Big data in health care
– a traditionalist's words of caution

Gustaf Edgren, MD PhD
Associate professor of Epidemiology and Hematologist
Karolinska Institutet/Karolinska University Hospital

Sweden: a nation of hamsters

- National registration number for all inhabitants since 1948
- Nationwide health registers starting in 1958 (cancer), 1960 (cause of death), 1964 (inpatient care), 1973 (medical birth), 2004 (drug prescriptions), 2003 (outpatient and primary care)
- Additional registers on: income, family structure, occupation, school grades, immigration/emigration, etc.

BIG DATA?

Outline

- Background
- Introductory example(s) of big data in health
- An epidemiologist's words of caution:
  - Common problems
  - Common solutions
- More examples

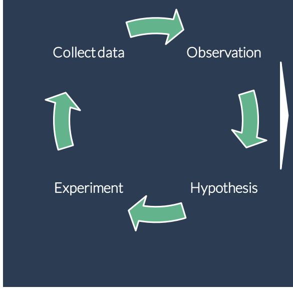
BACKGROUND

So, what's big data?

- The traditional definition of 'Big data' is typically technocratic – 3V:
  - Volume
  - Velocity
  - Variation
- > 100GB?, >100TB?, or >100PB?
- Does it matter?
- A practical definition is perhaps that the 3 V's together poses a technological challenge that can only be resolved with non-traditional methods?

The big data paradigm in health

Traditional paradigm



Big data paradigm

- Finds insights from complex, noisy, heterogeneous, longitudinal, and voluminous data.
- Inherently data driven
- Agnostic: Not hypothesis testing, but rather hypothesis generating – let the data speak for themselves!

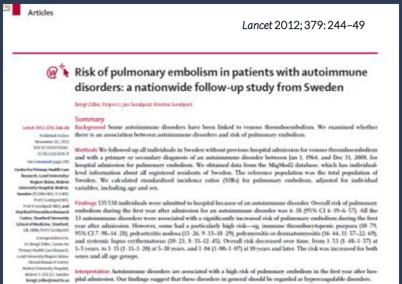
Similarities to 'omics' revolution

- One can easily see a parallel to the genomics revolution
- Transition from small scale datasets with one-dimensional data – all hypothesis driven
- Now: all massive-scale data with multidimensional data – all using agnostic methods
- Key lesson: replication is key

New paradigm? Perhaps, but…

- Health care 'big data' applications are typically based on routine administrative data (e.g. insurance claims data, discharge data, prescription data, etc)
- To allow <u>any</u> inference form such data, one must understand the data <u>context</u>:
  - How the data was collected
  - Why the data was collected
  - In what situations the data was collected
  - What incentive structures might have affected the data collection

Example 1



Articles

Lancet 2012; 379: 244–49

Risk of pulmonary embolism in patients with autoimmune disorders: a nationwide follow-up study from Sweden

Bengt Zöller, Xinjun Li, Jan Sundquist, Kristina Sundquist

Lancet 2012; 379: 244–49

Published Online November 29, 2011 DOI:10.1016/S0140-6736(11)61306-8

See Comment page 200

Center for Primary Health Care Research, Lund University/Region Skåne, Malmö University Hospital, Malmö, Sweden (B Zöller MD, X Li MD, Prof J Sundquist MD, Prof K Sundquist MD); and Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, CA, USA (Prof J Sundquist)

Correspondence to: Dr Bengt Zöller, Center for Primary Health Care Research, Lund University/Region Skåne, Clinical Research Centre, Malmö University Hospital, Malmö S-205 02, Sweden bengt.zoller@med.lu.se

Summary

Background Some autoimmune disorders have been linked to venous thromboembolism. We examined whether there is an association between autoimmune disorders and risk of pulmonary embolism.

Methods We followed up all individuals in Sweden without previous hospital admission for venous thromboembolism and with a primary or secondary diagnosis of an autoimmune disorder between Jan 1, 1964, and Dec 31, 2008, for hospital admission for pulmonary embolism. We obtained data from the MigMed2 database, which has individual-level information about all registered residents of Sweden. The reference population was the total population of Sweden. We calculated standardised incidence ratios (SIRs) for pulmonary embolism, adjusted for individual variables, including age and sex.

