Statistical methods in medical research

Arnoldo Frigessi, frigessi@medisin.uio.no

2.11.2009

Use of statistics in medical journals (Sept. 2009).

		Statistics
The Lancet	Paper 1	Repeated measures, Survival (Cox)
	Paper 2	Repeated measures
	Paper 3	Survival (Cox), logistic
	Paper 4	Survival (Cox)
NEJM	Paper 1	Survival (Cox)
	Paper 2	Survival (Cox)
	Paper 3	More simple statistics
	Paper 4	Survival (Cox)

JAMA	Paper 1	Simple statistics
	Paper 2	Simple statistics
	Paper 3	Repeated measures
	Paper 4	Clustered logistic, negative binomial
	Paper 5	Simple statistics
Nature Medicine	Paper 1	Linear regression
	Paper 2	Mann-Whitney, Kaplan-Meier
	Paper 3	Mann-Whitney
	Paper 4	T-test, ANOVA
	Paper 5	T-test, chi-square

The American Statistician

2007, vol. 61, no. 1, pp. 47 - 55

The Use of Statistics in Medical Research: A Comparison of *The New England Journal of Medicine* and *Nature Medicine*

Alexander M. Strasak; Qamruz Zaman; Gerhard Marinell; Karl P. Pfeiffer; Hanno Ulmer

Abstract

There is widespread evidence of the extensive use of statistical methods in medical research. Just the same, standards are generally low and a growing body of literature points to statistical errors in most medical journals. However, there is no comprehensive study contrasting the top medical journals of basic and clinical science for recent practice in their use of statistics.

All original research articles in Volume 10, Numbers 1-6 of *Nature Medicine* (*Nat Med*) and Volume 350, Numbers 1-26 of *The New England Journal of Medicine* (*NEJM*) were screened for their statistical content. Types, frequencies, and complexity of applied statistical methods were systematically recorded. A 46-item checklist was used to evaluate statistical quality for a subgroup of papers.

94.5 percent (95% CI 87.6-98.2) of *NEJM* articles and 82.4 percent (95% CI 65.5-93.2) of *Nat Med* articles contained inferential statistics. *NEJM* papers were significantly more likely to use advanced statistical methods ($p < 0.0001$). Statistical errors were identified in a considerable proportion of articles, although not always serious in nature. Documentation of applied statistical methods was generally poor and insufficient, particularly in *Nat Med*. Compared to 1983, a vast increase in usage and complexity of statistical methods could be observed for *NEJM* papers. This does not necessarily hold true for *Nat Med* papers, as the results of the study indicate that basic science sticks with basic analysis. As statistical errors seem to remain common in medical literature, closer attention to statistical methodology should be seriously considered to raise standards.

Why is statistics so important in medical research?

Two obvious reasons:

- Mechanistic understanding is (still) limited. Must trust observations / data

- We produce an enormous amount of data

Plan

- Some basic concepts in statistics: H1A1 flue
- Testing hypothesis and p-values: genomics
- Regression: observational studies
- Oppgave
- Discussion

Pandemi

- myndighetenes nettside om pandemisk influensa

Søk

Kontakt English

Pandemi:

- Begrense smitte
- Føler du deg syk?
- Pleie av syke hjemme
- Risikogrupper
- Spørsmål og svar
- Informasjonsmateriell
- Skole og barnehage
- Aktuelt
- Helsesektoren
- Planlegging
- Presse
- Lenker

Pandemi

På dette nettstedet vil helsemyndighetene gi løpende informasjon til befolkningen og helsetjenesten om **influensa A (H1N1)**. Her finner du **råd om hva du kan gjøre** for å **begrense smitte** og hva du bør gjøre dersom du tror du **kan være smittet**.

Skole og barnehage | Sykdom og symptomer | Risiko-grupper | Vaksine

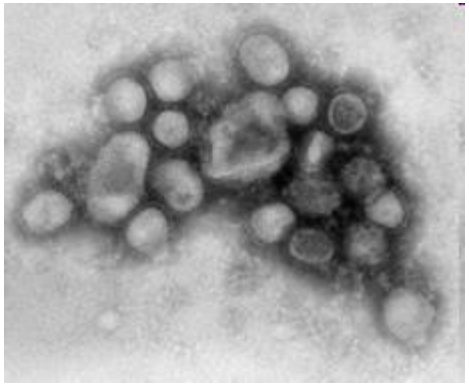
Aktuelt

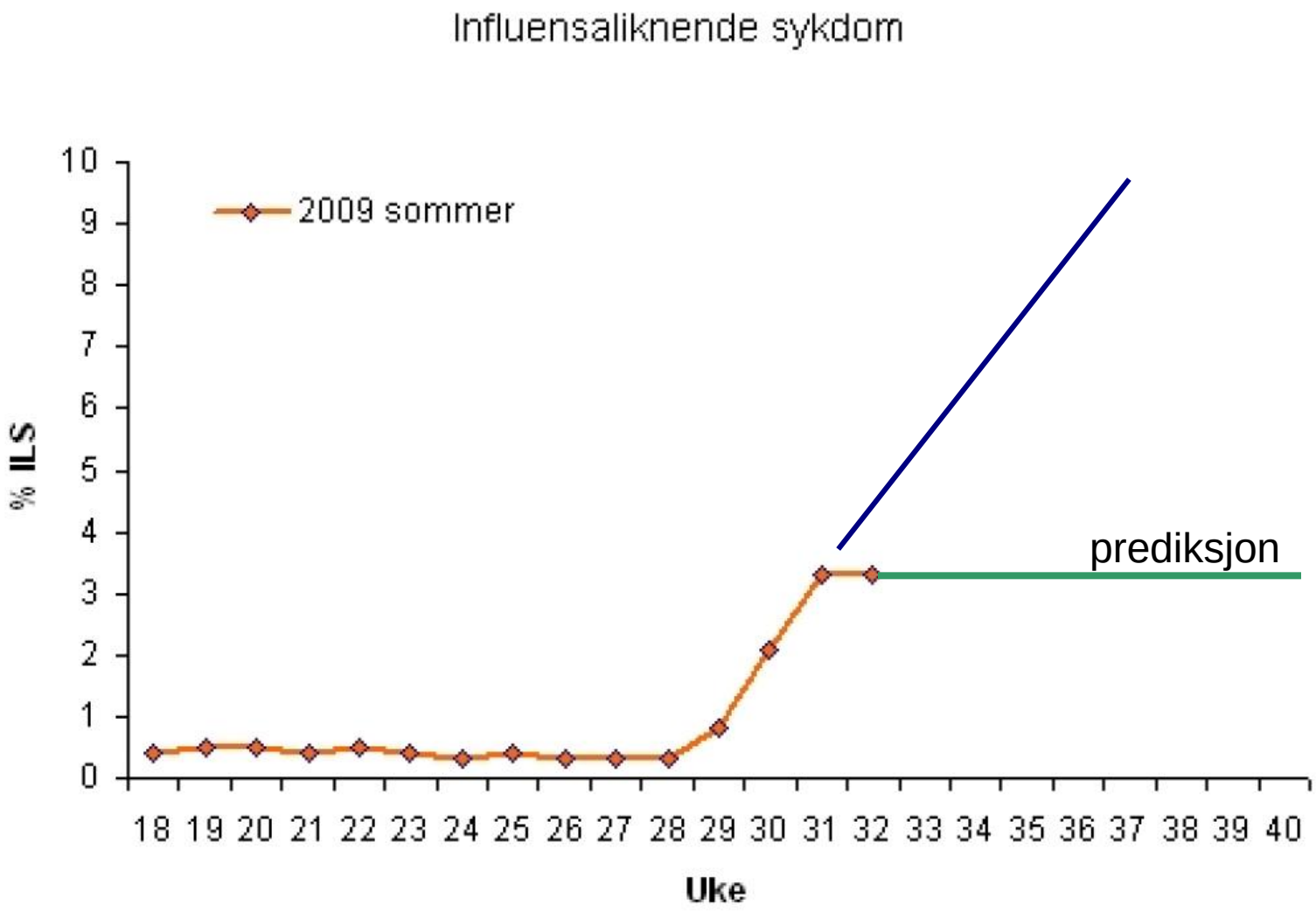
Oppstart av høstsemesteret i undervisningssektoren

[11.08.09] Helsemyndighetene anbefaler vanlig oppstart av skole og barnehage for alle, slik situasjonen er nå. Dette rådet gjelder også dem som tilhører risikogruppene.

Relatert

- Publikumstelefonen
- Status nå
- Planlegging for en pandemi





I uke 32 fikk 3,3 % av dem som gikk til legen diagnosen ”influensaliknende sykdom”.

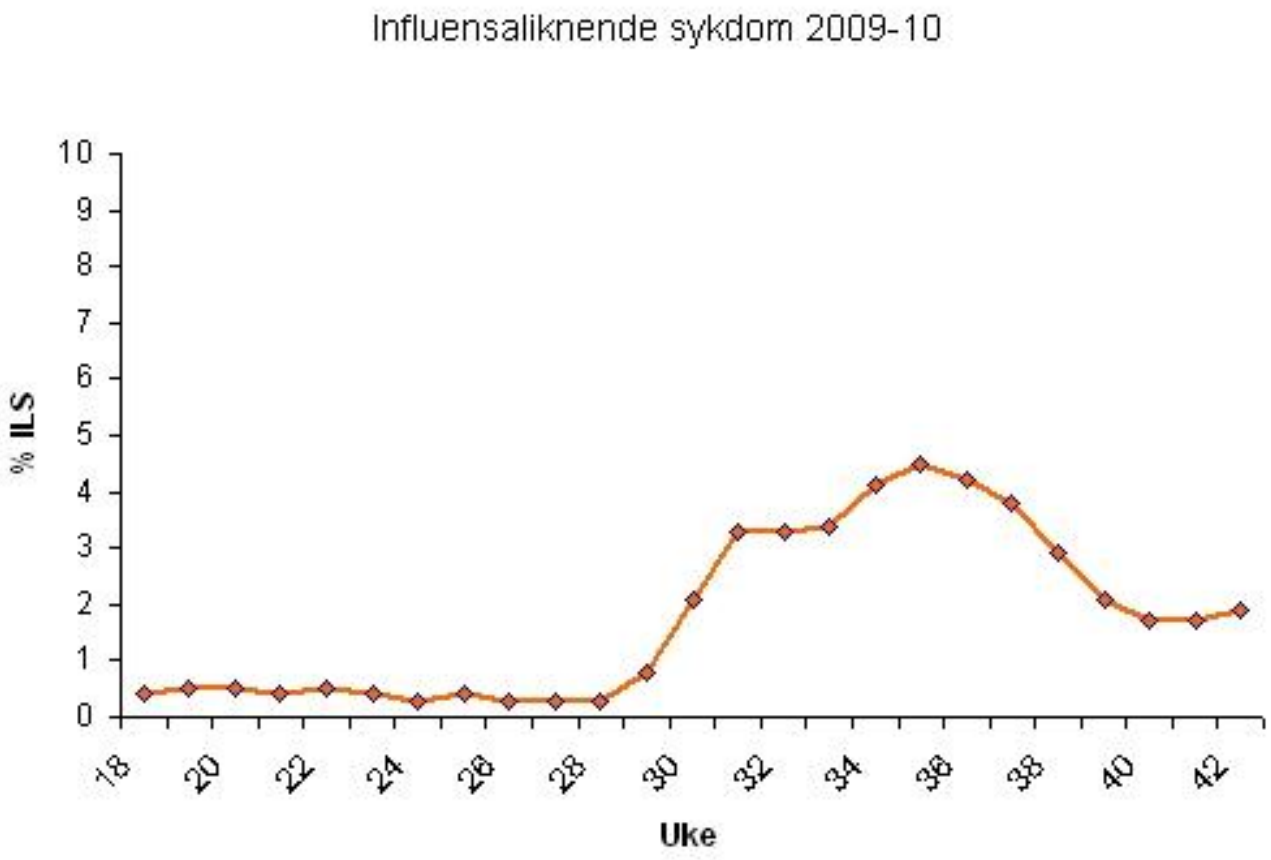
Øker …..

Men:

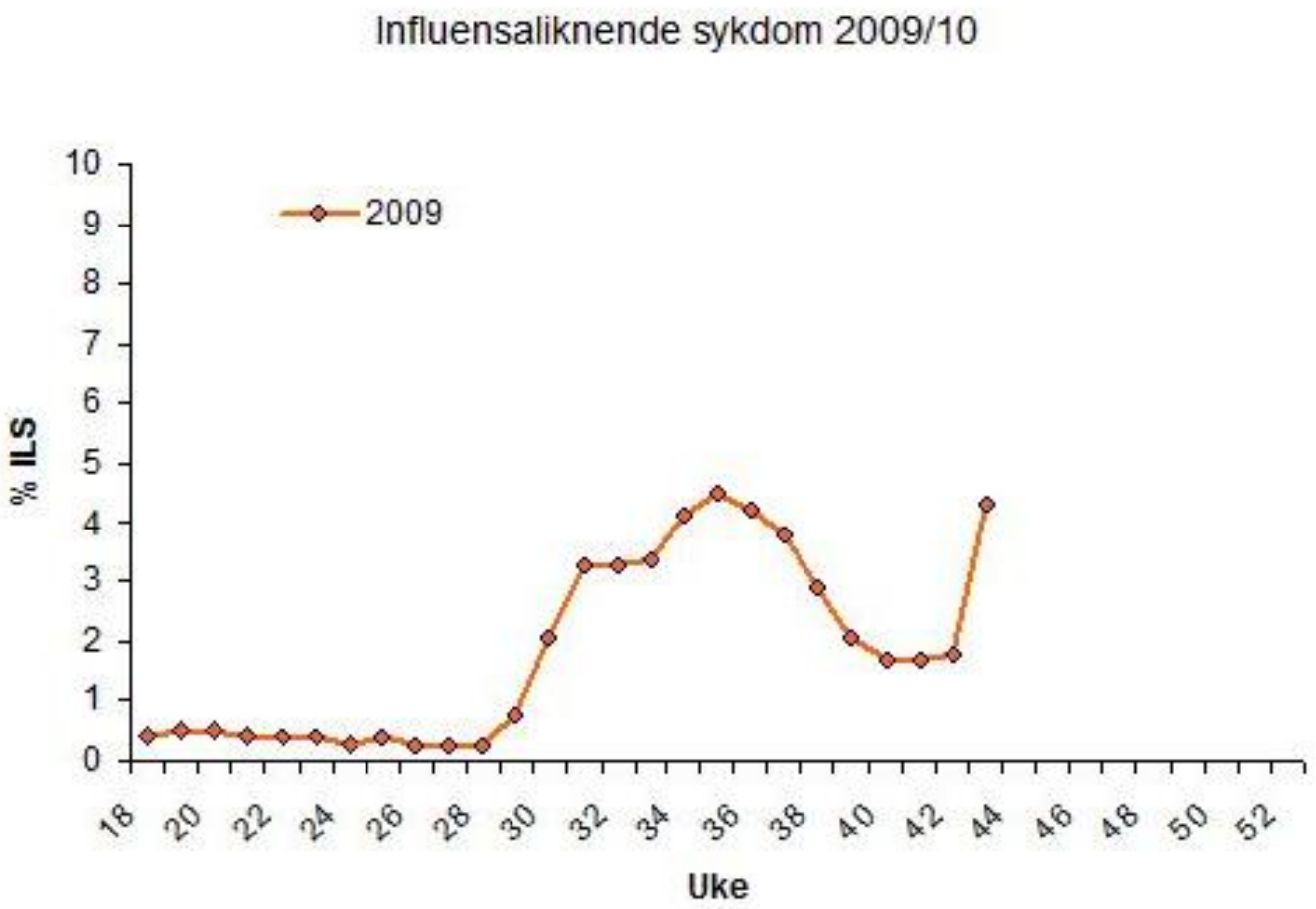
Den økte oppmerksomheten rundt epidemien medfører at flere enn normalt oppsøker lege på grunn av influensaliknende symptomer og det økte antallet slike konsultasjoner er derfor ikke en klar indikasjon på omfattende smitte innenlands.

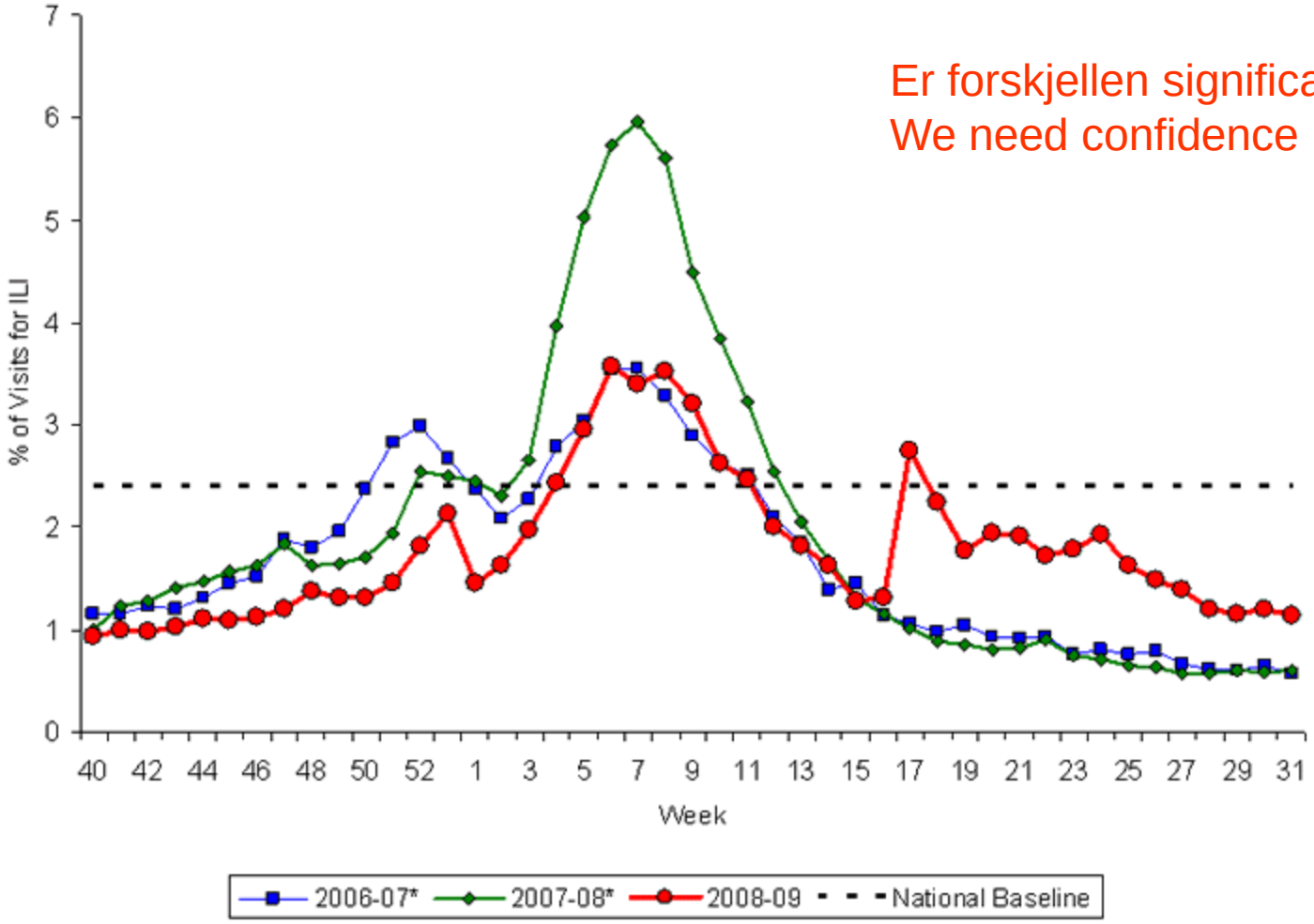
(konfundering)

Andel legekonsultasjoner (%) hvor diagnose ”influensaliknende sykdom” (ILS) ble satt

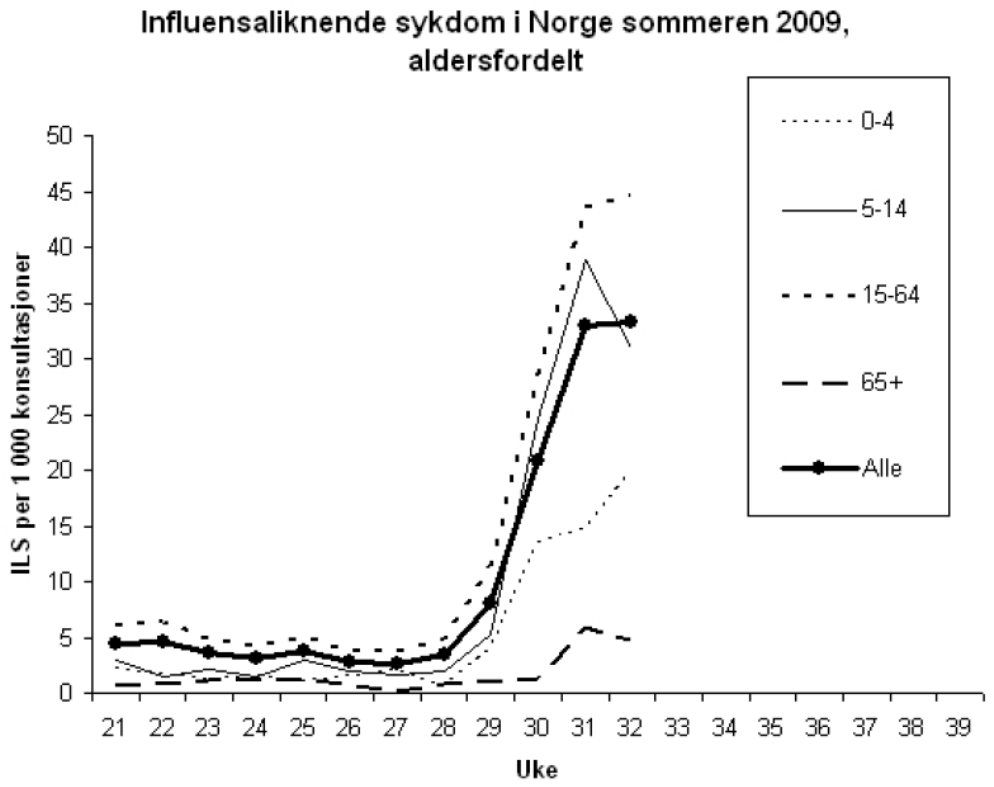


Kilde: FHIs vakttårnsystem (201 legekontorer over hele landet).


