Course EPIB-634: Survival Analysis & Related Topics [Winter 2007]
Assignment 2, due Feb 2

`material in www.epi.mcgill.ca/hanley/c634/ unless otherwise specified`
`( username: c634 ; password: 6 letters 2 numbers, H...J.## case-sensitive )`

• From "Questions" on pp166-167 at end of Chapter 8 of Rothman 2002

1 (crude risk ratio)

2 (pooled risk ratio)

4 OPTIONAL (SMR). If interested, see item on this in Resources for Stratified Data.

8 (contribution from a stratum in which all subjects were unexposed)

• "Homegrown"

1 In his question 8, Rothman asks you to *imagine* a stratum. Which strata in the *real* example used in Table 1 of the 1959 classic article by Mantel and Haenszel are of this type? The table is attached (full M-H article is available under 'Resources for Stratified Analyses ' in the course 634 website).

2 What are the contributions of a stratum to the summary odds ratio if (i) all of the 'controls' in the stratum are unexposed and all of the cases in the stratum are exposed (ii) the converse? Which strata in Nantel & Haenszel's Table 1 are of these types?

3 Compute a M-H summary odds ratio, and a test-based* 95%CI, from the 4 age-strata of "*housewives'* in Table 1 of the M-H article. Mantel and Haenszel have already done most of the work for you for the M-H summary OR measure, and for the test-based CI, even though the test-based CI wasn't developed until 1976. [see various computing sofware options available under Resources]

{ * as is common practice nowadays, omit the continuity correction in the X^2 statistic; this avoids the paradoxical situation where even if A equaled E[A] in *each* stratum -- so that each (stratum-specific) OR estimate (and thus the summary OR estimate would be unity -- the X^2 value would be nonzero! }

4 Mantel and Haenszel do not refer to the 1955 paper by Barnet Woolf (electronic copy also under Resources for Stratified Analyses).

(i) To see how well/poorly Woolf's method works if strata have *sparse* data (the frequencies used in the examples in Woolf's paper were quite large) compute the Woolf summary odds ratio and associated 95% CI from the same 4 'housewives' strata in q3.

(ii) If you had to stay with one CI method, which one seems the most versatile? (i.e. handles strata of all sizes, from ones like in the Woolf examples, down to strata consisting of matched pairs? Mantel and Haenszel discuss this issue on page 741, but there was no test-based or RBG CI (see q5) at that time.

5 OPTIONAL The 'Robins-Greenland-Breslow' formula (the last variance formula in Rothman's Table 8-4) for the variance of the log of the M-H summary OR estimate was developed later still -- in 1986. Their formula is not a lot of work if you program it (once) in a spreadsheet, but quite tedious by calculator, but a few people in a class a few years ago preferred to do it by calculator rather than program a few extra formulae in a spreadsheet. For the assignment this year simply compare the number of multiplications, divisions, additions and subtractions are involved to compute the test-based CI vs. the new-and-improved-RGB CI. If you have ready access to software that already has the RBG method implemented, you might want to see in a few datasets if all those extra keystrokes make any practical difference .

6 OPTIONAL Test-based CI's are most accurate near the *null* value (since the standard error is computed under the null hypothesis) and are less so for parameter values away from the null. Mantel and Hanszel's example of lung cancer and smoking provides an opportunity to see how much they differ in practice in an extreme non-null situation, and to see whether the 'new and improved' Robins-Greenland-Breslow CI is very different. Compare the two in this example. Is Stata's strong warning "We recommend that test-based confidence intervals be used only for pedagogical purposes and never be used for research" warranted in this example?

