

7장. 다변량 분석 준비: 연구대상자 정하기 (Setting Up a Multivariable Analysis: Subject)

2005. 10. 19.
의료관리학교실, 김소영

* 주요 내용

1. 다변량 분석에서 표본의 크기를 결정하는 방법 : ① 예측되는 결과의 크기에 따라, ② 이변량분석에서 필요한 표본의 크기를 기준으로, ③ 다중로지스틱회귀분석과 proportional hazard는 발생 건수가 최소 10건이상, 다중선형회귀분석은 발생 건수가 최소 20건 이상이 되도록
2. 표본의 크기에 비해 독립변수가 너무 많은 경우, 독립변수를 줄이는 방법 : ① 변수생략(by theoretical, measurement, empirical) ② 변수통합(by 불리언기호를 사용, 척도를 사용, 요인분석)
3. 연구도중 연구대상자가 탈락되는 경우에 처리하는 방법 : ① 삭제 ② censoring

1. 다변량 분석을 하기 위해서는 얼마나 많은 연구대상(subjects)이 필요하냐?

이변량 분석과 마찬가지로 다변량 분석에서도 표본의 크기(sample size)는 자료 분석을 통해서 증명하고자 하는 효과의 크기 및 자료의 타당성에 따라서 결정한다. 위험요소(risk factor)가 결과(outcome)와 강한 상관관계(OR=4.0)를 보이는 경우보다 위험요소가 결과와 약한 상관관계(OR=1.5)를 보이는 경우에 더 많은 표본이 필요하다. 이는 표본의 크기가 작을수록 신뢰구간이 커지기 때문이다. 또한 OR이 1.0에 가까워질수록 효과를 증명하기 위해서는 신뢰구간이 커져야 한다. 마찬가지로 표본의 크기가 클수록, 귀무가설을 증명할 수는 없을 지라도 1종 오류(상관관계가 있으나 놓칠 경우)는 줄일 수 있다. 또한 변이가 큰 변수에 대하여 집단간의 차이를 증명할 때에도 표본은 커야 좋다.

결과를 어떻게 예측하는가에 따라서 필요한 표본의 크기를 결정할 수 있다. 한편 다변량 분석의 결과는 예측하기 복잡하여 일반적으로 통계학자와의 상담이 필요하다. 이 장에서는 표본의 크기를 결정하는 몇 가지 유용한 방법을 소개하고자 한다.

첫째, 명망있는 책¹⁾에 실린 공식에 따라서 이변량 분석에 필요한 표본의 크기를 결정해라. 만약 종속변수가 결과율이라면 누적발생률(예, 3년의 결과발생률)을 구한다. 표본의 크기가 이변량 분석에서 효과를 증명

정의

결과에 대한 예측정도에 따라서 특정한 효과를 밝히기 위해 필요한 표본의 크기를 결정할 수 있다.

Tip

표본의 크기가 이변량 분석에서 효과를 증명하기에 부족할 경우, 다변량 분석에서도 효과를 증명할 수 없다.

1) 초보가를 위한 가이드로는 Hulley SB, Cummings SR의 Designing clinical research. Baltimore: Williams and Wilkins, 1988, pp. 139-50가 좋고 좀더 포괄적으로는 Friedman LM, Furber CE, Demets DL이 쓴 Fundamentals of clinical trials(2nd ed.). Littleton, MA: PSG Publishing Co., 1985, pp. 83-107이 있다.

하기에 부족할 경우, 다변량 분석에서도 효과를 증명할 수 없을 것이다.

둘째, 다중로짓회귀분석(multiple logistic regression)과 proportional hazards analysis의 경우, 모든 독립변수마다 최소한 10건이 발생해야 한다. 표본의 크기보다는 발생건수를 측정하는 이유는 우리가 궁극적으로 구하는 것은 특정한 결과(또는 효과)이기 때문이다. 의학분야의 연구에서 대부분의 발생률은 절반이하이다. 만약 한 연구에 적용대상이 1,000명이 넘을지라도 5명에서만 특정한 결과가 발생하였다면 효과를 증명하기가 어렵다. 또한 위의 결과는 오직 5명에 근거한 결론이기 때문에 발생한 사람이 많은 집단에서는 효과가 없다고 말할 수 없기 때문이다. 표본의 크기는 두 개의 비교 집단 중에서 표본의 크기가 더 작은 집단을 기준으로 정한다.

