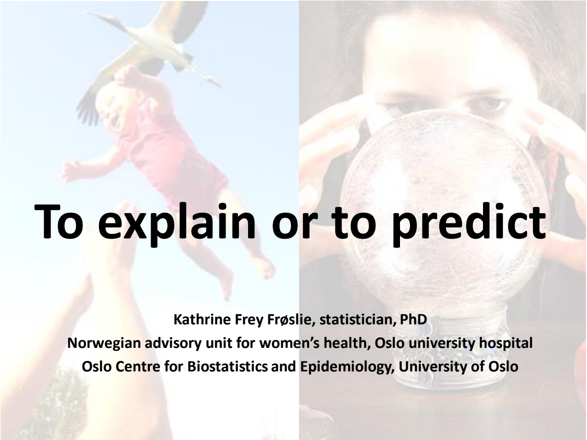


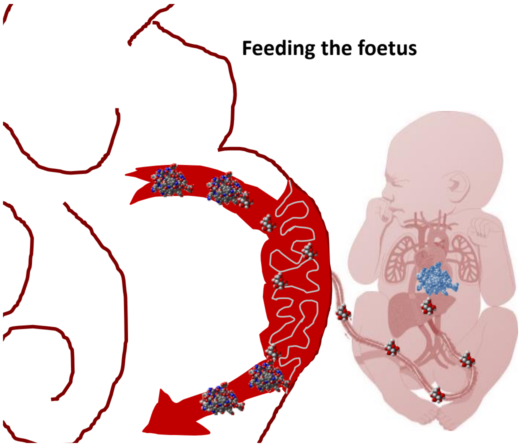
To explain or to predict

Kathrine Frey Frøslie, statistician, PhD
Norwegian advisory unit for women's health, Oslo university hospital
Oslo Centre for Biostatistics and Epidemiology, University of Oslo

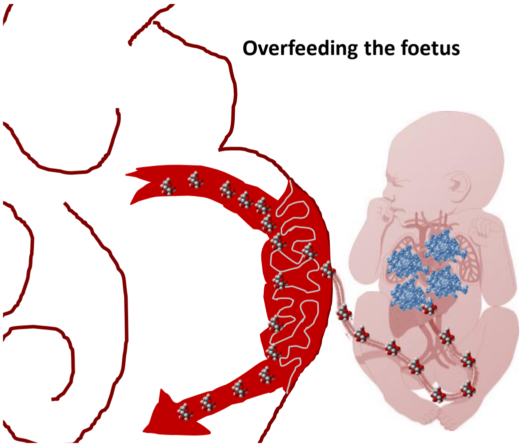
Aftenposten, 5/12-2001:

FØDT
Nils er høy, mørk og bredskuldret, og veide 4750g. Ingeborg er stolt storesøster. Ole Frøslie og Kathrine Frey Frøslie. Rikshospitalet, 3/12-2001

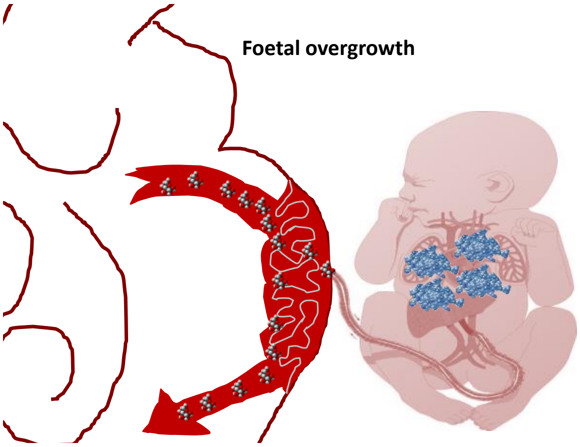
Feeding the foetus



Overfeeding the foetus



Foetal overgrowth



An insulin pump predicts future glucose values

http://www.healthline.com/diabetesmine/a-techie-type-1-reviews-the-new-medtronic-revel-insulin-pump-cgm-system#3

Artificial pancreas control system

$$\begin{cases}
\dot{G}_p = -(k_2 + kp_2)G_p + k_1 G_t - U_{ii} - E(G_p) - kp_3 I_d + \frac{f k_{abs} Q_{gut}}{BW} + kp_1 \\
\dot{G}_t = -k_1 G_t + k_2 G_p - \frac{Vm_0 + Vm_X X}{Km_0 + G_t} G_t \\
\dot{G}_i = -\frac{1}{\tau_{SG}}\left(G_i - \frac{G_p}{V_g}\right) \\
\dot{I}_d = -k_i (I_d - I_1) \\
\dot{I}_1 = -k_i \left(I_1 - \frac{I_p}{V_I}\right) \\
\dot{I}_p = -(m_2 + m_4) I_p + m_1 I_l + k_{a1} I_{SQ1} + k_{a2} I_{SQ2} \\
\dot{I}_l = -(m_1 + m_3) I_l + m_2 I_p \\
\dot{X} = -p_{2u}\left(X - \frac{I_p}{V_I} + I_b\right) \\
\dot{I}_{SQ1} = -(k_{a1} + k_d) I_{SQ1} + J(t) \\
\dot{I}_{SQ2} = -k_{a2} I_{SQ2} + k_d I_{SQ2} \\
\dot{Q}_{sto1} = -k_{gri} Q_{sto1} + D(t) \\
\dot{Q}_{sto2} = -k_{empt}(Q_{sto1} + Q_{sto2}) Q_{sto2} + k_{gri} Q_{sto1} \\
\dot{Q}_{gut} = -k_{abs} Q_{gut} + k_{empt}(Q_{sto1} + Q_{sto2}) Q_{sto2}
\end{cases}$$

$$\text{Where } E(G_p) = \begin{cases} ke_1(G_p - ke_2) & \text{if } G_p > ke_2 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } k_{empt}(Q_{sto1} + Q_{sto2}) = k_{min} + \frac{k_{max} - k_{min}}{2}\left(2 + \tanh\left(aa\left(Q_{sto1} + Q_{sto2} - b\,dose\right)\right) + \tanh\left(cc\left(Q_{sto1} + Q_{sto2} - d\,dose\right)\right)\right)$$

$$aa = \frac{2.5\,dose}{1-b} \quad cc = \frac{2.5\,dose}{d}$$

Patent application WO 2008157780 A1

Calculate optimal insulin dose by defining insulin injection as the linear combination of gain and state, which minimize a quadratic cost function.