Influensaliknende sykdom 2009/10
2009
% ILS
Uke
0
1
2
3
4
5
6
7
8
9
10
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52



*There was no week 53 during the 2006-07 and 2007-08 seasons, therefore the week 53 data point for those seasons is an average of weeks 52 and 1.



Aldersforskjell?

Figur 2. Influensaliknende sykdom (ILS) i Norge. Grafen viser aldersfordelingen på dem som har fått diagnosen "influensaliknende sykdom", angitt per 1000 konsultasjoner per uke.

Aldersforskjeller:

Foreløpig er de fleste pasientene barn og unge voksne (mange i 20-årene).

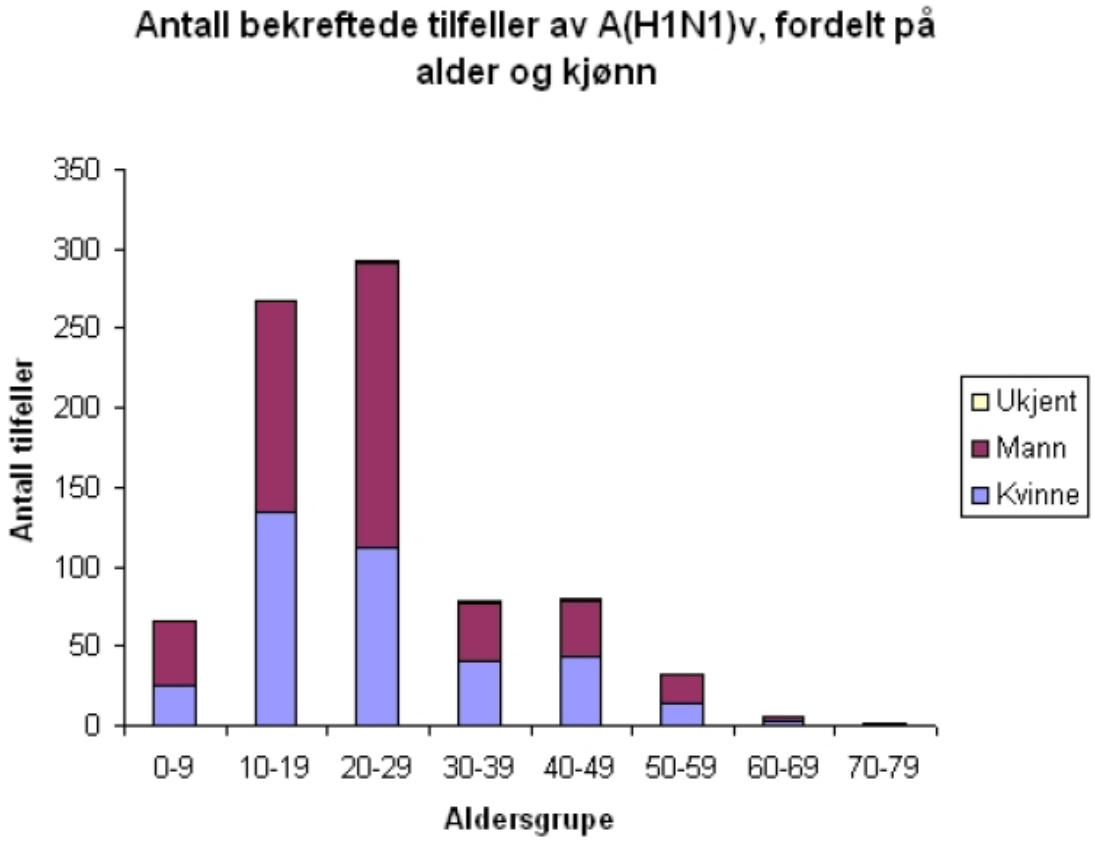
Median alder for bekreftede tilfeller har vært 12-17 år.

I USA er 60 % av pasientene mellom 5 og 24 år.

Det kan skyldes

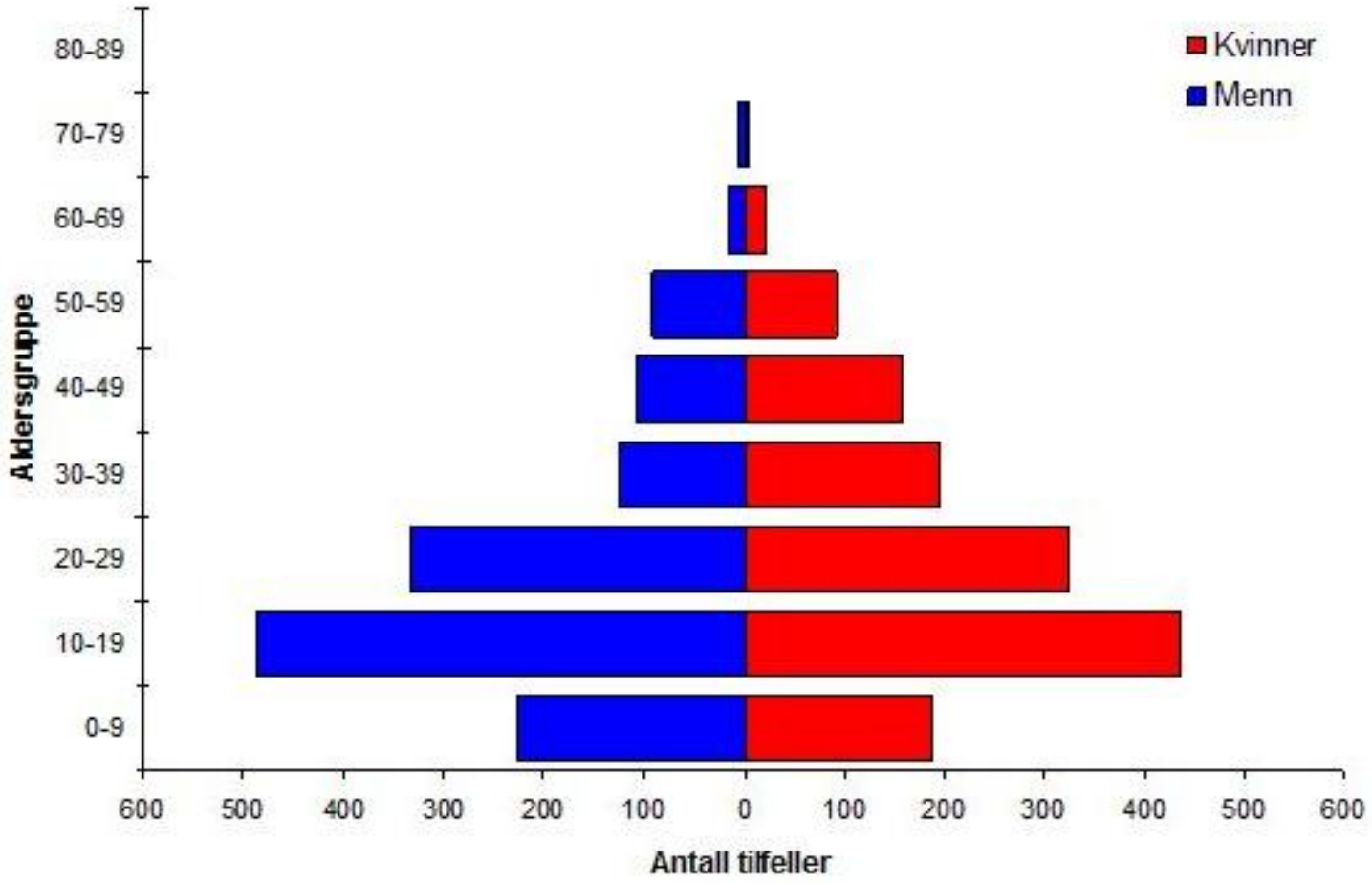
a) at det er de unge som er eksponert, for eksempel ved reiser,
b) at eldre har noe restimmunitet fra tidligere sesonger
c) at det er noe med dette viruset som gjør det mer sykdomsfremkallende for unge
d) at unge oftere får tatt prøve og dermed blir bekreftede tilfeller som registreres

Assosiasjon vs. Kausalitet



Mer mann enn kvinner?

Test for forskjell?


Kvinner
Menn
80-89
70-79
60-69
50-59
40-49
30-39
20-29
10-19
0-9
Aldersgruppe
600
500
400
300
200
100
0
100
200
300
400
500
600
Antall tilfeller

Planleggingsgrunnlag

Følgende verdier benyttes nå i planleggingen av beredskapen mot ny influensa A(H1N1):

Estimat

Variabel	Anslag	Usikkerhet
Andel av befolkningen som blir syk	30 %	10-50
Gjennomsnittlig sykdomsvarighet	7 dager	5-10
Andel av syke som søker lege	30 %	10-40
Andel av syke som ber om ativiralia	20 %	10-70
Andel av syke som legges inn i sykehus	1 %	0,5-2
Gjennomsnittlig innleggelsesvarighet	5 dager	2-7
Andel av innlagte som trenger intensivbehandling	20 %	10-30
Gjennomsnittlig oppholdstid i intensivavdeling	12 dager	5-20
Andel syke som dør, letalitet	0,05 %	0,1-0,01 %

Konfidensintervall

Smittsomhet:

Reproduksjonsantallet: Antallet nye smittede som en smittet gir opphav til. Dersom en gjennomsnittspasient smitter fire andre personer, sier vi at reproduksjonsantallet er 4. Da vil epidemien vokse raskt

Reproduksjonstallet

$R_0 < 1$ - infeksion er ikke epidemisk
$R_0 > 1$ - infeksion er epidemisk
værre hvis stor!

Andelen i populasjonen som trengs å vaksineres for å sikre immunitet i befolkning og hindre spredning er $1 - 1/R_0$

Values of R_0 of well-known infectious diseases

Disease	Transmission	R_0
Measles	Airborne	12-18
Pertussis	Airborne droplet	12-17
Diphtheria	Saliva	6-7
Smallpox	Social contact	5-7
Polio	Fecal-oral route	5-7
Rubella	Airborne droplet	5-7
Mumps	Airborne droplet	4-7
HIV/AIDS	Sexual contact	2-5
SARS	Airborne droplet	2-5
Influenza (1918 pandemic strain)	Airborne droplet	2-3

Reproduksjonstallet R_0 for H1N1:

Det anslås at en gjennomsnittspasient gir opphav til 1,4 – 3,5 nye pasienter, de fleste anslag rundt 1,5-1,7.

(Sesonginfluensa har en R_0 på 1,2-1,4.)

Anta $R_0 = 2$, da trenger vi å vaksinere $1 - 1/R_0 = 50\%$ av befolkning.

Estimering av reproduksjonstallet R_0 :

β -- Contact Rate (kontaktsansynlighet)
N -- Total Population (størrelse av populasjon)
1/γ -- Average Infectious Period (infektiviteteperiode i dager)

$R_0 = (\beta N)/\gamma$

fordi en smitsom individ har βN kontakter per dag og er smittsom for ca. 1/γ dager.

Vi må estimere β og 1/γ fra data!

Originally published in *Science* Express on 11 May 2009
Science 19 June 2009:
Vol. 324. no. 5934, pp. 1557 - 1561
DOI: 10.1126/science.1176062

REPORTS

Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings

Christophe Fraser,[1,*] Christl A. Donnelly,[1,*] Simon Cauchemez,[1] William P. Hanage, Maria D. Van Kerkhove,[1] T. Déirdre Hollingsworth,[1] Jamie Griffin,[1] Rebecca F. Baggaley,[1] Helen E. Jenkins,[1] Emily J. Lyons,[1] Thibaut Jombart,[1] Wes R. Hinsley,[1] Nicholas C. Grassly,[1] Francois Balloux,[1] Azra C. Ghani,[1] Neil M. Ferguson,[1,†] Andrew Rambaut,[2] Oliver G. Pybus,[3] Hugo Lopez-Gatell,[4] Celia M. Alpuche-Aranda,[5] Ietza Bojorquez Chapela,[4] Ethel Palacios Zavala,[4] Dulce Ma. Espejo Guevara,[6] Francesco Checchi,[7] Erika Garcia,[7] Stephane Hugonnet,[7] Cathy Roth,[7]

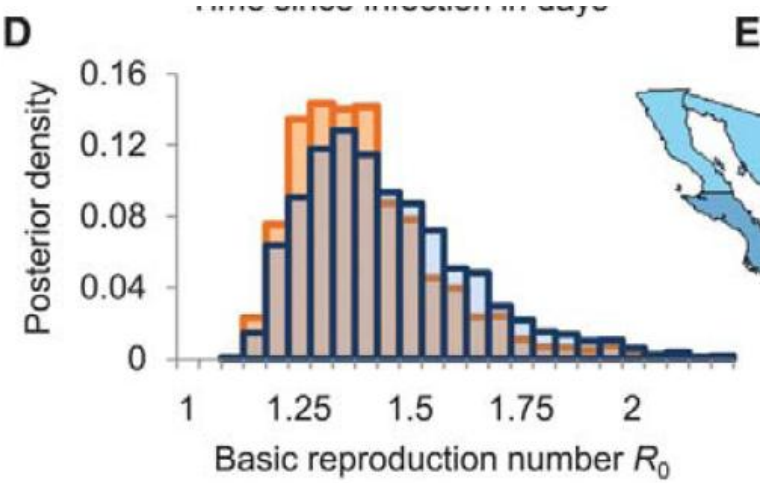
The WHO Rapid Pandemic Assessment Collaboration[‡]

By analyzing the outbreak in Mexico, early data on international spread, and viral genetic diversity, we make an early assessment of transmissibility and severity.

Our estimates suggest that 23,000 (range 6000 to 32,000) individuals had been infected in Mexico by late April, giving an estimated case fatality ratio (CFR) of 0.4% (range: 0.3 to 1.8%) based on confirmed and suspected deaths reported to that time.

Punktestimat og konfidensintervall

Three different epidemiological analyses gave basic reproduction number (R_0) estimates in the range of 1.4 to 1.6.



Characteristics of hospitalised cases

The proportion of hospitalised cases is significantly lower among patients who received Tamiflu prophylaxis compared with those who did not, RR=0.55 (0.34-0.89).

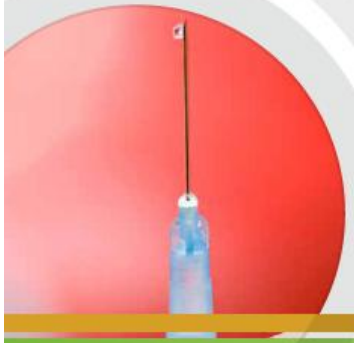
Relativ risiko (RR) er et forholdstall som angir hvor mye større sannsynlighet det er for en hendelse i én gruppe i forhold til en annen.

F.eks.

Gruppe A (med Tamiflu) vs. gruppe B (ikke Tamiflu).

Hvis RR for hospitalisering er 0.55, betyr det at det er ca. halvparten så sannsynlig at en person som bruker Tamiflu skal til sykehus, som en person som ikke får Tamiflu.

Konfidensintervall for RR = (0.34-0.89).



TECHNICAL DOCUMENT

Protocol for case-control studies to measure influenza vaccine effectiveness in the European Union and European Economic Area Member States

www.ecdc.europa.eu

2 Objectives

2.1 Primary objectives

The primary objectives are to:

- measure seasonal influenza vaccine effectiveness among people aged 65 years and above in EU/EEA countries; and
- measure pandemic influenza vaccine effectiveness in the target groups.

3 Methods

3.1 Study design

- Case-control study in each participating country, with various sets of controls.
- Multicentre case-control study in several countries, with various sets of controls.

Table 2: Sample size calculations

Power	Alpha	Controls/ case	Vaccine coverage in source population/controls	Detectable OR	Number of cases	Number of controls
0.90	0.05	1	0.5	0.6	345	345
0.80	0.05	1	0.5	0.6	262	262
0.90	0.05	1	0.5	0.5	194	194
0.80	0.05	1	0.5	0.5	148	148
0.90	0.05	1	0.5	0.4	116	116
0.80	0.05	1	0.5	0.4	89	89
0.90	0.05	1	0.5	0.3	72	72
0.80	0.05	1	0.5	0.3	56	56
0.90	0.05	1	0.6	0.6	341	341
0.80	0.05	1	0.6	0.6	259	259
0.90	0.05	1	0.6	0.5	188	188

The New York Times

A New Wave of Flu Could Be More Fatal for Europe

By JAMES KANTER and MATTHEW SALTMARSH
Published: August 17, 2009

BRUSSELS — Anxiety over a new flu strain may have eased over the summer, but millions of Europeans will soon receive a sharp reminder of its virulence as governments prepare for a large-scale vaccination effort aimed at keeping a second — and possibly deadlier — wave of infections at bay.

With another surge in cases of the H1N1 virus — initially known as "swine flu" — expected as soon as September, medical experts say the battle to tame the first pandemic flu in four decades is just getting under way in Europe.

NRK

Nordland

18.08.2009

Svineinfluensaen kan være på retur

SYKE: I begynnelsen av denne uka var det meldt inn 791 tilfeller av bekreftet svineinfluensa i Norge.

Foto: Trine Hamran

Fagdirektør Stein Tore Nilsen ved Stavanger universitetssjukehus tror svineinfluensaen er mindre omfattende enn fryktet.

NTB

Publisert 15.08.2009 16:42.

Published 10 August 2009, doi:10.1136/bmj.b3172
Cite this as: BMJ 2009;339:b3172

Research

Neuraminidase inhibitors for treatment and prophylaxis of influenza in children: systematic review and meta-analysis of randomised controlled trials

Matthew Shun-Shin, *academic foundation year 2 doctor*[1], **Matthew Thompson**, *senior clinical scientist*[2], **Carl Heneghan**, *clinical lecturer*[2], **Rafael Perera**, *university lecturer in medical statistics*[2], **Anthony Harnden**, *university lecturer in general practice*[2], **David Mant**, *professor of general practice*[2]

[1] Kadoorie Centre, John Radcliffe Hospital, Headington, Oxford OX3 9DU, [2] Oxford University Department of Primary Health Care, Rosemary Rue Building, Headington, Oxford OX3 7LF

Conclusions Neuraminidase inhibitors provide a small benefit by shortening the duration of illness in children with seasonal influenza and reducing household transmission. They have little effect on asthma exacerbations or the use of antibiotics. Their effects on the incidence of serious complications, and on the current A/H1N1 influenza strain remain to be determined.

