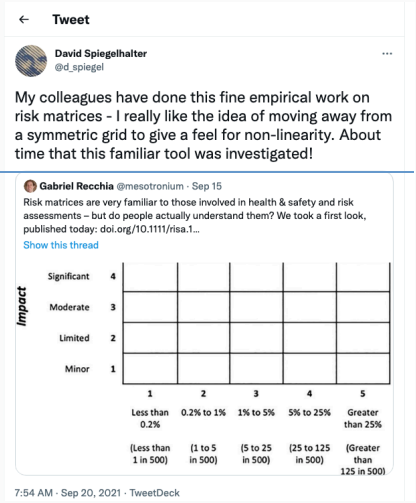
Methods of Applied Statistics I

STA2101H F LEC9101

Week 2

September 22 2021



Today

Start Recording

1. Office hours, Upcoming events, HW 2
   OH: Monday 7pm-8.30pm, Wednesday 4pm-5.30pm, Zoom until further notice
2. Steps in analysis; types of studies
3. Linear Regression Part 2: recap, testing groups of variables, checking model assumptions, collinearity, $p > n$
4. In the News

- SM – Statistical Models by Davison
- LM-1,2 – Linear Models with R by Faraway (1st and 2nd editions)
- ELM-1,2 – Extending the Linear Model with R by Faraway (1st and 2nd editions)
- CD – Principles of Applied Statistics by Cox & Donnelly

LM (both)

Upcoming

- Learning Individualized Treatment Rule for a Target Population
- Thursday 3.30 Link

About Guanhua Chen

I am an Assistant Professor of Biostatistics and Medical Informatics at the University of Wisconsin-Madison. I got my Ph.D. from the University of North Carolina at Chapel Hill in 2014 under the direction of Dr. Michael Kosorok. Before joining UW, I was an Assistant Professor of Biostatistics at Vanderbilt University.

Research Interest

Develop statistical learning methods for clinical and biomedical research. In particular, I am interested in analyzing heterogeneous, high-dimensional-omics data (genome, microbiome) and electronic health record data to advance precision medicine. My current research is supported by PCORI and NSF grants.

- Fridays at noon Toronto Data Workshop link

This term is a special series of talks featuring University of Toronto speakers on the relationship between data science and their other field of expertise.

Date	Speaker	Recording
Fri 24 September 2021, noon - 1pm	Karen Chapple, Geography, planning, cities	
Fri 1 October 2021, noon - 1pm	Special on 2021 Canadian Election	
Fri 8 October 2021, noon - 1pm	Fedor Dokshin, Sociology	

HW Question Week 2

STA2101F 2021

Due September 29 2021 11.59 pm

Homework to be submitted through Quercus

This question concerns the article "How people understand risk matrices..." by Sutherland et al., available on the course web site.

(a) What is a risk matrix? Find an example of a risk matrix for the assessment of risk related to either COVID-19 or wildfire. Provide the reference (link) to the example and a snapshot of one or two of the published risk matrices.
(b) The authors describe two randomized controlled experiments, in Sections 2 and 3, respectively. What are the units of analysis in these two experiments?
(c) In the first experiment, the authors describe derived variables: a "basic knowledge score", a "risk comparison score", a "prioritization score", the results of a "matrix preference test", and a "total numeracy score". Which of these are treated as response variables and which are treated as explanatory variables?
(d) What 'treatments' were randomly assigned to units in Experiment 1?

Steps in Analysis

- understand the physical background
- understand the objective
- make sure you know what the client wants
- put the problem into statistical terms

- How were the data collected:
  - are the data observational or experimental? etc.
  - is there nonresponse
  - are there missing values
  - how are the data coded
  - what are the units of measurement
  - beware of data entry errors

Components of investigation

CD §1.2

- start with a scientific question
- assess how data could shed light on this
- plan data collection
- consider of sources of variation and how careful planning can minimize their impact

- develop strategies for data analysis: modelling, computation, methods of analysis
- assess the properties of the methods and their impact on the question at hand