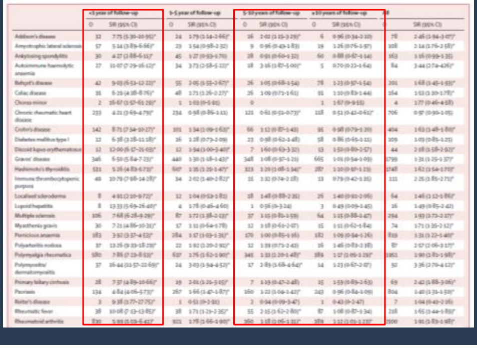
Findings 535 538 individuals were admitted to hospital because of an autoimmune disorder. Overall risk of pulmonary embolism during the first year after admission for an autoimmune disorder was 6·38 (95% CI 6·19–6·57). All the 33 autoimmune disorders were associated with a significantly increased risk of pulmonary embolism during the first year after admission. However, some had a particularly high risk—eg, immune thrombocytopenic purpura (10·79, 95% CI 7·98–14·28), polyarteritis nodosa (13·26, 9·33–18·29), polymyositis or dermatomyositis (16·44, 11·57–22·69), and systemic lupus erythematosus (10·23, 8·31–12·45). Overall risk decreased over time, from 1·53 (1·48–1·57) at 1–5 years, to 1·15 (1·11–1·20) at 5–10 years, and 1·04 (1·00–1·07) at 10 years and later. The risk was increased for both sexes and all age groups.

Interpretation Autoimmune disorders are associated with a high risk of pulmonary embolism in the first year after hospital admission. Our findings suggest that these disorders in general should be regarded as hypercoagulable disorders.

Example 1: What?

- An analysis of whether autoimmune disease increases risk of pulmonary embolism
- Important implications: common diseases, feared complication, effective prophylactic treatment
- Nationwide cohort study including entire Swedish population 1964-2008 (~15 million individuals, 500,000 with autoimmune disease)