WHO Emphasizes that Children and Adults with Severe H1N1 Disease Should Get Antivirals

JW

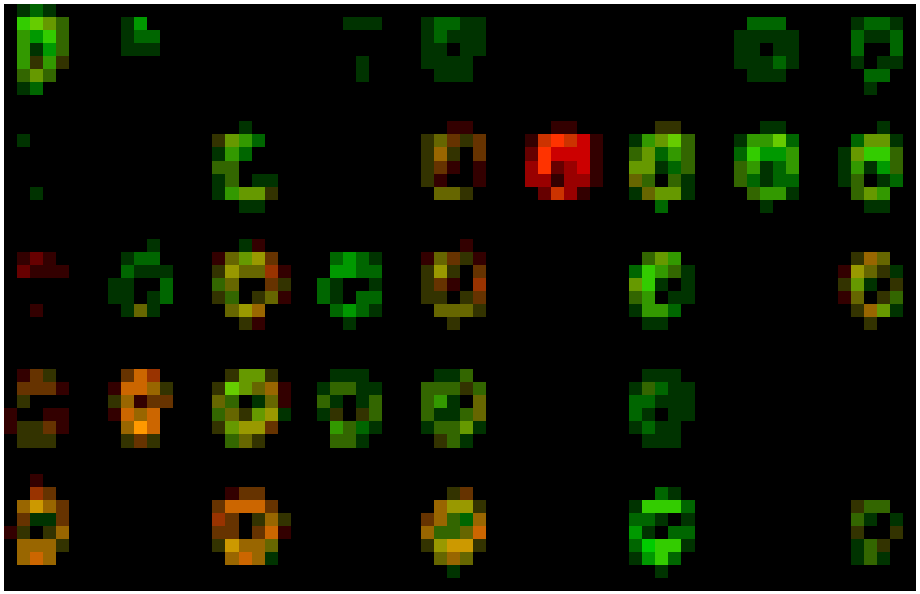
August 14th, 2009

The World Health Organization is **reiterating its position** that antivirals are appropriate for adults and children over age 1 who have severe novel H1N1 disease or who are at risk for H1N1 complications.

The WHO took this step in response to a ***BMJ* analysis**, published earlier this week, calling into question the wisdom of giving antivirals to children with mild seasonal influenza.

Pause!

Hypothesis testing in the era of genomic data



Outline

- Genomics and microarray data
- Finding differential expressed genes
- Two-sample comparisons (many!)
- Multiple testing: problems and solutions

What you will learn:

statistics:

- recap hypothesis testing
- classical t-test
- p-value
- type 1 and type 2 errors
- learn that there are assumptions behind using Student t distribution
- learn an alternative (permutation) which will make clear that:
    -- there is freedom in inventing useful new tests
    -- and we can compute p-values for them
- learn about multiple testing issue, and simple remedies

genomic data:

- introduction to microarray data and modern high throughput genomics

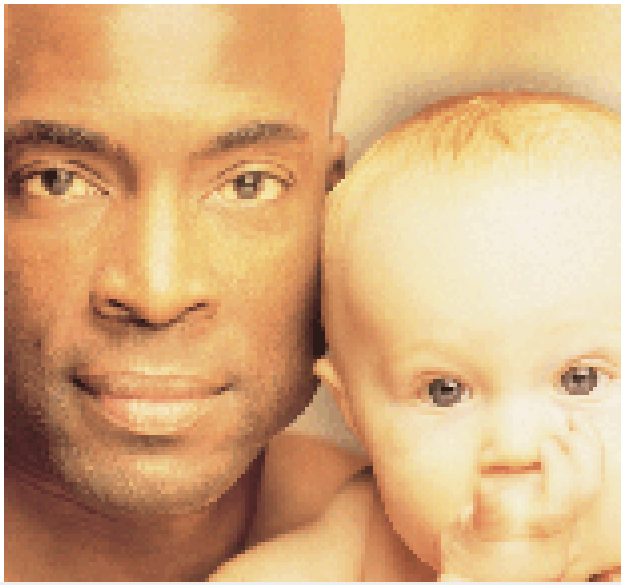
February, 2001:

Science magazine

Table of Contents

16 February 2001
Volume 291
Number 5507

The Human Genome



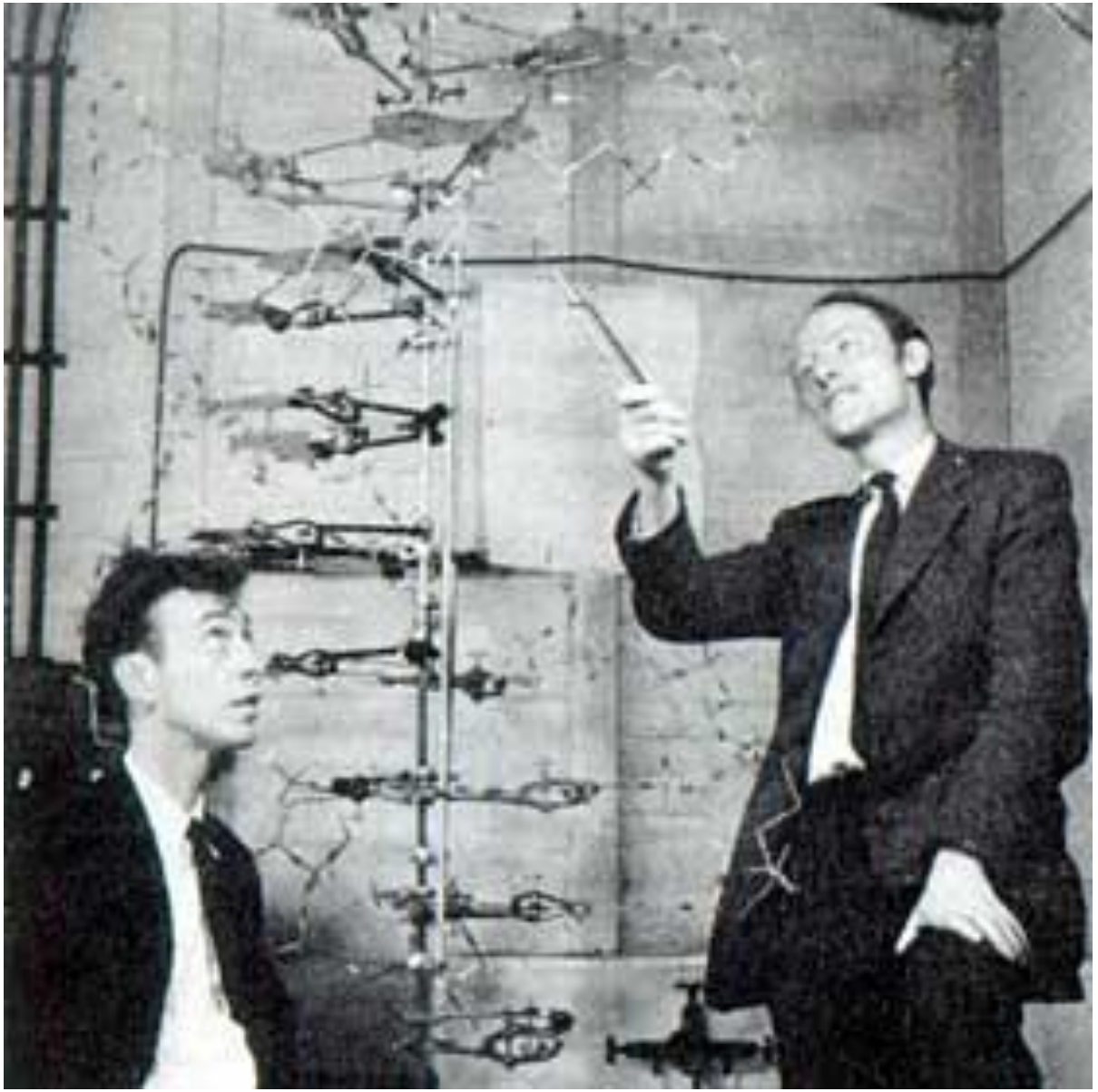
Aims:

1. Identify all genes in human DNA.
2. Store this information in databases.

Watson and Crick

and the

DNA molecule model



April 25, 1953: James Watson and Francis Crick's classic paper that first describes the double helical structure of DNA. They note that the structure "suggests a possible copying mechanism for the genetic material".

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., 11 (3) (1950).
[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik.* (*Stockholm*), **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without

is a residue on each chain every 3·4 A. in the *z*-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical *z*-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

Human Genome

- 3.4×10^9 bp long
- ~ 35,000 genes
- Average size of a gene ~3000 bp

Genes at work!

- Genes are inside cells
- Genes contribute to the production of proteins, which are building blocs (hormones, enzymes, etc)
- Proteins move between cells and produce effects

“Central Dogma of Molecular Biology”

- mRNA – single stranded RNA molecule
- Complementary to DNA
- Processed (spliced and polyadenylated) RNA transcript
- Carries the sequence of a gene out of the nucleus into the cytoplasm where it can be translated into a protein structure

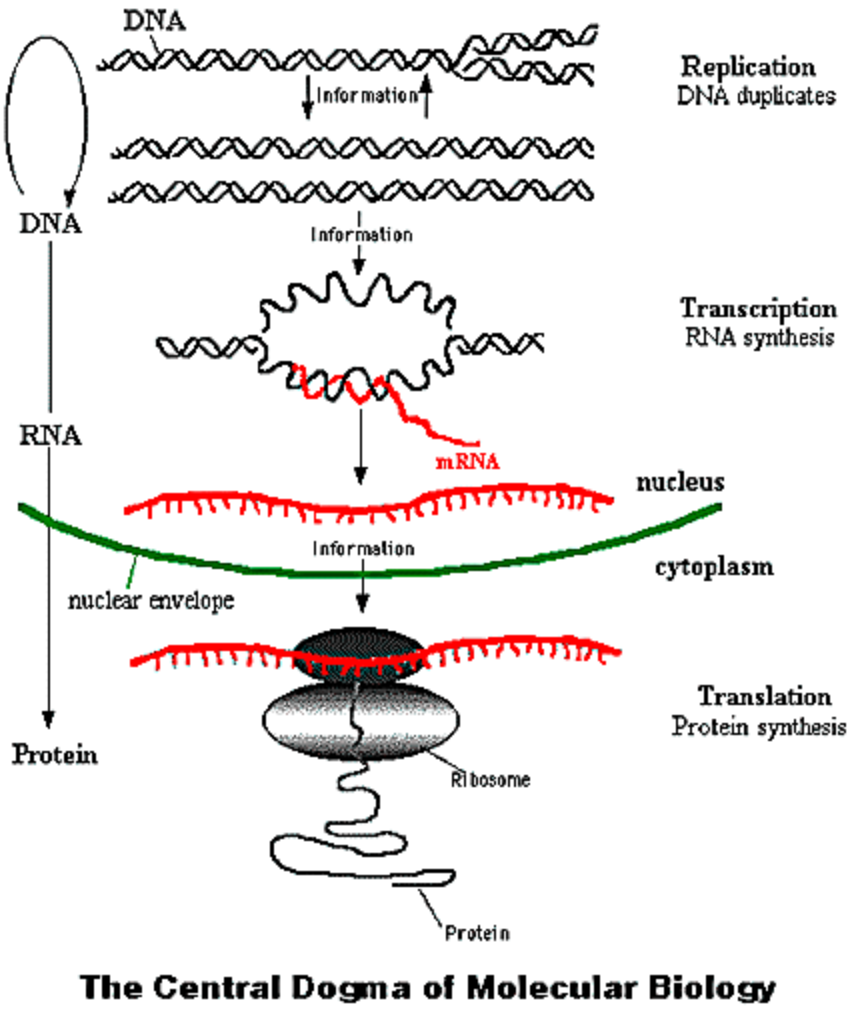


Image: Access Excellence, National Institutes of Heath

Microarray data and gene expressions

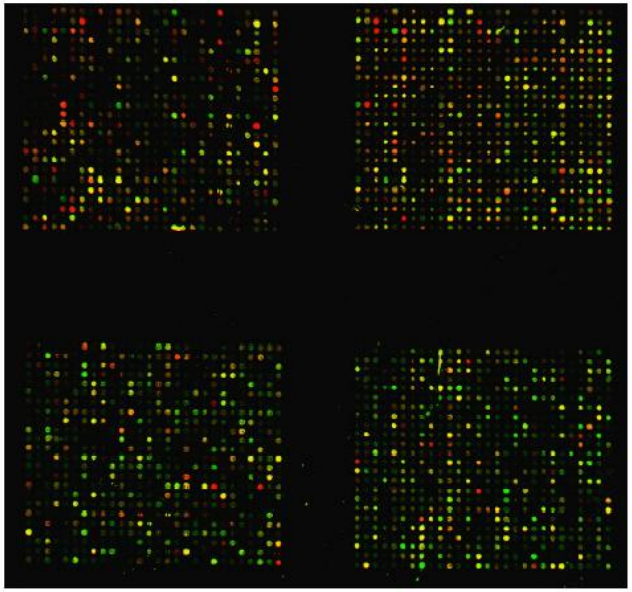
DNA

↓ transcription

mRNA

↓ translation

protein → cell phenotype → organism phenotype



- Microarrays measure gene expression at the transcription level
- Gene expression is a measure of how much a gene transcribes
- Gene expressions tell how much a gene might contribute to biological dynamics

What Can Be Done With Gene Expression Data?

- Identify genes associated with a biological state of interest
- Group genes with a similar pattern of behaviour
- Derive a biological pathway, a netrwork of genes jointly responsible for a biological dynamics

23andMe genetics just got personal.

When you purchase our complete service, this is what you'll get:

Your risk analyzed for 116 diseases and traits, including:

- Breast Cancer
- Rheumatoid Arthritis
- Type 2 Diabetes

See our full list of reports »

Your ancestral path, based on your DNA, in amazing detail:

- Ancestry Painting
- Global Similarity
- Maternal and Paternal Ancestry

See our full line of Ancestry features »

buy a kit

Or get a free account.

Our service starts with us mailing you a saliva collection kit.

Find a disease or trait that we cover:

Select a Disease or Trait

Popular Topics:

- Type 2 Diabetes
- Rheumatoid Arthritis
- Psoriasis
- Breast Cancer
- Colorectal Cancer
- Prostate Cancer
- Celiac Disease
- Crohn's Disease
- Hemochromatosis
- Restless Legs Syndrome
- Age-related Macular Degeneration
- Parkinson's Disease
- Coumadin® / Warfarin Sensitivity
- Plavix® Efficacy

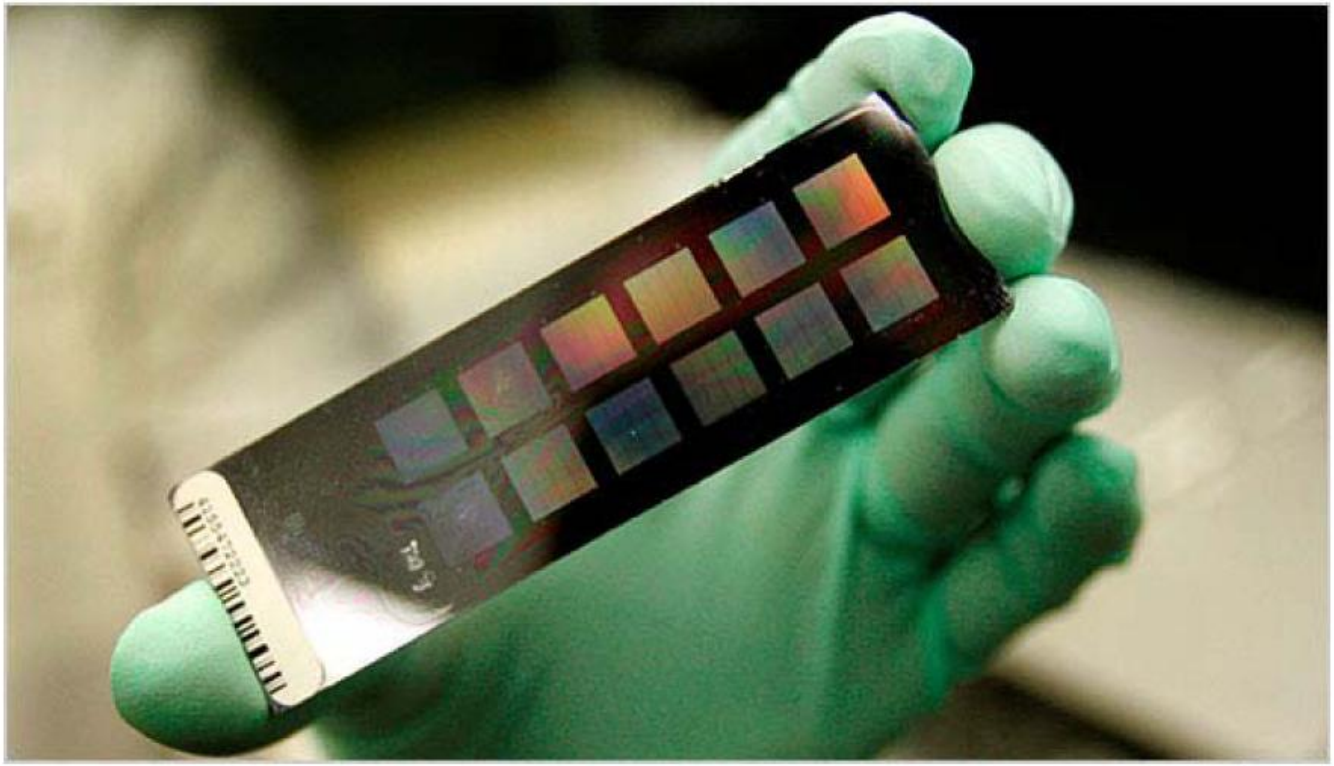
Browse all 116 health and traits topics »

https://www.23andme.com/

The New York Times

PERSONAL HEALTH

Buyer Beware of Home DNA Tests



Rob Carr/Associated Press

A slide used for DNA testing.

By JANE E. BRODY
Published: August 31, 2009

The New England
Journal of Medicine

VOLUME 347 DECEMBER 19, 2002 NUMBER 25

A GENE-EXPRESSION SIGNATURE AS A PREDICTOR OF SURVIVAL IN BREAST CANCER

MARC J. VAN DE VIJVER, M.D., PH.D., YUDONG D. HE, PH.D., LAURA J. VAN 'T VEER, PH.D., HONGYUE DAI, PH.D., AUGUSTINUS A.M. HART, M.SC., DORIEN W. VOSKUIL, PH.D., GEORGE J. SCHREIBER, M.SC., JOHANNES L. PETERSE, M.D., CHRIS ROBERTS, PH.D., MATTHEW J. MARTON, PH.D., MARK PARRISH, DOUWE ATSMA, ANKE WITTEVEEN, ANNUSKA GLAS, PH.D., LEONIE DELAHAYE, TONY VAN DER VELDE, HARRY BARTELINK, M.D., PH.D., SJOERD RODENHUIS, M.D., PH.D., EMIEL T. RUTGERS, M.D., PH.D., STEPHEN H. FRIEND, M.D., PH.D., AND RENÉ BERNARDS, PH.D.