- communicate the results: accurately
- visualization strategies, conveyance of uncertainties

but not pessimistically

- choice of material/individuals to study – “units of analysis”
- “For studies of a new phenomenon it will usually be best to examine situations in which the phenomenon is likely to appear in the most striking form, even if this is in some sense artificial”
- statistical analysis needs to take account of the design (even if statistician enters the project at the analysis stage)
- need to be clear at the design stage about broad features of the statistical analysis – more publicly convincing **and** “reduces the possibility that the data cannot be satisfactorily analysed”
- “it is unrealistic and indeed potentially dangerous to follow an initial plan unswervingly ... it may be a crucial part of the analysis to clarify the research objectives”

Experimental and observational studies

CD Ch. 1

- experiment is a study in which all key elements are under the control of the investigator
- in an observational study key elements cannot be manipulated by the investigator.
- “It often, however, aids the interpretation of an observation study to consider the question: what would have been done in a comparable experiment?”
- Example: hormone replacement therapy and heart disease
- observational study – strong and statistically significant reduction in heart disease among women taking hormone replacement therapy
- women’s health study (JAMA, 2002, p.321) – statistically significant increase in risk among women randomized to hormone replacement therapy

Hydroxychloroquine

ORIGINAL ARTICLE

Observational Study of Hydroxychloroquine in Hospitalized Patients with Covid-19

Joshua Geleris, M.D., Yifei Sun, Ph.D., Jonathan Platt, Ph.D., Jason Zucker, M.D., Matthew Baldwin, M.D., George Hripcsak, M.D., Angelena Labella, M.D., Daniel K. Manson, M.D., Christine Kubin, Pharm.D., R. Graham Barr, M.D., Dr.P.H., Magdalena E. Sobieszczyk, M.D., M.P.H., and Neil W. Schluger, M.D.

Article Figures/Media Metrics

14 References 300 Citing Articles

June 18, 2020
N Engl J Med 2020; 382:2411-2418
DOI: 10.1056/NEJMoa2012410
Chinese Translation 中文翻译

Abstract

June 2020

ORIGINAL ARTICLE

A Randomized Trial of Hydroxychloroquine as Postexposure Prophylaxis for Covid-19

David R. Boulware, M.D., M.P.H., Matthew F. Pullen, M.D., Ananta S. Bangdiwala, M.S., Katelyn A. Pastick, B.Sc., Sarah M. Lofgren, M.D., Elizabeth C. Okafor, B.Sc., Caleb P. Skipper, M.D., Alanna A. Nascene, B.A., Melanie R. Nicol, Pharm.D., Ph.D., Mahsa Abassi, D.O., M.P.H., Nicole W. Engen, M.S., Matthew P. Cheng, M.D., et al.

