예방의학회지 제33권 제1호(2000년 3월)
Korean J Prev Med 2000;33(1):31-35

Capture-recapture 방법을 이용한 광주광역시 지역암등록 자료의 완전성 추정

임정수, 권순석, 김상용, 박경수, 손석준, 최진수

전남대학교 의과대학 예방의학교실

Completeness Estimation of the Population-based Cancer Registration with Capture-Recapture Methods

Jeong-Soo Im, Sun-Seog Kweon, Sang-Yong Kim, Kyeong-Soo Park, Seok-Joon Sohn, Jin-Su Choi

Department of Preventive Medicine, Chonnam University Medical School

Objectives : This study aimed to estimate the completeness of cancer registration with Capture-recapture method.

Methods : The study was conducted in the population based cancer registry of Kwangju, Korea, for which there are three main sources of *notification: reports by Korean Central Cancer Registry*, reports by pathology data, and the others reports by radiology data, death certificates, etc. The defined cases in three sources were matched by 13 digits Resident Register Number. To derive an estimates, log-linear models were applicated.

Results : Overall completeness was estimated to be around 93%. There was some variation with age(consistently high levels below age group 60-74 years, a minimum of 88.6% above 75 years). Among the most common cancer sites, estimates of completeness were highest for thyroid cancer(97.1%), while lower estimates of completeness were derived for stomach cancer(92.3%), liver cancer(92.6%).

Conclusions : Careful application of Capture-recapture method may provide an alternative to traditional approaches for estimating the completeness of cancer registration in Kwangju city.

Korean J Prev Med 2000;33(1):31-35

Key Words: Capture-recapture method, Completeness, Cancer Registration

서 론

지역암등록은 한정된 지역 내에서 발생하는 모든 새로운 암을 등록하고자 한다. 그래서 가능한 모든 자료원(source)을 통하여 완전하게 찾는 것이 꼭 필요하다. 등록의 완전성을 평가하기 위해서 국제암연구기구(IARC)에서 자주 사용하는 방법은 DCO index(Death Certificate Only index)를 측정하는 것과 사망률/발생률 비(Mortality/Incidence ratio)를 측정하는 방법 등이 있다(Skeet, 1991; Esteban 등, 1995). DCO index는 사망진단서를 통해서만 암 발생을 확인한 예(case)가 차지하는 비율인데, 치명률에 의존하여 완전성을 측정하는 간접적인 방법이다. 이 방법은 암종의 치명률 차이에 따라 값이 변할 뿐만 아니라, 등록 초창기에는 값의 신뢰도가 떨어진다는 한계가 있다. 이러한 DCO index의 한계는 사망률/발생률 비에도 적용된다.

따라서 치명률에 의존하지 않고 등록의 완전성을 직접 추정하는 방법이 필요하다. 직접적인 추정은 독립적인 자료원이 있는 상황에서 가능한데, 이때 완전성은 독립적인 자료원에 의해 확인된 예에서 등록된 예의 비율에 의해 추정될 수 있다. Capture-recapture 방법은 질환 등록의 완전성을 직접 추정하는 도구로 최근 역학분야에서 널리 활용되고 있으며(Neugebauer, 1984; Robles 등, 1988; Cochi 등, 1989; Hook 등, 1992; Laporte 등, 1992; McCarty 등, 1993; Brenner 등, 1994; Ernest 등, 1995), 국내에서도 소아천식 유병률을 파악하고 의료보험 상병자료의 완전성을 평가한 논문(하미나 등, 1997) 등이 있다. 이 방법은 일반적으로 서로 다른 자료원에서 겹치는 부분에 대한 정보를 이용하여 각 자료가 불완전하게 파악하고 있는 환자의 수를 보정된 수치로 파악하고자 시도되는 방법이다.

