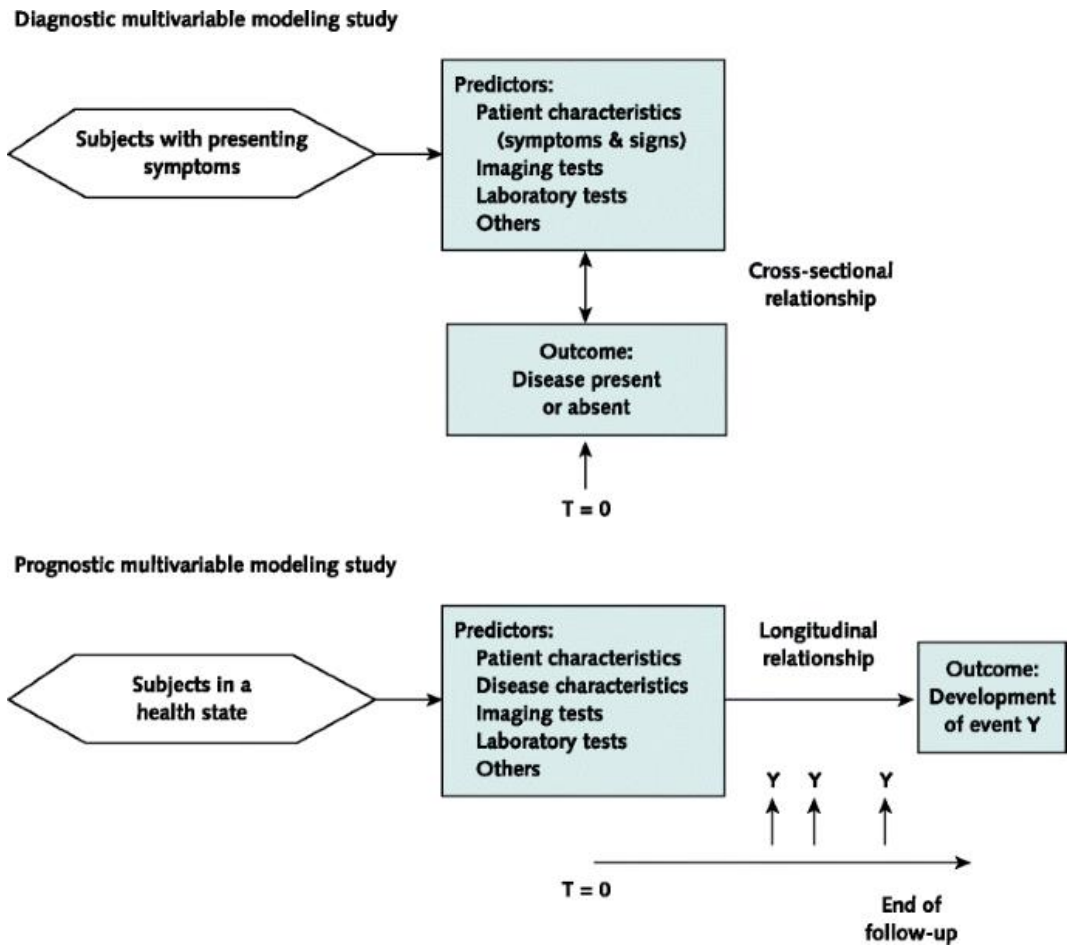
- Uses statistical models to test causal explanations derived from theories.
- Causal inference principles.



Predictive modeling

Uses statistical models to predict future or otherwise unknown observations.

Requires relatively large hold-out datasets.



Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

- Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description

- In many disciplines there is near-exclusive use of statistical modeling for causal explanation and the assumption that models with high explanatory power are inherently of high predictive power

To Explain or to Predict?

- Prediction
  - What will happen next?
- Explanation
  - Why did it happen?

Explanatory modeling

- A set of underlying (causal) factors measured by variables X are hypothesised to cause an underlying effect measured by variable Y
- Focus: statistically testing the causal hypotheses
- Data come from well designed experiments or observational studies

Predictive modeling

- Goal: Predict Y from new observations of given input variables X

- This includes temporal forecasting: Observations until time t (the input) are used to forecast future values at time t+k, k>0 (the output)

- A predictive model is any model that produces predictions, regardless of its underlying modelling approach: Bayesian or Frequentist, Parametric or Non-Parametric, Data Mining Algorithm or Statistical Model, etc.