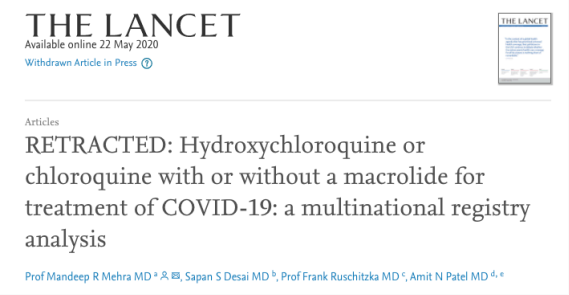
Article Figures/Media Metrics

18 References 128 Citing Articles Letters 11 Comments

August 6, 2020
N Engl J Med 2020; 383:517-525
DOI: 10.1056/NEJMoa2016638

August 2020

Hydroxychloroquine



THE LANCET
Available online 22 May 2020
Withdrawn Article in Press

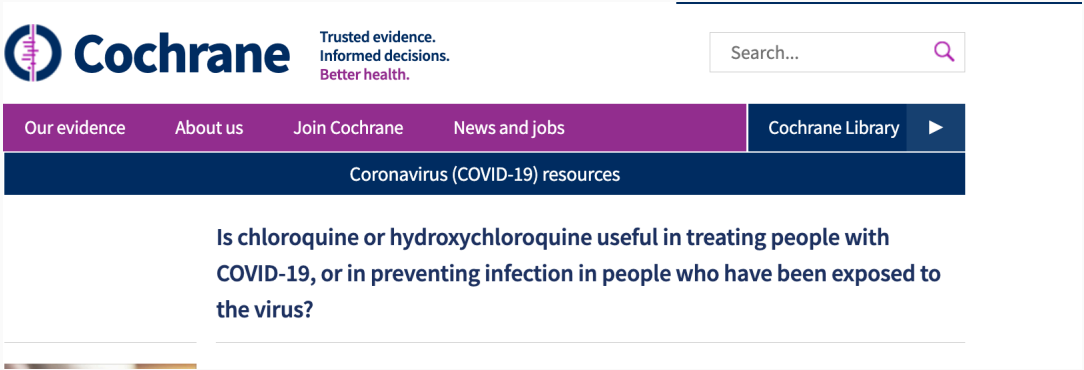
Articles

RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis

Prof Mandeep R Mehra MD [a], Sapan S Desai MD [b], Prof Frank Ruschitzka MD [c], Amit N Patel MD [d, e]

retracted

... hydroxychloroquine



Cochrane Reviews

... hydroxychloroquine

Cochrane Reviews ▾ Trials ▾ Clinical Answers ▾ About ▾ Help ▾

Cochrane Database of Systematic Reviews | Review - Intervention

Chloroquine or hydroxychloroquine for prevention and treatment of COVID-19

Bhagteshwar Singh, Hannah Ryan, Tamara Kredo, Marty Chaplin, Tom Fletcher Authors' declarations of interest

Version published: 12 February 2021 Version history

https://doi.org/10.1002/14651858.CD013587.pub2

Collapse all Expand all

Abstract

Hydroxychloroquine does not reduce deaths from COVID-19, and probably does not reduce the number of people needing mechanical ventilation.

Hydroxychloroquine caused more unwanted effects than a placebo treatment, though it did not appear to increase the number of serious unwanted effects.

We do not think new studies of hydroxychloroquine should be started for treatment of COVID

... hydroxychloroquine

Cochrane Database of Systematic Reviews | Editorial

Contested effects and chaotic policies: the 2020 story of (hydroxy) chloroquine for treating COVID-19

Susan Gould, Susan L Norris Authors' declarations of interest

Version published: 25 March 2021

https://doi.org/10.1002/14651858.ED000151

Cochrane Library

Trusted evidence.
Informed decisions.
Better health.

Title Abstract Ke

Cochrane Reviews | Trials | Clinical Answers | About | Help

Cochrane Database of Systematic Reviews | Review - Intervention

Ivermectin for preventing and treating COVID-19

Maria Popp, Miriam Stegemann, Maria-Inti Metzendorf, Susan Gould, Peter Kranke, Patrick Meybohm, Nicole Skoetz, Stephanie Weibel Authors' declarations of interest

Version published: 28 July 2021 Version history

https://doi.org/10.1002/14651858.CD015017.pub2

"Overall, the reliable evidence available does not support the use of ivermectin for treatment or prevention of COVID-19 outside of well-designed randomized trials."

Collapse all Expand all

Abstract

"Moderate-certainty evidence finds that large reductions in COVID-19 deaths are possible using ivermectin. ... The apparent safety and low cost suggest that ivermectin is likely to have a significant impact on the SARS-CoV-2 pandemic globally."

Linear regression recap

- generic form of linear regression, in matrix notation $y = X\beta + \epsilon$
- least squares estimate of β is $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$
- $\hat{\beta}$ has expected value β and variance-covariance matrix $\sigma^2(X^{\mathrm{T}}X)^{-1}$

- this is the maximum likelihood estimate if $\epsilon \sim N(0, \sigma^2 I)$
- $\hat{\beta} \sim N(\beta, \sigma^2(X^{\mathrm{T}}X)^{-1})$
- $\tilde{\sigma}^2 = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})/(n - p)$ — called s^2 in SM
- leads to t-tests for individual components β_j — and confidence intervals

- X is an $n \times p$ matrix of explanatory variables, which may be
  - measured in the sample (SM Ex 8.3),
  - fixed by design (SM Ex 8.4),
  - introduced to make the model more flexible (SM Ex 8.2)
  - X often called the design matrix

SM – Davison

in R, `model.matrix`

Aside: Lazy Notation

- $y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \ldots, n$
- $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{y}, \boldsymbol{\epsilon}$ vectors of length n
- $y = X\beta + \epsilon, \qquad$ also vectors of length n — the lazy way