The New York Times
ON THE WEB

Health

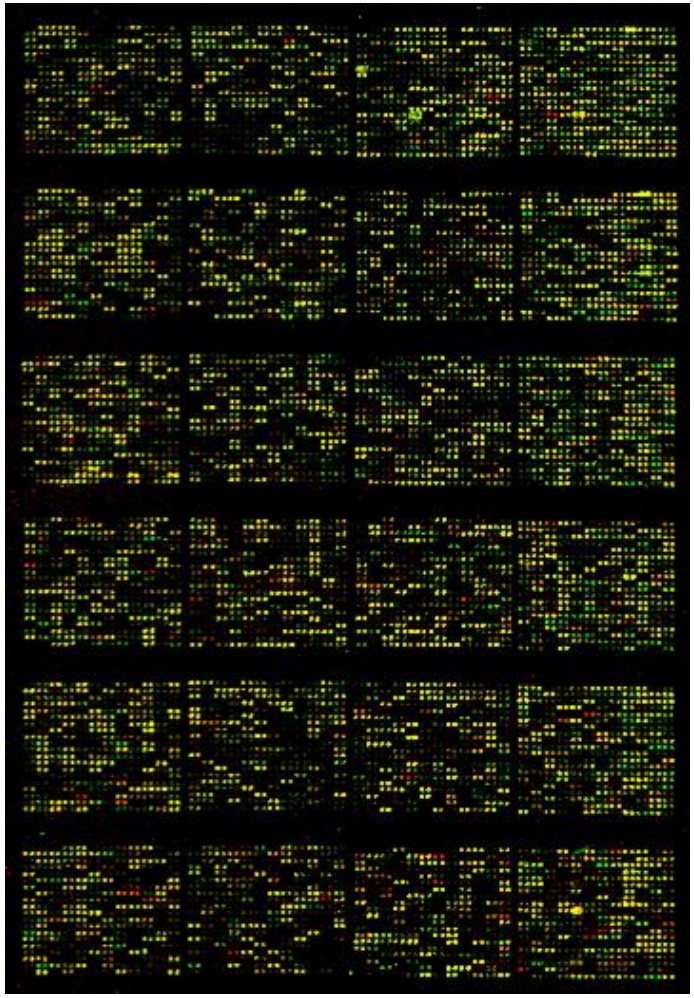
December 19, 2002

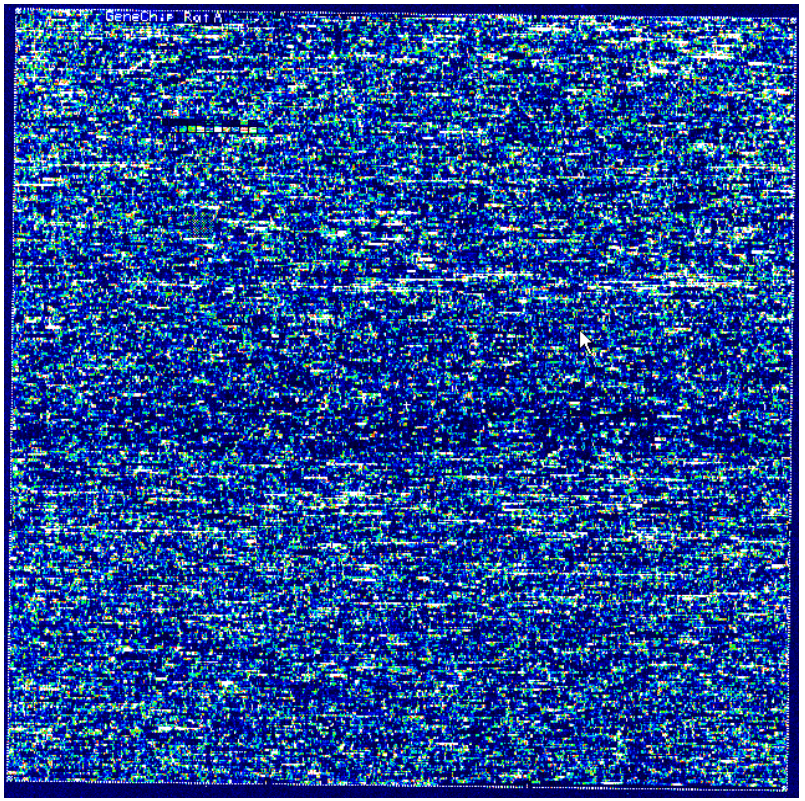
Breast Cancer: Genes Are Tied To Death Rates

Researchers find genetic signature in breast tumors that seems to be powerful predictor of whether cancer will spread and kill or whether it can easily be cured by surgery; study also indicates that it may soon be possible to make distinction between women who need more aggressive therapy and those who do not; findings are reported in The New England Journal of Medicine.

Microarrays (DNA Chips)

- Enables monitoring of expression levels for thousands of genes simultaneously.
- knowledge derived may be used for drug development and gene therapy.
- There are many microarray technologies.





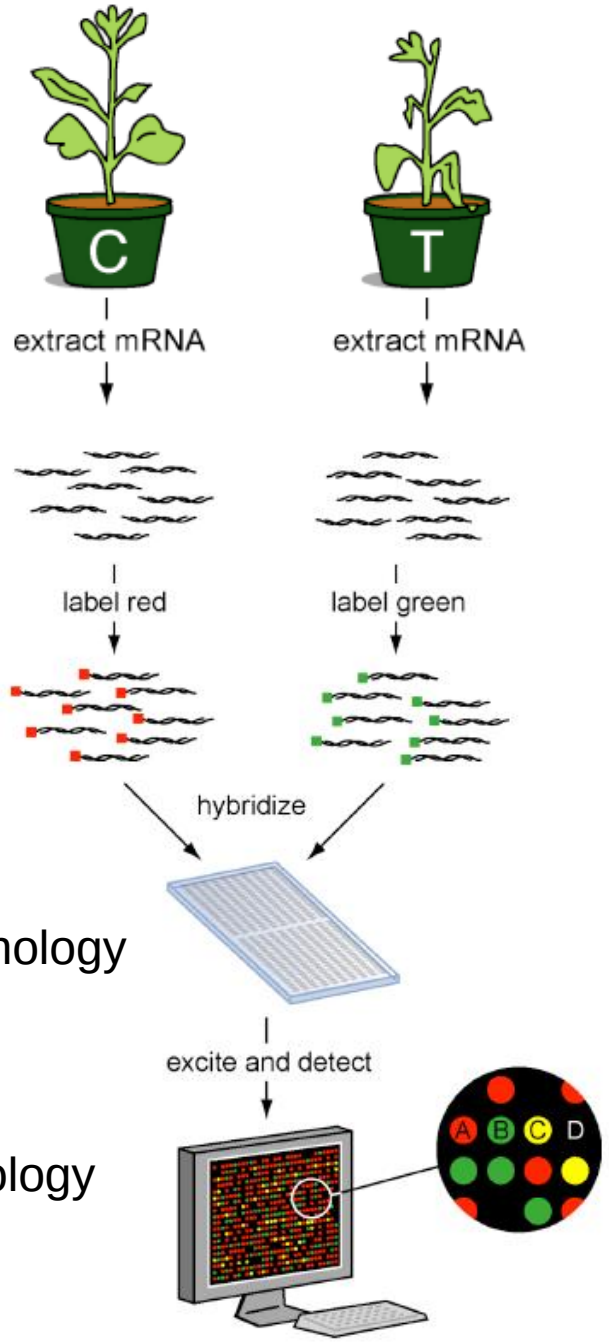
Simple explanation:

- each spot corresponds to a specific gene
- each spot contains a gene specific “glue”: it attracts mRNA specifically from that gene
- by measuring how much mRNA has attached to a spot, we can measure how much the gene was active

Microarray Hybridization

Control

Treated



Colour red

Colour green

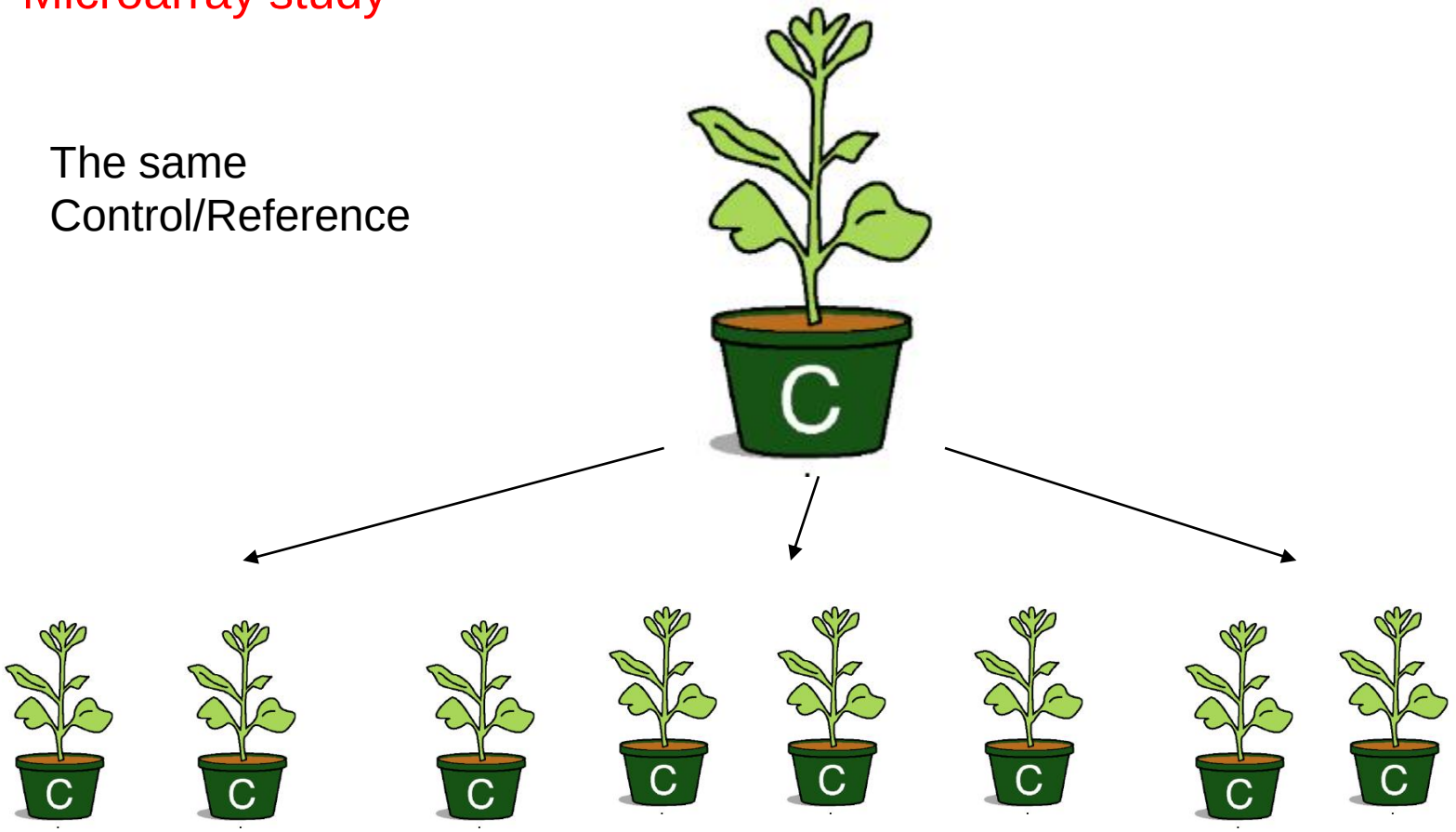
Microarray technology

Laser technology

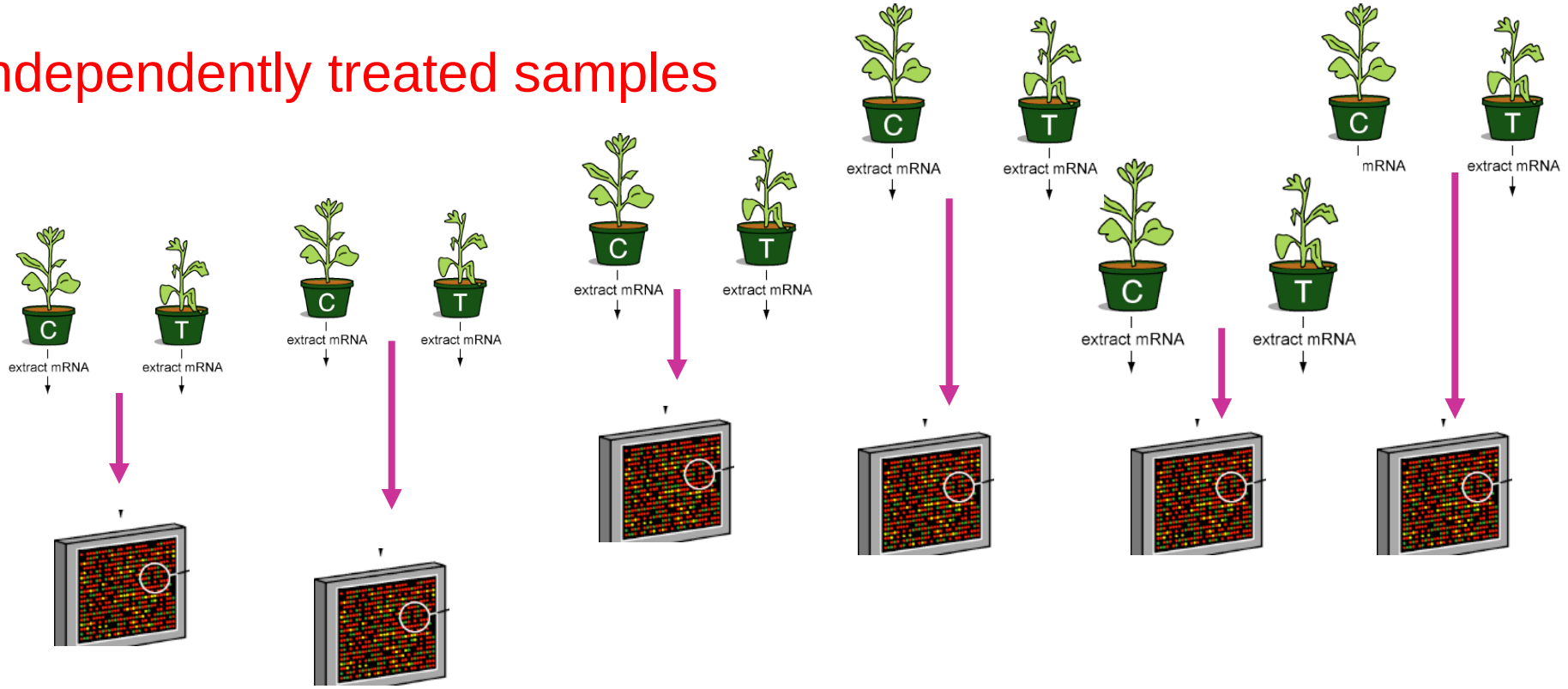
Measure how much green and red material has glued on each spot, and therefore how much each gene was active

Microarray study

The same
Control/Reference



Independently treated samples

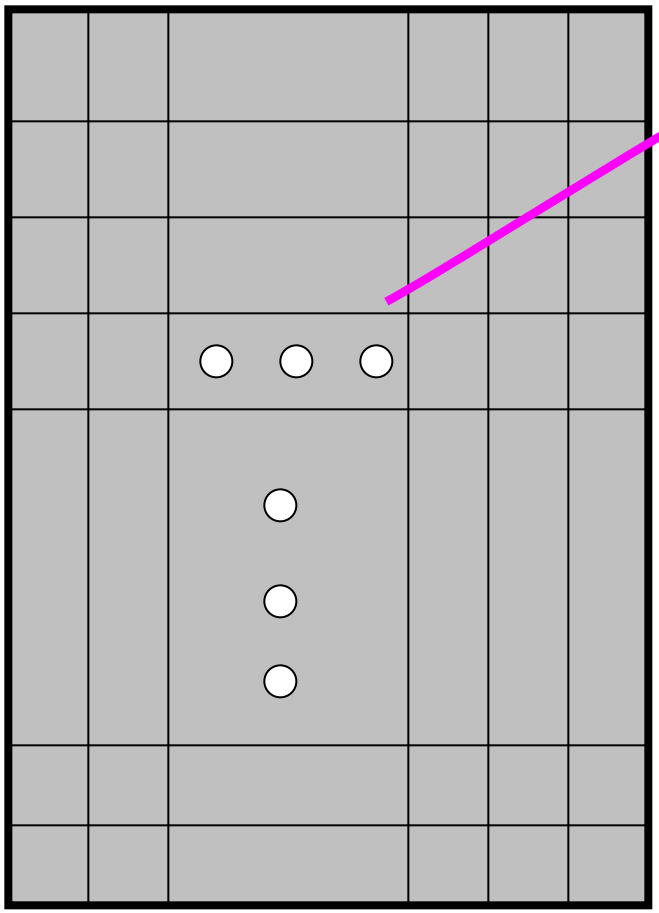


Many treated samples,
hybridised together with the same reference/control,
in separate hybridisations,
each producing a measure of the treated sample compared to the reference

Microarray Data

Individuals/samples

Genes

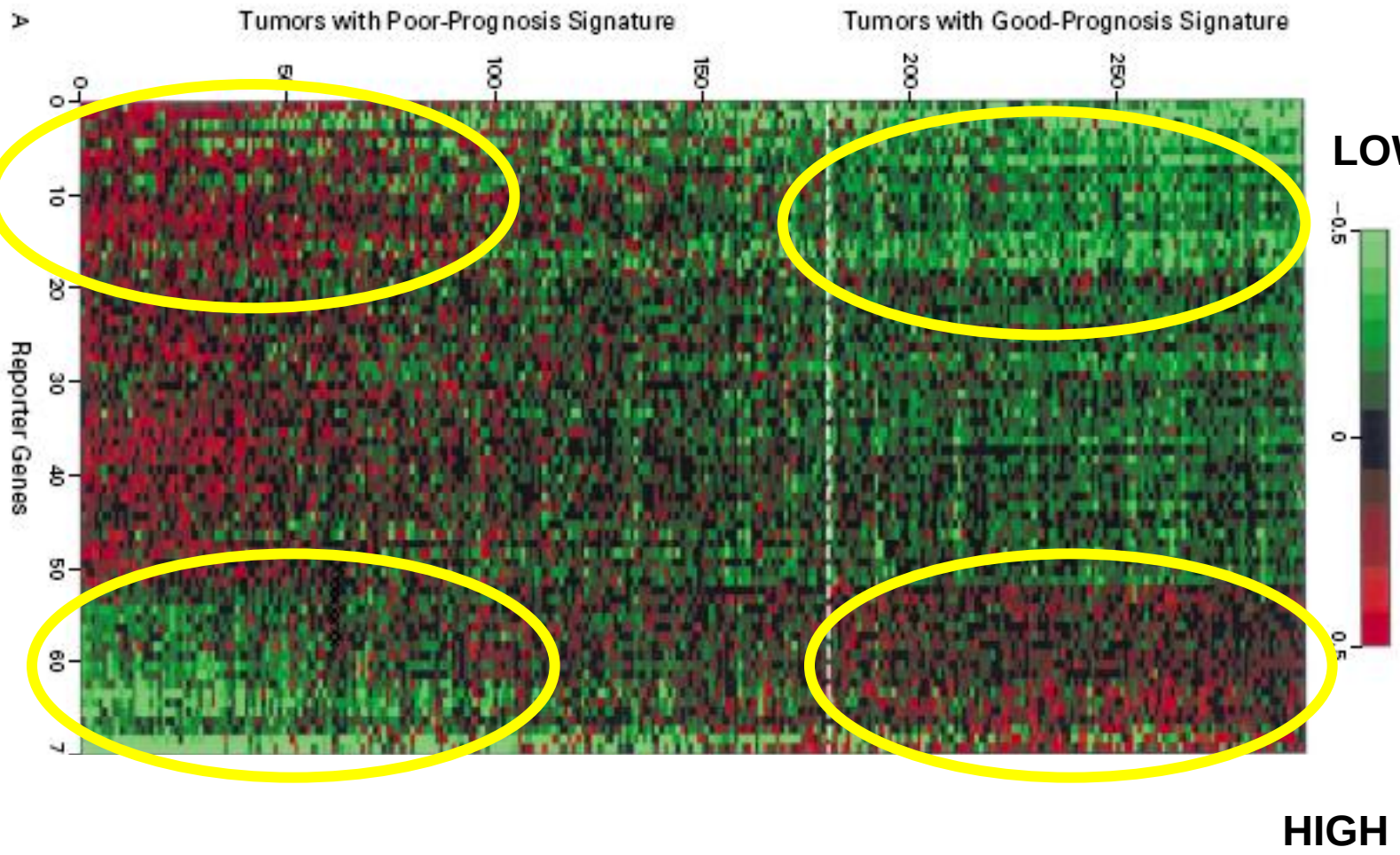


$$M_{gene} = \log_2\left(\frac{treated_{gene}}{reference_{gene}}\right)$$

Each element in this matrix
corresponds to a gene (row)
and a sample (column).
It measures how much that gene
was active in the sample
(compared to the comon reference)

Expressions

295 patients.