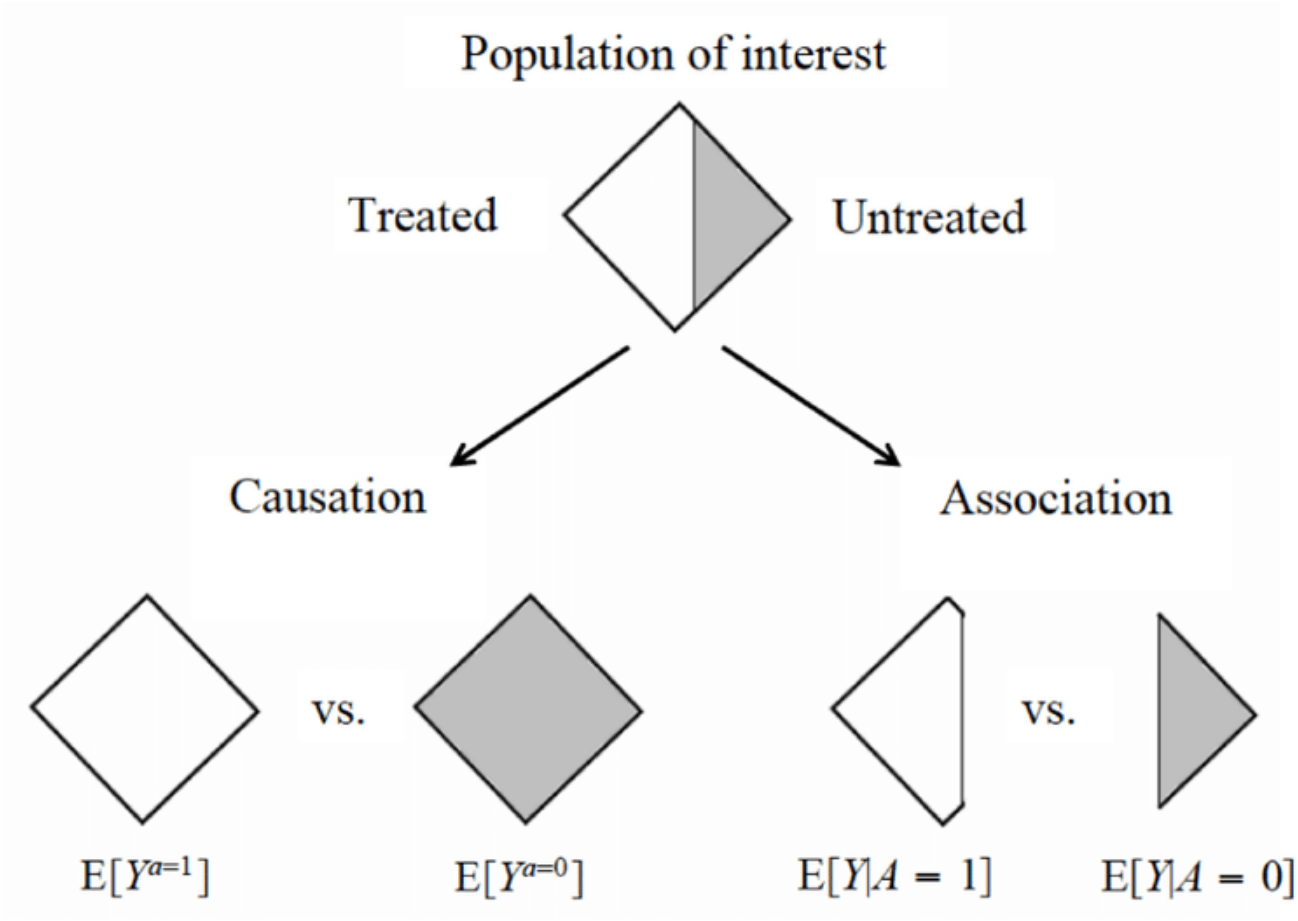
Two modeling paths

- In each step, there are differences in the choice of methods, criteria, data, information to consider when the goal is predictive vs explanatory
- Conceptual / practical differences invariably lead to a difference between a final explanatory model and a predictive model
- A priori determination of main study goal is essential to conduct adequate modelling.