Example 1: Results



	<1 year of follow-up		1–5 year of follow-up		5–10 years of follow-up		≥10 years of follow-up		All	
	O	SIR (95% CI)	O	SIR (95% CI)	O	SIR (95% CI)	O	SIR (95% CI)	O	SIR (95% CI)
Addison's disease	32	7·75 (5·30–10·95)*	24	1·79 (1·14–2·66)*	16	2·02 (1·15–3·29)*	6	0·96 (0·34–2·10)	78	2·46 (1·94–3·07)*
Amyotrophic lateral sclerosis	57	5·14 (3·89–6·66)*	23	1·54 (0·98–2·32)	9	0·96 (0·43–1·83)	19	1·26 (0·76–1·97)	108	2·14 (1·75–2·58)*
Ankylosing spondylitis	30	4·27 (2·88–6·11)*	45	1·27 (0·93–1·70)	28	0·91 (0·60–1·32)	60	0·88 (0·67–1·14)	163	1·16 (0·99–1·35)
Autoimmune haemolytic anaemia	27	11·07 (7·29–16·12)*	34	3·73 (2·58–5·22)*	18	3·16 (1·87–5·00)*	5	0·70 (0·22–1·64)	84	3·44 (2·74–4·26)*
Behçet's disease	42	9·03 (6·53–12·22)*	55	2·05 (1·55–2·67)*	26	1·05 (0·68–1·54)	78	1·23 (0·97–1·54)	201	1·68 (1·45–1·93)*
Celiac disease	31	6·29 (4·28–8·76)*	48	1·71 (1·26–2·27)*	26	1·09 (0·71–1·61)	51	1·10 (0·83–1·44)	154	1·53 (1·30–1·78)*
Chorea minor	2	16·67 (1·57–61·29)*	1	1·03 (0–5·91)	0		1	1·67 (0–9·55)	4	1·77 (0·46–4·58)
Chronic rheumatic heart disease	233	4·21 (3·69–4·79)*	234	0·98 (0·86–1·11)	121	0·61 (0·51–0·73)*	118	0·51 (0·42–0·61)*	706	0·97 (0·90–1·05)
Crohn's disease	142	8·71 (7·34–10·27)*	101	1·34 (1·09–1·63)*	66	1·12 (0·87–1·43)	95	0·98 (0·79–1·20)	404	1·63 (1·48–1·80)*
Diabetes mellitus type I	12	6·38 (3·28–11·18)*	16	1·28 (0·73–2·09)	23	0·98 (0·62–1·48)	58	0·86 (0·65–1·11)	109	1·03 (0·85–1·25)
Discoid lupus erythematosus	12	12·00 (6·17–21·03)*	12	1·94 (1·00–3·40)*	7	1·60 (0·63–3·32)	13	1·50 (0·80–2·57)	44	2·18 (1·58–2·92)*
Graves' disease	346	6·50 (5·84–7·23)*	440	1·30 (1·18–1·43)*	348	1·08 (0·97–1·21)	665	1·01 (0·94–1·09)	1799	1·31 (1·25–1·37)*
Hashimoto's thyroiditis	511	5·26 (4·83–5·73)*	687	1·35 (1·25–1·47)*	323	1·20 (1·08–1·34)*	287	1·10 (0·97–1·23)	1748	1·62 (1·54–1·70)*
Immune thrombocytopenic purpura	49	10·79 (7·98–14·28)*	34	2·02 (1·40–2·82)*	15	1·32 (0·74–2·18)	13	0·79 (0·42–1·35)	111	2·25 (1·85–2·71)*
Localised scleroderma	8	4·81 (2·10–9·72)*	12	1·04 (0·53–1·81)	18	1·48 (0·88–2·35)	26	1·40 (0·91–2·05)	64	1·46 (1·12–1·86)*
Lupoid hepatitis	8	13·33 (5·69–26·40)*	4	1·78 (0·46–4·60)	1	0·56 (0–3·24)	3	0·49 (0·09–1·45)	16	1·49 (0·85–2·42)
Multiple sclerosis	106	7·68 (6·28–9·29)*	87	1·72 (1·38–2·13)*	37	1·15 (0·81–1·59)	64	1·15 (0·88–1·47)	294	1·93 (1·73–2·17)*
Myasthenia gravis	30	7·21 (4·86–10·31)*	17	1·11 (0·64–1·78)	12	1·18 (0·61–2·07)	15	1·11 (0·62–1·84)	74	1·71 (1·35–2·15)*
Pernicious anaemia	183	3·92 (3·37–4·53)*	284	1·17 (1·03–1·31)*	170	1·00 (0·85–1·16)	182	1·09 (0·94–1·26)	819	1·31 (1·22–1·40)*
Polyarteritis nodosa	37	13·26 (9·33–18·29)*	22	1·92 (1·20–2·91)*	12	1·39 (0·71–2·43)	16	1·46 (0·83–2·38)	87	2·57 (2·06–3·17)*
Polymyalgia rheumatica	580	7·86 (7·23–8·53)*	617	1·76 (1·62–1·90)*	341	1·33 (1·20–1·48)*	389	1·17 (1·06–1·29)*	1951	1·90 (1·81–1·98)*
Polymyositis/dermatomyositis	37	16·44 (11·57–22·69)*	24	3·03 (1·94–4·52)*	17	2·83 (1·65–4·54)*	14	1·23 (0·67–2·07)	92	3·36 (2·70–4·12)*
Primary biliary cirrhosis	28	7·37 (4·89–10·66)*	19	2·01 (1·21–3·15)*	7	1·03 (0·41–2·13)	15	1·53 (0·85–2·53)	69	2·42 (1·88–3·06)*
Psoriasis	134	4·84 (4·06–5·73)*	267	1·66 (1·47–1·87)*	160	1·22 (1·04–1·43)*	243	0·96 (0·84–1·09)	804	1·40 (1·31–1·50)*
Reiter's disease	3	9·38 (1·77–27·75)*	1	0·51 (0–2·91)	2	0·94 (0·09–3·47)	1	0·43 (0–2·47)	7	1·04 (0·42–2·16)
Rheumatic fever	38	10·08 (7·13–13·85)*	38	1·71 (1·21–2·35)*	55	2·15 (1·62–2·80)*	87	1·08 (0·87–1·34)	218	1·65 (1·44–1·89)*
Rheumatoid arthritis	830	5·93 (5·53–6·42)*	921	1·78 (1·66–1·90)*	560	1·18 (1·08–1·31)*	389	1·12 (1·01–1·23)*	2700	1·91 (1·83–1·98)*