HIGH

LOW

70 genes

LOW

HIGH

Looking for differentially expressed gene

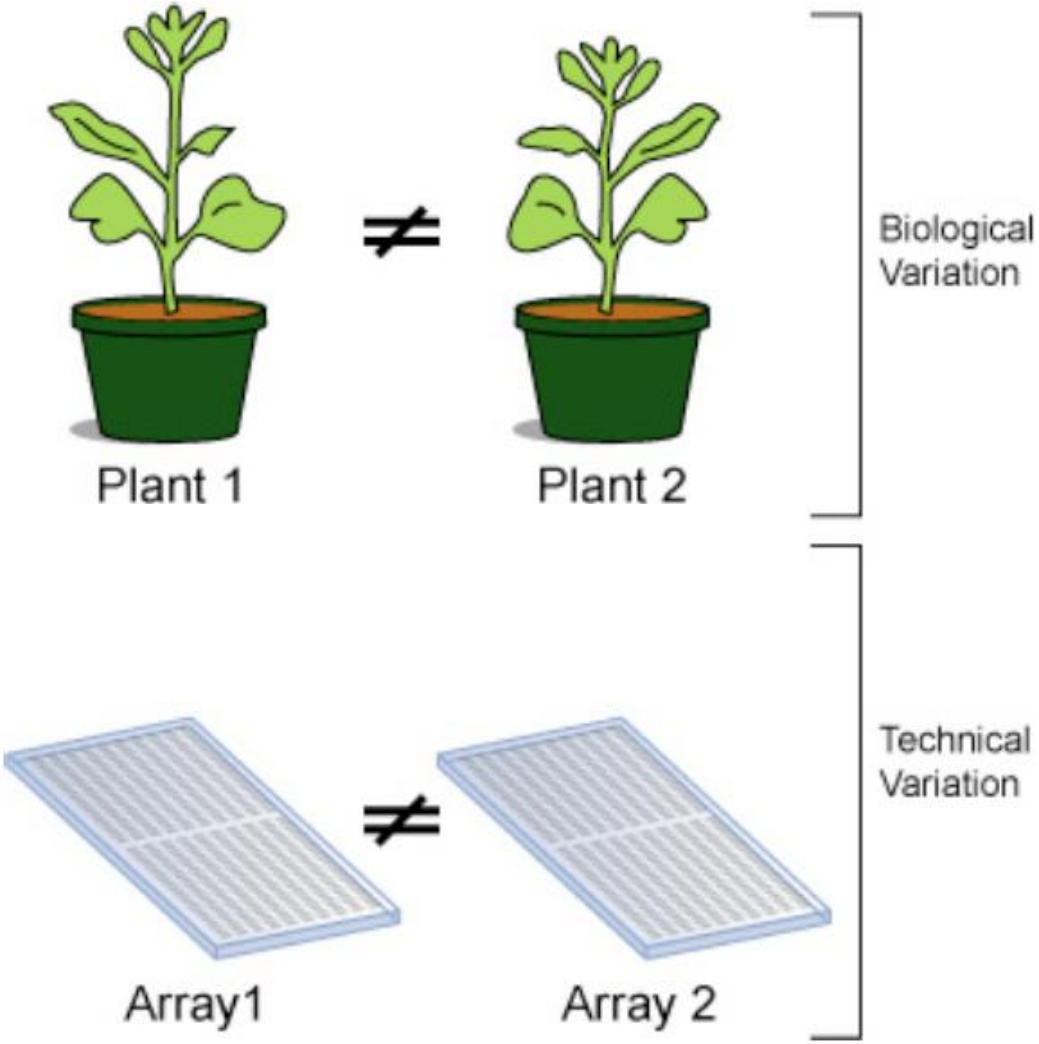
Biological variation

Individuals/Samples in different conditions, have different gene expressions, because the activity of their genes varies.

The measured expression levels vary from individual to individual used in the study.

Technical variation

Due to human error, there can be slight variation in microarray technology and hybridization during the experiment. Measurements of colour is also prone to measurement error.



Sources of variation in gene expression studies.

We are interested in the biological variation, given data that contain both biological and technical variation.

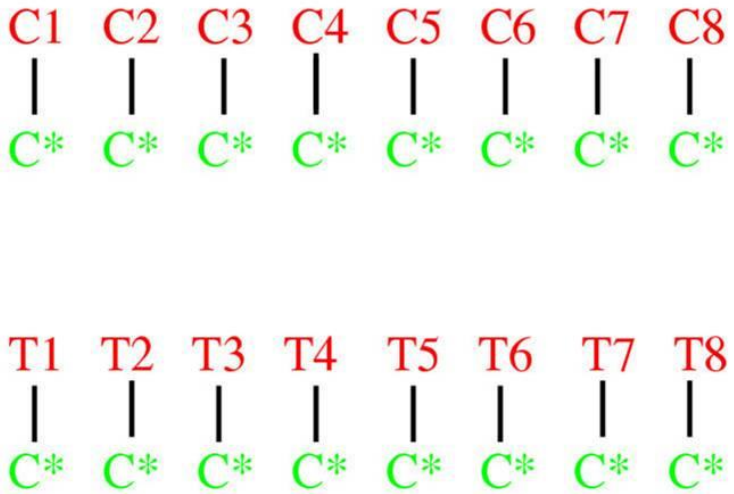
Pause!

Apo AI experiment

(Matt Callow et al., Genome Research 2000)

Goal: To identify genes with altered expression in the livers of Apo AI knock-out mice (T) compared to inbred control mice (C).

- 8 treatment and 8 control mice
- 16 hybridizations, each against a common reference
- Number of genes, m: ~ 6,000



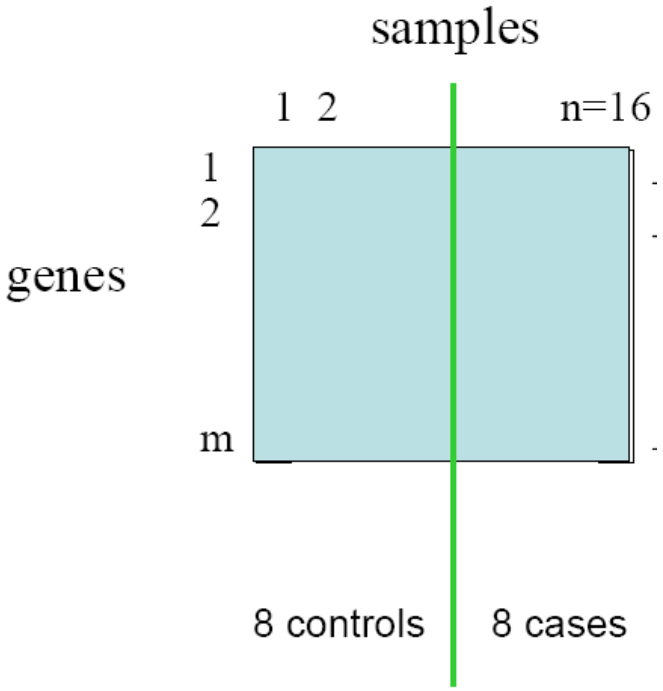
Microarray Expression Profiling Identifies Genes with Altered Expression in HDL-Deficient Mice

Matthew J. Callow,[1,4] Sandrine Dudoit,[2] Elaine L. Gong,[1] Terence P. Speed,[3] and Edward M. Rubin[1]

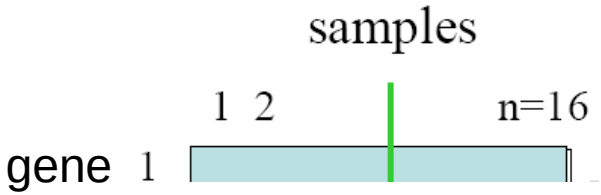
[1]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, California, 94720, USA; [2]Department of Biochemistry, Stanford University, Stanford, California 94305, USA; [3]Department of Statistics and Program in Biostatistics, University of California, Berkeley, California 94720-3860, USA

Abstract (NB! simplified version!)

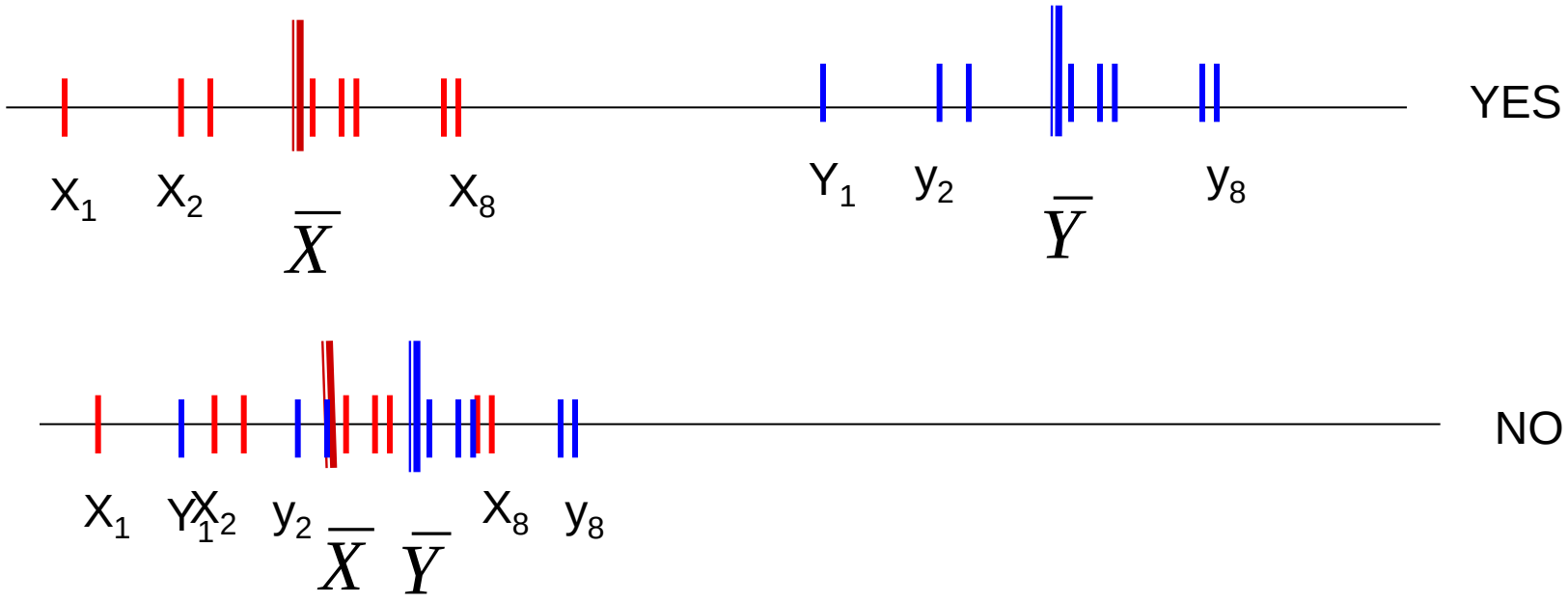
Based on the assumption that severe alterations in the expression of genes known to be involved in high-density lipoprotein (HDL) metabolism may affect the expression of other genes, we screened an array of >5000 mouse genes for altered gene expression in the livers of two lines of mice with dramatic decreases in HDL plasma concentrations. Labeled cDNA from apolipoprotein AI (apoAI)-knockout mice, scavenger receptor BI (SR-BI) transgenic mice, and control mice were cohybridized to microarrays. Two-sample *t* statistics were used to identify genes with altered expression levels in the knockout or transgenic mice compared with control mice. In the SR-BI group we found nine genes that were significantly altered on the basis of an adjusted *P* value < 0.05. In the apoAI-knockout group, eight genes were altered compared with the control group (adjusted $P < 0.05$). Several of the genes identified in the SR-BI transgenic suggest altered sterol metabolism and oxidative processes. These studies illustrate the use of multiple-testing methods for the identification of genes with altered expression in replicated microarray experiments.

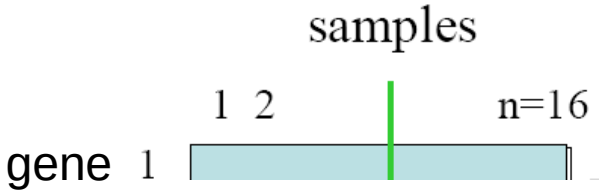


First we compare
one gene at the time
(one row), to see if this
gene is differently expressed
in controls vs. cases.

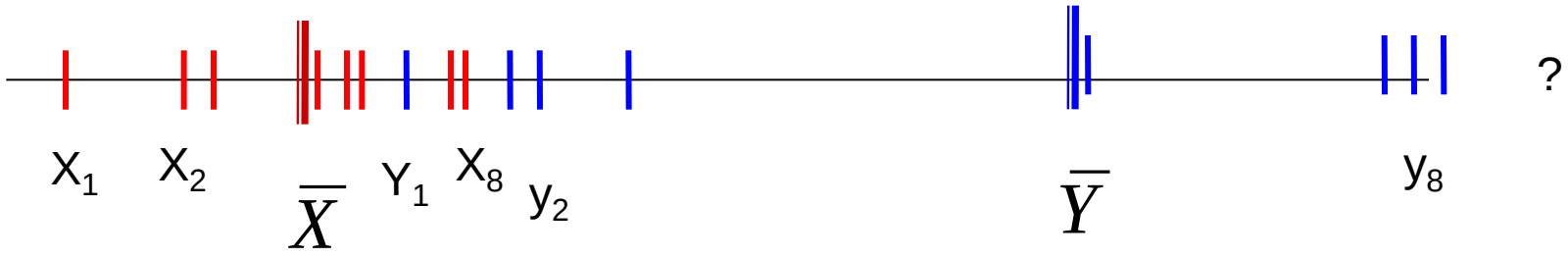


For every gene on the microarray we want to decide whether the expressions in the two experimental groups are (significantly) different from each other or not.

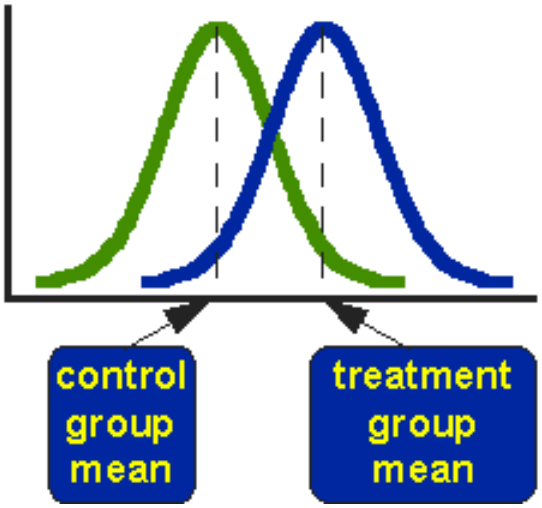




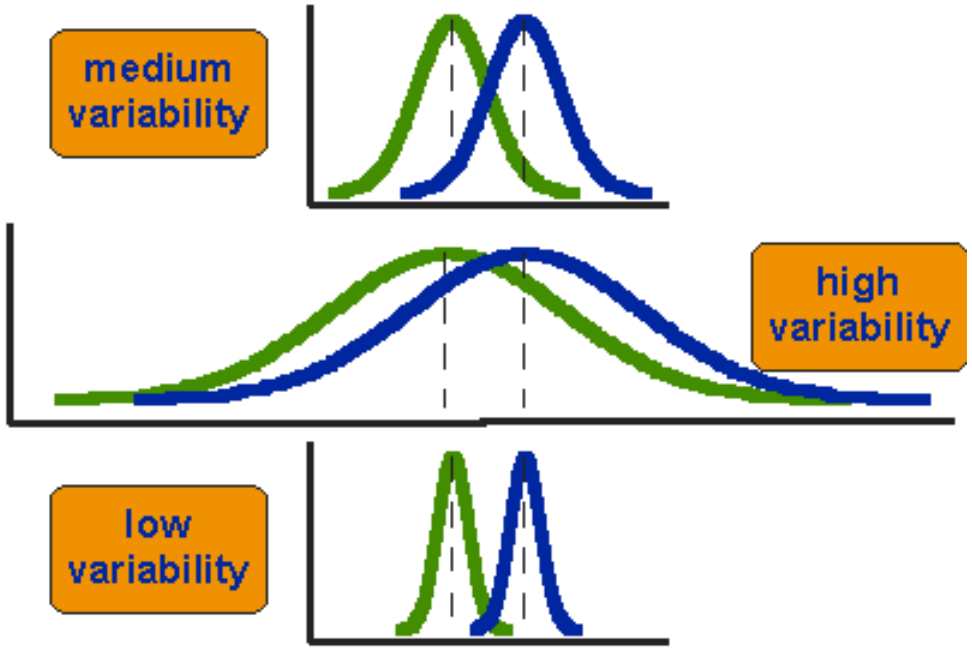
Only if the distance between the averages of our measurements is large compared to the variation of out measurements, can we assume that the gene has different expression in the two experimental groups.



Two sample t-test!



Are the means different?



The significance of the difference in means depends on the variances.

http://trochim.human.cornell.edu/kb/stat_t.htm

HYPOTHESIS TESTING

- A null hypothesis is assumed (H_0).
- Usually a neutral/conservative one
- We set up an alternative hypothesis (H_A), often the complement of H_0.
- In our case: the mean gene expressions in the two groups (indexed 1 and 2) are equal

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_A : \mu_1 \neq \mu_2$$

- We reject the null hypothesis if the data have small probability when the null hypothesis is true, ”under H_0”.

- We compute the p-value: the probability to obtain a result at least as extreme as the actually observed one, if the null hypothesis were true, ”under the null hypothesis”.

- The null hypothesis is rejected if the p-value is very small.

- The significance level is the threshold under which the p-values is considered as small. Often 0.05, or 0.01. If a null hypothesis is rejected, we say the result is significant.

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_A : \mu_1 \neq \mu_2$$

- We reject the null hypothesis if the data have small probability when the null hypothesis is true, ”under H_0”.

- The data are summarised in a test statistics, which summarises the information in the data relative to the hypothesis in question.

- For a given situation, there are many possible test statistics; some are better than others.

- Better means that they control the possible errors better

Hypothesis Truth vs. Decision

Truth \ Decision	not rejected	rejected
true Ho	ok	Type I error [false positive]
non-true Ho	Type II error [false negative]	ok

Type I error:

the error of rejecting the null hypothesis when it is actually true (false positive)

Significance level α is chosen so that

$P(\text{Type I error}) = P(\text{reject Ho} \mid \text{Ho is true}) < \alpha$

The p-value is the smallest significance level for which the null hypothesis Ho would be rejected, given the actual data.

Type II error:

the error of failing to reject a null hypothesis when it is in fact false (false negative).

$$P(\text{Type II error}) = P(\text{accept Ho} \mid \text{Ho is false})$$
$$= P(\text{accept Ho} \mid H_A \text{ is true})$$

1 - P(Type II error) = β = power of the test

We wish the power of the test to be as high as possible, while we keep the P(type I error) under control!

What is a good test?

A test that has high power 1-P(accept Ho | H_A is true), while it keeps the P(reject Ho | Ho is true) under control!

Example: Ho: $\mu_1 = \mu_2$
H_A: $\mu_1 \neq \mu_2$

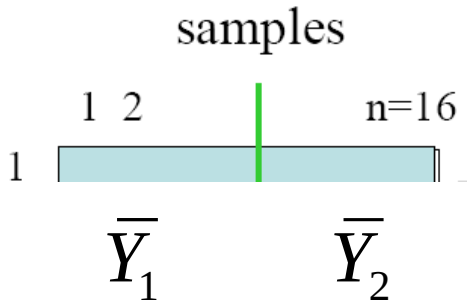
This is not easy: to compute the power we need to use the assumptions H_A which usually are not fully specified as they depend on more or unknow parameters.

Uniformly most powerful test (UMP)

A test with the greatest *power* for all values of the parameters being tested.

t-test

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_A : \mu_1 \neq \mu_2$$



$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}$$

- sample means
- N_1 and N_2 are the sample sizes
- s_1^2 s_2^2 are sample variances.

- The t-test is UMV in many situations!

t-test

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_A : \mu_1 \neq \mu_2$$

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}$$

Reject the null hypothesis that the two means are equal if

$$T < -t_{(\alpha/2,\upsilon)} \quad \text{or} \quad T > t_{(\alpha/2,\upsilon)}$$

where

$$t_{(\alpha/2,\upsilon)}$$

is the critical value of the t distribution with

$$\upsilon = \frac{(s_1^2/N_1 + s_2^2/N_2)^2}{(s_1^2/N_1)^2/(N_1 - 1) + (s_2^2/N_2)^2/(N_2 - 1)}$$

degrees of freedom.

t-test

$$\mathrm{H}_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad \mathrm{H}_\mathrm{A} : \mu_1 \neq \mu_2$$

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}$$

Under certain assumptions on the data (here expressions), T has under the null hypothesis a Student t-distibution with

$$v = \frac{(s_1^2/N_1 + s_2^2/N_2)^2}{(s_1^2/N_1)^2/(N_1 - 1) + (s_2^2/N_2)^2/(N_2 - 1)}$$

degrees of freedom.