Descriptive Modeling

- Aimed at summarizing or representing the data structure in a compact manner
- Unlike explanatory modeling, in descriptive modeling the reliance on an underlying causal theory is absent or incorporated in a less formal way
- Unlike predictive modeling, descriptive modeling is not aimed at prediction
- Fitting a regression model can be descriptive if it is used for capturing the **association between the dependent and independent variables** rather than for causal inference or for prediction

Association versus Causation



Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.

Association versus Causation

Association

- Definition: The concurrence of two variables more often than would be expected by chance

- Types of association:
  - Spurious association
  - Indirect association
  - Direct (causal) association
    - One to one causal association
    - Multi-factorial association

Association

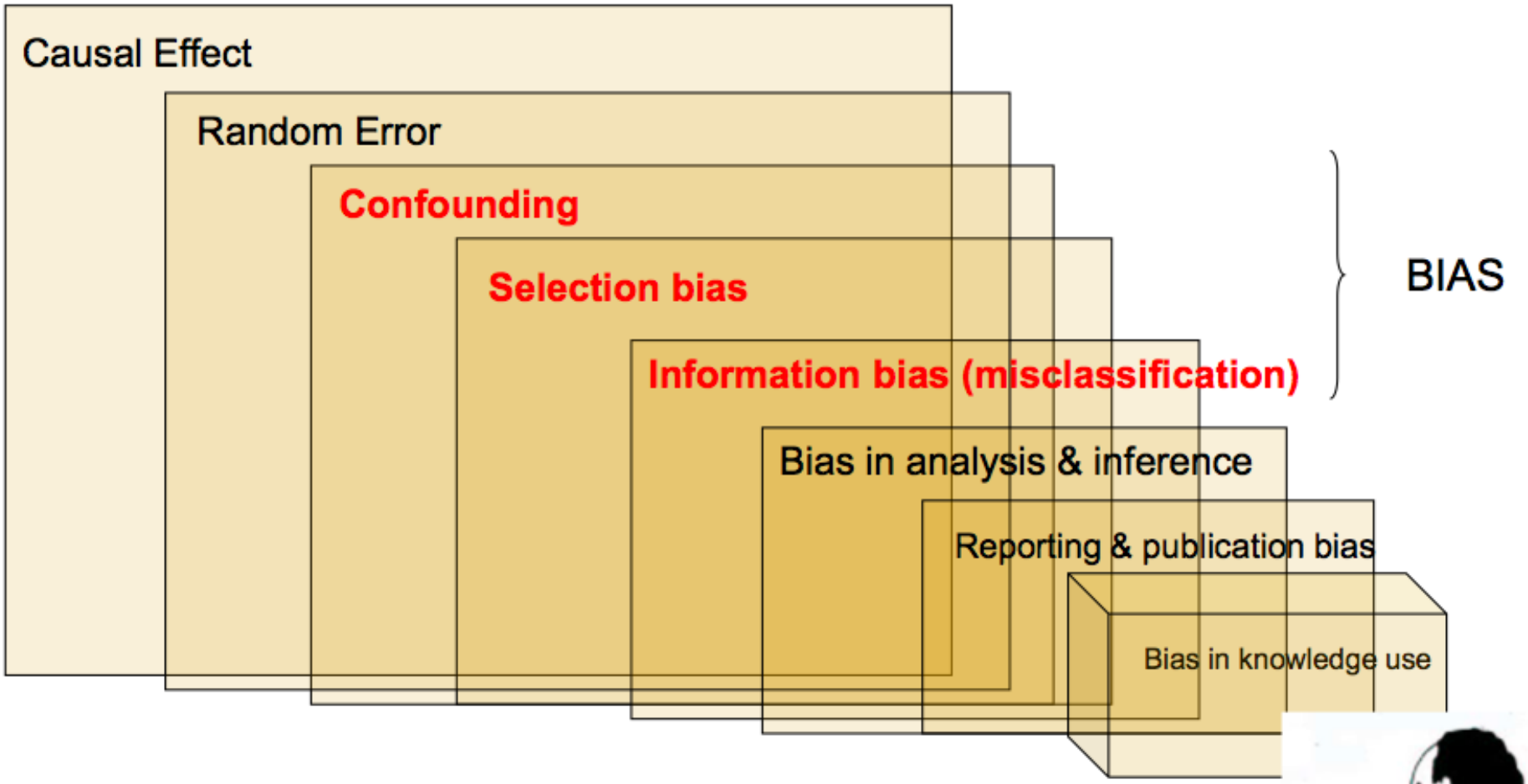
- A researcher in his observational study found that the average serum homocysteine among men patients of ischemic heart disease was 15 mcg/dl (normal=10-12 mcg/dl)

- What can we say about that?