{later we will see a *conditional* approach to finely stratified data, using conditional logistic regression}

7 Refer to rows 2 and 3 of Table 3 to the Ayas et al article from last week, on Percutaneous Injuries, and to the paragraph beginning "To assess the relationships..." in the last column of page 1057 of the article.
(i) Using the data supplied to us by Dr Ayas (they are in a spreadsheet under Resources for Stratified Analysis) carry out *their* analysis (point estimates and CIs) for the "OR"s in row 3 (injuries in ICU) and row 2 (injuries reported to OH). *As per assignment 1, they should be called rate ratios rather than an odds ratios. Also, your CI's for the "OR" may not match theirs -- I asked him and he said "I used the formula in Rosner's book to calculate the CI - I wrote the program by hand in SAS."*

(ii) For row 3, apply the (stratified) Mantel-Haenszel chi-square <u>*test*</u> to the 8 strata (interns) to test whether the underlying rate ratio is 1. *Effectively, if you choose the numbers of injuries in extended periods as the "a" cells in your 8 sets of data, what you are doing is calculating how many of the 8 injuries would be expected in the "Extended Periods" if the total of 8 were distributed at random, i.e. just proportional to the numbers of opportunities only, and not influenced by whether these opportunities were in extended or non-extended periods.*
(iii) from what you have been told about chi-square tests, do you think that it is going to be "accurate" in this particular application to row 3? to row 2?

8 **Collaborative Project (**to reduce the 'counting of numerators' and 'estimating of denominators' work per person), with one report per team

This exercise is based on the article A POPULATION-BASED STUDY OF MEASLES, MUMPS, AND RUBELLA VACCINATION AND AUTISM [Full article is under Resources, abstract given below]

The *crude* rate ratio of 1.44 was never mentioned in the entire article, just the *'adjusted'* rate ratio of 0.92. The article does give enough data to calculate the crude rate ratio, but not enough of the stratified data to be able to see exactly why the adjustment makes such a big difference. However there are some bits of scattered information, including some footnotes, in the article that allow us to reconstruct the situation fairly closely. *[JH asked the authors just for the numerators and PT's , broken down by age and vaccination status, but this was a very 'hot potato' and they made the excuse to him that person-time counts and age-specific frequencies cannot leave Denmark, because these data are confidential! (JH re-iterated that he was not asking for anything that could possibly identify people, but anyway... the authors did invite JH to spend a sabbatical at their (quire sophisticated) epi research centre, and maybe he will some day!]*

(i) Explain to a *journalist* why the big difference between these two rate ratios in this example.

(ii) Complete the exercise at the bottom of the page entitles "316 Cases Randomly Generated from above Child-Time Distribution and with all Age-Specific Dx RR's = 1" To do so, use each entire vertical slice (1 year wide) as a 'stratum' and use the Mantel-Haenszel summary rate ratio (Sum over 8 vertical slices*) and calculate a (test-based or other) CI to accompany it. (see overleaf)

(Rothman2002 p 153, and already used for Q 7)

MH summary rate ratio

$$\text{MH summary rate ratio} = \frac{\text{Sum [\#Exposed Cases} \times \text{Unexposed P-T / total P-T in stratum]}}{\text{Sum [\#Unexposed Cases} \times \text{Exposed P-T / total P-T in stratum]}}$$

* *Ideally, should summarize over all (36!) age-year cells, to also keep calendar year 'constant'. Since focus <u>here</u> is on how the calculations, ignore (collapse over) calendar time.*

Spreadsheets were made for repetitive calculations like these! If you want a quickstart, the spreadsheet under "resources for stratified data is a useful template to build on. There is also Stata, or R, or SAS (you can modify the SAS or the Stata example in the Resources link)

Since the counting and classification of all 312 cases is quite tedious, divide up the work among you (among as large a group as you care to organize yourselves .. entire class if you like .. i.e. each person might take a few rectangles). Since you will also have to estimate the unexposed and the exposed PT amounts (the 2 denominators) in each vertical slice, maybe you want to have 'numerator gangs' and 'denominator gangs. Since this was Denmark, they had the denominators, and so you can make an estimate of the percentage of each colour (rough is fine here since this is for 634 and not for the NEJM)! But imagine you had to do this in a place that didn't have such records, so that you had to resort to sampling the base of children. It would be like sticking pins (probes) randomly into the base and classifying each as to whether it landed in the dark (exposed, already vaccinated) or light (not vaccinated) person-moments. (one of the early mathematicians used this strategy to make an estimate of Pi -- by sticking pins at random in a square that enclosed a circle, and estimating how many fell inside the circle]