약 130만명의 광주광역시 시민을 대상으로 하는 광주광역시 지역암등록은 1996년도에 시작되었다. 그런데 1996년도에 등록된 암환자의 사망신고자료는 이들이 대부분 사망하는 5년 이후에나 얻을 수 있기 때문에 암등록의 완전성을 추정하는 지표인 DCO index나 사망률/발생률 비는 신뢰성이 부족한 실정이다(광주광역시 지역암등록사업단, 1999). 따라

접수 : 1999년 9월 13일, 채택 : 1999년 12월 1일
교신저자 : 임정수 (전남대학교 의과대학 예방의학교실, 전화번호 : 062-220-4170, 팩스번호 : 062-233-0305)

서 저자들은 DCO index나 사망률/발생률 비를 대신하여 capture-recapture 방법을 이용하여 광주광역시 지역암등록의 완전성을 추정하고자 하였다.

연구방법

1. 자료원 및 짝짓기 과정

광주광역시 지역암등록 자료는 수련병원이 중앙암등록본부에 자발적으로 보고하는 입원 환자의 암발생 자료와 지역암등록사업단이 광주광역시 의료기관을 방문하여 발생 암환자에 대한 자료를 확인하여 등록시키는 방문조사 자료로 구성되어 있다. 그리고 이러한 노력에도 불구하고 누락된 경우에 대해서는 사망신고 자료로부터 사망원인이 암인 사람을 골라내어 암 확인 후 등록시키고 있다. 방문조사 자료는 해당 병원 외래 및 입원 환자의 보험청구자료, 병리 판독 자료, CT, MRI 등의 방사선 판독 자료, 치료방사선과 자료 등에서 암환자를 색출한 후 이들의 의무기록을 열람하여 암여부를 확인, 등록한 자료이다(광주광역시 지역 암등록사업단, 1999).

본 연구에서는 capture-recapture 방법을 수행하여 광주광역시 지역암등록의 완전성을 추정하기 위하여 다음과 같은 세 개의 서로 다른 자료원을 이용하였으며, 1997년도에 광주광역시에서 발생한 모든 암환자(ICD-O C00-C97)를 대상으로 하였다.

첫 번째 자료원은 1999년 5월 31일까지 중앙암등록본부에 보고된 자료에서 중복을 제외하고, 초진 일자가 1997년 1월 1일부터 12월 31일 사이에 있는 수련병원 입원 환자의 암발생 자료이다(이하 중앙암등록자료, Central cancer registry data).

두 번째 자료원은 1997년 1월 1일부터 12월 31일까지 광주광역시 소재 의료기관의 해부병리 및 임상병리과와 5개 해부병리진단기관에 의뢰된 병리검사 소견 중 중복을 제외하고, 암으로 판명되었고, 그리고 의뢰한 해당 병원의 초진 일자가 1997년 1월 1일부터 12월 31일 사이에 있는 외래 및 입원 환자의 암발생 자료이다(이하 병리자료, Biopsy data).

세 번째 자료원은 첫 번째와 두 번째 자료원을 제외한 나머지 자료원, 즉 사망신고자료, 보험청구자료, 방사선과 판독자료, 치료방사선과 자료 등을 이용하여 광주광역시 소재 의료기관의 의무기록을 확인한 결과 암으로 판명되었고, 그리고 해당 병원의 초진 일자가 1997년 1월 1일부터 12월 31일 사이에 있는 외래 및 입원 환자의 암발생 자료이다(이하 기타 자료).

이렇게 서로 다른 자료원에서 수집된 암환자들을 주민등록번호 13자리를 이용하여 세 자료원에서 동시에 확인될 수 있도록 짝짓기 하였다.