Association

- Hyperhomocystenemia causes IHD?
- Hypothesize that,
  - Hyperhomocystenemia may have a role in etiology of IHD
  - Hyperhomocystenemia is associated with IHD
- For advancing in the understanding there has to be a ‘comparison’
  - Comparison would generate another summary measure which shows **the extent of** this ‘Association’ or ‘Effect’ or ‘risk’ (RR, OR, AR)

From association to causation in epidemiology



RR_{causal} "truth" → $RR_{association}$

the long road to causal inference

Open your mind,

Open your mind,

What exactly does your mind need to consider something **causal**?

Association versus Causation (conceptually speaking)

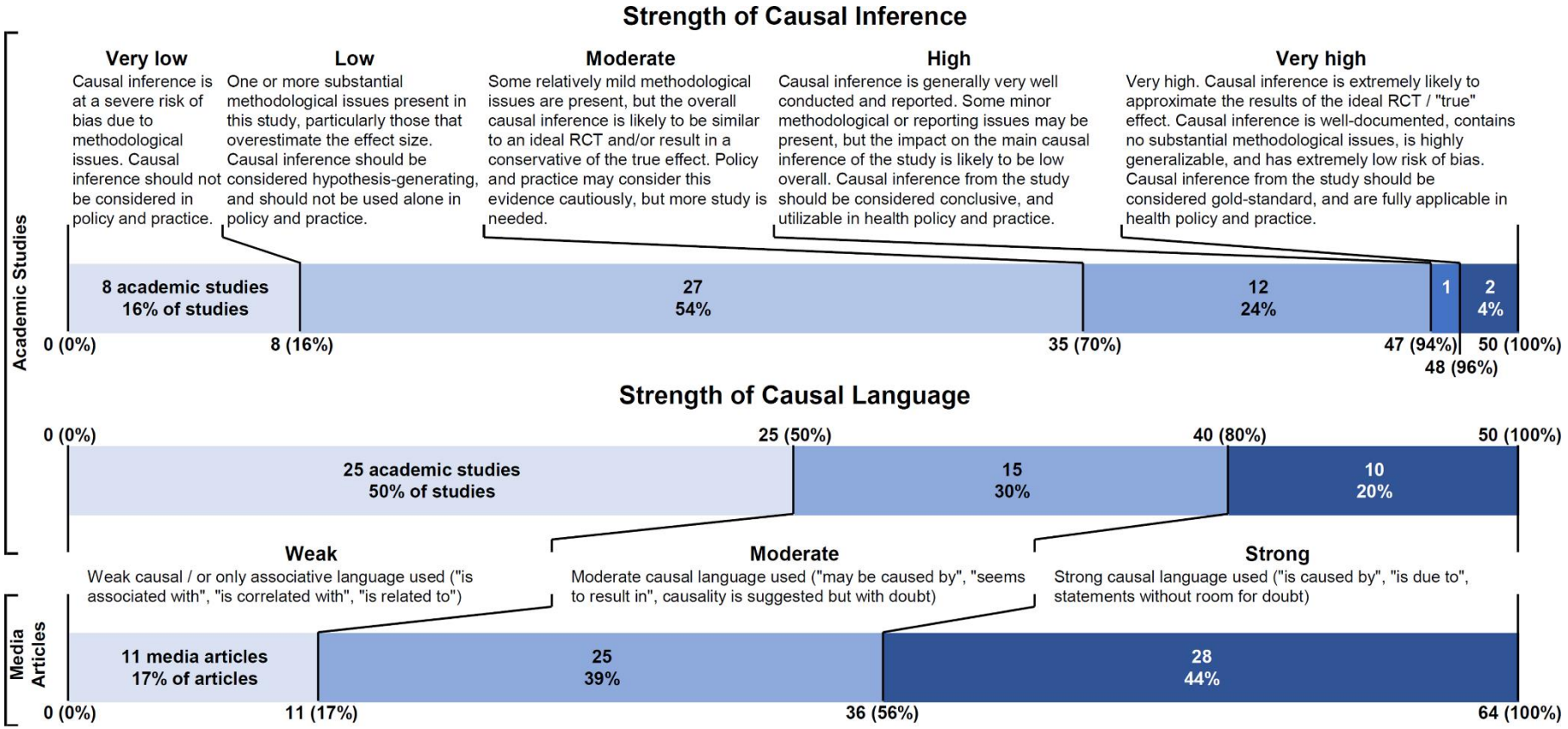
Association

- Correlation
- Regression
- Odds / relative risk
- Dependence
- Likelihood
- Conditional

Causation

- Influence
- Effect
- Confounding
- Explanation
- Intervention
- Randomization
- Attribution

The importance of language



Haber N, Smith ER, Moscoe E, Andrews K, Audy R, et al. (2018) Causal language and strength of inference in academic and media articles shared in social media (CLAIMS): A systematic review. PLOS ONE 13(5): e0196346. https://doi.org/10.1371/journal.pone.0196346

The importance of language

- Avoiding “causal” language with observational study designs is common publication practice, often justified as being a more cautious approach to interpretation
- There is a substantial disconnects between causal implications used to link an exposure to an outcome and the action implications made.
- This undercuts common assumptions about what words are often considered non-causal and that policing them eliminates causal implications.

Causal and Associational Linking Language From Observational Research and Health Evaluation Literature in Practice: A systematic language evaluation Noah A. Haber, et al. medRxiv 2021.08.25.21262631; doi: https://doi.org/10.1101/2021.08.25.21262631

Example

- **Question:** Do patient outcomes differ between physicians who work part time clinically vs full-time clinicians?
- **Objective:** To examine the association between the number of days worked clinically per year by physicians and patient mortality

- **Methods:** This cross-sectional analysis was completed on a 20% random sample of Medicare fee-for-service beneficiaries 65 years and older who were admitted to the hospital with an emergency medical condition and treated by a hospitalist in 2011 through 2016.