There used to be a definition of an epidemiologist as a doctor who can count. Now, maybe its a doctor who can compute. Those who feel that such counting is too time-consuming, and would in real life have a research assistant do the counting for them, can if they prefer take this approach: In the past, we had one observer (whose reproducibility we haven't checked!) manually classify and count the cases, and calculate the distribution of person-time, for each of the 36 age-calendar-year 'cells' shown on the diagram. These raw data, along with statements that will do several analyses, can be found in the file "counts / person times measured from diagram, together with SAS program" under "Resources for Stratified Analysis" in the 634 course page.. The file also contains the sas code to run several analysis (they are labeled using title statements), Some are so as to interpret and contrast the outputs from the different regression analyses. We will come to the regression analyses later. For now, you might just want to run the Mantel-Haenszel analysis in SAS or Stata, or (manually or by software) add up the numbers of child-years and the numbers of events in the same age-slice to collapse the finer calendar time-age strata into 'age-only' strata.

Ultimately, whether your team does it all manually from the diagram, or somehow from the computer file, you need the numbers of exposed and unexposed cases, and the corresponding denominators, for each of the 8 age-slices, i.e. 32 numbers in all. From there, the analysis has the same structure as we have had with the PIs among the 8 mds.

9 Which spelling is correct?

Mantel Mantell Maentel Mantal Mental ???

Haensel Haenzsel Hansell Hansel Henzsell Haenszel ???

A POPULATION-BASED STUDY OF MEASLES, MUMPS, AND RUBELLA VACCINATION AND AUTISM

KREESTEN MELDGAARD et al.

ABSTRACT

Background

It has been suggested that vaccination against measles, mumps, and rubella (MMR) is a cause of autism.

Methods

We conducted a retrospective cohort study of all children born in Denmark from January 1991 through December 1998. The cohort was selected on the basis of data from the Danish Civil Registration System, which assigns a unique identification number to every live-born infant and new resident in Denmark. MMR-vaccination status was obtained from the Danish National Board of Health. Information on the children's autism status was obtained from the Danish Psychiatric Central Register, which contains information on all diagnoses received by patients in psychiatric hospitals and outpatient clinics in Denmark. We obtained information on potential confounders from the Danish Medical Birth Registry, the National Hospital Registry, and Statistics Denmark.

Results

Of the 537,303 children in the cohort (representing 2,129,864 person-years), 440,655 (82.0 percent) had received the MMR vaccine. We identified 316 children with a diagnosis of autistic disorder and 422 with a diagnosis of other autistic-spectrum disorders. After adjustment for potential confounders, the relative risk of autistic disorder in the group of vaccinated children, as compared with the unvaccinated group, was 0.92 (95 percent confidence interval, 0.68 to 1.24), and the relative risk of another autistic-spectrum disorder was 0.83 (95 percent confidence interval, 0.65 to 1.07). There was no association between the age at the time of vaccination, the time since vaccination, or the date of vaccination and the development of autistic disorder.

Conclusions

This study provides strong evidence against the hypothesis that MMR vaccination causes autism.

(N Engl J Med 2002;347:1477-82.)

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 !)