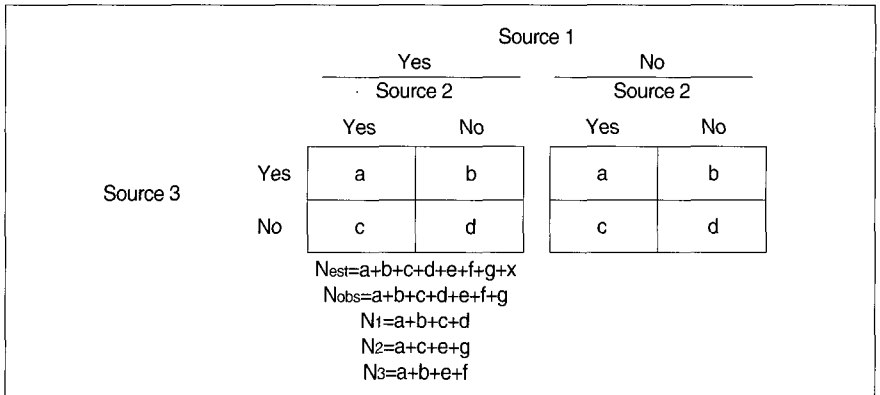
2. 세 자료원 모델을 이용한 완전성 추정

세 자료원에서 확인되어 숫자가 파악된 암환자들은 $2^3-1=7$개의 조합으로 존재할 것이며, 세 자료원 어디에도 포함되지 않는 암환자는 미지의 수 x로 존재할 것이다. 저자들은 미지의 값 x를 추정하기 위해 log-linear 모델을 이용하였고, 그 결과로 Figure 1과 같은 8가지의 모델과 이에 따른 추정값 공식을 구할 수 있었다.

8개의 모델은 k개의 자료원에서는 많으면 k-1의 상호작용(interaction)까지 있다는 일반적인 log-linear 모델의 원칙을 이용하여 3원 상호작용(three way interaction)은 없다고 가정하고 구한 것이며, 이 모델에 따른 추정값은 BMDP 4F를 이용하여 구했다. 이후 가장 최상의 모델을 선택하기 위하여 Schwarz(1978)가 주장한 BIC(Bayesian Information Criterion) 값이 가장 적은 모델을 선택하였다. BIC는 다음과 같은 식에 의해 구해진다.

$$BIC = G^2 - (\log N_{obs}/2\pi)\,(d.f.)$$

G^2는 우도비와 관련된 통계값이며, N_{obs}는 관측된 총 수, d.f.는 모델의 자유도이다(Ernest 등, 1995). 이렇게 미지의 수 x를 구하고, 이 값을 이용하여 지역암등록 자료의 전체 추정값과 완전성을 95% 신뢰구간 안에서 추정하였다.



d.f.*	Model	Estimator(X_{est})
3	Independent	$N_{est} - N_{obs}$ N_{est} is the solution of $(N_{est}-N_1)(N_{est}-N_2)(N_{est}-N_3)=N_{est}^2(N_{est}-N_{obs})$
2	1-2 interactions	(c+d+g)(f)/(a+b+e)
2	1-3 interactions	(b+d+f)(g)/(a+c+e)
2	2-3 interactions	(e+f+g)(d)/(a+b+c)
1	1-2, 1-3 interactions	gf/e
1	1-2, 2-3 interactions	gd/c
1	1-3, 2-3 interactions	df/b
0	1-2, 1-3, 2-3 interactions	(adfg)/(bce)

Figure 1. The data layout and the application of log-linear, three-source model developed for derivation of an estimate. * d.f., degree of freedom.

연구 결과

세 자료원에 의해 개별적으로 또는 동시에 확인된 전체 암발생건수는 2,142건이었다. 중앙암등록자료로 확인된 암발생건수는 1,581건, 병리자료로 확인된 암발생건수는 1,359건, 그리고 기타 자료로 확인된 암발생건수는 1,320건이었다. 둘 이상의 자료원에 의해 70% 이상의 암발생이 확인되었으며, 세 자료원 모두에 의해 확인된 암발생도 27.3%를 차지하였다.

capture-recapture 방법으로 추정된 전체 암발생건수는 2,307건(95% 신뢰구간, 2,282-2,332)으로 나타났으며, 등록의 완전성은 92.9%(95% 신뢰구간, 91.9-93.9%)로 추정되었다. 남자가 여자보다 자료의 완전성이 더 높게 나타났으며, 연령대별로 비교하니 75세 이상 연령대에서 자료의 완전성이 두드러지게 감소하였다(Table 1, Figure 3).

세 자료원에 의해 확인된 암발생건수가 100건을 넘는 다발성 암종에 대해 capture-recapture 방법으로 암발생건수를 추정해보니 자료의 완전성은 암종 모두가 90%를 넘었으며, 갑상선암이 97% 이상의 높은 완전성을 보인데 비하여 위암과 간암은 92% 정도의 상대적으로 낮은 완전성을 보였다(Table 2).

