Background

You are asked in conduct a study on the possible role of MMR vaccination on the etiology of autism (cf. Madsen study in Practicum session). Suppose that 315 cases of autism were diagnosed in the dynamic population,[1] 1991-1998, followed to December 1999 ("the base"). The base can be split into 36 age-year 'cells': each of the 28 full-year cells (see diagram on next page[2]) contain 67,000 infant-years of experience, and each of the other 8 contains half this amount.

Unfortunately, there is no electronic vaccination registry that documents this base. The vaccination records are maintained by the regional heath authorities in card files, one card per family. They can determine if a given child was(1) or was not(0) vaccinated before a specified age/date. The Agency will, for a fee, extract this information and deliver you a .csv data file. However, because this is a tedious manual process, it costs \$10/probe (each 'probe' returns a 1 if at the child-moment in question, the child had already been vaccinated and 0 if not. Furthermore, because of confidentiality concerns, and because different abstractors work independently on the case series and the base series, in the file delivered to you, you cannot check any overlap in the person-moments in the base-series portion of the file with those in the case-series portion, or with probes into other cells.

The exercise should be performed under two scenarios: (1) assuming you have a limited budget that does not allow you to pay more than \$20,000 for data collection (2) assuming you are rich and have an unlimited budget for data collection (in effect, that you know/can reconstruct the exact numbers of vaccinated and unvaccinated child-years underlying the cases arising from each cell of the base.) For both, assume that the only potential confounding variables are the child's age and year of birth.[3]

Finally, you are asked to investigate the efficiency of the poor-person's 'sampling of the base' approach relative to the rich-person's greedy 'I know the vaccination status at *each* of the $67,000 \times 32 \times 365.25 = 783,096,000$ child-days' approach.

[1] See (available on course site) the expository IJE articles on this topic by Vandenbroucke.

[2] Ignore the 'EXERCISE' at the bottom right – unless of course you would like to stick several pins at random into each square and count what proportions of them land on the vaccinated and unvaccinated person moments. [That's one of the ways the ancient mathematicians estimated the value of π – they drew a circle inside a square, and counted the proportion of random shots that fell inside the circle.]

[3] Because of the adverse publicity, the vaccinated coverage was thought to have fallen sharply over the decade and may no longer be sufficient for herd immunity.

Scenario 1

Decide on the size of the base-series (a.k.a. the 'denominator-series'). The same base:case ratio will be used within every 'cell.'

Provide a power/precision statement to justify this base:case ratio.

(After setting the base:case ratio in the calling function) run the `R` Code to extract the data and save it in a `.csv` file in your directory. Although it is difficult to specify the exact moment when a case occurred, the vaccination status for each 'base-probe' is the status `(1=yes, 0=no)` at that moment. Each base 'probe' (identified as `case=0`) is connected (matched) in age and calendar time to a specific case by a '`set.no`' from 1 to 315.

Compute the crude rate ratio and the Mantel-Hanszel summary rate ratio. Comment on the difference. (btw: what would be the effect of merging matched sets from the same 'age-year' cell?)

Analyze the data by unconditional logistic regression. Do you think this (or conditional logistic regression) is the more appropriate approach? Also, if you decided to include age and year, describe how you represented them, and comment on their fitted coefficients.

Analyze the data by conditional logistic regression. How different (from that in unconditional logistic regression) is the point estimate? the CI?

Scenario 2

Analyze the data using Poisson regression, as the authors did, and via the Mantel-Haenszel RateRatio estimator.

You can use the R code provided to put the data for the Poisson regression analysis into a data frame with 72 records: 2 'exposure' levels (vaccinated/unvaccinated) × 36 cells. You can use the R code provided later in the same file to set up the data for the M-H analysis: 36 records, 1 per age-year cell.

Efficiency of the 'sampling of person-moments' (i.e., 'case:base') approach

Run step 5 of the `R` code[4] to display on a graph the point and interval estimates that different investigators, using different samples of the base, would obtain. Using your results from scenario 2, superimpose on this graph the point and interval estimate obtained from the Poisson regression, and the point estimate obtained from the Mantel-Haenszel estimator. Comment on the lessons learnt.

[4] In beta-testing in the Windows OS, this function sometimes produced an error if the `.csv` file (named `AutismCaseBaseStudy1.csv`) it wished to create for each sample already existed from a previous run. If this happens, delete the `.csv` fill and re-run the `simulateDifferentBaseCaseSamples` function.

Why can the crude Rate Ratio (RR) be 1.45 if RR=1 at all ages and in all years?