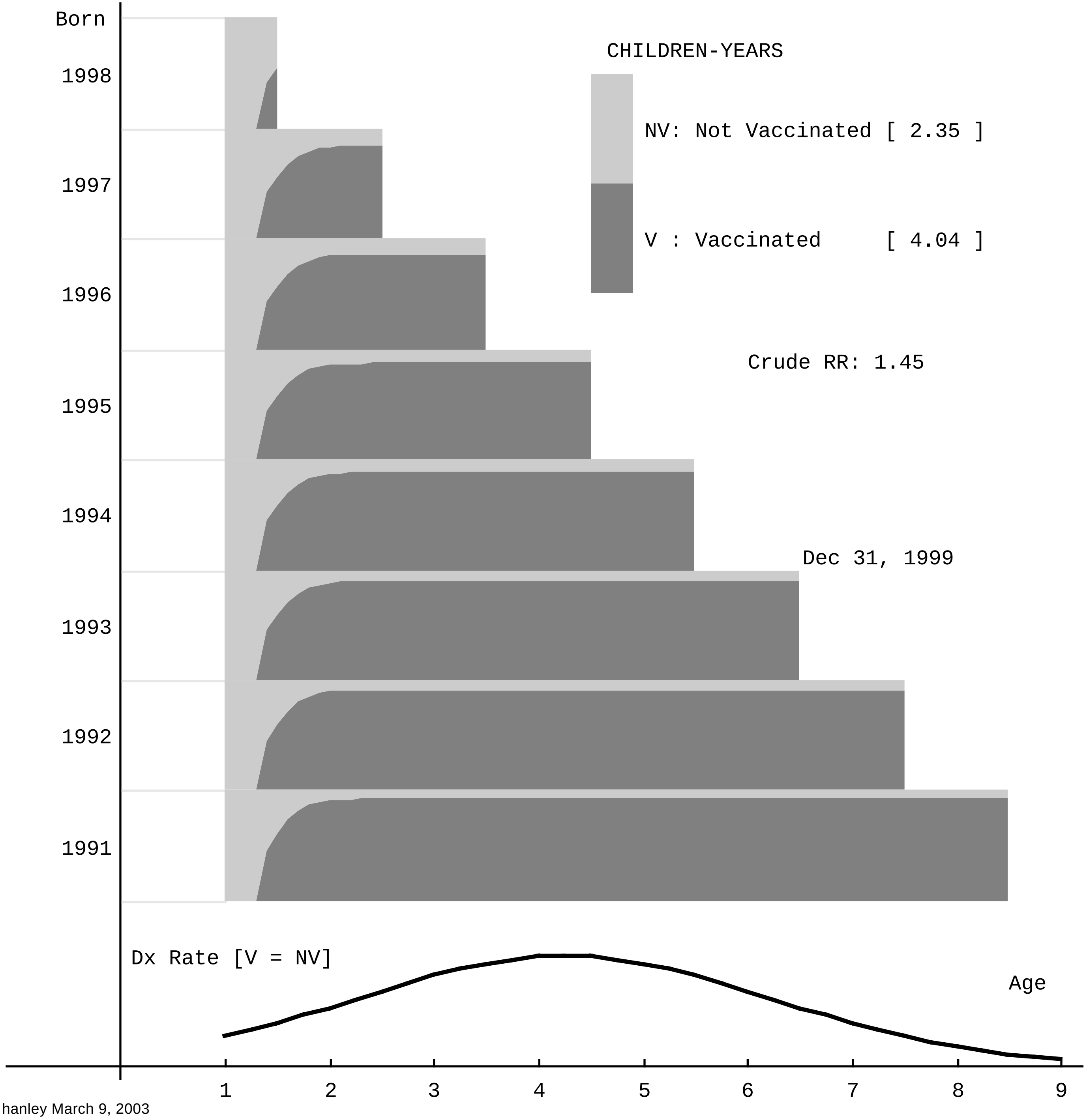
[2.35] & [4.40] : Ave. age of Unvaccinated & Vaccinated child-years

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.