- **Results:** Among 392.797 hospitalizations of patients treated by 19.170 hospitalists (7482 female [39.0%], 11 688 male [61.0%]; mean [SD] age, 41.1 [8.8] years), patients treated by physicians with more days worked clinically exhibited lower mortality.

Kato H, Jena AB, Figueroa JF, Tsugawa Y. Association Between Physician Part-time Clinical Work and Patient Outcomes. JAMA Intern Med. 2021;181(11):1461–1469. doi:10.1001/jamainternmed.2021.5247

Example

- **Interpretation:**
  - Included hospitals should increase the working days of physicians in order to reduce 30-day mortality of patients.
  - If you work more your patients had lower 30-day mortality.

- **Conclusions and Relevance:** In this cross-sectional study, hospitalized Medicare patients treated by physicians who worked more clinical days **had lower 30-day mortality**.
- Given that physicians with reduced clinical time must often balance clinical and nonclinical obligations, improved support by institutions may be necessary to maintain the clinical performance of these physicians.

Kato H, Jena AB, Figueroa JF, Tsugawa Y. Association Between Physician Part-time Clinical Work and Patient Outcomes. JAMA Intern Med. 2021;181(11):1461–1469. doi:10.1001/jamainternmed.2021.5247

unclear analysis aims causality, prediction or description: it makes a difference	**evidence of absence fallacy** not significant = no effect? if only things were that easy	**data dredging** "look what I found that turned out significant"	**noisy data fallacy** that what doesn't kill it makes it stronger? i wish!
dichotomania categorize everything, chapter 1 of data torture handbook	**table 2 fallacy** interpret each regression coefficient as adjusted for confounding? yeah, that is probably wrong	**regression to the mean** yes, measurements at follow-up are different than at baseline. doesn't mean much	**ignoring dependent observations** the data are probably nested, and that is important
poor reporting yes, there are reporting guidelines for that	**small sample size** gets these confidence intervals nice and wide	**ignoring missing data** can't miss what you never had? think again	**data driven variable selection** post selection inference, yikes!
only apparent predictive performance that is optimistic. could have tried a cross-validation or a bootstrap, perhaps?	**point estimate is the effect** why did you even bother calculating the confidence interval?	**collider** is it a cause? is it a confounder? no, it is a collider here to ruin your inferences!	**multivariate model** what you mean to say is: multi*variable* model