Compute *J(q)* as:

$$J(q) = \frac{1}{100}\sum_{i=1}^{100} t2tgt(q)_i + tbtgt(q)_i$$

The optimal parameter *q** is defined as :

$$q^* = \underset{q}{\operatorname{argmin}}\, J(q)$$

Thom R. Prédire N'est Pas Expliquer (1991)

Shmueli G. To explain or to predict (2010)

Abdelnoor M, Sandven I: Etiologisk versus prognostisk strategi i klinisk epidemiologisk forskning (2006)

Overview

The aim of epidemiology
The research process
Regression analysis

To explain or to predict: Explain

Mechanisms
Causality
DAGs
Exposure & outcome
Confounder, Mediator, Collider

To explain or to predict: Predict

Diagnostic tests, Forecasting
Personalized medicine
Statistical learning, big data, black box
Prediction error

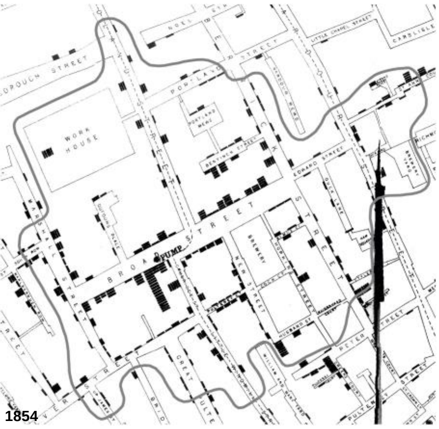
EPIDEMIOLOGY

The study of the occurrence and distribution of health-related states or events in specified populations, including the study of DETERMINANTS influencing such states, and the application of this knowledge to control the health problems.

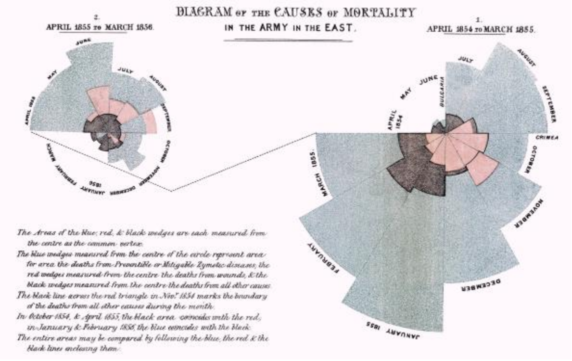
Study includes surveillance, observation, hypothesis testing, analytic research, and experiments. Distribution refers to analysis by time, place, classes or subgroups of persons affected in a population or in a society. Determinants are all the physical, biological, social, cultural, economic and behavioral factors that influence health. Health-related states and events include diseases, causes of death, behaviors, reactions to preventive programs, and provision and use of health services. Specified populations are those with common identifiable characteristics. Application… to control… makes explicit the aim of epidemiology – to promote, protect, and restore health.

The primary "knowledge object" of epidemiology as a scientific discipline are causes of health-related events in populations. In the last 70 years, the definition has broadened from concern with communicable disease epidemics to take in all processes and phenomena related to health in populations. Therefore epidemiology is much more than a branch of medicine treating epidemics.

Porta M: A Dictionary of Epidemiology, Fifth Edition, 2008



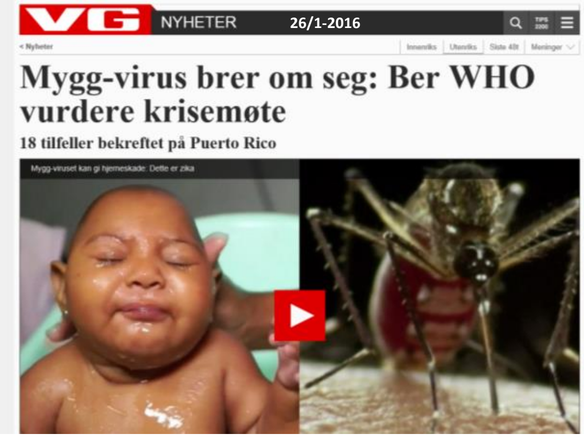
John Snow, 1854



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red, in January & February 1856, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Causes of death Preventible Wounds Other

Florence Nightingale, 1858

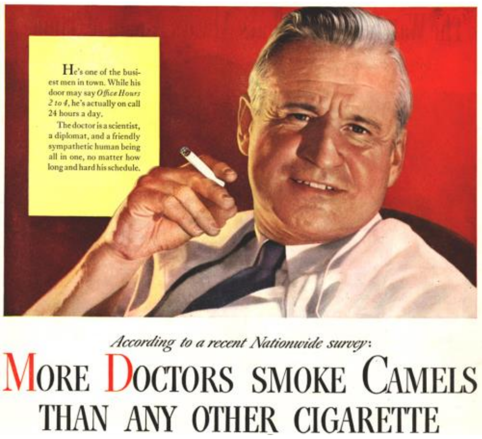


26/1-2016

Mygg-virus brer om seg: Ber WHO vurdere krisemøte

18 tilfeller bekreftet på Puerto Rico

Mygg-viruset kan gi hjerneskade: Dette er zika


He's one of the busiest men in town. While his door may say *Office Hours 2 to 4*, he's actually on call 24 hours a day.
The doctor is a scientist, a diplomat, and a friendly sympathetic human being all in one, no matter how long and hard his schedule.
According to a recent Nationwide survey:
MORE DOCTORS SMOKE CAMELS
THAN ANY OTHER CIGARETTE

DOLL R, HILL AB. Smoking and carcinoma of the lung. BMJ 1950;4682:739-748.