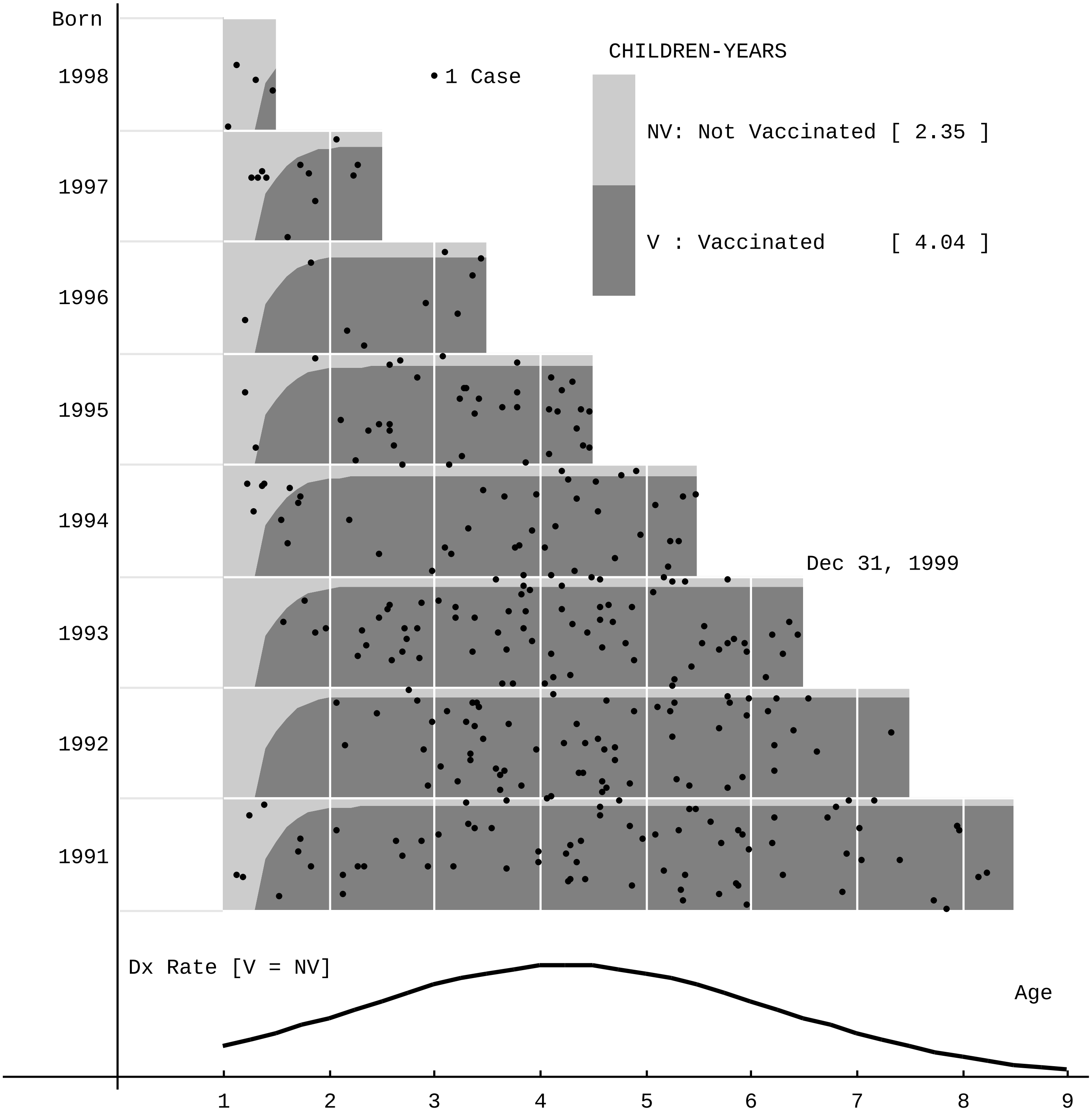
The idea of being able to see/count each of the millions of vertical/horizontal dots emphasizes that the denominators in this study are "child-moments" (and, most importantly, that the 2,129,864 child-years can AND SHOULD be subdivided not just into the 482,360 unvaccinated and 1,647,504 vaccinated child-years, 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.