# Probability and Risk

Improving public understanding of probability and risk with special emphasis on its application to the law. Why Bayes theorem and Bayesian networks are needed

**Norman Fenton**

Norman Fenton is Professor in Risk Information Management at Queen Mary University of London and also a Director of Agena, a company that specialises in risk management for critical systems.

**Martin Neil**

Martin is Professor in Computer Science and Statistics at QMUL and a Director of Agena Ltd.

**About Me**

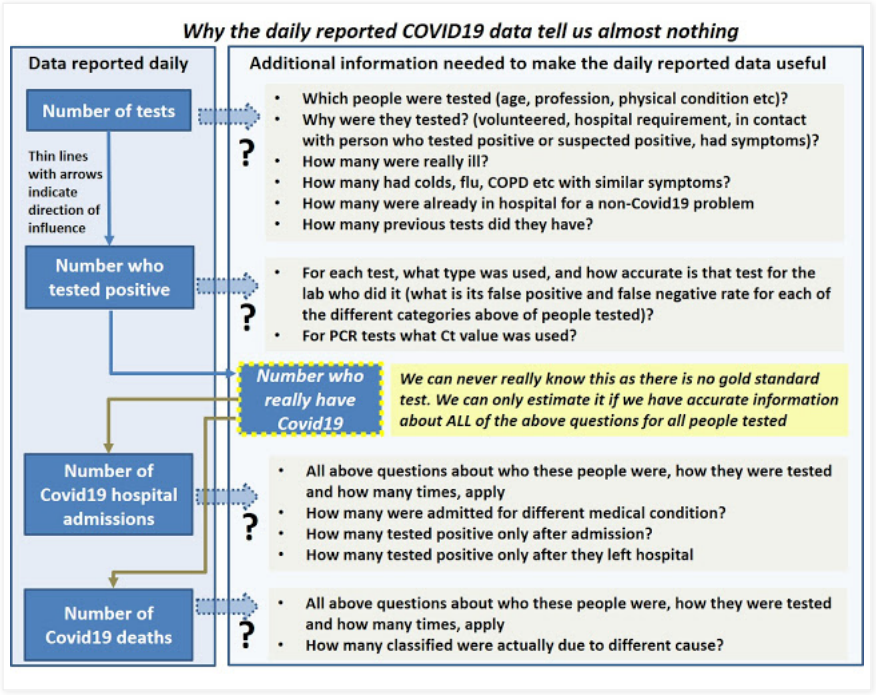
**Norman Fenton**

Norman's experience in risk assessment covers application domains such as legal reasoning (he has been an expert witness in major criminal and civil cases), software project risk, medical decision-making, vehicle reliability, football prediction, transport systems, and financial services. Norman has published over 130 articles and 5 books on these subjects

View my complete profile

**Sunday, 11 October 2020**

## Why we know so little about COVID-19 from testing data - and why some extra easy-to-get data would make a big difference




*This blog post provides some context for a short article (with Martin Neil, Scott McLachlan and Magda Osman) that was published in LockdownSkeptics and which has received quite a bit of attention.*

The daily monitoring of COVID-19 cases (such as the very crude analysis we have been doing) are intended ultimately to determine what the 'current' population infection rate really is and how it is changing.

However, in the absence of a gold-standard test for COVID-19, it is always uncertain whether a person has the virus (let alone whether a person can infect someone else). Obviously this means that the population infection rate (sometimes referred to as the *community infection prevalence rate*) on a given day is also unobservable. The best we can do is estimate it from data that are observable. To get a feel for how complex this really is to do properly - and why current estimates are unreliable, here is a (massively simplified, yes really - see **Notes about simplified assumptions** below) schematic** showing the information we need to get these estimates.