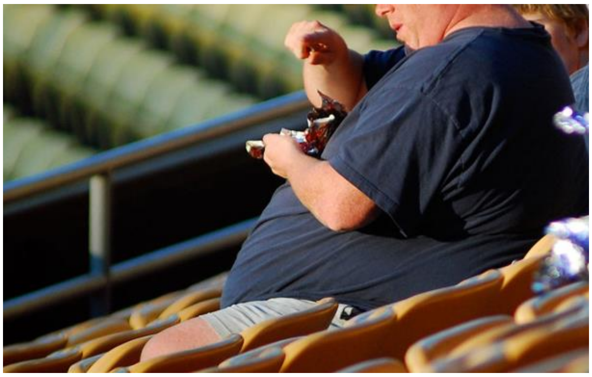
"To summarize, it is not reasonable, in our view, to attribute the results to any special selection of cases or to bias in recording. In other words, it must be concluded that there is a real association between carcinoma of the lung and smoking."

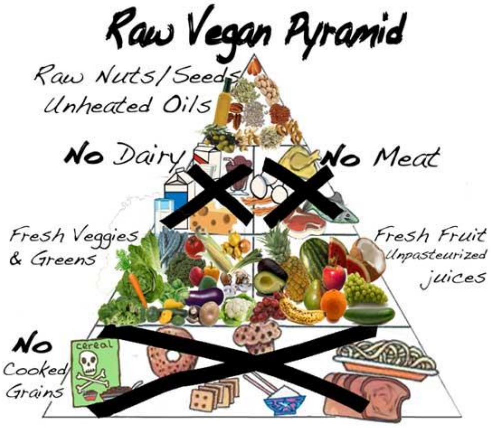
"...it is concluded that smoking is an important factor in the cause of carcinoma of the lung."

Richard Doll, Austin Bradford Hill, 1950

Ill: Åshild Irgens

Marit B Veierød, 2015: Melanoma incidence on the rise again





To explain

Understanding

Mechanisms

Expert knowledge

What is the best estimate for the association between the main exposure and the main outcome?

Fokus på $\hat{\beta}$

Etiology

Causality

Causality

What is causality?

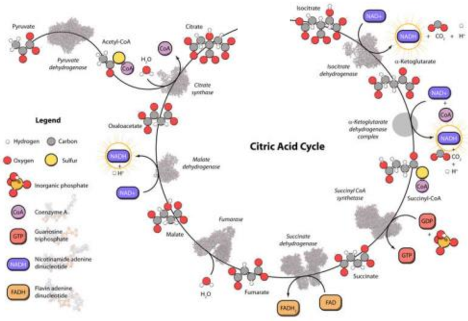
The Counterfactual concept

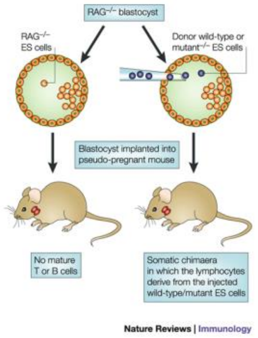
Philosophic background

Interventions and causality

Consequences of actions

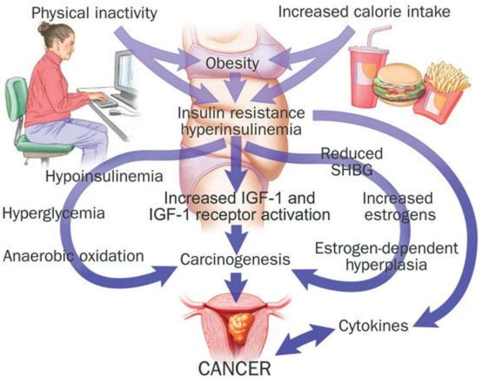
Ultimate goal: Action!



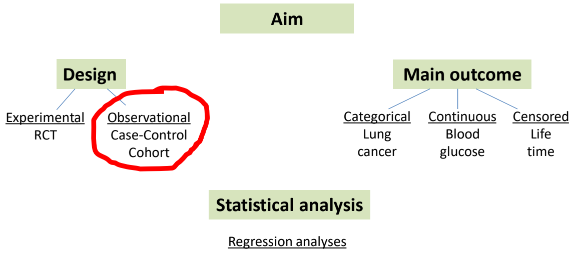

RAG−/− blastocyst
RAG−/− ES cells
Donor wild-type or mutant−/− ES cells
Blastocyst implanted into pseudo-pregnant mouse
No mature T or B cells
Somatic chimaera in which the lymphocytes derive from the injected wild-type/mutant ES cells

Nature Reviews | Immunology

Implantation of blastocysts from RAG-deficient mice into pseudo-pregnant mothers results in the development of viable mice that lack mature B and T cells. However, if normal embryonic stem (ES) cells are injected into RAG-deficient blastocysts, somatic chimaeras are formed, which develop mature B and T cells.



Problem of interest



Aim

Design

Experimental
RCT

Observational
Case-Control
Cohort

Main outcome

Categorical
Lung cancer

Continuous
Blood glucose

Censored
Life time

Statistical analysis

Regression analyses

Based on expert knowledge of the topic under investigation, we want to estimate the association between an exposure and an outcome as unbiasedly as possible.

Knowledge of the topic makes it plausible that the estimated association can be interpreted as an effect, i. e. causal, i.e. as a quantification of mechanisms.

The expert knowledge about the topic is formalised in a graph of the variables studied, a Directed Acyclic Graph (DAG).

In a DAG, one variable is defined as the main outcome, and one variable is defined as the main exposure.