If equal variances are assumed, then the formula reduces to:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{1/N_1 + 1/N_2}}$$

where

$$s_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$$

In this case the degrees of fredom are simply:

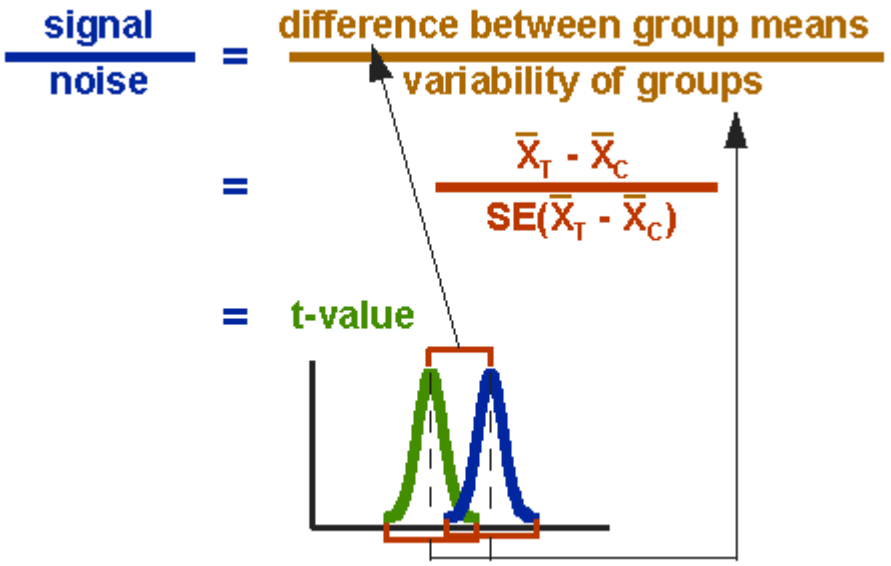
$$v = N_1 + N_2 - 2$$

A little intuition about the formula of the t-test statistics

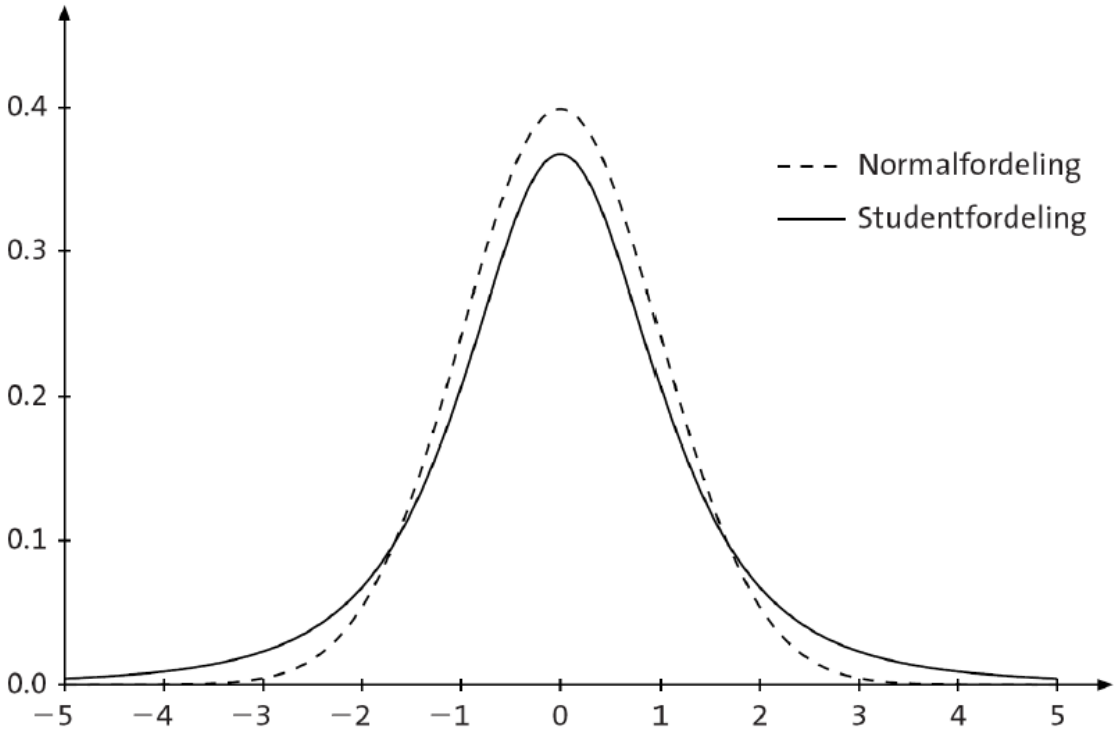
The formula for the t-test is a ratio.
The top part of the ratio is just the difference between the two means.
The bottom part is a measure of the dispersion of the data.

This formula is an example of the signal-to-noise situation:
the difference between the means is the signal that, in this case, we think our treatment may have produced into the data;
the bottom part of the formula is a measure of variability that is essentially noise that may make it harder to see the difference.

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{1/N_1 + 1/N_2}}$$



t-fordeling (3 f.g.) vs normalfordeling



Figur 8.3 Standardnormalfordelingen er tegnet inn sammen med Studentfordelingen med 3 frihetsgrader. En ser at den siste fordelingen er mer spredt ut enn den første

Ligner mer og mer normalfordeling når antall frihetsgrader er stort

Tabell over Studentfordelingen

Tabellen gir sannsynligheten $P(t \geq t_0)$ der t er studentfordelt og t_0 er et tall i tabellen.

Eksempel: For Studentfordelingen med 7 frihetsgrader has: $P(t \geq 1.895) = 0.05$

	Sannsynlighet for å overstige angitt grense					
Frihetsgrader	**0.25**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947

P(T>t)

Can I use the t-distribution? Assumptions.

Observations (individual expressions) should be

- independent
- normal distributed (at least approximately)

The more samples, the less important is individual normality, by the central limit theorem.

If few samples, and we cannot assume normality then

- transform data (log scale etc.)
- Use other tests, that do not require normality; for example non-parametric tests (Wilcoxon)

Example: Expression of gene APO1 in 12 controls and 9 treated samples.

	Control (Y_1)	Treated (Y_2)
1	9.65	6.11
2	5.17	4.70
3	6.48	6.87
4	7.58	7.20
5	6.50	8.49
6	6.09	7.07
7	5.75	6.58
8	7.99	7.02
9	5.63	6.62
10	8.05	
11	8.88	
12	6.28	

	Sannsynlighet for å overstige angitt grense					
Frihetsgrader	0.25	0.10	0.05	0.025	0.01	0.005
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845

$\overline{Y}_1 = 7.004 \qquad \overline{Y}_2 = 6.740$

$s_1^2 = 1.402 \qquad s_2^2 = 1.005$

Can assume same variance.

$s_p = 1.25$

$T = 0.48$

$df = 12 + 9 - 2 = 19$

p - value for two sided test of

$\text{Ho} : \mu_1 = \mu_2$ against

$\text{H}_\text{A} : \mu_1 \neq \mu_2$

is given by $2 \cdot \text{P}(\text{T}_{19} \geq 0.48) > 0.5 \approx 0.64.$

Cannot reject H_0.

Effect measure

We will typically be interested in saying something about the size of the effect of the new treatment.

We want some measure of effect.

In our example, we compared the two treatments with regard to average level of cholesterol, so the typical effect measure will be the difference between the mean values, $\overline{X}_1 - \overline{X}_2$.

We will present this, together with a measure of uncertainty, typically a 95% confidence interval.

NB: 95% confidence intervall for the difference of the two means

(Shared variance)

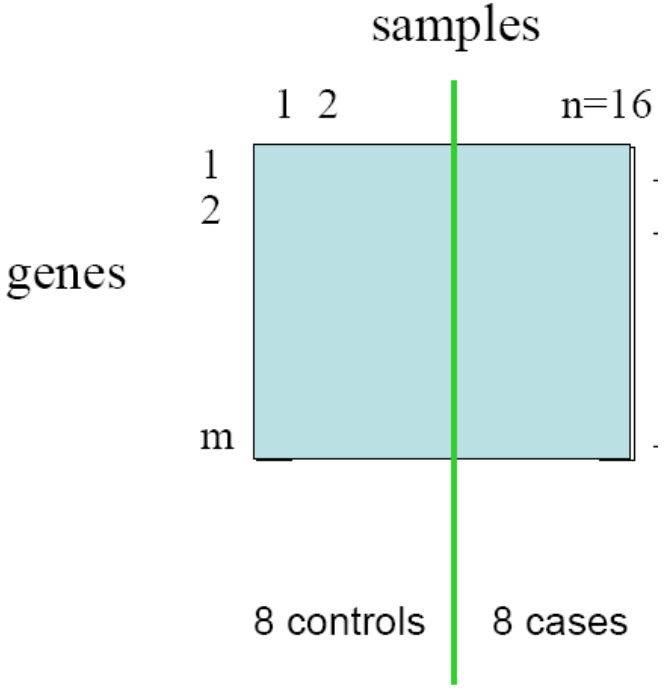
$$\left(\bar{Y}_1 - \bar{Y}_2\right) \pm t_{\alpha/2} \cdot s_p \cdot \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

Pause!

Microarray Expression Profiling Identifies Genes with Altered Expression in HDL-Deficient Mice

Matthew J. Callow,[1,4] Sandrine Dudoit,[2] Elaine L. Gong,[1] Terence P. Speed,[3] and Edward M. Rubin[1]

[1]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, California, 94720, USA; [2]Department of Biochemistry, Stanford University, Stanford, California 94305, USA; [3]Department of Statistics and Program in Biostatistics, University of California, Berkeley, California 94720-3860, USA



Statistical Analysis

For each experiment the expression log ratios were displayed in a 5600 times 16 matrix with rows corresponding to genes and columns corresponding to samples. To test the null hypothesis H*j* of equal mean expression for gene *j* in the control and knockout mice, a two-sample *t* statistic was used

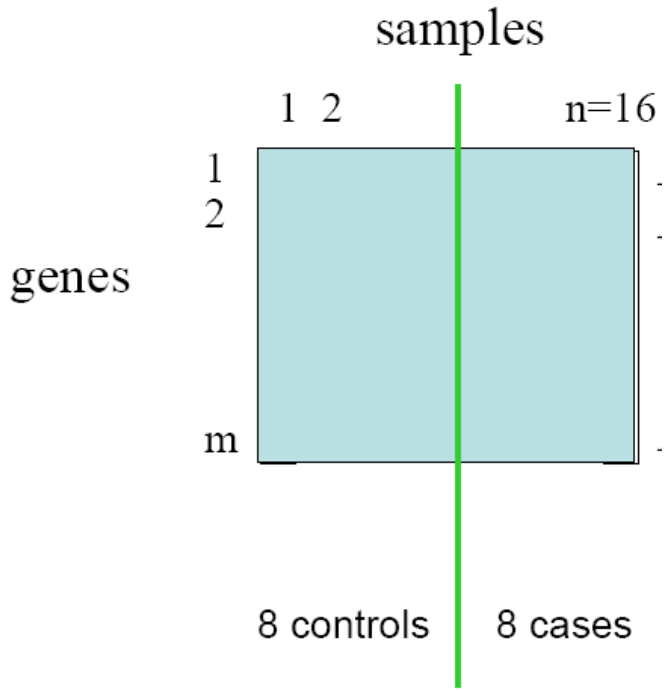
$$t_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{\sqrt{\frac{s^2_{1j}}{n_1} + \frac{s^2_{2j}}{n_2}}}$$

(unequal variances)

$$t_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{\sqrt{\frac{s^2_{1j}}{n_1} + \frac{s^2_{2j}}{n_2}}}$$

where $\bar{x}_{1j}$ and $\bar{x}_{2j}$ denote the average log ratio of element j in the control and [redacted] knockout group and s^2_{1j} and s^2_{2j} denote the variances of element j's log ratios in the control and [redacted] knockout hybridizations, respectively. The number of arrays in the control and [redacted] knockout group is denoted by n_1 and n_2, respectively.

Assessing the strength of the evidence against the null hypotheses of equal expression in the control and knockout mice is typically done by calculating *p*-values for each hypothesis; that is, by calculating for each gene the chance of getting a *t* statistic as extreme, or more extreme, than the observed statistic under the null hypothesis.



However, with a typical microarray data set comprising thousands of genes, an immediate concern is **multiple testing** because the probability that at least one null hypothesis is erroneously rejected (type-I error) can increase sharply with the number of hypotheses tested. To account for multiple testing we computed adjusted *p*- values for each gene (Westfall & Young, 1993; Shaffer 1995).

What is Multiple Testing?

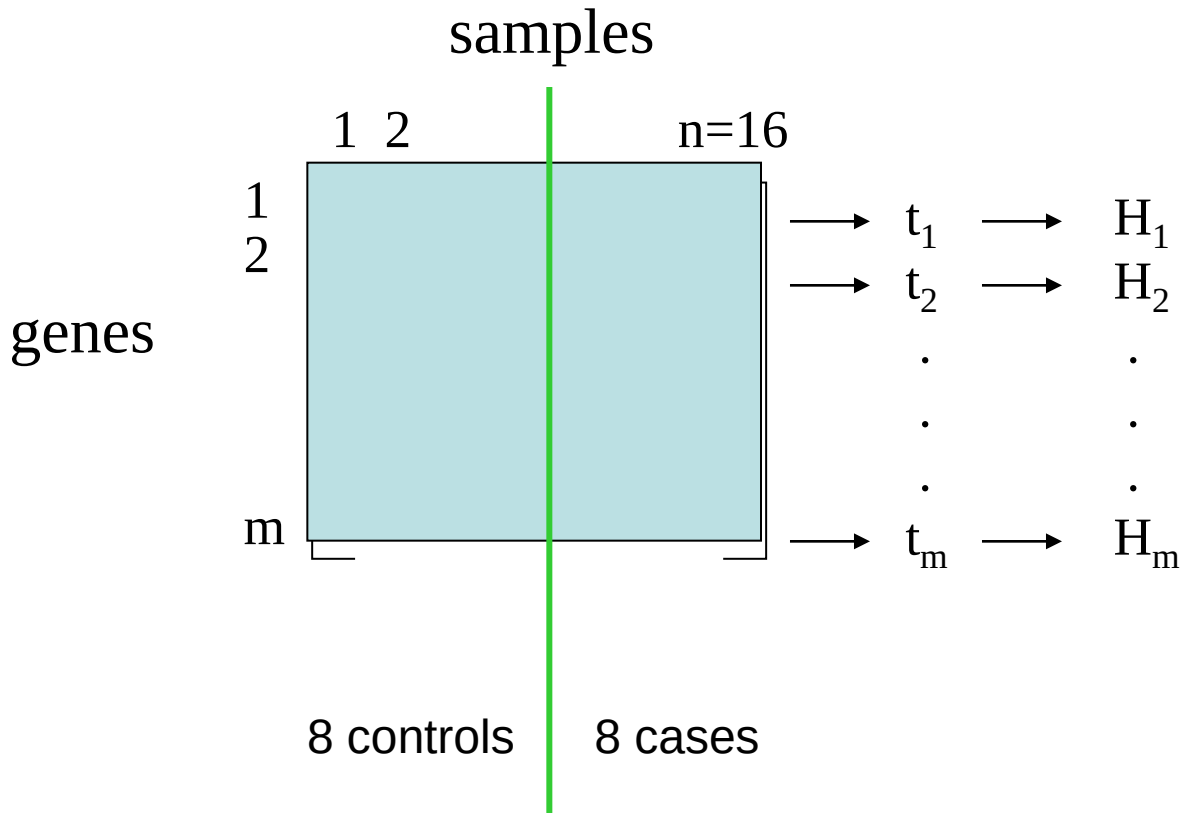
p-value: the probability to obtain a result at least as extreme as the actually observed one, under the null hypothesis:

$P(T \geq t_{data} \mid Ho)$

To compute this, it is required that we are able to assume that T is t-Student distributed

For our experiments, issues complicating *P* value calculations include an unknown null distribution of the test statistics. A suitable permutation distribution of the test statistics, as in Algorithm 4.1 of Westfall and Young (1993), was used to deal with these problems. In this algorithm, the permutation distribution of the *t* statistics was obtained by permuting the columns of the data matrix. Note that we are not assuming that the *t* statistics follow a *t* distribution or even a normal distribution, rather, we use a permutation distribution to estimate the null distribution of the *t* statistics.

What is Permutation Testing?


samples
1 2
n=16
genes
1
2
m
t_1
t_2
t_m
H_1
H_2
H_m
8 controls
8 cases

Computing *p*-values by permutations

1. Under the null hypothesis there is no difference between the treated and the control mice: the expressions originate from a common distribution.
2. Therefore, UNDER Ho, it was arbitrary to divide the 16 mice in treated and controls: we could have split them differently in two groups of 8 each.
3. Actually every possible split is “correct” under the null hypothesis that all mice have identical expressions (except for measurement variation)
4. Under Ho, we can produce very many “equivalent” data sets, by permuting the samples: each such permuted data set is “correct” under the null hypothesis.

Computing *p*-values by permutations

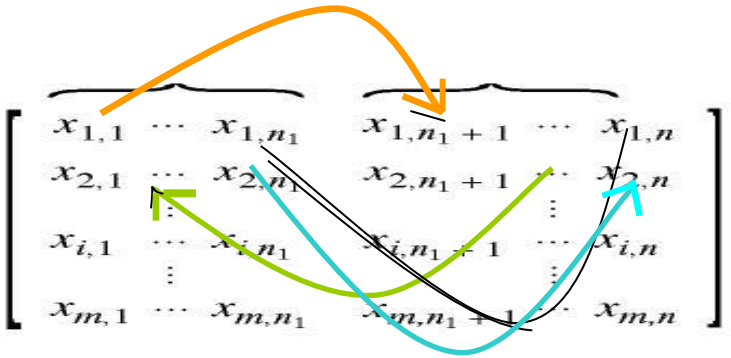
5. We produce many permuted data sets (b=1,2,…, B). For each we compute the t-test statistics T_b.
6. Next we plot histogram of these T_b., which represents the distribution of the t-test statistics under the null hypothesis.
7. We use this histogram to read of the p-value.
8. In this way we have not used the t-distribution to compute the p-value but the distribution obtained from the data by permutation.
9. We used less assymptions and got the p-value. Good!

10. Magic?

Computing *p*-values by permutations

After B permutations,

b = 1, ⋯ , *B* , we compute

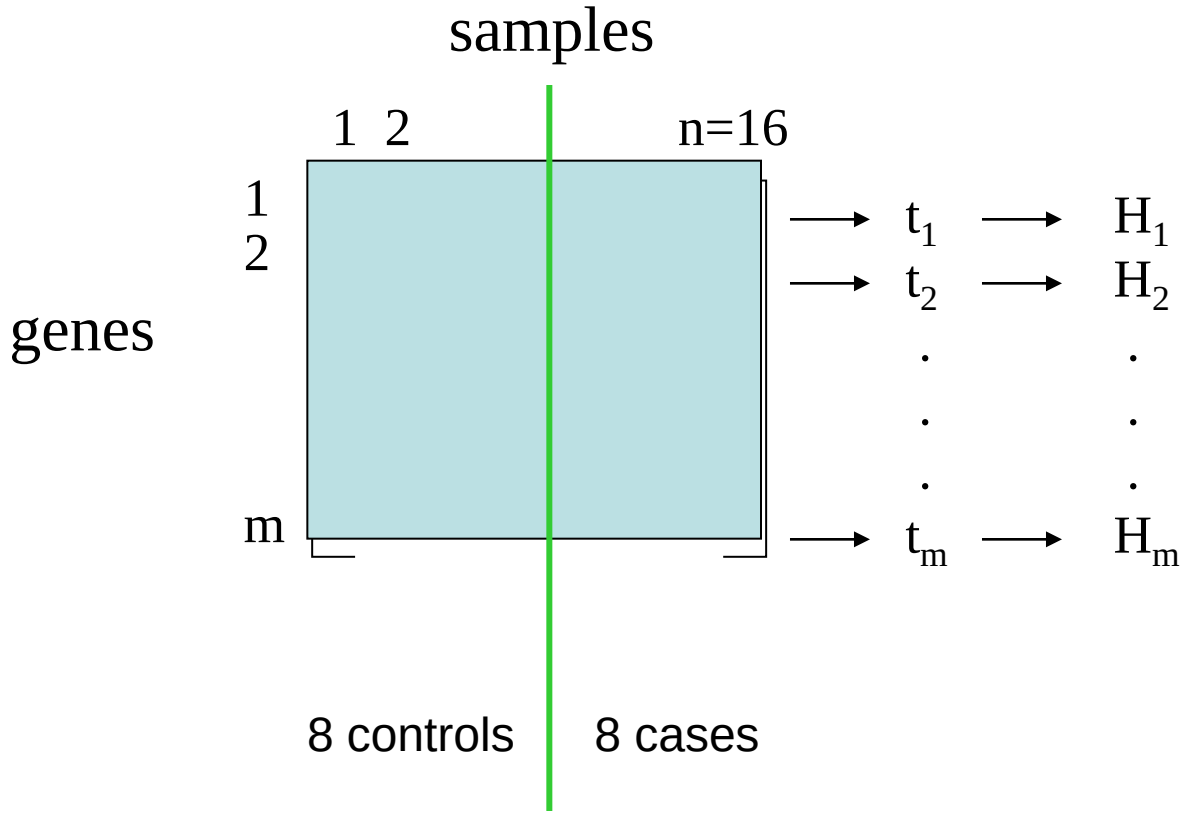


$$p\text{-}value = 1/B \cdot \#\{b: |T_b| \geq |T_{observed}|\}$$

How many permutations, *b* = 1, ⋯ , *B* ?