Tip
다중로짓회귀분석과 proportional hazards analysis의 경우, 모든 독립변수마다 최소한 10건이 발생해야 한다.

변수 당 10개의 발생건수는 단지 참고치일 뿐이다. 따라서 한 연구의 표본크기가 이 기준에 맞지 않아도 의미있는 연구결과를 낼 수도 있다(가령, New England Journal of Medicine에 실릴 수 있다). 반대로 각 독립변수 당 발생 건이 10건 이상일지라도 연구주제에 대한 답을 구할 수 없을 수도 있다. 그러나 일반적으로 특정한 독립변수가 연구결과와 상관관계가 있다는 것을 증명하기 위해서는, 독립변수 중에서 발생률이 더 적은 경우에 따라서 표본의 크기를 정해야 한다.

이제 앞서 제시한 원칙을 적용한 예를 살펴보고자 한다. Schwarcz 등은 HIV 감염자에서 폐렴구균성 폐렴(PCP, pneumocystic pneumonia)에 대한 폐렴구균 예방접종의 효과를 평가하기 위하여 로짓회귀분석을 하였다. 총 표본의 크기는 326명의 환자로 이들은 모두 폐렴구균성 폐렴(PCP)에 걸렸다. 326명 중에서 114명(35%)은 PCP로 진단받기 전에 폐렴구균 예방접종을 받았고 나머지 212명(65%)는 받지 않았다. 이 연구에서 6개의 독립변수를 정했다. 앞서의 원칙에 따르면 대상자가 최소한 60명(6개의 독립변수×최소 10건)이 필요한데, 발생한 대상수가 더 적은 쪽이 114명이므로 이는 충분한 수로 보인다. 표 7.1을 살펴보면, 로짓회귀분석에서 두 가지 변수(백인이 아닌 사람, 비보험)는 폐렴구균 예방접종을 받은 군에서 PCP 발생률이 낮고 유의한 상관관계를 보인다(p-value<0.05).

TABLE 7.1
Frequency of primary PCP prophylaxis among patients whose AIDS-defining diagnosis was PCP.

Characteristic	Did not receive PCP prophylaxis n (%)	Received primary PCP prophylaxis n (%)	Adjusted P value	Adjusted odds ratio	95% confidence limits
Total	212 (65.0)	114 (35.0)			
Age group					
<35 years	61 (69.3)	27 (30.7)		1.0 (ref.)	
≥35 years	151 (63.5)	87 (36.6)	0.55	1.19	0.68, 2.07
Ethnicity					
Nonwhite	74 (77.9)	21 (22.1)		0.49	0.28, 0.87
White	138 (59.7)	93 (40.3)	0.01	1.0 (ref.)	
Sex					
Male	206 (64.6)	113 (35.4)		0.81	0.06, 10.12
Female	6 (85.7)	1 (14.3)	0.87	1.0 (ref.)	
Sexual orientation					
Gay/bisexual man	165 (62.7)	110 (37.3)		3.19	0.78, 13.03
Heterosexual	27 (87.1)	4 (12.9)	0.11	1.0 (ref.)	
Injection drug use					
Yes	35 (72.9)	13 (27.1)		1.11	0.49, 2.54
No	177 (63.7)	101 (36.3)	0.80	1.0 (ref.)	
Insurance					
None	52 (82.5)	11 (17.5)		0.35	0.17, 0.73
Public/private	151 (59.5)	103 (40.6)	0.005	1.0 (ref.)	

Adapted with permission from Schwarcz, S.K., et al. "Prevention of *Pneumocystis carinii* pneumonia: Who are we missing?" AIDS 1997; 11:1263-8. Copyright Rapid Science Publishers Ltd.

그러나 성별(sex)의 경우 여성의 표본은 7명뿐(총 표본의 2%)으로 OR 값이 0.81임에도 불구하고 P-value>0.87, 신뢰구간이 0.06~10.12로 유의하다고 볼 수 없다. 성적지향(sexual orientation)에서도 비슷한 결과를 보인다. 즉 이성애자(heterosexual)의 표본은 30명뿐(총 표본의 10%)으로 신뢰구간이 0.78~13.03으로 매우 넓다.