- a generic observation $y \in \mathbb{R}$ for a generic vector of covariates $x \in \mathbb{R}^1$ often written

$$y = x^{\mathrm{T}}\beta + \epsilon$$

or even $x\beta + \epsilon$

- "where we hope there is no confusion"

Aside: ancillarity

SM Ex.12.4; Cox 2006 §4.3.1

- in a statistical model (i.e. likelihood function) that factorizes, in a way that separates the parameters, there are strong theoretical (and often practical) reasons for using only the relevant factor for inference
- in our example, even if (y, X) have a $(p+1)$-dimensional distribution (maybe even jointly multivariate normal), it would typically be the case that β, which by definition is $\mathbb{E}(y \mid X)$, is not a parameter of the distribution of X.
- So we would have something like

$$f(y, X; \beta, \sigma^2, \theta) = f_1(y \mid X; \beta, \sigma^2) f_2(X \mid \theta)$$

- and we base our inference for β and σ^2 (which are the parameters of interest, by assumption) on $f_1(y \mid X; \beta, \sigma^2)$
- in technical terms X is ancillary for (β, σ^2)
- If we did decide to include the variability in X as part of our analysis, our inference about β would be much less precise, and needlessly so, because we are worrying about explanatory variable values that were not seen in our data set

- residual sum of squares

$$SS(\hat{\beta}) = RSS_\Omega = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})$$

$SS(\hat{\beta})$ SM p.366; RSS_Ω LM-2 p.16; LM-1, p.15

- Decomposition of variance

$$\begin{aligned} y^{\mathrm{T}}y &= (y - \hat{y})^{\mathrm{T}}(y - \hat{y}) + \hat{y}^{\mathrm{T}}\hat{y} \\ &= (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}) + \hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\hat{\beta} \\ &= \text{Residual SS + Regression SS} \end{aligned}$$

LM Fig 2.1

- Typically first column of X is $(1, \ldots, 1)^{\mathrm{T}}$, so $y = \beta_0 + X_2\beta_2 + \epsilon$, say; then decomposition becomes

- residual sum of squares

$$SS(\hat{\beta}) = RSS_\Omega = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})$$

$SS(\hat{\beta})$ SM p.366; RSS_Ω LM-2 p.16; LM-1, p.15

- Decomposition of variance: $y^{\mathrm{T}}y = (y - \hat{y})^{\mathrm{T}}(y - \hat{y}) + \hat{y}^{\mathrm{T}}\hat{y}$ LM Fig 2.1

$$= (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}) + \hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\hat{\beta}$$

$$= \text{Residual SS + Regression SS}$$

- Typically first column of X is $(1, \ldots, 1)^{\mathrm{T}}$, so $y = \beta_0 + X_2\beta_2 + \epsilon$, say; then decomposition becomes

$$\begin{aligned}(y - \bar{y}\mathbf{1})^{\mathrm{T}}(y - \bar{y}\mathbf{1}) &= (y - X_2\hat{\beta}_2)^{\mathrm{T}}(y - X_2\hat{\beta}_2) + \hat{\beta}_2^{\mathrm{T}}(X_2^{\mathrm{T}}X_2)\hat{\beta} \\ \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - x_{2i}^{\mathrm{T}}\hat{\beta}_2)^2 + \hat{\beta}_2^{\mathrm{T}}(X_2^{\mathrm{T}}X_2)\hat{\beta}\end{aligned}$$

... comparing models

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}) + \hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta}$$

$$\text{Total SS} = \text{Residual SS} + \text{Regression SS}$$

RSS, $SS(\hat{\beta})$

- LHS is residual SS fitting only the 1-vector
- comparison of LHS to $SS(\hat{\beta})$ reflects importance of other βs, i.e. importance of explanatory variables
- $$F = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)} \sim$$
- here $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, but we don't care about β_1 $(\beta_0, \beta_1, \ldots, \beta_p)$

... comparing models

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= (y-X\hat{\beta})^{\mathrm{T}}(y-X\hat{\beta})+\hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta}\\ \text{Total SS} &= \text{Residual SS}+\text{Regression SS}\end{aligned}$$