고 찰

본 연구에서 사용한 capture-recapture 방법은 과거 동물생태학에서 일정지역 안에 있는 동물의 수를 추정하기 위해 사용한 방법으로, 이 방법을 적용하기 위해서는 다음의 세 가지 중요한 가정이 따라야 한다. 첫 번째 가정은 어떤 자료원에 등록될 확률 즉, 포획능력(catchability)이 각각의 개체에서 동일해야 한다는 것이고, 두 번째 가정은 두 자료원 모형에서는 각 자료원이 서로 독립적이어야 한다는 것이다. 그리고 세 번째 가정은 연구대상이 닫혀 있어야 한다는 것으로, 연구기간 동안 새로 들어오거나 탈락이 없어야 한다는 것이다(Ernest 등, 1995; Laure 등, 1996).

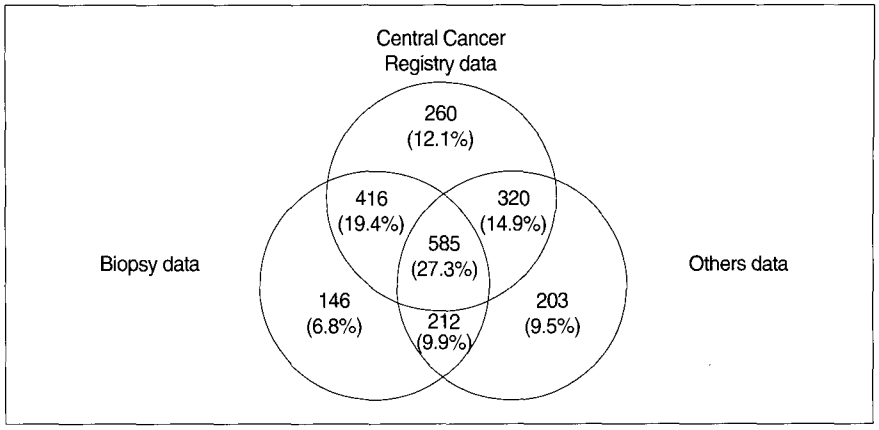


Figure 2. Cancer cases diagnosed and reported to the cancer registry of Kwangju by sources of notification.

Table 1. Capture-recapture estimates of completeness of the Kwangju cancer registry by sex

Sex	Observed cases	Estimated cases (95% CI †)	Estimated completeness (95% CI †)
Total	2,142	2,307(2,282-2,332)	92.9%(91.9-93.9)
Male	1,216	1,299(1,281-1,317)	93.6%(92.4-94.9)
Female	926	1,014(995-1,032)	91.3%(89.7-93.0)

† 95% CI, 95% confidence interval

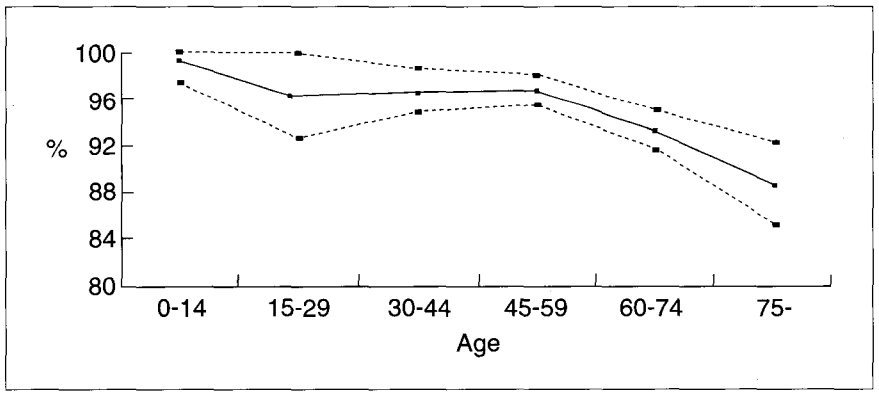


Figure 3. Capture-recapture estimates of completeness of the Kwangju cancer registry by age. The solid line is point estimate. The dashed line is 95% confidence interval.