일반적으로 표준오차와 신뢰구간이 클수록 표본의 크기가 적절하지 못하다는 것을 의미한다. 또 다른 표현으로 logistic regression과 proportional hazards analysis에서 표본의 크기가 충분하다는 것은 연구결과가 하나로 수렴될 수 있을 만큼의 표본이 있다는 것을 의미한다.

Tip
다중선형회귀분석에서도 표본의 수가 독립변수 당 최소한 20개 이상의 표본(발생건수가 아니라)을 갖아야 할 것이다.

다중선형회귀분석에서 표본의 크기는 로짓회귀분석이나 proportional hazards regression에서처럼 중요한 문제는 아니다. 이는 결과변수가 간격변수이기 때문에 모든 연구대상자가 결과변수의 발생건수가 되기 때문이다. 그러나 다중선형회귀분석에서도 표본의 수가 충분하지 못하면 표준오차가 커지므로 독립변수 당 최소한 20개 이상의 표본(발생건수가 아니라)을 갖아야 할 것이다. 물론 표본의 수가 적다고 모든 분석이 가치가 없는 것은 아니므로 계수(coefficients)를 해석할 때는 주의를 기울여야 한다.

2. 표본의 크기에 비해 독립변수가 너무 많으면 어떻게 할까?

만일 표본의 크기에 비해 독립변수가 너무 많다면, 표본의 수를 늘리거나 독립변수의 수를 줄여야 한다. 표본의 수를 늘리는 것이 보다 바람직하지만, 대부분의 경우에 그러기는 불가능하다. 따라서 대부분의 연구자들은 독립변수를 줄이는 쪽을 택한다. 표 7.2는 독립변수를 줄이는 방법들이다.

TABLE 7.2 Methods for decreasing the number of independent variables.	
1. Omitting variables	
Theoretical considerations	
Measurement considerations	
Two variables highly correlated. Omit the one with:	
Larger amount of missing data	
Greater measurement error	
Theoretically less important	
Empirical findings	
Variable unrelated to outcome in bivariate analysis	
Variable unrelated to outcome in multivariable analysis	
Variable unrelated to main independent variable and outcome in bivariate analysis	
Variable has minimal impact on main effect in multivariable analysis	
2. Combine variables into a single variable or scale	
"And/or" constructions	
Summation scales	
Factor analysis	

가. 변수를 생략하는 방법

(1) 이론적으로 고찰(theoretical considerations)하는 방법

연구결과를 분석하는 중에 연구대상 질병에 대하여 병리생리학적으로 중요하지 않은 독립변수가 있다면 생략할 수 있다. 예를 들어, 관상동맥질환의 원인에 대한 연구에서 운전 중 안전벨트 착용유무를 독립변수로 잡을 수 있다. 안전벨트 착용자가 건강문제에 더 관심이 많고 좀더 적극적으로 건강행태(저지방 식이, 운동 등)를 실천할 경우, 안전벨트의 착용은 관상동맥질환의 likelihood를 줄이는 데에 유의한 상관관계를 보일 수도 있다. 그러나 결단코 안전벨트의 착용은 관상동맥질환의 원인이 되지도 그것을 예방할 수 있게 할 수도 없다. 따라서 독립변수에 포함할 하등의 이유가 없다.

Tip
어떤 변수를 생략할지를 결정하기 위하여 이론(theory), 측정값(measurement), 경험적인 결과를 고려해야 한다.

(2) 측정치를 고찰(measurement considerations)하는 방법

만일 두 개의 변수가 또 다른 변수와 중간정도의 상관관계를 보일 경우, 연구자는 두 개의 변수 중 하나를 생략하고 싶을지도 모른다. 왜냐하면 두 개의 변수가 다른 변수와 비슷한 상관성을 보일 경우, 하나의 변수에 대한 해석만으로도 다른 변수를 설명할 수 있을 것이기 때문이다. 그럼에도 불구하고 두 개의 변수가 완벽하게 상관성을 보이지 않을 경우, 하나만 분석함으로써 중요한 정보를 놓칠 수 있다. 즉, 정보 손실율은 상관성의 정도와 관련된다. 하지만 독립변수를 모두 포함하기에는 표본의 크기가 충분하지 못할 경우, 생략할 독립변수를 결정해야만 한다. 생략하지 않고 남겨놓을 변수는 multicollinearity에서와 마찬가지로, 이론적으로 더 중요하고 자료손실(missing data)이 적고, 측정오차(measurement error)가 적은 것으로 결정한다.