RSS, $SS(\hat{\beta})$

- LHS is residual SS fitting only the 1-vector
- comparison of LHS to $SS(\hat{\beta})$ reflects importance of other βs, i.e. importance of explanatory variables
- $$F=\frac{(TSS-RSS)/(p-1)}{RSS/(n-p)} \sim F_{p-1,n-p}$$
- here $\beta=(\beta_1,\beta_2,\ldots,\beta_p)$, but we don't care about β_1 $(\beta_0,\beta_1,\ldots,\beta_p)$

... comparing models

- same argument can be derived for comparing submodels
- for example, testing $(\beta_2, \beta_3, \beta_4) = (0, 0, 0)$

- fit full model $\longrightarrow RSS_{full}$; fit reduced model $\longrightarrow RSS_{red}$
- $$F = \frac{(RSS_{red} - RSS_{full})/(p - q)}{RSS_{full}/(n - p)}$$
- see LM 3.1, SM §8.2 (p.367) for connection to likelihood ratio test

- when would we want to do this?

... comparing models

```
head(prostate)
#     lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
4 -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
5  0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
6 -1.0498221  3.2288  50 -1.386294   0 -1.38629       6     0  0.76547

model1 <- lm(lpsa ~ ., data = prostate)
```

... comparing models

```
> summary(model1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766157   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045142   0.157465   0.287  0.77503
pgg45        0.004525   0.004421   1.024  0.30886
---
Residual standard error: 0.7084 on 88 degrees of freedom

F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

... comparing models

```
model3 <- lm(lpsa ~ lcavol + lweight  + age + lbph + svi, data = prostate)
anova(model3,model1)
Analysis of Variance Table

Model 1: lpsa ~ lcavol + lweight  + age + lbph + svi
Model 3: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     91 45.526
2     88 44.163  3    1.3625 0.905 0.4421
```

does this make sense?

Factor variables

- *F*-tests are used when the columns to be removed form a group
- if a covariate is a factor, i.e. categorical, then `lm` will construct a set of dummy variables as part of the model matrix
- these variables should either all be in, or all be out

in most cases

- ```
  prostate$gleason_factor <- factor(prostate$gleason)
  levels(prostate$gleason_factor)
  ```
  ```
  [1] "6" "7" "8" "9"
  ```
  ```
  model3 <- lm(lpsa ~ .-gleason, data=prostate)
  ```
  ```

... factor variables

```
model4 <- lm(lpsa ~ .-gleason, data=prostate)
summary(model4)
> Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.913282   0.840838   1.086  0.28044
lcavol           0.569988   0.090100   6.326 1.09e-08 ***
lweight          0.468791   0.169610   2.764  0.00699 **
age             -0.021749   0.011361  -1.914  0.05890 .
lbph             0.099685   0.058984   1.690  0.09464 .
svi              0.745879   0.247398   3.015  0.00338 **
lcp             -0.125112   0.095591  -1.309  0.19408
pgg45            0.004990   0.004672   1.068  0.28848
gleason_factor7  0.267607   0.219419   1.220  0.22595
gleason_factor8  0.496820   0.769267   0.646  0.52011
gleason_factor9 -0.056215   0.500196  -0.112  0.91078
```

... factor variables

```
Analysis of Variance Table

Model 1: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45 + gleason_factor
Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     86 42.724
2     89 44.204 -3   -1.4805 0.9934 0.3999
```

... factor variables

- with designed experiments, covariates are often factors set at pre-determined levels
- see, e.g. Example 8.4 in SM

also Ch 14 in LM-2; Ch 13 in LM-1

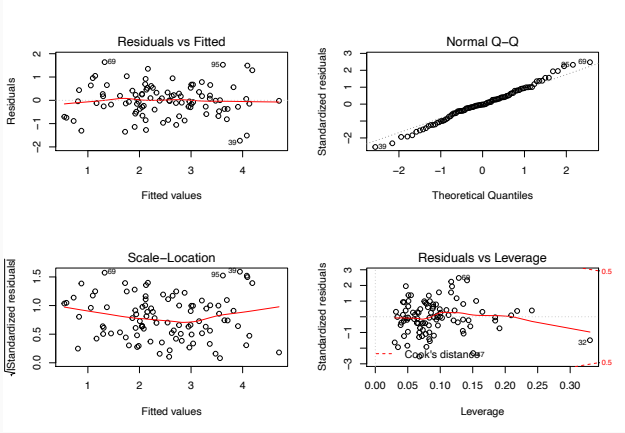
- if the design is perfectly balanced, then X has orthogonal columns, and $X^{\mathrm{T}}X$ is diagonal
- so $\hat{\beta}_j$'s are uncorrelated, and hence independent (under normality assumption)

- more generally we might have $X^{\mathrm{T}}X$ block diagonal, e.g.

importance?