Table 2. Capture-recapture estimates of completeness of the Kwangju cancer registry by cancer site

Site	ICD-O	Observed cases	Estimated cases (95% CI †)	Estimated completeness (95% CI †)
Stomach	16	406	440(429-451)	92.3%(89.9-94.7)
Colon-Rectum	18-20	216	230(222-237)	94.1%(91.2-97.1)
Liver & IHBD	22	276	298(289-307)	92.6%(89.8-95.6)
Bronchus & Lung	34	249	265(257-273)	94.0%(91.3-96.9)
Breast	50	122	127(123-132)	95.9%(92.6-99.4)
Uterus	53-55	122	130(124-134)	94.6%(91.0-98.5)
Thyroid	73	100	103(100-106)	97.1%(94.0-100.0)

† 95% CI, 95% confidence interval

첫 번째 가정이 위배되면 추정에 편이가 발생하게 되어 과대 또는 과소 추정을 가져온다. 본 연구에서도 전체 암환자, 남자, 여자, 연령대별로 각각 최상의 다른 모델을 선택하여 값을 추정하였는데, 남자의 추정값 1,299건과 여자의 추정값 1,014건의 합 2,313건은 전체 암환자 추정값 2,307건과 비슷하게 나온 데 비해 각 연령대별 추정값의 합 2,267건은 전체 추정값보다 적게 나왔다. 이러한 차이는 연령에 따라 자료원에 대한 포획능력이 다르기 때문에 발생한 것으로 사료된다. 연령뿐만 아니라, 암종에 따라서도 포획능력에 차이가 나타날 수 있다. 그런데, 의학연구와 같은 인간을 대상으로 한 연구에서 포획능력이 각각의 개체에서 동일해야한다는 가정을 만족시키기는 대단히 어렵다. 그 해결책은 각 연구특성에 따라 계층별로 log-linear 모델을 수립하는 것이 될 것이다.

두 번째 가정을 세 자료원 모형에서는 상호작용을 보정하는 log-linear 모델을 이용하여 해결할 수 있는데, 이 방법은 자료원 사이의 의존성을 통계적 모델을 통해 보정하는 것이다.

세 번째 가정은 본 연구에 큰 영향을 미치지 않았을 것으로 판단된다. 비록 연구기간에 출입과 탈락이 있다고 하더라도 지역암등록은 연구기간 이후에도 지속적으로 암환자를 포획하여 초진 일자에 따라 등록함으로 이 기간 동안 발생한 이동은 이후에 보완될 가능성이 크다. 그리고, 암은 매우 중한 질병임으로 이동은 매우 적을 것으로 사료된다.

비록 암종, 연령이 자료원의 포획능력에 영향을 미쳐 과대 또는 과소 추정되는 한계에도 불구하고, capture-recapture 방법은 치명률에 의존해 간접적으로 완전성을 추정하는 DCO index나 사망률/발생률 비에 비하여 치명률에 의존하지 않고 독립적으로 완전성을 추정할 수 있다는 점에서 그 의의가 있다고 사료된다. 즉, 서로 다른 예후를 가진 여러 암종의 완전성 추정에 있어서 DCO index 등은 암종의 치명률에 따라 그 값이 약 10% 정도의 차이가 나타난 반면 capture-recapture 방법은 그 값이 거의 일정하게 나타난다는 장점이 있는 것이다(Brenner, 1994).