(3) 경험적으로 고찰(empirical considerations)하는 방법

연구자는 경험적인 결과에 비추어 생략할 변수를 결정할 수도 있다. 가장 흔히 이용하는 경험적인 방법은 이변량 분석의 결과에서 통계적으로 유의한 상관관계를 갖는 변수만을 선택하는 것이다. 표본이 매우 크지 않는 한, 통계적인 유의성을 검증할 때는 0.05 p-value값을 cut off값으로 하여 p-value가 0.05미만인 변수만을 선택하는 것이 좋다. 일반적으로 연구자들은 이변량 분석에서 p-value < 0.15이거나 p-value < 0.20인 변수만을 다변량 분석에 포함시킨다. 단 억압효과(suppressor effect)를 볼 경우에는, 변수가 이변량 분석의 결과와 상관관계가 없을 수 있음을 유념해야 한다.(1.3장-11p-을 다시 떠올려 보면 zidovudine을 쓸 경우에, 다변량 분석에 따르면 치료에 효과가 있는 것으로 나타났으나 이변량 분석에 따르면 유의하지 않았다.)

변수선택절차(variable selection procedures) 또한 흔하게 이용하는 경험적 방법이다. 이 절차는 8.9장(11p)에서 설명할 것이다.

연구자의 주 관심사가 하나의 주요 변수와 그 결과 사이의 관계인 경우, 관심 있는 독립변수인 동시에 이변량 분석의 결과변수인 것만을 변수에 포함시키고 싶을 수 있다. 그러나 이 두 조건을 만족시키는 변수는 혼란변수이거나 억압변수일 수도 있음을 기억해야 한다. 따라서 이런 경우에 쓸 수 있

는 변형된 방법은 주요한 효과의 크기를 변화시키는 독립변수만을 포함시키는 것이다. 예를 들어, Chamber 등은 fluoxetine(Prozac)이 태아에 미치는 영향(birth outcome)을 연구하기 위하여 임신 말기에 fluoxetine을 복용한 산모의 출생아 70명과 임신 초기에 fluoxetine을 복용한 산모의 출생아 98명 중에서 조산율을 비교하였다. 그 결과 조산아는 14명이었다. 이 연구에서 변수를 제한하기 위하여 주요한 효과 - 임신 중 fluoxetine의 초기 복용과 말기 복용에 따른 효과 - 의 측정값에 10% 이상의 변화를 주는 독립변수만을 선정했다. 11개의 변수만이 이 기준에 부합하여 이것을 연구에 포함시켰다. 뿐만 아니라 두 개의 변수(모성연령, fluoxetine 복용량)는 이론적인 이유로 변수에 포함시켰다.

총 13개의 독립변수와 14명의 조산아 발생 건수로 변수 당 10건 이상이 발생해야 한다는 가이드라인을 만족시킬 수가 없었다. fluoxetine에 늦게 노출된 경우 조산의 위험이 증가하는 것으로 나타났다(OR=4.8). 독립변수는 많고 상대적으로 발생 건수는 적었으므로 95% 신뢰구간은 넓었다(1.1~20.8). 신뢰구간에서 1이 포함되지 않으므로 임신 말기 fluoxetine 복용은 조산아의 발생률을 증가시키는 것과 관련이 있음을 시사한다. 그러나 손익을 따졌을 때, 신뢰구간 중에서 1.1의 OR값과 20.8의 OR은 큰 차이를 보인다.

나. 변수를 통합하는 방법

(1) “그리고/또는(and/or)” 구조를 사용하는 방법

두 개 이상의 연관성 있는 변수는 ‘그리고/또는’ 구절을 사용하여 합할 수 있다. 예를 들어, 임신 중에 fluoxetine을 복용하는 것이 태아에 미치는 영향(birth outcome)을 연구하는 경우에, 연구자는 고혈압, 전자간증, 자간증을 하나의 변수로 묶었다. 즉, 고혈압, 전자간증, 자간증 중 하나 이상의 질환을 앓고 있는 경우에는 변수가 모두 ‘yes’로 코딩되고 셋 중 어느 질환도 있지 않은 경우는 변수가 ‘no’로 코딩된다.