**Book "Risk Assessment and Decision Analysis with Bayesian Networks"**

- Book blog page
- Buy (Amazon)
- Buy (CRC Press)

**Key readings**

- Bayes and causal modelling in decision making, uncertainty and risk
- Irrational restrictions on Bayes in the Law
- Probability Fallacies and the Law

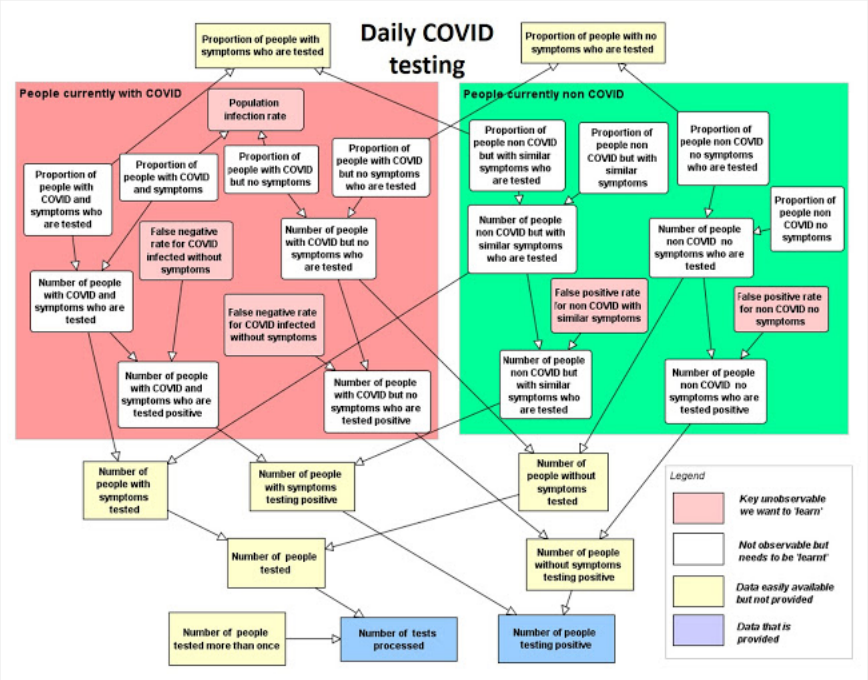
**Labels**

- AgenaRisk
- Bayes and probability theory
- case study
- COVID
- legal reasoning
- likelihood ratio
- medical
- New paper
- review
- risk assessment

**Links**

- BAYES-KNOWLEDGE Blog
- Agena: Bayesian networks
- Book: Risk Assessment with Bayesian Networks
- Bayes and the Law
- Pi Football (Using Bayesian nets to predict football results)
- Probability: Fallacies, Myths and Puzzles
- Risk Assessment and Decision Analysis at Queen Mary

**Blog archive**




Note that ***all*** the variables we need to 'know' for accurate estimation (the rectangle boxes coloured light red and white) are ***unobservable***. Hence, we are totally reliant on the other things (the variables represented by yellow and blue rectangles) which ***are*** observable.

But here is the BIG problem: the only accessible daily data we have (e.g. from https://coronavirus.data.gov.uk/) are the two blue rectangles: ***number of tests processed*** and ***number of people testing positive***. This means that any estimates of the things we really want to know are poor and highly uncertain (including the regular updates we have been providing based on this data). Yet, in principle, we should easily be able to get daily data for all the yellow rectangles and, if we did, our estimates would be far more accurate. Given the critical need to know these things more accurately, it is a great shame that these data are not available.