[I use the default $(N_1 + N_2)^2$]

Multiple Testing



m=6000, often 35000, can be 1 million!
Many t-test!

p-value = P($T \geq t_{data}$ | Ho)

Of the 6000 t-test statistics computed, roughly 300 will have t-test statistics with values ”more extreme” than the 0.05 significance,

$$T > t_{(\alpha/2,\nu)}$$

just by chance. These are false positives.

Many tests: what is the problem? Simulation to illustrate it.

Example: assume we have 30 000 independent genes on a microarray and not a single gene is truly differentially expressed.

If we reject the null hypothesis at level 0.05, we still expect $30000 \cdot 0.05 = 1500$ to have by chance a p-value below 0.05.

We create a simulated data set, where nothing is differentially expressed, and then we compute the t statistics and the p-values. No gene should be found as differentially expressed.

Simulation of 6,000 genes with 8 treatments and 8 controls: All the gene expression values were simulated *i.i.d* from a *N* (0,1) distribution, i.e. NOTHING is differentially expressed in our simulation.

We show the 10 smallest p-values, obtained by permutation:

Simulation

"gene" index	t value	*p*-value
2271	4.93	2×10^{-4}
5709	4.82	3×10^{-4}
5622	-4.62	4×10^{-4}
4521	4.34	7×10^{-4}
3156	-4.31	7×10^{-4}
5898	-4.29	7×10^{-4}
2164	-3.98	1.4×10^{-3}
5930	3.91	1.6×10^{-3}
2427	-3.90	1.6×10^{-3}
5694	-3.88	1.7×10^{-3}

Clearly we can't just use standard p-value thresholds of 0.05 or 0.01.

Multiple testing: Counting errors

Testing m genes: $H_1, H_2, \cdots, H_m$.

How many wrong decisions do we do?

m_0 = # of null hypotheses which are true (unknown)

R = # of rejected null hypotheses (known)

We want to control the number of errors we do.

There are two type of errors: Type I and Type II.

We count them separately:

Hypothesis Truth vs. Decision

Truth \ Decision	Number of null hypothesis not rejected	Number of null hypothesis rejected	totals
Number of true null hypothesis	U	V	m_0
Number of non-true hypothesis	T	S	m_1
totals	m - R	R	m

V = # Type I errors [false positives]

T = # Type II errors [false negatives]

Global control of Type I Error (False Positive)

We wish to control the total number of false positives, among the m tests.
Two main ways of doing multiple testing control:

	# not rejected	# rejected	totals
# true H	U	V (F +)	m_0
# non-true H	T	S	m_1
totals	m - R	R	m

Family-wise Error Rate

FWER = P(V ≥ 1 | null hypothesis) < α

The probability of one or more false positives is controlled.

False Discovery Rate $\mathrm{FDR} = E(\frac{V}{R} \mid \text{null hypothesis}) < \alpha$

The expected percentage of false positives, among the number of rejected genes (discovered genes)

Control of the FWER

Bonferroni adjusted p-values

each p-value p_j is multiplied by the number of tests

Bonferroni-adjusted p-value

$$p_j^{\mathrm{BON}} = p_j \cdot m$$

If we reject hypothesis H_j when $p_j^{\mathrm{BON}} < \alpha$

then overall FWER is smaller or equal to α:

FWER = P(V ≥ 1 | null hypothesis) < α

Control of the FDR

FDR (FALSE DISCOVERY RATE) adjusted p-values can also be computed.

Formula not given, but something like

$$p_j^{\mathrm{FDR}} = p_j \cdot (\text{a number smaller than } m)$$

If we reject hypothesis H_j when $p_j^{\mathrm{FDR}} < \alpha$

then overall FDR is smaller or equal to α:

$$\mathrm{FDR} = E\left(\frac{V}{R} \mid \text{null hypothesis}\right) < \alpha$$

We read again from Callow...

Assessing the strength of the evidence against the null hypotheses of equal expression in the control and knockout mice is typically done by calculating *p*-values for each hypothesis; that is, by calculating for each gene the chance of getting a *t* statistic as extreme, or more extreme,than the observed statistic under the null hypothesis.

However, with a typical microarray data set comprising thousands of genes, an immediate concern is **multiple testing** because the probability that at least one null hypothesis is erroneously rejected (type-I error) can increase sharply with the number of hypotheses tested. To account for multiple testing we computed adjusted *p*- values for each gene (Westfall & Young,1993; Shaffer 1995).

The adjusted *p*-value corresponding to the test of a null hypothesis H_j for a single gene *j* can be defined as the level of the entire test procedure at which H_j would just be rejected, given the values of all test statistics involved.

Adjusted by Westfall & Young,1993 ???

- Westfall & Young (1993) adjusted p-values

$$p_{r_j}^* = \max_{k = 1 \ldots j} \{ \text{prob} \, (\max_{l \in \{r_k \ldots r_m\}} |T_l| \geq |t_{r_k}| \mid H_0^{COM})\}$$

... complicated; but must be another way of doing the adjustement.
Indeed:

Wikipedia:
General methods of alpha adjustment for multiple comparisons:
Bonferroni correction
Boole–Bonferroni bound
Holm–Bonferroni method
Westfall-Young step-down approach of Westfall and Young
method of Benjamini and Hochberg

We read again from Callow...

Results: Genes that Change in apoAI-Knockout Mice

Adjusted *P* values were computed to obtain a more precise assessment of the statistical significance of the results and to account for multiple comparisons. Table 1A lists those genes in the analysis with the largest (absolute value) *t* statistics and their adjusted *P* values. We identified five genes with an adjusted *P* value <0.05.

		Adjusted *P* value
SR-B1	+11	0.0036
Glutathione-*S*-transferase	+2.5	0.0023
β-globin	+1.7	0.0177
Cytochrome P450 2B10	−5.0	0.0319
EST AI746730	−3.5	0.0407

Three of these genes showed increased expression and two showed decreased expression. We also identified a further 11 elements that represent seven genes that had adjusted *P* values 0.2.

The use of microarray expression profiling is gaining popularity through the general availability of ESTs and commercial availability of microarray printers and scanners. We have described a strategy that addresses the issues of detecting small but potentially significant changes in expression using microarrays and report a statistically simple approach to data analysis comparing two conditions. In our analysis we have also addressed the issue of multiple testing by calculating adjusted *P* values for comparisons of control and genetically modified mice. This analysis has yielded a number of genes with reproducible alterations in expression. These genes have drawn attention to cellular pathways and processes that warrant further investigation in studies of HDL metabolism.

gene index	t statistic	unadj. p ($\times 10^{-4}$)
2139	-22	1.5
4117	-13	1.5
5330	-12	1.5
1731	-11	1.5
538	-11	1.5
1489	-9.1	1.5
2526	-8.3	1.5
4916	-7.7	1.5
941	-4.7	1.5
2000	+3.1	1.5
5867	-4.2	3.1
4608	+4.8	6.2
948	-4.7	7.8
5577	-4.5	12

ALL
differentially
expressed!
Many genes!
Many false positives?

gene index	t statistic	unadj. p ($\times 10^{-4}$)	Bonferroni adjust.
2139	-22	1.5	.53
4117	-13	1.5	.53
5330	-12	1.5	.53
1731	-11	1.5	.53
538	-11	1.5	.53
1489	-9.1	1.5	.53
2526	-8.3	1.5	.53
4916	-7.7	1.5	.53
941	-4.7	1.5	.53
2000	+3.1	1.5	.53
5867	-4.2	3.1	.76
4608	+4.8	6.2	.93
948	-4.7	7.8	.96
5577	-4.5	12	.99

NO gene differentailly expressed!

gene index	t statistic	unadj. p ($\times 10^{-4}$)	Bonferroni adjust.	("almost") FDR adjust.
2139	-22	1.5	.53	2×10^{-4}
4117	-13	1.5	.53	5×10^{-4}
5330	-12	1.5	.53	5×10^{-4}
1731	-11	1.5	.53	5×10^{-4}
538	-11	1.5	.53	5×10^{-4}
1489	-9.1	1.5	.53	1×10^{-3}
2526	-8.3	1.5	.53	3×10^{-3}
4916	-7.7	1.5	.53	8×10^{-3}
941	-4.7	1.5	.53	0.65
2000	+3.1	1.5	.53	1.00
5867	-4.2	3.1	.76	0.90
4608	+4.8	6.2	.93	0.61
948	-4.7	7.8	.96	0.66
5577	-4.5	12	.99	0.74

8 Diff. expr. genes

Reading a paper with a serious mistake!

Science **314**, 268 (2006);

The Consensus Coding Sequences of Human Breast and Colorectal Cancers

Tobias Sjöblom,[1*] Siân Jones,[1*] Laura D. Wood,[1*] D. Williams Parsons,[1*] Jimmy Lin,[1] Thomas D. Barber,[1†] Diana Mandelker,[1] Rebecca J. Leary,[1] Janine Ptak,[1] Natalie Silliman,[1] Steve Szabo,[1] Phillip Buckhaults,[2] Christopher Farrell,[2] Paul Meeh,[2] Sanford D. Markowitz,[3] Joseph Willis,[4] Dawn Dawson,[4] James K. V. Willson,[5] Adi F. Gazdar,[6] James Hartigan,[7] Leo Wu,[8] Changsheng Liu,[8] Giovanni Parmigiani,[9] Ben Ho Park,[10] Kurtis E. Bachman,[11] Nickolas Papadopoulos,[1] Bert Vogelstein,[1‡] Kenneth W. Kinzler,[1‡] Victor E. Velculescu[1‡]

Abstract:

The elucidation of the human genome sequence has made it possible to identify genetic alterations in cancers in unprecedented detail. To begin a systematic analysis of such alterations, we determined the sequence of well-annotated human protein-coding genes in two common tumor types. Analysis of 13,023 genes in 11 breast and 11 colorectal cancers revealed that individual tumors accumulate an average of ~90 mutant genes but that only a subset of these contribute to the neoplastic process. Using stringent criteria to delineate this subset, we identified 189 genes (average of 11 per tumor) that were mutated at significant frequency. The vast majority of these genes were not known to be genetically altered in tumors and are predicted to affect a wide range of cellular functions, including transcription, adhesion, and invasion. These data define the genetic landscape of two human cancer types, provide new targets for diagnostic and therapeutic intervention, and open fertile avenues for basic research in tumor biology.

Statistical methods:

As 13,023 genes were evaluated for mutations, it was necessary to correct these probabilities for multiple comparisons. For this purpose, we used the FDR algorithm described by Benjamini and Hochberg (S*16*). The genes were ranked in ascending order, assigning a 1 to the gene with the lowest **probability of having the observed number of mutations** in it, a 2 to the gene with the next lowest probability, etc.

But:

p-value = P(observe the number of mutations, **or more** | Ho)

Comments on "Significance of candidate cancer genes as assessed by the CaMP score" by Parmigiani et al.

Holger Höfling*† Gad Getz‡ Robert Tibshirani§

June 26, 2007

First, the authors incorrectly apply the FDR formula. The formula requires the tail probabilities [Prob(X ≥ T)] as input, but Sjöblom *et al.* instead use the point probabilities [Prob(X = T)]. Consequently, their probabilities are smaller than they should be and therefore falsely appear to be more significant. When *P* values rather than point probabilities are used, the number of candidate genes falls from 122 to 6 in breast cancer and from 69 to 28 in colorectal cancer.

What have you learned?

statistics:

- recap hypothesis testing
- classical t-test
- p-value
- type 1 and type 2 errors
- learn that there are assumptions behind using Student t distribution
- learn an alternative (permutation) which will make clear that:
    -- there is freedom in inventing useful new tests
    -- and we can compute p-values for them
- learn about multiple testing issue, and simple remedies

genomic data:

- introduction to microarray data and modern high throughput genomics

Pause!

Observational studies

Observational studies

Epidemiology is mainly based upon observational studies.

Main problem: Confounding – an observed association between an exposure and an outcome is really caused by another factor.

Epidemiology, 2005

Folate Supplementation and Twin Pregnancies

Stein Emil Vollset,[†] Håkon K. Gjessing,[‡] Anne Tandberg,[§] Thorbjørn Rønning,[†] Lorentz M. Irgens,[*†] Valborg Baste,[†] Roy M. Nilsen,[†] and Anne Kjersti Daltveit[*†]*

Background

Pregnant women and women who plan to become pregnant are advised to increase their intake of folate to prevent neural tube defects.

Several studies have reported an association between use of folate and multiple births.

Folic acid has also been used to increase litter size in the swine industri.

This possible association with multiple births has been used as an argument against fortification of foods.

Medical Birth Registry, births 1998 – 2001.

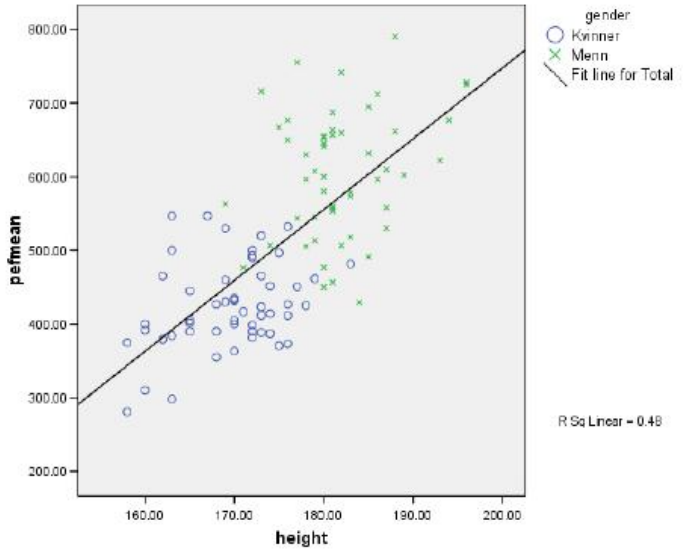
Observed association between folate use and twin births:
OR = 1.76 (1.57 – 1.97).

Logistic regression, adjustment for maternal age and parity:
OR = 1.59 (1.41 – 1.78).

Further adjustment for in vitro fertilization:
OR = 1.04 (0.91 – 1.18).

Linear regression example:

Association between height and lung function.



Regression equation

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Pefmean = -1174.9 + 9.61×height + ε

SPSS utskrift:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	height[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: pefmean

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,693[a]	,480	,475	84,02849

a. Predictors: (Constant), height

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	671200,6	1	671200,571	95,060	,000[a]
	Residual	727261,0	103	7060,787		
	Total	1398462	104			

a. Predictors: (Constant), height

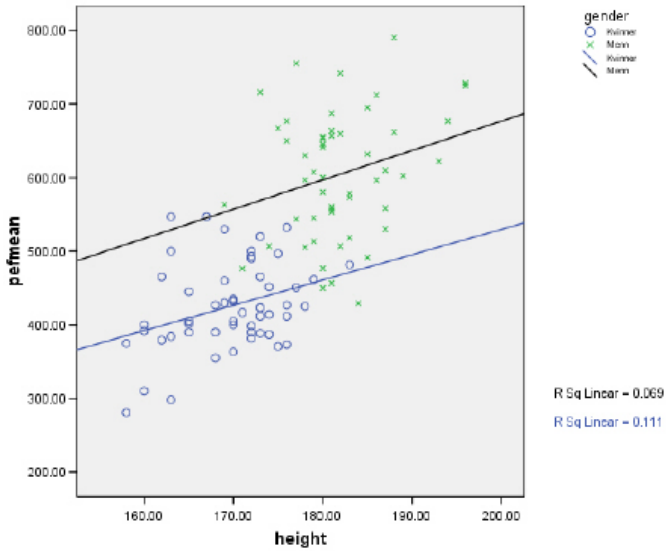
b. Dependent Variable: pefmean

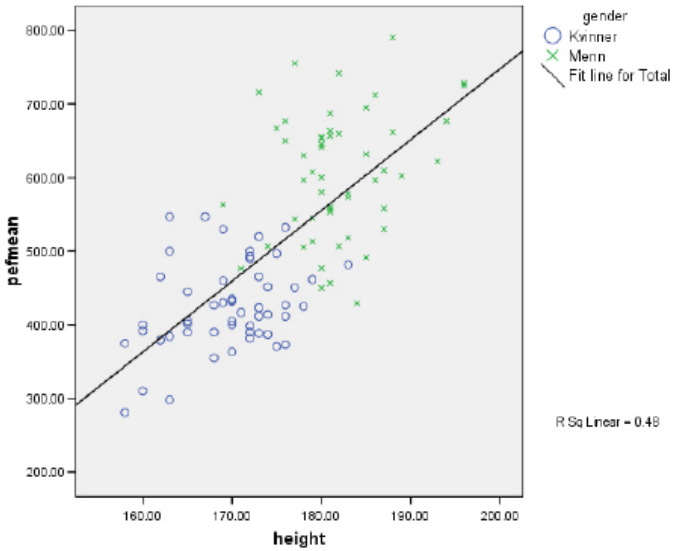
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-1174,895	173,411		-6,775	,000	-1518,815	-830,975
	height	9,612	,986	,693	9,750	,000	7,657	11,568

a. Dependent Variable: pefmean

What about the effect of gender?





$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Pefmean = -337.3 + 3.7×height + 133.7×gender + ε

For given gender, the effect of height is estimated to be an increase in lung function of 37 litres with a 10 cm increased height.

Opposite: For a given height, the effect of gender is estimated to be 133.7 litres.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	gender, height[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: pefmean

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,796[a]	,634	,627	70,80671

a. Predictors: (Constant), gender, height

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	887075,4	2	443537,714	88,467	,000[a]
	Residual	511386,2	102	5013,590		
	Total	1398462	104			

a. Predictors: (Constant), gender, height

b. Dependent Variable: pefmean

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-337,280	194,028		-1,738	,085	-722,133	47,574
	height	3,707	1,225	,267	3,027	,003	1,278	6,136
	gender	133,700	20,375	,579	6,562	,000	93,286	174,115

a. Dependent Variable: pefmean

Hvordan lese SPSS utskriften:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,796[a]	,634	,627	70,80671

a. Predictors: (Constant), gender, height

$r^2 = 0.63$

Forklart variasjon

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-337,280	194,028		-1,738	,085	-722,133	47,574
	height	3,707	1,225	,267	3,027	,003	1,278	6,136
	gender	133,700	20,375	,579	6,562	,000	93,286	174,115

a. Dependent Variable: pefmean

$\hat{a} = -337.28$

$\hat{b}_1 = 3.71$

$\hat{b}_2 = 133.70$

p-verdier

95% KI

Pause!

Og oppgave!

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Intensive Lipid Lowering with Simvastatin and Ezetimibe in Aortic Stenosis

Anne B. Rossebø, M.D., Terje R. Pedersen, M.D., Ph.D.,
Kurt Boman, M.D., Ph.D., Philippe Brudi, M.D., John B. Chambers, M.D.,
Kenneth Egstrup, M.D., Ph.D., Eva Gerdts, M.D., Ph.D.,
Christa Gohlke-Bärwolf, M.D., Ingar Holme, Ph.D.,
Y. Antero Kesäniemi, M.D., Ph.D., William Malbecq, Ph.D.,
Christoph A. Nienaber, M.D., Ph.D., Simon Ray, M.D.,
Terje Skjærpe, M.D., Ph.D., Kristian Wachtell, M.D., Ph.D.,
and Ronnie Willenheimer, M.D., Ph.D., for the SEAS Investigators*

ABSTRACT

BACKGROUND

Hyperlipidemia has been suggested as a risk factor for stenosis of the aortic valve, but lipid-lowering studies have had conflicting results.

METHODS

We conducted a randomized, double-blind trial involving 1873 patients with mild-to-moderate, asymptomatic aortic stenosis. The patients received either 40 mg of simvastatin plus 10 mg of ezetimibe or placebo daily. The primary outcome was a composite of major cardiovascular events, including death from cardiovascular causes, aortic-valve replacement, nonfatal myocardial infarction, hospitalization for unstable angina pectoris, heart failure, coronary-artery bypass grafting, percutaneous coronary intervention, and nonhemorrhagic stroke. Secondary outcomes were events related to aortic-valve stenosis and ischemic cardiovascular events.