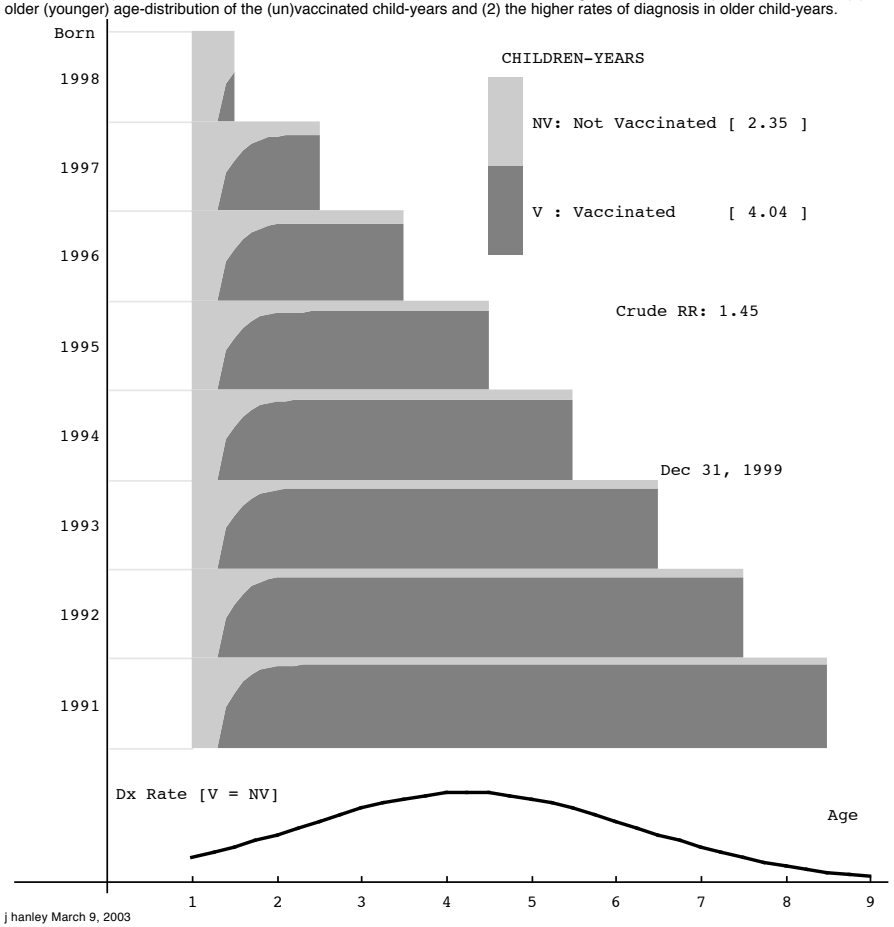
82% vaccinated. Note that 20% of vaccinated, 53% of unvaccinated children were born after 1996
(would take me too long to set up exactly the 28% and 51% that authors report !)
[2.35] & [4.40] : Ave. age of Unvaccinated & Vaccinated child-years

In diagram, all born June 30, so x & 1/2 years of f-u; in calculations, born uniformly throughout year. same no. born each year.

Think of timecourse of each of the >67,000 children in each birth cohort as a separate horizontal line; most lines switch from light to dark i.e.. the children become vaccinated. Because of limitations of printer, the 537,000 lines fuse together, they would be visible separately with a million dpi laser printer and a fine microscope, or in a printout with more than 537,000 separate horizontal lines.