**Notes about simplified assumptions**

There are many such assumptions, but here I list just the most critical ones:

- We make a crucial distinction between people who *do* and *do not* have COVID symptoms - for the important reason that a) the former are more likely to be tested than the latter, and b) the testing accuracy rates will be different in each case. However, we don't (but really should) also distinguish between people who *have* and *have not* been in recent contact with a person tested positive, because again a) the former are more likely to be tested; and b) the testing accuracy rates will be different in each case. It could also be reasonably argued that we should also distinguish between different age categories.
- We are making the massively simplified assumption that the testing process is somehow 'constant'. Not only are there many different types of tests, but for the most common - PCR testing - there are massive variations depending on what 'Ct value' is used (i.e. the number of cycles) and small changes can lead to radically different false positive rates. If there are government changes to the ct value guidelines then this can cause apparent (but non-real) massive changes in the 'population infection rate' from one day to the next.
- While we have allowed for the fact that some people are tested multiple times (hence the observable, but never reported, variable **number of people tested more than once**) this actually massively over-simplifies a very complex problem. If a person tests positive where the ct value was above 40, then (because it is known that ct values even above 30 lead to many false positives) the recommendation is to retest, but we do not know if and when this happens and how many retests are performed. Similarly, some people may receive multiple negative tests before a single positive test and such people would count only as one of the people testing positive.

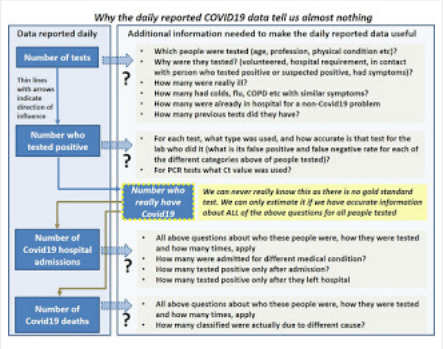
**The schematic is actually a representation of what is called a Bayesian network; the direction of the arrows is important because every variable (box) that has arrows going into it is calculated as an arithmetic or statistical function of the variable which are its 'parents'.

As all unobserved variables like ***population infection rate*** are never known for certain they will always be represented as a probability distribution (which could be summarised, for example as "a 95% chance of being between 0.1% and 20%" or something like that). As we enter observed data (such as ***number of people testing positive***) we can calculate the updated probability of each unobserved variable; so, for example, the population infection rate might change to "a 95% chance of being between 0.1 and 10%". The more data we enter for the observable variables the more accurate the estimates for the unobserved variables will be. Unlike traditional statistical methods, Bayesian inference works 'backwards' (in the reverse direction of the arrows) as well as forwards.

Time to demand the evidence to support continued C...
Why we know so little about COVID-19 from testing ...
COVID19 Hospital admissions data: evidence of expo...
► September (8)
► August (1)
► July (3)
► June (1)
► May (3)
► April (4)
► March (3)
► January (1)
► 2019 (22)
► 2018 (31)
► 2017 (9)
► 2016 (15)
► 2015 (22)
► 2014 (9)
► 2013 (7)
► 2012 (8)
► 2011 (11)

We have published many papers and reports applying Bayesian network analysis to COVID data. For this and related work see, for example:

- Impact of false positives in Covid testing
- Covid19 hospital admissions data: evidence of exponenial increase?
- Don't panic: limits to what we know about Covid-19 PC testing, inferred infection rates and alse positive rates
- A privacy-preserving Bayesian network model for personalised COVID19 risk assessment and contact tracing
  Covid-19: Infection rates are higher, fatality rates lower than widely reported
- Coronavirus: country comparisons are pointless unless we account for these biases in testing
- Why most studies into COVID19 risk factors may be producing flawed conclusions - and how to fix the problem
- Causal explanations, error rates, and human judgment biases missing from the COVID-19 narrative and statistics
- Covid-19 risk for the black and minority ethnic community: why reports are misleading and create unjustified fear and anxiety
- UK Covid19 death rates by religion: Jews by far the highest and atheists by far the lowest 'overall' - but what does it mean?




at 12:20

Labels: COVID

## 2 comments:

**Finn McCool** 12 October 2020 at 17:58

Great article.
It's an uphill task trying to explain conditional probability to friends and relatives.

Reply

Replies

**Norman Fenton** 13 October 2020 at 05:49

Thanks

**Reply**

Enter your comment...

Comment as: Google Accour

Publish Preview

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)