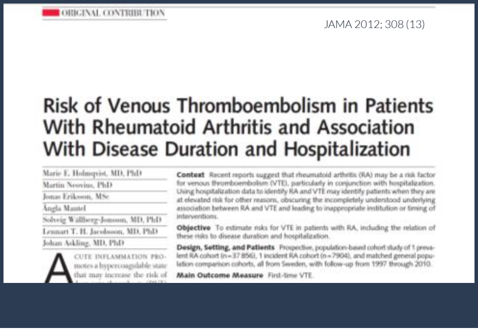
Example 1: Interpretation

- The risk of VTE is very high for individuals with a new diagnosis of autoimmune disease;
- the risk then decreases gradually, ostensibly with successful treatment?
- Perhaps patients with autoimmune disease should routinely be given thrombosis prophylaxis?
- Important and powerful illustration of the power of big(ish) data?

No!

Why?

Example 2



ORIGINAL CONTRIBUTION

JAMA 2012; 308 (13)

Risk of Venous Thromboembolism in Patients With Rheumatoid Arthritis and Association With Disease Duration and Hospitalization

Marie E. Holmqvist, MD, PhD
Martin Neovius, PhD
Jonas Eriksson, MSc
Ängla Mantel
Solveig Wållberg-Jonsson, MD, PhD
Lennart T. H. Jacobsson, MD, PhD
Johan Askling, MD, PhD

Context Recent reports suggest that rheumatoid arthritis (RA) may be a risk factor for venous thromboembolism (VTE), particularly in conjunction with hospitalization. Using hospitalization data to identify RA and VTE may identify patients when they are at elevated risk for other reasons, obscuring the incompletely understood underlying association between RA and VTE and leading to inappropriate institution or timing of interventions.

Objective To estimate risks for VTE in patients with RA, including the relation of these risks to disease duration and hospitalization.

Design, Setting, and Patients Prospective, population-based cohort study of 1 prevalent RA cohort (n=37 856), 1 incident RA cohort (n=7904), and matched general population comparison cohorts, all from Sweden, with follow-up from 1997 through 2010.

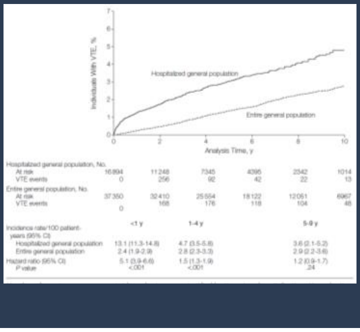
Main Outcome Measure First-time VTE.

ACUTE INFLAMMATION PROmotes a hypercoagulable state that may increase the risk of

Example 2: What?