‘그리고/또는’ 구절은 변수가 multicollinear(예: 자간증을 가진 환자는 모두 전자간증이 있는 경우)이거나 표본집단의 대다수가 동일한 독립변수값을 갖을 때(예: 자간증은 상대적으로 드문 질환이므로 대다수는 변수가 ‘no’로 코딩되는 경우)도 도움이 된다. 독립변수는 어떠한 독립변수의 값이 표본집단의 대다수에서 동일한 경우는 표준오차와 신뢰구간이 커지는 경향이 있다.

(2) 통합 척도(summation scales)를 사용하는 방법

단순히 기능 상태나 태도 등을 평가하는 것만으로 복합적인 상태를 설명하기 어려울 경우 연구자는 척도(scales)를 사용하기도 한다. 연구결과를 분석하는 단계에서야 척도의 필요성을 느낄 때도 있지만 대부분의 경우에 척도의 사용여부는 연구 설계단계에서 결정한다.

척도를 통합하기 위해서는 우선, 모든 연구문제(all questions)는 같은 방향(same direction)으로 코딩해야 한다. 예를 들면, 점수가 높을수록 연구결과는 더 좋은 것을 의미하도록 설계하는 것이 좋다. 다른 예로, 모든 변수는 동일 범주의 순위척도를 쓰는 게 좋다. 왜냐하면 한 변수는 0~10 척도를 쓰고 다른 변수는 0~5 척도를 쓰는 경우 전자는 2배의 가중치가 매겨지기 때문이다. 변수를 순위척도로 나눈 경우, 일단 변수들을 통합했다면 하나의 순위척도에 포함되는 변수들의 평균값이 한 변수의 값을 대신한다.

척도를 만들 때 자료가 손실되지 않도록 주의를 기울여야 한다. 만약 하나의 척도를 구성하는 자

료의 대부분이 결측값(missing value)인 경우, 그 척도의 값은 결측 처리해야 한다. 그러나 반 이상이 타당한 경우, 결측값은 척도를 구성하는 표본들의 평균값으로 대체한다. 평균값으로 대체한 후, 척도들로 나타내고 각 척도의 값은 그 척도에 해당하는 변수들의 수로 나눈다.

척도를 구성하기 위해서 결측값을 처리하는 또 다른 방법은 각 대상(subject)을 결측값이 아닌 변수의 수로 나누는 것이다. 예를 들어, 연구 대상(subject)이 하나의 척도를 구성하는 5개의 변수 중 4개에서만 타당한 값(valid response)을 보였을 경우, 4개의 값을 합하여 4로 나눈다. 마찬가지로 5개 모두 타당한 값을 보일 경우, 5개를 합하여 5로 나눈다. 이와 같은 방법을 사용하기 위해서는 척도를 구성하는 변수 중 최소한 반 이상의 자료는 무결해야 한다.

이와 같이 통합된 척도를 사용하는 경우는 변수들의 상관관계가 높을 때만 유효하다. 변수들 사이의 상관성을 측정하는 흔한 방법은 alpha이다. alpha는 신뢰도계수(reliability coefficient)라고도 하며 일반적으로 alpha가 0.65보다 커야 변수들이 신뢰할 만한 척도를 만든다.

Tip
일반적으로 alpha가 0.65보다 커야 변수들이 신뢰할 만한 척도를 만든다.

(3) 요인분석법

요인분석법은 행동과학에서 다중으로 연관된 독립변수(multiple related independent variables)의 수를 줄이고자 할 때 잘 사용하는 방법이다. 요인분석을 통해 다중으로 연관된 독립변수를 몇 개의 잠재요인(underlying factor)으로 요약할 수 있다. 이 절차를 통해 요인들 사이의 상관관계를 줄인다. 각 요인은 변수에 가중치를 써서 통합(weighted combination)한 값이다. 변수의 개수를 요인의 수만큼 줄이기 위하여 각 변수의 대상값(subject's value)에 근거하여 요인별 점수(factor score)를 개발한다. 각 독립변수들은 요인들과 상관관계를 갖지만, 그 정도는 제각각이다. 상관관계의 정도에 따라 요인을 구분한다.