Expert knowledge is used to define other variables as either a confounder, a mediator, or a collider.

DIRECTED ACYCLIC GRAPH (DAG)

See CAUSAL DIAGRAM.

CAUSAL DIAGRAM

(Syn: causal graph, path diagram) **A graphical display of causal relations** among variables, in which each variable is assigned a fixed location on the graph (called a *node*) and in which each direct causal effect of one variable on another is represented by an arrow with its tail at the cause and its head at the effect. Direct noncausal associations are usually represented by lines without arrowheads.

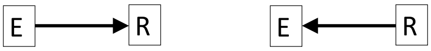
Graphs with only directed arrows (in which all direct associations are causal) are called *directed graphs*.

Graphs in which no variable can affect itself (no feedback loop) are called *acyclic*.

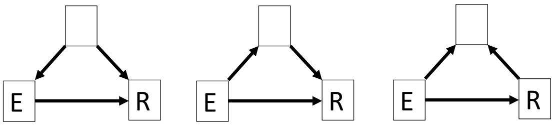
Algorithms have been developed to determine from causal diagrams which sets of variables are sufficient to control for confounding, and for when control of variables leads to bias.

Porta M: A Dictionary of Epidemiology, Fifth Edition, 2008

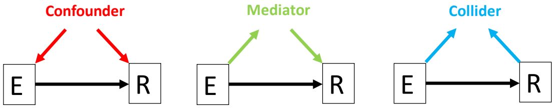
Exposure and response



Confounder, mediator, collider



Confounder, mediator, collider



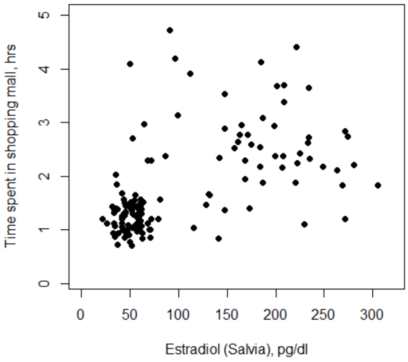
Note: The presence of either of these may affect the association of interest. Hence, including either of these in regression analyses may change effect estimates.

Confounding

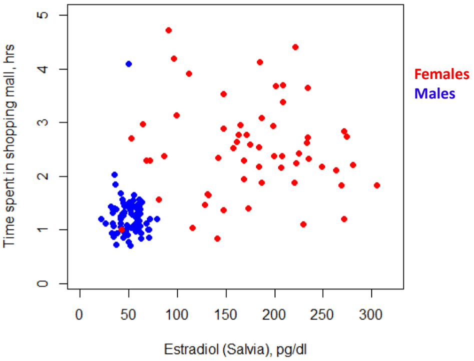
Confounding is bias of the estimated effect of an exposure on an outcome due to the presence of a common cause of the exposure and the outcome.

Confounding can be reduced by proper adjustment. Exploring data is not sufficient to identify whether a variable is a confounder, and such evaluation of confounding may lead to bias. Other evidence like pathophysiological and clinical knowledge and external data is needed. DAGs are useful tools when considering confounding variables.

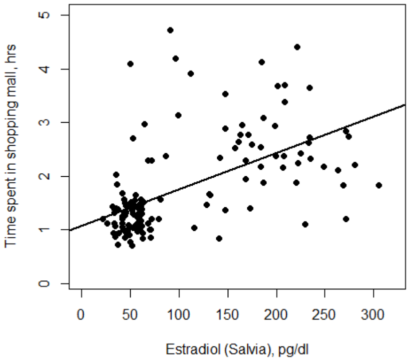
Ex: Shopping time vs estradiol level. Simulated data.



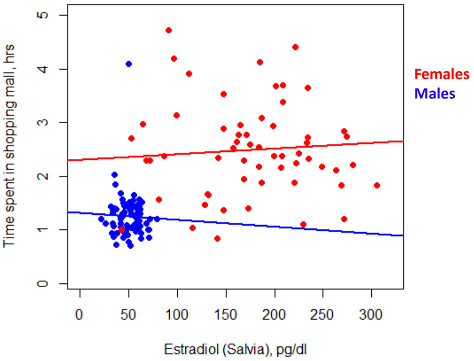
Ex: Shopping time vs estradiol level. Simulated data.



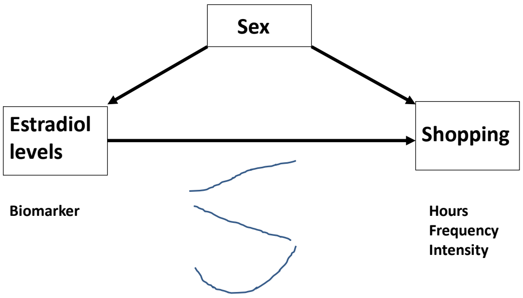
Ex: Shopping time vs estradiol level. Simulated data.



Ex: Shopping time vs estradiol level. Simulated data.



Confounder:
Sex or gender?



Sex

Estradiol levels

Shopping

Biomarker

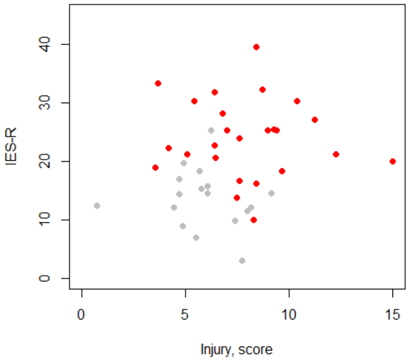
Hours
Frequency
Intensity

COLLIDER

A variable directly affected by two or more other variables in the causal diagram.

Porta M: A Dictionary of Epidemiology, Fifth Edition, 2008

Ex: Post-traumatic stress after terror attack. Simulated data.