- An analysis of whether rheumatoid arthritis (an autoimmune disease) increases risk of thromboembolism
- Important implications: common disease, feared complications, effective prophylactic treatment
- Large-scale cohort study of >40,000 patients with rheumatoid arthritis

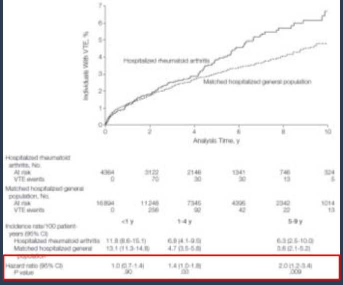
Example 3: Results (1)



Hospitalized general population, No.						
At risk	16 894	11 248	7345	4395	2342	1014
VTE events	0	256	92	42	22	13
Entire general population, No.						
At risk	37 350	32 410	25 554	18 122	12 051	6967
VTE events	0	168	176	118	104	48

	<1 y	1-4 y	5-9 y
Incidence rate/100 patient-years (95% CI)			
Hospitalized general population	13.1 (11.3-14.8)	4.7 (3.5-5.8)	3.6 (2.1-5.2)
Entire general population	2.4 (1.9-2.9)	2.8 (2.3-3.3)	2.9 (2.2-3.6)
Hazard ratio (95% CI)	5.1 (3.9-6.6)	1.5 (1.3-1.9)	1.2 (0.9-1.7)
P value	<.001	<.001	.24

Example 3: Results (2)



Hospitalized rheumatoid arthritis, No.						
At risk	4364	3122	2146	1341	746	324
VTE events	0	70	30	30	13	5
Matched hospitalized general population, No.						
At risk	16 894	11 248	7345	4395	2342	1014
VTE events	0	256	92	42	22	13

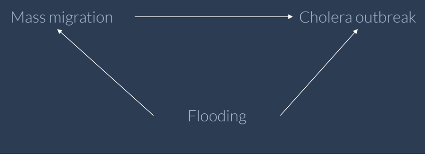
	<1 y	1-4 y	5-9 y
Incidence rate/100 patient-years (95% CI)			
Hospitalized rheumatoid arthritis	11.8 (8.6-15.1)	6.8 (4.1-9.5)	6.3 (2.5-10.0)
Matched hospitalized general population	13.1 (11.3-14.8)	4.7 (3.5-5.8)	3.6 (2.1-5.2)
Hazard ratio (95% CI)	1.0 (0.7-1.4)	1.4 (1.0-1.8)	2.0 (1.2-3.4)
P value	.90	.03	.009

So, lessons learned?

- Even something seemingly trivial, such as assessing whether one patient group is at increased risk of a second disease is not all that trivial
- First, we must recall some key things that pertain to the analysis of any medical data…
  - Confounding (by indication)
  - Reverse causation
  - Surveillance bias
  - Misclassification of diagnosis
  - Will Roger's phenomenon

1. Indication is key (confounding)

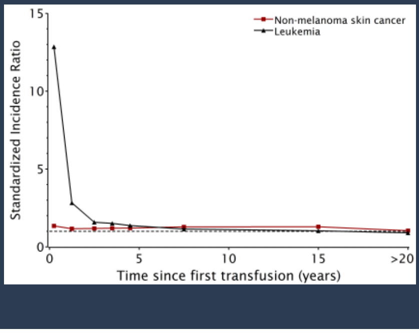
- Confounding is a central element in epidemiology and medical research
- It results from the confusion of the effect of one variable on a particular outcome with a third variable:



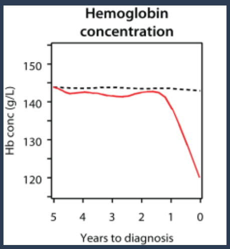
1. Indication is key (2)

- In any analysis of clinical data 'confounding by indication' is a an ever-present problem
- It results from the conscious choice of different treatments for patients with different prognosis
  - Sicker patients get more aggressive treatment
  - *Or, really sick patients can't get aggressive treatment because they are too sick*

2. Sequence isn't all that easy



2. Post hoc ergo propter hoc…



2. Reverse causation

- It will make it appear from your data that different events happen in one particular order
- However, due to incompleteness, that order will often be obfuscated, leading to incorrect characterization of sequence/causation

3. Surveillance bias

- As a rule of thumb: the more you look, the more you will find
- In any situation when patients undergo intense scrutiny (i.e. screening), disease rates automatically go up
- Implicitly, and disturbingly, complication rates therefore often go up with the quality of care