임상연구자에게 있어서 요인분석의 주요한 문제점은 원래의 변수를 잃는다는 것이다. 예를 들어, 한 연구자가 요인 분석을 사용하여 폐렴회복에 관련된 변수에 대한 요인별 점수(factor score)를 개발했다고 하자. 첫 번째 요인은 환자 comorbidity(연령, 기저폐질환, 울혈성 심부전의 과거력)를 나타내고, 두 번째 요인은 virulence of attack(방사선 소견의 심각성 정도, 높은 열, 특정 조직의 회복)을, 세 번째 요인은 치료효과(첫 번째 투약속도, 적절한 항생제 선택)를 나타낸다고 가정해보자. 이것은 자료를 폐렴의 병리생리학적 관점에서는 자료를 만족스럽게 통합한 예이다.

proportional hazard 모델을 이용하여 환자의 comorbidity, virulence of attack 요인이 많으면 생존율이 나빠지고 효과적으로 치료를 하면 생존율이 좋아지는 것의 상관관계를 밝힌다고 가정해보자. 그러나 이러한 요인의 요인별 점수(factor score)를 매기는 것은 수학적으로 매우 복잡하기 때문에 유용한 정보가 될 수 없다. 그뿐만 아니라 요인이 결과에 대하여 얼마나 중요한지는 알 수 없고 다만 요인이 결과에 미치는 중요성만을 알 수 있을 뿐이다. 또한 어떤 변수는 요인과 깊은 상관관계를 갖으면서도 연구결과(outcome)와는 상관관계가 적을 수도 있다. 왜냐하면 요인분석은 변수들과의 관계에 근거하여 독립변수들을 한 집단으로 묶기 때문에 결과변수와는 아무런 관계를 갖지 않을 수도 있기 때문이다. 이러한 이유로 요인분석은 임상적인 자료를 분석할 때는 상대적으로 덜 사용한다. 행동과학처럼 측정하기 복잡한 문제(자기 존중감, 자율성 등)를 다룰 때는 한 개의 변수를 잃어도 요인분석으로 훨씬 좋은 결과를 낼 수 있는 경우에 사용한다.

3. 연구 도중에 연구 대상(subject) 중 일부분이 탈락된 경우에 어떻게 할 것인가?

종적인 임상연구에서 연구가 완수되지 못하는 경우는 흔하다. 연구 대상자가 더 이상 참여할 의사가 없는 경우, 이사를 가버린 경우, 부작용이 생겨서 연구를 중단하는 경우, 연구자가 알 수 없는 이유로 추적관찰자 수가 줄어든 경우 등이 그 예(표 7.3)가 될 수 있다. 임상 연구는 다른 연구와는 다르게 연구 중단에 탈락한 대상자를 처리할 일이 생긴다. 탈락자를 처리하는 방법 중 하나는 그들을 연구에서 삭제하는 것이다. 그러나 이 경우에 처음 표본의 상당 부분이 손실될 수 있다. 만약 연구기간이 5년인 연구가 3-4년 진행되었다면, 연구 대상자는 중요한 정보가 되므로 이를 탈락할 경우 이는 연구의 오류(bias)로 작용할 수 있다. 이런 문제점을 만들지 않기 위해서는 연구가 종료될 때까지 연구대상자를 추적할 수 있어야 한다. 그런 추적기술을 일명 censoring이라고 부른다. censoring은 생존분석에서 주요한 요소이다. censoring을 통하여 탈락한 대상자를 다시 추적 관찰하기도 하고 대상자마다 추적기간을 달리하여 분석을 하기도 한다. 실제로 proportional hazards analysis에서 연구목적에 맞지 않는 결과를 보이는 관찰 대상자에 대해서도 연구종료 시점에서 censoring을 실시한다.

Tip
임상 연구는 다른 연구와는 다르게 연구 중단에 탈락한 대상자를 처리할 일이 생긴다.

Tip
Censoring을 통해 연구 종료시점까지 연구 대상자를 추적한다.

TABLE 7.3
Reasons for censoring observations.

Reason for censoring	Examples
1. Lost to follow-up.	Subject moves, doesn't wish to participate, stops attending a particular clinic.
2. Subject has an outcome that precludes the outcome you are studying: (also known as alternative outcomes or competing risks.)	Death from coronary artery disease in a study of cancer incidence.
3. Subject is withdrawn from study.	Development of side effects, not ethical to continue treatment or placebo.
4. Varying dates of enrollment.	Study enrolls subjects with different starting dates over a 2-year period.
5. End of study.	All subjects who have not experienced outcome are considered censored at the end of the study.