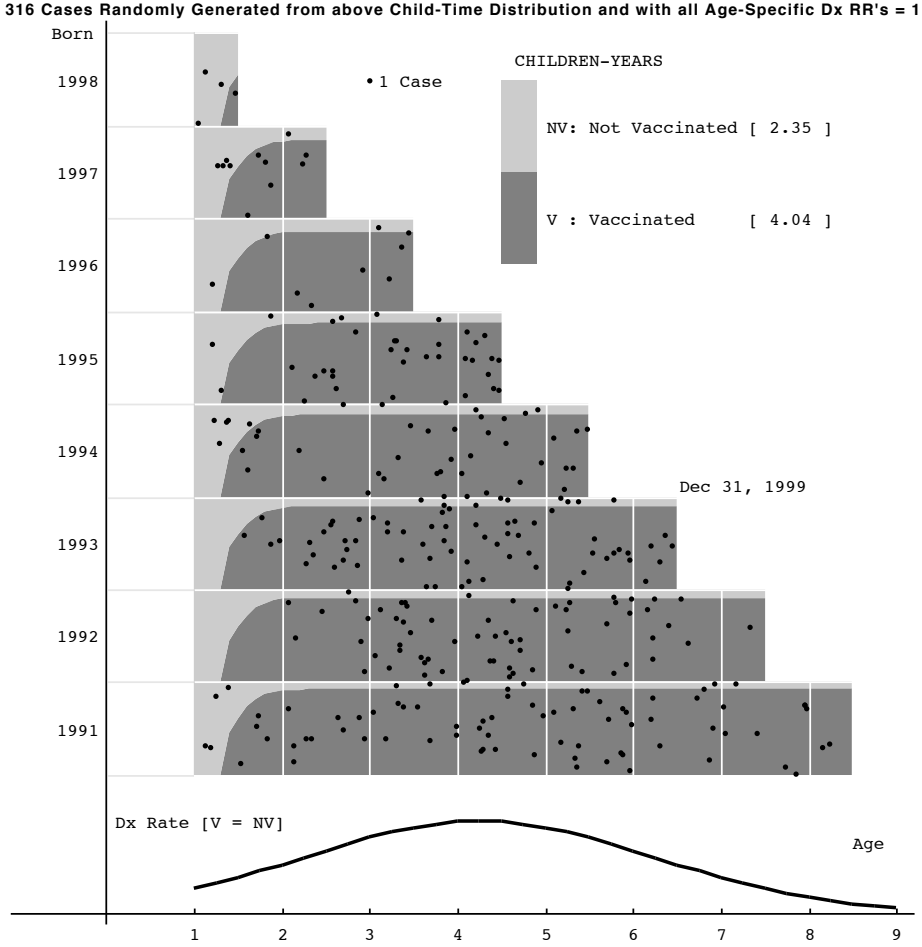
The idea of being able to see/count each of the millions of vertical/horizontal dots emphasizes that the denominators in this study are "child-moments" (and, most importantly, that the 2,129,864 child-years can AND SHOULD be subdivided not just into the 482,360 unvaccinated and 1,647,504 vaccinated child-years, but -- to allow comparison of like with like --, the number of unvaccinated and vaccinated child-years within narrower age-ranges (see 1×1 age-calendar "cells" in Lexis Diagram on next page)

The Child-Time distribution is estimated using above data and assumptions, and from clues in text about the fall-off in vaccination rates over the decade. Likewise, the rate ("incidence") of *diagnosis* of autism as a function of age (same whether V or NV) is chosen to be reasonably realistic: even if the rate curve is not exactly as shown, confounding is still produced by the confluence of (1) the older (younger) age-distribution of the (un)vaccinated child-years and (2) the higher rates of diagnosis in older child-years.



j hanley March 9, 2003

316 Cases Randomly Generated from above Child-Time Distribution and with all Age-Specific Dx RR's = 1



The locations of the 316 cases in this modification of the Lexis diagram were randomly generated by ...

1 Calculating the "rate of diagnosis by age" curve (arbitrary scale) at ages=1.25 to 8.25 in steps of 0.5 (i.e. at 15 age-points; to simplify your job of counting cases in the various age cells, the diagram shows coarser, 1 year , i.e., birthday, boundaries)

2 Multiplying these "rates" by the numbers of children "in view" at each of these that ages, to get, for each of the 15 vertical age-slices of "child-time", a number proportional to the expected number of cases in that vertical child-time slice; then scaling the 15 expected numbers summing to 316.0: expect an average of 19.0 to be diagnosed between 1 and 1.5 years of age, 23.5 b/w ages 1.5 and 2, ... 31.1, 33.2, 38.8, 35.5, 36.6, 28.4, 25.9, 16.6, 13.3, 6.71, 4.76, 1.58, ... 0.992 between ages 8 and 8.5.

3 For each age-slice, randomly generating a count from a Poisson distribution with the corresponding expected value. Repeat until the sum of the observed number of cases is in fact 316, as it was in the actual study. This gave 19 between 1 and 1.5 years of age, 19 between ages 1.5 and 2, and so on, .. 23, 27, 37, 35, 42, 31, 27, 24, 13, 7, 5, 5, ... 2 between ages 8 and 8.5.

4 For each of these cases, randomly choose a year of birth (i.e. randomly along the vertical scale, *without regard to whether the location will be in a unvaccinated or a vaccinated child-time cell.*) and a more refined age at diagnosis (randomly within the 0.25 age-band on each side of 1.25, or 1.75, or etc. ,*without regard to light/dark*). If the random location is in the darker(lighter) area, the case involves a child who was (un)vaccinated at the time of diagnosis.

EXERCISE: From the diagram, (manually) count the vaccinated and unvaccinated cases (numerators) in each vertical age-slice. Estimate (roughly) the (relative) sizes of the corresponding vaccinated and unvaccinated child-years (denominators) [hint: the proportions vaccinated by the end of the study range from 0.92 (1991 cohort) to 0.88 (1994), to 0.84 (1997), to 0.55 (1998)]. Using these numerators and denominators, calculate *an age-adjusted* RR.