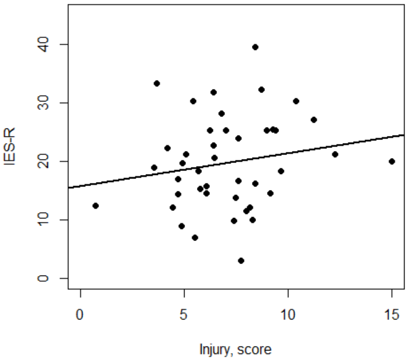


Grouped after length of hospital stay:

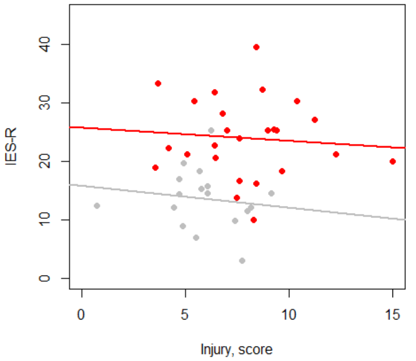
> 2 weeks

< 2 weeks

Ex: Post-traumatic stress after terror attack. Simulated data.



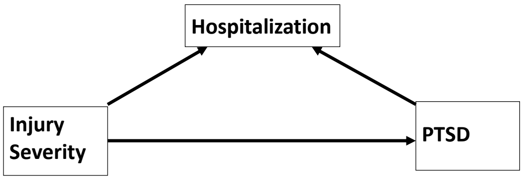
Ex: Post-traumatic stress after terror attack. Simulated data.



Grouped after length of hospital stay:

> 2 weeks

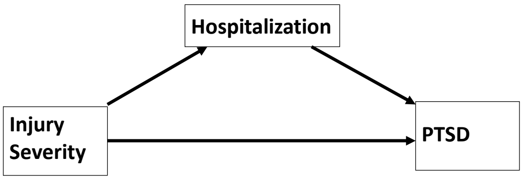
< 2 weeks



Hospitalization

Injury Severity

PTSD

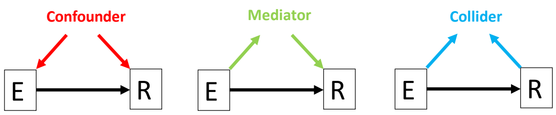


Hospitalization

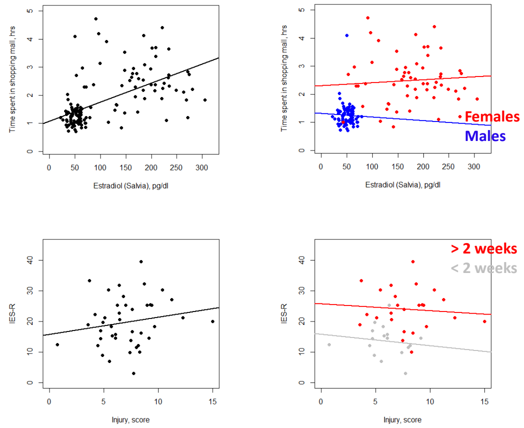
Injury Severity

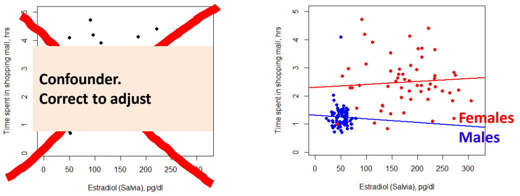
PTSD

Confounder, mediator, collider



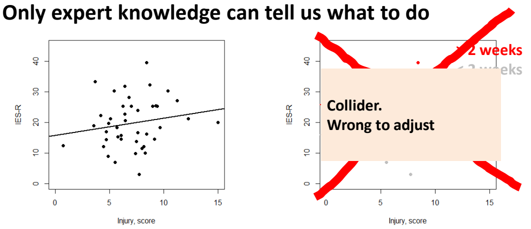
Note: The presence of either of these may affect the association of interest. Hence, including either of these in regression analyses may change effect estimates.





Confounder.
Correct to adjust

Only expert knowledge can tell us what to do



Collider.
Wrong to adjust

Oppgave 3 fra eksamen H2016

Regression analysis recipe (kind of)

When the ultimate goal is to understand mechanisms, and to estimate associations between an exposure and an outcome as unbiasedly as possible:

Use expert knowledge to identify exposure & outcome, confounders, colliders & mediators. Use DAGs to clarify and communicate.

Find the crude association between exposure and outcome

Measured variables:
    Adjust for confounders
    Do not adjust for colliders
    Sometimes adjust for mediators

Unmeasured variables:
    Sensitivity analysis

Alas, real world may not be so simple.

Alas, real world may not be so simple:

Expert knowledge may be lacking or inconclusive, regarding which variables and arrows to include or leave out in the DAG.

Additional variables may come into consideration as confounders e.g. for indirect effects, resulting in a very complex DAG.

It may be hard to tell (based on present knowledge) the direction of a causal effect, e.g. whether a variable is a confounder or a mediator, or a mediator or a collider.

Feed-back may be of concern.

Time-dependent covariates may exist.

Topics not covered

Inverse probability weighting

Dynamic path analysis
Time-dependent confounders
Marginal structural models

Latent variables
Structural equation modelling

To predict

Does not have to explain/understand mechanisms, as long as it predicts well.

Fokus på $\hat{y}$

Diagnostic tests

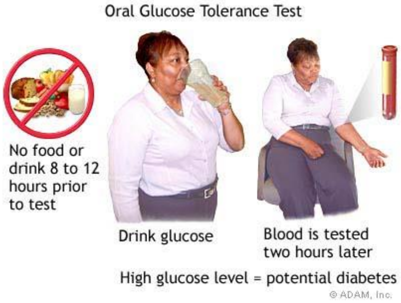
Ex Diabetes diagnosis
Melanoma screening based on picture and blood samples

Expected devlopment in disease (e.g. prognosis after sepsis)

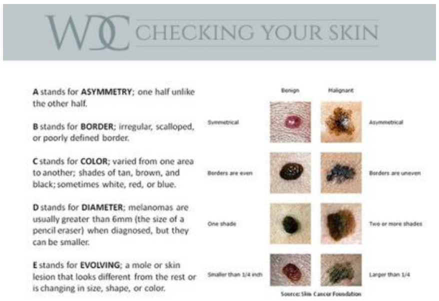
Weather forecast

Geology

Diagnostic tests


Oral Glucose Tolerance Test
No food or drink 8 to 12 hours prior to test
Drink glucose
Blood is tested two hours later
High glucose level = potential diabetes
© ADAM, Inc.