3. If you look, you shall find

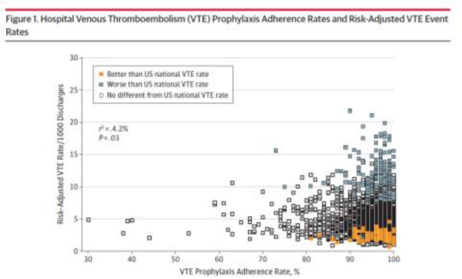
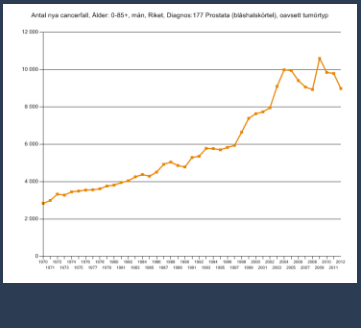
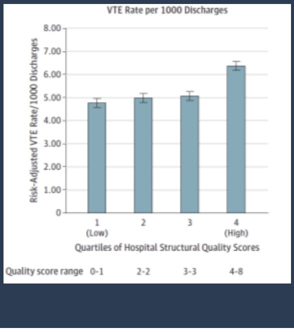


Figure 1. Hospital Venous Thromboembolism (VTE) Prophylaxis Adherence Rates and Risk-Adjusted VTE Event Rates

3. If you look you shall find



3. Surveillance bias



4. Doctors are sloppy/lazy/greedy

- Most EHR data are based on diagnosis and procedure codes
- Such coding systems are generally inherently accurate and well structured
- However, the accuracy of the coding use is dependent on:
  1. Quality of care
  2. Misunderstanding of coding practices
  3. Financial incentives for coding use
  4. Diagnostic progress – i.e. misdiagnoses

5. Will Rogers' phenomenon

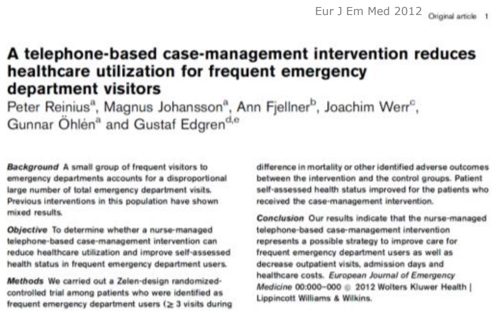
When the Okies left Oklahoma and moved to California, they raised the average intelligence in both states

CASE STUDIES

Application 1: ER frequent fliers

- In Sweden (as in most other parts of the world) a small part of the population account for a large fraction of healthcare spending
- It is possible to predict who will have large healthcare needs in the near future
- We hypothesized that interventions targeted at the small high-risk population may prevent unnecessary healthcare consumption

Hälso- och sjukvårdsförvaltningen
STOCKHOLMS LÄNS LANDSTING



Eur J Em Med 2012 Original article 1

A telephone-based case-management intervention reduces healthcare utilization for frequent emergency department visitors

Peter Reinius[a], Magnus Johansson[a], Ann Fjellner[b], Joachim Werr[c], Gunnar Öhlén[a] and Gustaf Edgren[d,e]

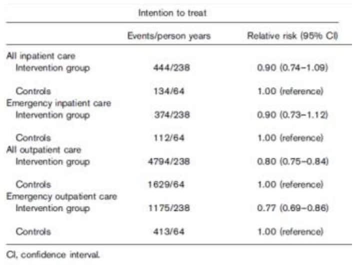
Background A small group of frequent visitors to emergency departments accounts for a disproportional large number of total emergency department visits. Previous interventions in this population have shown mixed results.

Objective To determine whether a nurse-managed telephone-based case-management intervention can reduce healthcare utilization and improve self-assessed health status in frequent emergency department users.

Methods We carried out a Zelen-design randomized-controlled trial among patients who were identified as frequent emergency department users (≥ 3 visits during difference in mortality or other identified adverse outcomes between the intervention and the control groups. Patient self-assessed health status improved for the patients who received the case-management intervention.