A CHecklist for statistical Assessment of Medical Papers (the CHAMP statement): explanation and elaboration

Mohammad Ali Mansournia ,[1,2] Gary S Collins,[3,4] Rasmus Oestergaard Nielsen ,[5,6] Maryam Nazemipour,[7] Nicholas P Jewell,[8,9] Douglas G Altman,[3] Michael J Campbell[10]

Mansournia MA, *et al*. *Br J Sports Med* 2021;**0**:1–9. doi:10.1136/bjsports-2020-103652

Table 1 CHecklist for statistical Assessment of Medical Papers

Design and conduct
1. Clear description of the goal of research, study objective(s), study design and study population
2. Clear description of outcomes, exposures/treatments and covariates, and their measurement methods
3. Validity of the study design
4. Clear statement and justification of sample size
5. Clear declaration of design violations and acceptability of the design violations
6. Consistency between the paper and its previously published protocol

Data analysis
7. Correct and complete description of statistical methods
8. Valid statistical methods used and assumptions outlined
9. Appropriate assessment of treatment effect or interaction between treatment and another covariate
10. Correct use of correlation and associational statistical testing
11. Appropriate handling of continuous predictors
12. CIs do not include impossible values
13. Appropriate comparison of baseline characteristics between the study arms in randomised trials
14. Correct assessment and adjustment of confounding
15. Avoiding model extrapolation not supported by data
16. Adequate handling of missing data

Reporting and presentation
17. Adequate and correct description of the data
18. Descriptive results provided as occurrence measures with CIs and analytical results provided as association measures and CIs along with p values
19. CIs provided for the contrast between groups rather than for each group
20. Avoiding selective reporting of analyses and p-hacking
21. Appropriate and consistent numerical precisions for effect sizes, test statistics and p values, and reporting the p values rather than their range
22. Providing sufficient numerical results that could be included in a subsequent meta-analysis
23. Acceptable presentation of figures and tables

Interpretation
24. Interpreting the results based on association measures and 95% CIs along with p values, and correctly interpreting large p values as indecisive results, not evidence of absence of an effect
25. Using CIs rather than post hoc power analysis for interpreting the results of studies
26. Correctly interpreting occurrence or association measures
27. Distinguishing causation from association and correlation
28. Results of prespecified analyses are distinguished from the results of exploratory analyses in the interpretation
29. Appropriate discussion of the study methodological limitations
30. Drawing only conclusions supported by the statistical analysis and no generalisation of the results to subjects outside the target population

What is Biostatistics

- Statistics applied to biomedical problems
- Decision making in the face of uncertainty or variability
- Design and analysis of experiments; detective work in observational studies (in epidemiology, outcomes research, etc.)
- Attempt to remove bias or find alternative explanations to those posited by researchers with vested interests
- Experimental design, measurement, description, statistical graphics, data analysis, inference, prediction

To optimize its value, biostatistics needs to be **fully integrated into biomedical research** and we must recognize that experimental design and execution are all important.

Statistical scientific method

- Statistics is not a bag of tools and math formulas but an evidence-based way of thinking

  - It is all important to
  - understand the problem
  - properly frame the question to address it
  - understand and optimize the measurements
  - understand sources of variability
  - much more

Problem	Units & Target Population (Process) Response Variate(s) Explanatory Variates Population Attribute(s) Problem Aspect(s) – causative, descriptive, predictive
Plan	Study Population (Process) (Units, Variates, Attributes) Selecting the response variate(s) Dealing with explanatory variates Sampling Protocol Measuring process Data Collection Protocol
Data	Execute the Plan and record all departures Data Monitoring Data Examination for internal consistency Data storage
Analysis	Data Summary numerical and graphical Model construction build, fit, criticize cycle Formal analysis
Conclusion	Synthesis plain language, effective presentation graphics Limitations of study discussion of potential errors