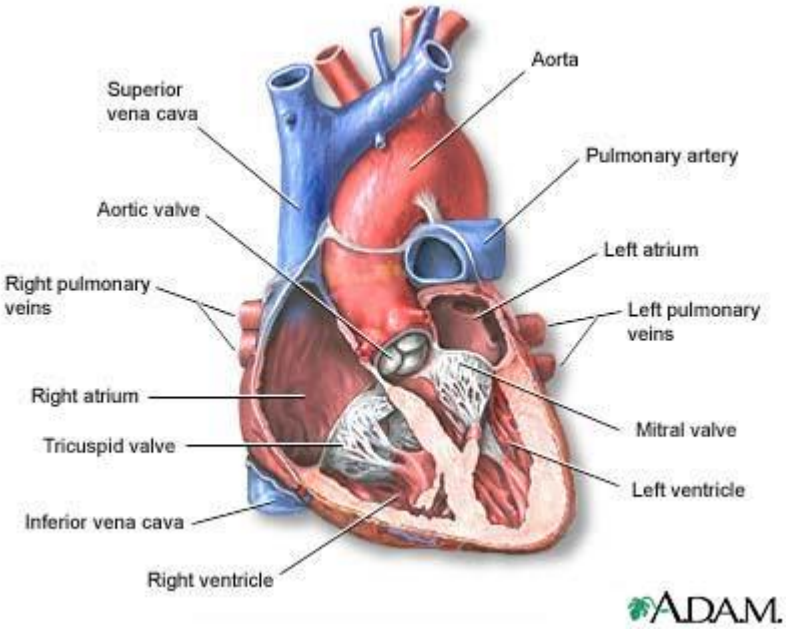
RESULTS

During a median follow-up of 52.2 months, the primary outcome occurred in 333 patients (35.3%) in the simvastatin–ezetimibe group and in 355 patients (38.2%) in the placebo group (hazard ratio in the simvastatin–ezetimibe group, 0.96; 95% confidence interval [CI], 0.83 to 1.12; P=0.59). Aortic-valve replacement was performed in 267 patients (28.3%) in the simvastatin–ezetimibe group and in 278 patients (29.9%) in the placebo group (hazard ratio, 1.00; 95% CI, 0.84 to 1.18; P=0.97). Fewer patients had ischemic cardiovascular events in the simvastatin–ezetimibe group (148 patients) than in the placebo group (187 patients) (hazard ratio, 0.78; 95% CI, 0.63 to 0.97; P=0.02), mainly because of the smaller number of patients who underwent coronary-artery bypass grafting. Cancer occurred more frequently in the simvastatin–ezetimibe group (105 vs. 70, P=0.01).

CONCLUSIONS

Simvastatin and ezetimibe did not reduce the composite outcome of combined aortic-valve events and ischemic events in patients with aortic stenosis. Such therapy reduced the incidence of ischemic cardiovascular events but not events related to aortic-valve stenosis. (ClinicalTrials.gov number, NCT00092677.)

- Cholesterol-lowering medication vs placebo. 1873 patients
- No effect on cardiac events related to aortic-valve stenosis!
- But increased risk of cancer among those on active medication (105 vs. 70, P = 0.01).



Hence, the statistical result is in opposition to the mechanistic understanding. Both the lack of positive effects, and the increased rate of cancer was unexpected.

Lessons from this example

- Even a good scientific understanding of disease mechanisms is not sufficient to predict treatment effects with any certainty
  - Mechanistic understanding of the heart is good. Still the result was not as expected
  - An enormous effort has gone into understanding how cancer arises. Still surprising result of the trial
- Editorial in New England J. Med.: Increased cancer: true signal or play of chance?
  - “Ezetimibe interferes with the gastrointestinal absorption not only of cholesterol, but also of other molecular entities that could conceivably affect the growth of cancer cells.”
- Why trust the statistics in this case? This was a randomized double blind trial

Design aspects

- A detailed description is given of the design, with an excellent overview in Figure 1.
- Discuss:
  - Exclusions
  - Drop-out
  - “Discontinued placebo, followed per protocol”. What does this mean?

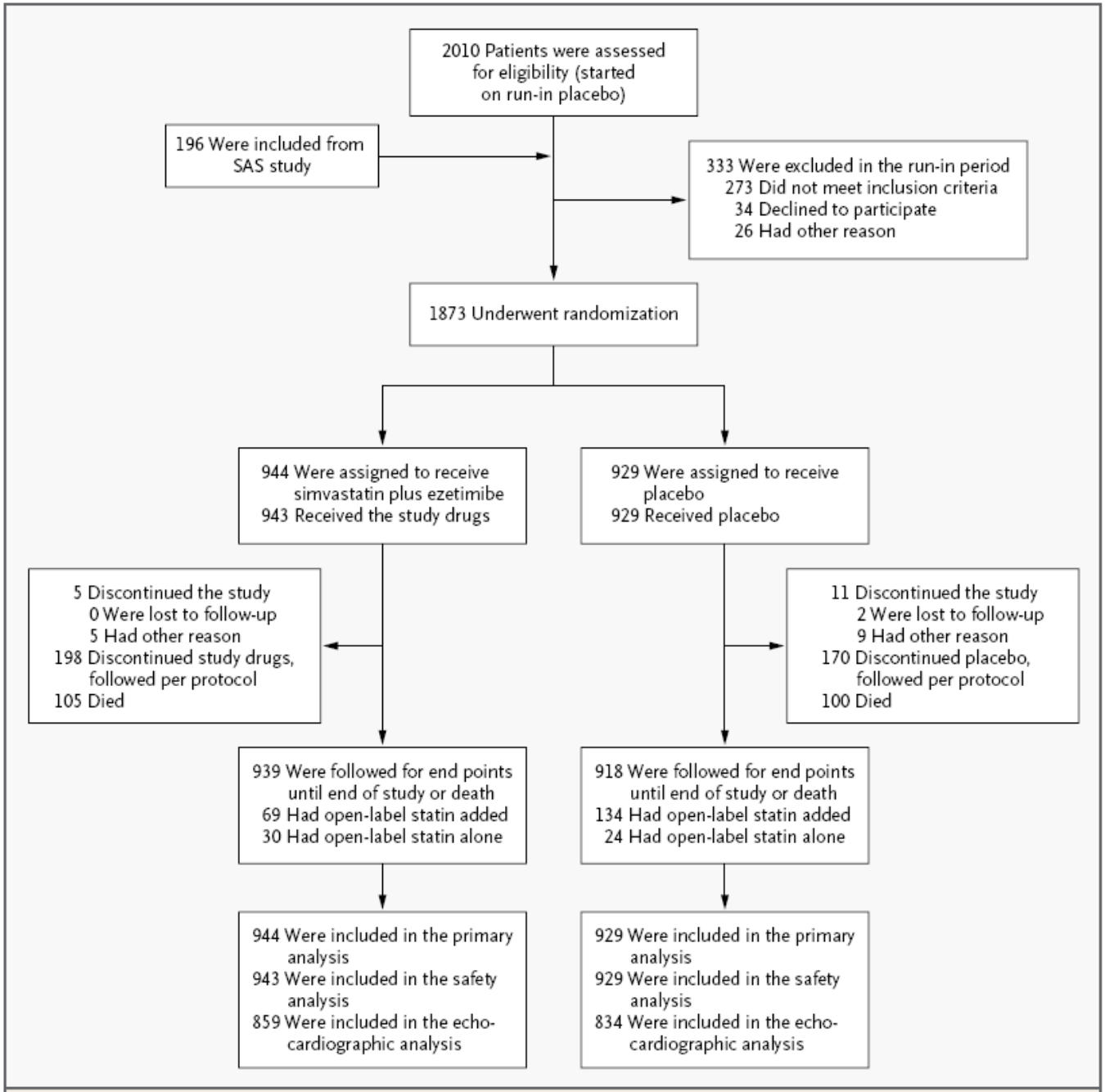


Figure 1. Enrollment and Outcomes.

SAS denotes the Simvastatin in Aortic Stenosis Study.

P-values

- What do the P-values mean in this paper? What is their function?
- The P=0.02 for cancer incidence: Should this be interpreted differently than the other P-values Why?

Table 1. Baseline Characteristics of the Patients.*

Characteristic	Placebo (N = 929)	Simvastatin–Ezetimibe (N = 944)	P Value†
Age — yr	67.4±9.7	67.7±9.4	0.46

Table 1. Baseline Characteristics of the Patients.*

Characteristic	Placebo (N = 929)	Simvastatin–Ezetimibe (N = 944)	P Value†
Age — yr	67.4±9.7	67.7±9.4	0.46
Female sex — no. (%)	360 (38.8)	363 (38.5)	0.92
White race — no. (%)‡	928 (99.9)	940 (99.6)	NA
Blood pressure — mm Hg			
Systolic	144.0±20.0	145.6±20.4	0.08
Diastolic	82.0±10.0	82.0±10.6	0.98
Hypertension — no. (%)	476 (51.2)	489 (51.8)	0.82
Smoking status — no. (%)			0.59
Current	171 (18.4)	189 (20.0)	
Former	344 (37.0)	333 (35.3)	
Never	414 (44.6)	422 (44.7)	
Body-mass index	26.8±4.3	26.9±4.3	0.58
Atrial fibrillation — no. (%)§	90 (9.7)	87 (9.2)	0.75
Atrioventricular block — no. (%)	23 (2.5)	21 (2.2)	0.76
Benign prostatic hyperplasia — no. of men (%)	63 (11.1)	74 (12.7)	0.47
Neoplasm (benign, malignant, or unspecified) — no. (%)	103 (11.1)	79 (8.4)	0.05
Drug therapy — no. (%)			
Angiotensin-converting–enzyme inhibitor	149 (16.0)	139 (14.7)	0.44
Angiotensin-receptor blocker	98 (10.5)	95 (10.1)	0.76
Calcium antagonist	160 (17.2)	157 (16.6)	0.76
Beta-blocker	268 (28.8)	242 (25.6)	0.12
Aspirin or other platelet inhibitor	260 (28.0)	236 (25.0)	0.16
Anticoagulant agent	49 (5.3)	58 (6.1)	0.43
Diuretic (including spironolactone)	229 (24.7)	209 (22.1)	0.21
Digitalis glycoside	22 (2.4)	28 (3.0)	0.47
Laboratory values			
Glucose — mg/dl	96.2±15.5	96.3±14.7	0.95
Creatinine — mg/dl	1.06±0.17	1.06±0.18	0.82
Estimated glomerular filtration rate — ml/min per 1.73 m^2¶	68.2±12.0	68.5±12.6	0.54
High-sensitivity C-reactive protein — mg/liter			0.76

Table 2. Prespecified Primary and Secondary Composite Outcomes and Death.*

Outcome	Placebo (N=929)	Simvastatin plus Ezetimibe (N=944)	Hazard Ratio (95% CI)†	P Value
	number (percent)			
Primary outcome				
Patients with any event‡	355 (38.2)	333 (35.3)	0.96 (0.83–1.12)	0.59
Death from cardiovascular causes	56 (6.0)	47 (5.0)	0.83 (0.56–1.22)	0.34
Aortic-valve replacement surgery	278 (29.9)	267 (28.3)	1.00 (0.84–1.18)	0.97
Congestive heart failure as a result of progression of aortic stenosis	23 (2.5)	25 (2.6)	1.09 (0.62–1.92)	0.77
Nonfatal myocardial infarction	26 (2.8)	17 (1.8)	0.64 (0.35–1.17)	0.15
Coronary-artery bypass grafting	100 (10.8)	69 (7.3)	0.68 (0.50–0.93)	0.02
Percutaneous coronary intervention	17 (1.8)	8 (0.8)	0.46 (0.20–1.06)	NA
Hospitalization for unstable angina	8 (0.9)	5 (0.5)	0.61 (0.20–1.86)	NA

Death				
From any cause	100 (10.8)	105 (11.1)	1.04 (0.79–1.36)	0.80
From cardiovascular causes	56 (6.0)	47 (5.0)	0.83 (0.56–1.22)	0.34
Myocardial infarction	10 (1.1)	5 (0.5)	0.49 (0.17–1.42)	
Stroke	6 (0.6)	5 (0.5)	0.82 (0.25–2.70)	
Sudden death	20 (2.2)	20 (2.1)	0.99 (0.53–1.83)	
Related to cardiac surgery (perioperative)	7 (0.8)	7 (0.7)	0.99 (0.35–2.83)	
Heart failure	5 (0.5)	6 (0.6)	1.21 (0.37–3.95)	
Other	8 (0.9)	4 (0.4)	0.49 (0.15–1.63)	
From noncardiovascular causes	44 (4.7)	56 (5.9)	1.26 (0.85–1.86)	0.26
Cancer¶	23 (2.5)	39 (4.1)	1.67 (1.00–2.79)	0.05
Infection	14 (1.5)	7 (0.7)	0.50 (0.20–1.23)	

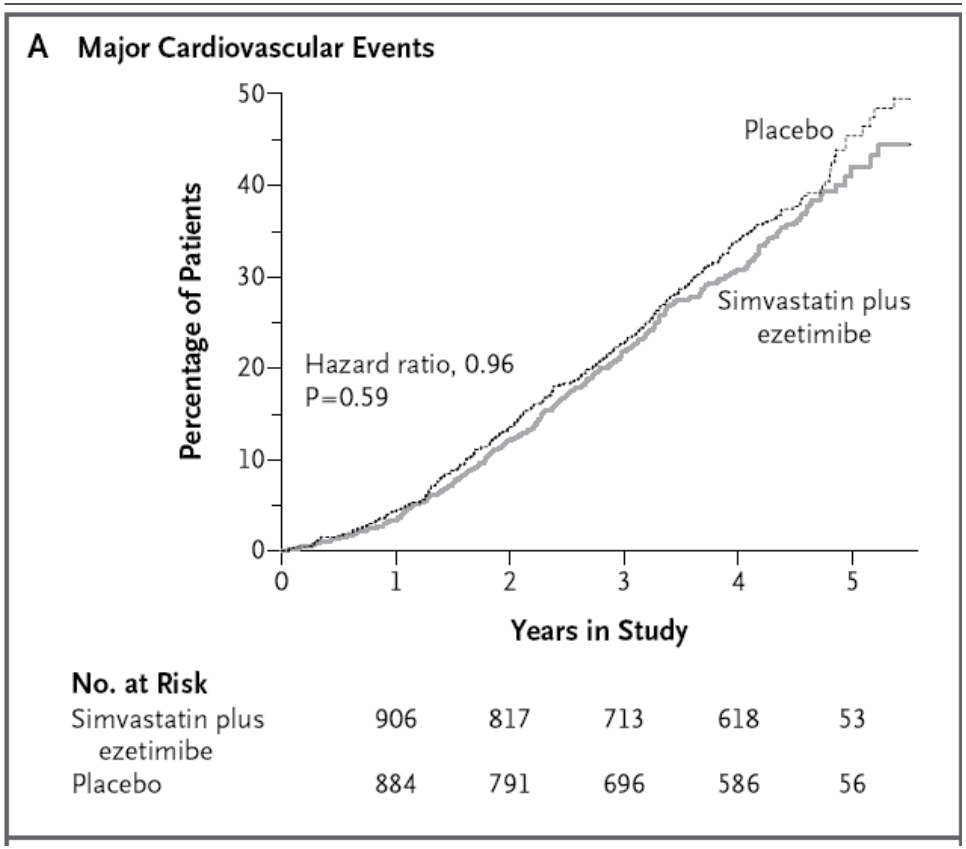
Effect measures

- There are at least two levels of effect: *cholesterol modification* and *events* (primary and secondary outcomes and death)
- Discuss the relationship between these two levels. What type of effect measures do you find at the two levels?

Tables and figures

- Discuss the set up of figures from a statistical point of view. What information is conveyed in the figures?

Figure 3. Kaplan–Meier Curves for Primary and Secondary Outcomes and Death.



You need to know that the hazard ratio is to be interpreted as a relative risk. HR=RR=0.96; the risk is reduced of 4%, but not significant.

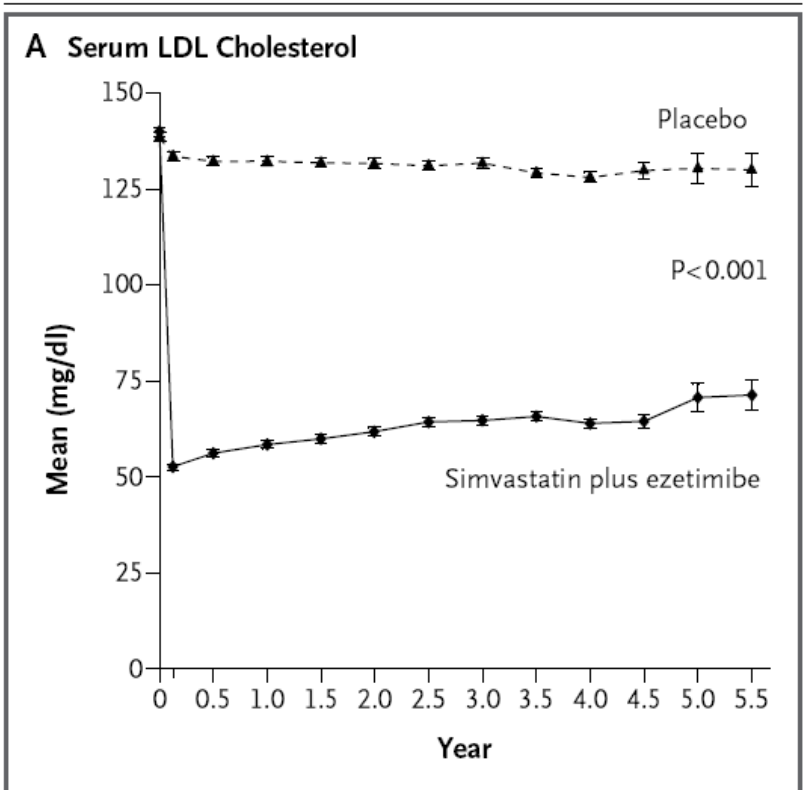
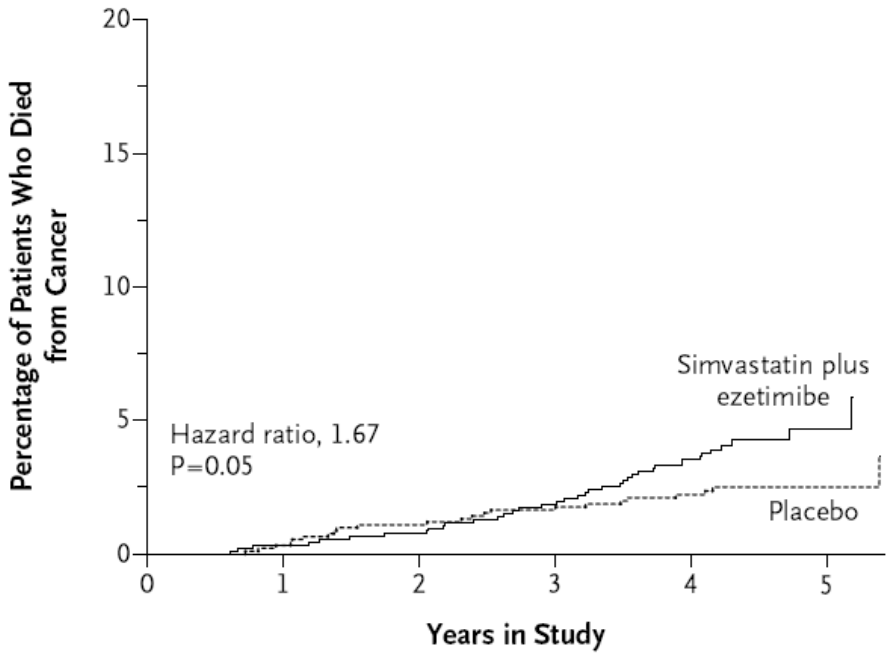


Figure 2. Serum Low-Density Lipoprotein (LDL) Cholesterol Levels (Panel A) and the Change from Baseline in Peak Aortic-Jet Velocity (Panel B).

The I bars represent standard errors. In Panel A, the first data points correspond to results at the 8-week visit.



No. at Risk					
Simvastatin plus ezetimibe	930	912	884	855	89
Placebo	916	890	865	835	94

Figure 4. Kaplan–Meier Curves for Cancer-Related Mortality.

P=0.06 as calculated with log-rank continuity correction.

Discussion!

Hypothesis testing
vs.
Hypothesis generating

Darwin: 'How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service'.

S. Karlin: 'The purpose of models is not to fit the data, but to sharpen the questions'.

Statistics does not prove a hypothesis,
does not find the gene of breast cancer:
it helps to learn from data, complex data.

Validation and
mechanistic, substantive understanding are needed

The End!