Diagnostic tests



A stands for **ASYMMETRY**; one half unlike the other half.

B stands for **BORDER**; irregular, scalloped, or poorly defined border.

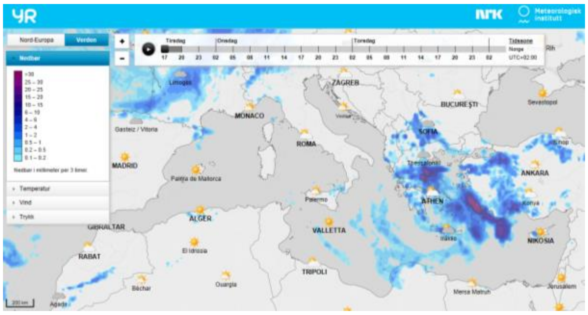
C stands for **COLOR**; varied from one area to another; shades of tan, brown, and black; sometimes white, red, or blue.

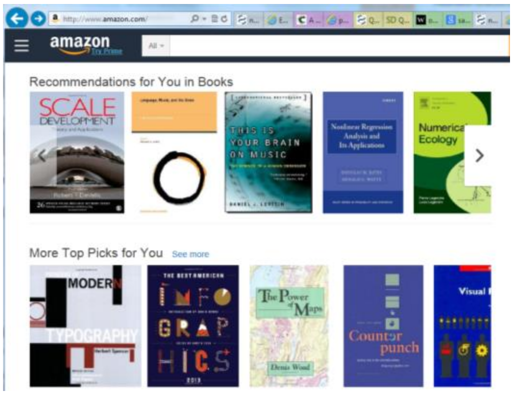
D stands for **DIAMETER**; melanomas are usually greater than 6mm (the size of a pencil eraser) when diagnosed, but they can be smaller.

E stands for **EVOLVING**; a mole or skin lesion that looks different from the rest or is changing in size, shape, or color.

Source: Skin Cancer Foundation

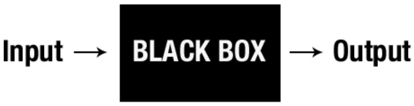
Forecasting





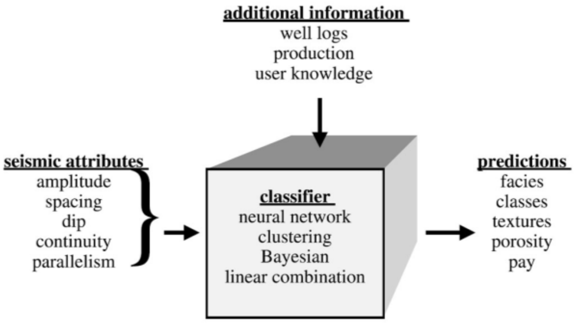
Recommendations for You in Books

More Top Picks for You See more



Input →
BLACK BOX
→ Output

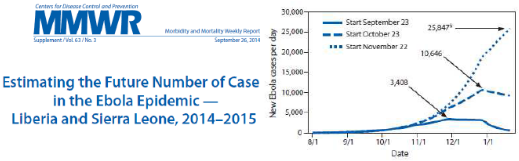
What is inside the black box?



The big issue in prediction models:

Prediction error

Modelling studies making headlines
26 September 2014



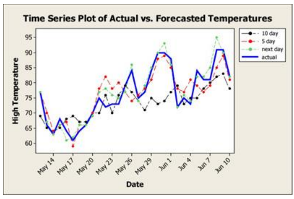
Estimating the Future Number of Case in the Ebola Epidemic — Liberia and Sierra Leone, 2014–2015

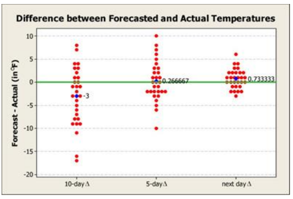
"If the virus continues to spread at the current rate, Liberia and Sierra Leone alone will have reported about 550,000 Ebola cases by 20 January …But if the official numbers so far represent only 40% of the real burden []..that would mean a total of ***1.4 million Ebola cases in those two countries by 20 January****. "*

Science: WHO, CDC publish grim new Ebola projections

NYT: Now Ebola Cases Could Reach 1.4 Million Within Four Months, C.D.C. Estimates

[slide source: Birgitte Freiesleben deBlasio, Oslo Center for Biostatistics and Epidemiology]


Time Series Plot of Actual vs. Forecasted Temperatures


Difference between Forecasted and Actual Temperatures

http://www.minitab.com/en-us/Published-Articles/Weather-Forecasts--Just-How-Reliable-Are-They-/

The big issue in prediction models:

Prediction error

The best model = the optimal predictor is the one which minimises the prediction error, i.e. the model that predicts new values (or classifies undiagnosed patients) best possible

Optimal predictor = the best model

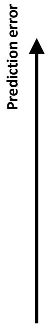
1) Define prediction error.

Optimal predictor = the best model

1) Define prediction error.