- assumptions on errors: $\epsilon_i \sim_{i.i.d.} N(0, \sigma^2)$ on structure $\mathbb{E}(y \mid X) = X\beta$
- normality; constant variance; independent

```
plot(model1)
```



... Model checking

- residuals: $\hat{\epsilon}_i =$
- $\text{Var}(\hat{\epsilon}) =$
- i.e. don't all have the same variance

- hat matrix $H =$

- standardized residuals: $r_i =$

- Cook's distance $C_i =$

https://data.library.virginia.edu/diagnostic-plots/

Collinearity

- simple model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \ldots n$
- if $x_1 \perp x_2$, then interpretation of β_1 and β_2 clear
- if $x_1 = x_2$ then β_1 and β_2 not separately identifiable
- usually we're somewhere in between, at least in observational studies
- may be very difficult to dis-entangle effects of correlated covariates
- example: health effects of air pollution
- measurable increase in mortality on high-pollution days
- measurable increase in mortality on high-temperature days
- high temperatures and high levels of pollutants tend to co-occur +++
- mathematically, $X^T X$ is nearly singular, or at least ill-conditioned, so calculation of its inverse is subject to numerical errors
- if $p > n$ then $X^T X$ not invertible, no LS solution

ridge, Lasso

On the 'Breast cancer epidemic' poster

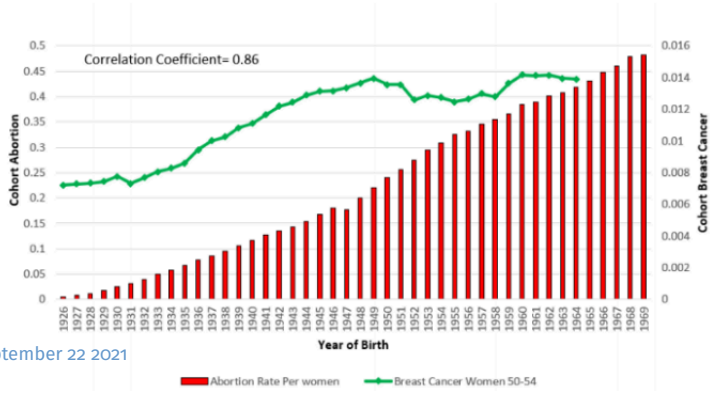
A poster and presentation at the RSS conference featured flawed analysis.

Anthony B. Masters 6 days ago · 4 min read ★

First: Correlation is not causation

Tyler Vigen

Figure 5. England & Wales. Cumulated Cohort Incidence for Birth cohorts of Women. Breast Cancer within ages 50-54 and cumulated cohort abortion rates. Correlation coefficient 0.86.



Second, there is no discussion of other well-known risk factors for breast cancer.

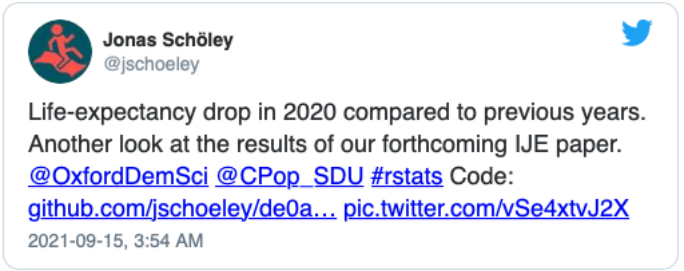
Instead of calling upon extensive research about breast cancer, the authors cite themselves. Large prospective cohort studies do not imply heightened risks of breast cancer after abortion.

Third, there is no exploration of absolute and relative risks.

There is an established increased risk of breast cancer using contraceptive pills. Breast cancer is rare among young women. Increased risk during this time means a low number of additional cases.

Fourth, methods and sources are unclear

In the News



link