지역암등록 통계의 완전성을 파악하는 지표에는 DCO index나 사망률/발생률 비 외에 비슷한 지리적 위치와 민족적 특성을 갖는 지역암등록 자료와 발생률과 비교하는 방법, 그리고 연령별 암발생률 곡선의 변화를 관찰하는 방법 등이 있다(Jensen, 1991). 광주광역시 지역암등록사업 결과 1997년도 광주광역시의 인구 10만명당 세계 표준인구 연령보정 암발생률은 남자가 272.5건, 여자가 142.5건이었다(광주광역시 지역암등록사업단, 1999). 이러한 광주광역시의 세계 표준인구 암발생률은 서울의 남자 289.4와 여자 174.2(안돈희 등, 1999)에 비해 여자에서 비교적 낮은 수치를 보이고 있으나, 부산, 대구, 강화(Parkin 등, 1997)에 비해서 비슷하거나 상대적으로 높음을 볼 수 있다(Table 3). 또한 연령별 암발생률 곡선에 있어서도 여자 75세 이상을 제외하고는 계속 증가되는 경향을 보이고 증가율의 감소도 75세 이상에서만 관측되고 있다(Figure 4). 이러한 결과로 미루어 보아 광주광역시 지역암등록은 상당히 높은 완전성을 보일 것으로 사료된다.

세 자료원에 의해 추정된 등록의 완전성은 92.9%로 추정되었는데, 남자가 여자보다 자료의 완전성이 더 높게 나타났으며 연령대별로 비교하니 75세 이상 연령대에서 완전성이 감소하는 경향을 보였다. 이러한 남녀의 완전성의 차이는 다른 지역과의 암발생률을 비교할 때, 그리고 연령별 발생률 곡선의 변화를 관찰할 때 목격된 결과와 동일하다. 또한 75세 이상 연령대에서 완전성이 감소하는 경향도 위의 두 경우에서 목격된 결과와 역시 동일하다.

본 연구에서는 서로 독립적인 세 자료원의 capture-recapture 방법을 통해 암환자 수와 지역암등록 자료의 완전성을 추정하였다. capture-recapture 방법은 암등록 자료의 완전성을 추정하는데 있어서 특별한 조사(survey)가 필요없다. 단지 개개의 암 환자에 대해 일상적으로 수집되는 자료들을 자료원별로 묶어 추적하면 capture-recapture 방법을 이용하는 유용한 수단이 될 수 있다(Brenner, 1994). 따라서 역사가 짧은 광주광역시 지역암

Table 3. Comparison of cancer incidence per 100,000 between registries

Registry	Year	Male		Female	
		Crude rate	ASR * (world)	Crude rate	ASR * (world)
Kwangju	1997	181.5	272.5	137.4	142.5
Seoul	1992-1994	183.0	289.4	58.1	174.2
Pusan	1996-1997	177.7	255.2	142.7	144.5
Taegu	1997	151.6	222.4	131.5	131.1
Kangwha	1986-1992	238.9	201.7	146.3	110.7

* ASR, Age standardized rate

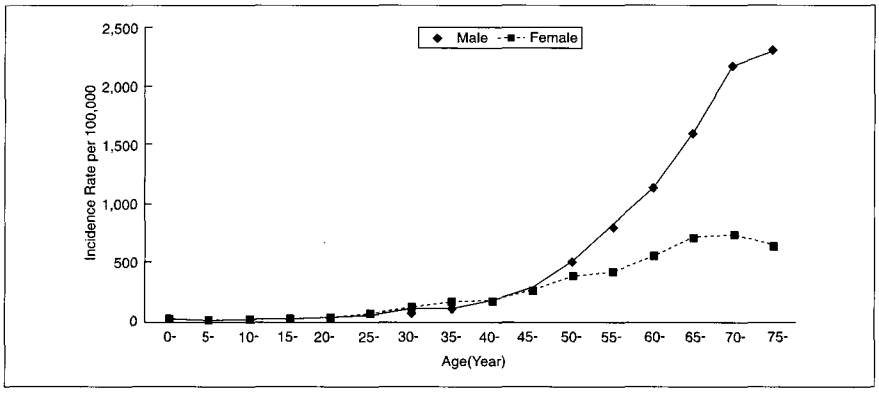


Figure 4. Age-specific cancer incidence rates in Kwangju, 1997.

등록사업에 있어서 capture-recapture 방법은 신뢰성이 부족한 DCO index나 사망률/발생률 비를 대신하여 지역암등록 자료의 완전성을 추정하기 위한 다른 방법으로 이용될 수 있으리라 판단된다.