Conclusion Our results indicate that the nurse-managed telephone-based case-management intervention represents a possible strategy to improve care for frequent emergency department users as well as decrease outpatient visits, admission days and healthcare costs. *European Journal of Emergency Medicine* 00:000–000 © 2012 Wolters Kluwer Health | Lippincott Williams & Wilkins.

Hälso- och sjukvårdsförvaltningen
STOCKHOLMS LÄNS LANDSTING



	Intention to treat	
	Events/person years	Relative risk (95% CI)
All inpatient care		
Intervention group	444/238	0.90 (0.74–1.09)
Controls	134/64	1.00 (reference)
Emergency inpatient care		
Intervention group	374/238	0.90 (0.73–1.12)
Controls	112/64	1.00 (reference)
All outpatient care		
Intervention group	4794/238	0.80 (0.75–0.84)
Controls	1629/64	1.00 (reference)
Emergency outpatient care		
Intervention group	1175/238	0.77 (0.69–0.86)
Controls	413/64	1.00 (reference)

CI, confidence interval.

33

Application 2: New cancer syndromes?

- We all know there is a heritable component for many cancers
- For example, one estimate of the heritability of breast cancer is 28%
- This is an average – estimating how much the variability in risk is influenced by genetic factors
- However, rare highly penetrant genes should not have a strong effect on heritability measures

Background (2)

- There are many well-characterized, monogenic cancer syndromes described in the literature, e.g.:
  - BRCA1/2: breast and ovarian cancer
  - Familial adenomatous polyposis (FAP) and lynch syndrome: colon polyps and/or cancer
  - Li-Fraumeni syndrome: sarcomas, breast, brain and adrenal gland cancers, as well as leukemia
  - MEN-1/2: endocrine cancers
  - Dicer1 syndrome: lung, kidney, ovaries, etc

Background (3)

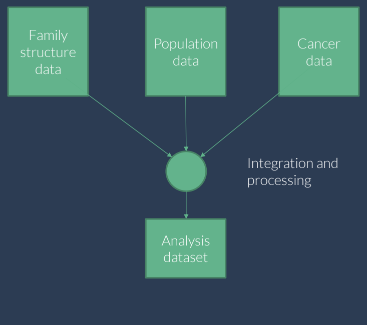
- Most of the known cancer syndromes are caused by highly penetrant single mutations
- Most have also been ”detected” by careful individual researchers or clinicians with access to a particularly informative patient material
  - Selective in terms of cancer sites
  - Unlikely to pick up genes with low penetrance
- No (to me known) systematic attempts have been made to find new rare, monogenic cancer syndromes*

*GWAS:s don’t count

Approach (1)

- Very simple analytical approach:
  1. Use multi-generation register to find all ”maximized” family trees (i.e. top people and their descendants)
  2. Use cancer register to identify all cancers in each family tree (i.e. observed)
  3. Use Poisson regression (or suitable alternative) to calculate expected number of cancers (given age, calendar year, sex, etc.)
  4. Reiterate step 2-3 for one cancer at a time, combinations of 2 cancers, 3 cancers, etc.

Data process



Family structure data

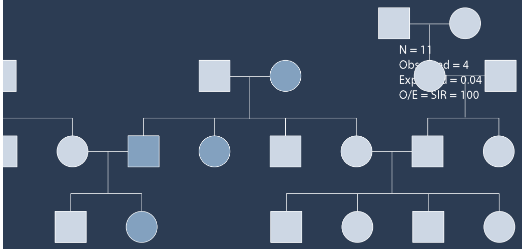
Population data

Cancer data

Integration and processing

Analysis dataset

Example family



Concluding thoughts

- Most important lesson: You have to get as close to the bottom of the data context as possible!
- Data scientists and data context experts have to work together
  - Who needs to be involved?
- Variation in data availability:
  - Behaviour, almost no data – 'hard outcomes,' a lot of data
  - How can we improve that?

gustaf.edgren@ki.se