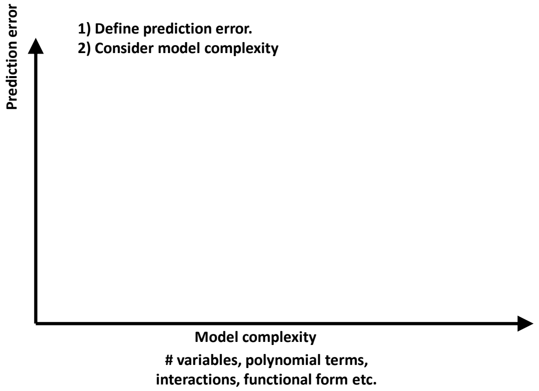
Optimal predictor = the best model

1) Define prediction error.
2) Consider model complexity



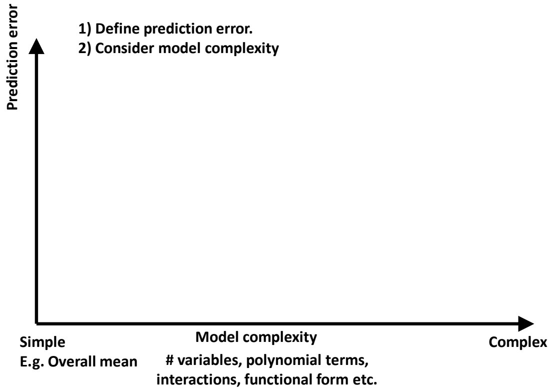
Optimal predictor = the best model

1) Define prediction error.
2) Consider model complexity


Prediction error
Model complexity
# variables, polynomial terms,
interactions, functional form etc.

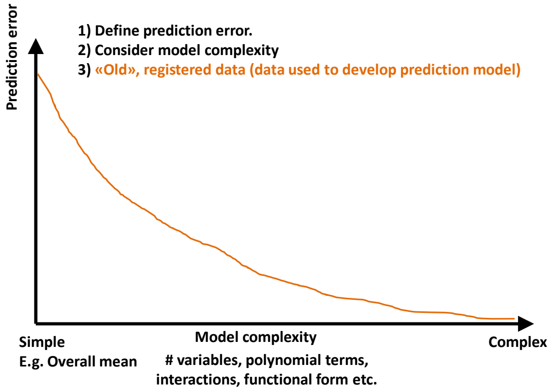
Optimal predictor = the best model

1) Define prediction error.
2) Consider model complexity


Prediction error
Simple
E.g. Overall mean
Model complexity
# variables, polynomial terms,
interactions, functional form etc.
Complex

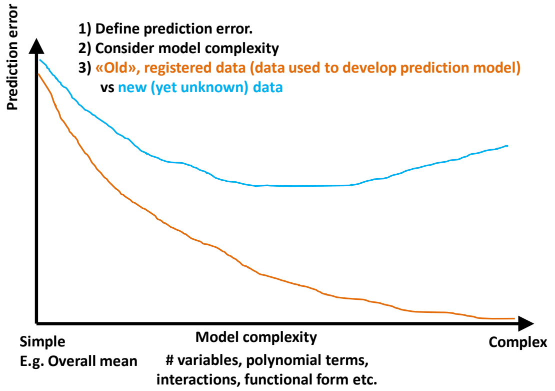
Optimal predictor = the best model

1) Define prediction error.
2) Consider model complexity
3) «Old», registered data (data used to develop prediction model)

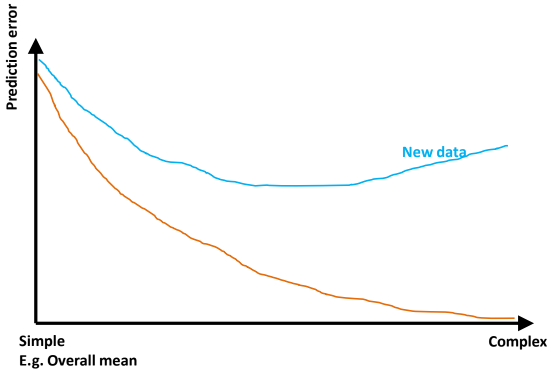


Optimal predictor = the best model

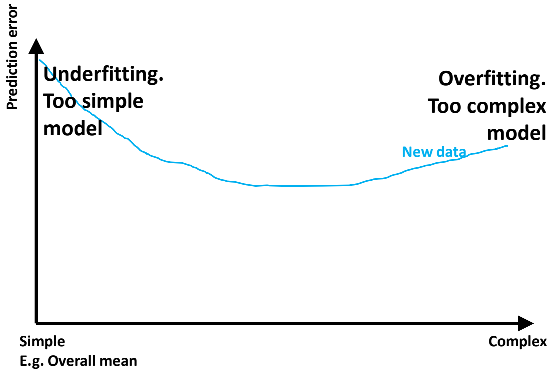
1) Define prediction error.
2) Consider model complexity
3) «Old», registered data (data used to develop prediction model)
   vs new (yet unknown) data



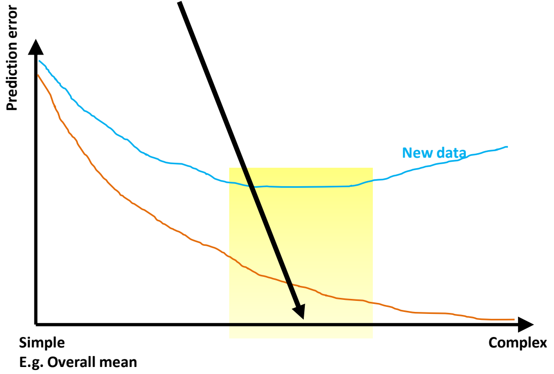
Optimal predictor = the best model



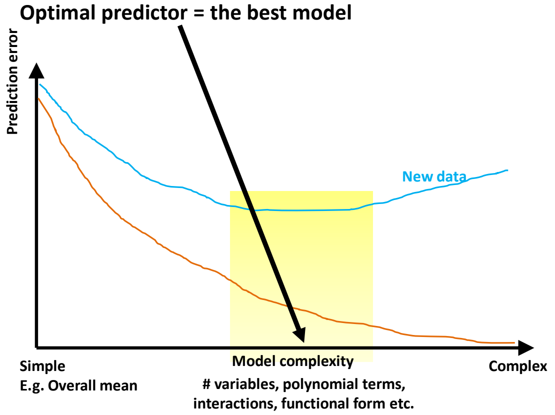
Optimal predictor = the best model



Optimal predictor = the best model



Optimal predictor = the best model



"Best model" dependent on how we define the error criterion, and how we weight/penalize bias and variance.