결 론

본 연구는 capture-recapture 방법을 이용하여 광주광역시 지역암등록의 완전성을 추정하고자 하였다. 자료원으로 중앙암등록 자료, 병리 자료, 기타 자료를 이용하였으며, 서로 다른 자료원에서 수집된 암발생건을 주민등록번호 13자리를 이용하여 짝짓기 하였다. 세 자료원 어디에도 해당되지 않는 미지의 환자 수를 추정하기 위해 log-linear 모형을 이용하였고, 추정된 미지의 값을 이용하여 지역암등록 자료의 전체 추정값과 완전성을 95% 신뢰구간 안에서 계산하였다. 연구결과 자료의 완전성은 92.9%로 추정되었다. 남자가 여자보다 자료의 완전성이 더 높았으며, 연령대별로 비교하니 75세 이상 연령대에서 자료의 완전성이 두드러지게 감소하였다. 다발성 암종 모두에서 자료의 완전성은 90%를 넘었는데, 갑상선암이 97% 이상의 높은 완전성을 보인데 비하여 위암과 간암은 92% 정도의 상대적으로 낮은 완전성을 보였다.

참고문헌

광주광역시 지역암등록사업단. 광주광역시 지역암등록사업 연례 보고서. 1999. 49쪽

안돈희. 5대 대도시 지역암등록 사업. 1999

중앙암등록본부. 암등록지침서. 1998

하미나, 권호장, 강대희, 조수헌, 유근영 등. 소아천식을 통해서 본 의료보험 상병자료의 완전성 추정: Capture-Recapture 분석방법의 적용. 예방의학회지 1997; 30(2): 428-436

Brenner H, Stegmaier C, Ziegler. Estimating completeness of cancer registration in Saarland/Germany with capture-recapture method. *Eur J Cancer* 1994; 30A(11): 1659-1663

Cochi SL, Edmonds LE, Dyer K, et al. Congenital rubella syndrome in the United States, 1970-1985. *Am J Epidemiol* 1989; 129: 349-361

Ernest B, Hook, Ronald R, Regal. Capture-Recapture Method in Epidemiology: Methods and Limitations. *Epidemiol Rev* 1995; 17(2): 243-264

Esteban D, Whelan S, Laudico A, Parkin DM. Quality control. In: Manual for Cancer Registry Personnel, IARC; 1995.

Jensen OM, Storm HH. Purposes and uses of cancer registration. In: Jensen OM, Parkin DM, editors. Cancer Registration: Principles and Methods(IARC Scientific Publications No. 95). Lyon, IARC; 1991.p.7-21.

Hook EB, Regal RR. The value of capture-recapture methods even for apparent exhaustive surveys. The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. *Am J Epidemiol* 1992; 135: 1060-1067

Laporte RE, Tull ES, McCarty DJ. Monitoring the incidence of myocardial infarctions: applications of capture-mark-recapture technology. *Int J Epidemiol* 1992; 21: 258-263

Laure P, Beverley B, Joseph L. Case counting in epidemiology: limitations of methods based on multiple data sources. *Int J Epidemiol* 1996; 25: 474-478

McCarty DJ, Tull ES, Moy CS, Kwoh CK, LaPorte RE. Ascertainment corrected rates: applications of capture-recapture methods. *Int J Epidemiol* 1993; 22: 559-565

Neugebauer R. Application of a capture-recapture method(the Bernoulli census) to historical epidemiology. *Am J Epidemiol* 1984; 120: 626-634

Parkin DM, Whelan SL, Ferlay J, Raymond L, Young J. Cancer incidence in Five Continents Vol. Ⅶ(IARC Scientific Publications No. 143). IARC; 1997.p.406-409.

Robles SC, Marrett LD, Clarke EA, Risch HA. An application of capture-recapture methods to the estimation of completeness of cancer registration. *J Clin Epidemiol* 1988; 41: 495-501

Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978; 6: 461-464

Skeet RG. Quality and quality control. In: Jensen OM, Parkin DM, editors. Cancer Registration: Principles and Methods(IARC Scientific Publications No. 95). Lyon, IARC; 1991.p. 101-107.