Variable selection?

No DAG to identify the roles of the variables; exposure, confounder, mediator etc.

Model selection rather than variable selection.

Not necessary to assume linearity by convenience for the interpretation of effect estimates, as prediction of future values is more important than interpretation of parameters.

How to obtain prediction error based on new data?

Divide the data set into training set (1/3) and test set (2/3)

Leave-one-out cross-validation

K-fold cross-validation

A brilliant tutorial is found at
http://www.autonlab.org/tutorials/overfit10.pdf

Cross-validation for detecting and preventing overfitting
Andrew W. Moore

"Best" model?

Variable selection
- (Forward/Backward/stepwise)
- Akaike's information criterion (AIC)
- Bayes' information criterion (BIC)
- Focused information criterion (FIC)

Variable shrinkage
- PCA
- Ridge regression — cf. Penalization in the curve fitting
- Lasso

Functional forms of variables?
Nonlinear models?

Input →
BLACK BOX
→ Output

What is inside the black box?

Regression analysis comes with a huge toolbox and add-ons.

Different regression tools and different approaches to (regression modelling) must be chosen for different scientific questions, even though the data set in the study «contains several variables to be included in the analysis».

Topics not covered

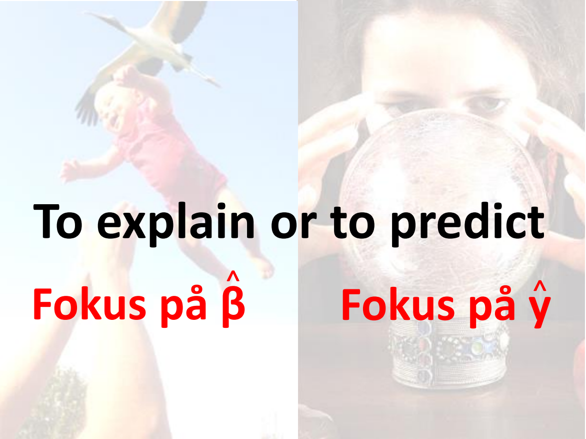
Model selection
- (Forward/Backward/stepwise)
- Akaike's information criterion (AIC)
- Bayes' information criterion (BIC)
- Focused information criterion (FIC)

Variable shrinkage
- PCA
- Ridge regression
- Lasso

Cross-validation

Discussion exercises

What are the main differences between
a regression analysis for the estimation of associations, and
a regression analysis for prediction purposes?



To explain or to predict

Fokus på $\hat{\beta}$

Fokus på $\hat{y}$

Artificial pancreas control system

$$\begin{cases}\dot{G}_p = -(k_2 + kp_2)G_p + k_1 G_t - U_{ii} - E(G_p) - kp_3 I_d + \frac{f k_{abs} Q_{gut}}{BW} + kp_1 \\ \dot{G}_t = -k_1 G_t + k_2 G_p - \frac{Vm_0 + Vm_X X}{Km_0 + G_t} G_t \\ \dot{G}_s = -\frac{1}{\tau_{SG}}\left(G_s - \frac{G_p}{V_g}\right) \\ \dot{I}_d = -k_i (I_d - I_1) \\ \dot{I}_1 = -k_i \left(I_1 - \frac{I_p}{V_i}\right) \\ \dot{I}_p = -(m_2 + m_4) I_p + m_1 I_l + k_{a1} I_{SQ1} + k_{a2} I_{SQ2} \\ \dot{I}_l = -(m_1 + m_3) I_l + m_2 I_p \\ \dot{X} = -p_{2u}\left(X - \frac{I_p}{V_i} + I_b\right) \\ \dot{I}_{SQ1} = -(k_{a1} + k_d) I_{SQ1} + J(t) \\ \dot{I}_{SQ2} = -k_{a2} I_{SQ2} + k_d I_{SQ2} \\ \dot{Q}_{sto1} = -k_{gri} Q_{sto1} + D(t) \\ \dot{Q}_{sto2} = -k_{empt}(Q_{sto1} + Q_{sto2}) Q_{sto2} + k_{gri} Q_{sto1} \\ \dot{Q}_{gut} = -k_{abs} Q_{gut} + k_{empt}(Q_{sto1} + Q_{sto2}) Q_{sto2}\end{cases}$$

Where $E(G_p) = \begin{cases} ke_1(G_p - ke_2) & \text{if } G_p > ke_2 \\ 0 & \text{otherwise}\end{cases}$

and $k_{empt}(Q_{sto1} + Q_{sto2}) = k_{min} + \frac{k_{max} - k_{min}}{2}\left(2 + \tanh\left(aa(Q_{sto1} + Q_{sto2} - b\,dose)\right) + \tanh\left(cc(Q_{sto1} + Q_{sto2} - d\,dose)\right)\right)$

$aa = \frac{2.5\,dose}{1-b} \quad cc = \frac{2.5\,dose}{d}$

Patent application WO 2008157780 A1

Calculate optimal insulin dose by defining insulin injection as the linear combination of gain and state, which minimize a quadratic cost function.

Compute *J(q)* as:

$$J(q) = \frac{1}{100}\sum_{i=1}^{100} t2tgt(q)_i + tbtgt(q)_i$$

The optimal parameter *q** is defined as :

$$q^* = \underset{q}{\operatorname{argmin}}\, J(q)$$

To explain and to predict