韓國疫學會誌 : 第24 卷 第1號
Korean Journal of Epidemiology
Vol. 24, No. 1, June 2002, 7-16

Intermediate 변수의 영향을 통제하는 통계적 방법론에 대한 연구
- 건강근로자효과를 통제하기 위한 새로운 접근

남정모[1], 김진흠[2], 강대룡[3], 안연순[4], 이후연[1], 이대희[1]

연세대학교 의과대학 예방의학교실[1], 수원대학교 통계정보학과[2], 연세대학교 보건대학원[3]
한국산업안전공단 산업안전보건연구원[4]

서 론

건강근로자효과(healthy worker effect)를 다루는 문제는 산업역학에서 중요한 방법론적인 쟁점이다. 즉, 특정 위험작업의 노출정도와 사망(또는 질병발생)의 관련성을 밝히는 연구에서 건강근로자효과는 그 관련성의 강도를 크게 왜곡하는 치우침을 야기할 수 있다. 여기서 특정 위험작업의 노출정도에 대한 정의는 높은 농도의 유해물질, 높은 작업강도(탄광, 주물 등), 그리고 고도의 스트레스 직종(야간작업, 운전작업) 등을 포함하는 포괄하는 것으로 생각할 수 있다.

이러한 건강근로자효과는 크게 건강근로자고용효과(healthy worker hire effect)와 건강근로자생존효과(healthy worker survival effect)로 나누어 설명할 수 있다. 건강근로자고용효과는 건강한 사람이 그렇지 못한 사람에 비해 특정 위험작업에 노출되는 직업에 고용될 가능성이 높기 때문에 발생하는 초기선택과정(initial selection process)으로 설명할 수 있다. 반면 건강근로자생존효과는 t 시점에서 건강한 근로자가 그렇지 못한 근로자 보다 계속해서 고용되어 그 직업에 남아있을 가능성이 높고 또한 t 시점의 고용여부는 다음 시점의 위험작업에 노출되는 정도를 결정하는 계속적인 선택과정(continuing selection process)으로 설명할 수 있다[1, 2].

건강근로자고용효과는 노출정도와 사망(또는 질병발생)의 관련성에 대한 크기를 과소추정하는 방향으로 치우침을 야기한다. 예를 들어 특정 위험작업에 노출된 직업군의 일반인구집단에 대한 표준화사망비가 1보다 작아지는 경우를 생각할 수 있다. 이러한 치우침을 제거하고자 일반인구집단이 아닌 초기선택과정이 비슷한 내부 대조군을 산업역학 분야에서 많이 사용하고 있다[3]. 건강근로자고용효과와 마찬가지로 건강근로자생존효과도 사망에 대한 노출변수의 효과를 일반적으로 약화시킨다. 그러나 그 과정은 건강근로자고용효과에 비해 훨씬 복잡하다. 또한 건강근로자생존효과는 내부 대조군을 사용하여도 여전히 노출정도와 사망의 관련성에 대한 치우침을 야기한다.

건강근로자고용효과는 특정직업에 고용되는 초기선택과정이므로 고용되는 시점에서 건강상태의 차이로 발생하는 혼란효과로 생각할 수 있다. 따라서 이론적으로는 각 개인의 건강상태를 혼란변수로 간주하여 그 효과를 통제하면 건강근로자고용효과를 제거할 수 있다. 그렇지만 각 개인의 초기 건강상태를 정확하게 측정한다는 것은 매우 어렵다. Checkoway등[4]은 건강근로자생존효과를 고용시 연령(age at hire), 고용기간(duration of employment), 고용이후의 시간(time-since-hire), 그

접수 : 2001년 1월 11일 채택 : 2002년 2월 19일
교신저자 : **남정모** 연세대학교 의과대학 예방의학교실
주소 : 서울시 서대문구 신촌동 134번지
전화 : 02-361-5358 팩스 : 02-392-8133
E-mail : cmnam@yumc.yonsei.ac.kr
연구비 : 이 연구는 연세대학교 의과대학 교수연구비(1999)에 의해 수행되었음

리고 위험에 노출되는 연령(age at risk) 4가지 요인으로 발생하는 혼란효과로 설명하였다. 한편 Steenland와 Stayner[3]는 건강근로자생존효과를 시간의존형 변수인 고용상태(employment status)가 혼란변수로 작용하여 발생하는 문제로 설명하였다. Robins는 건강근로자생존효과를 시간의존형 변수인 고용상태가 노출정도와 사망의 관계에서 혼란변수의 작용과 동시에 중간 매개변수(intermediate variable)로 작용하는 현상으로 설명하였다. 따라서 고용상태를 단순히 콕스의 비례위험모형(Cox's proportional hazards model)에 시간의존형 독립변수로 추가한다고 그 치우침이 제거되지 않는다고 주장하였으며 이러한 치우침을 제거하고자 G-알고리즘 등을 제안하였다[5-8]. 또한 Nam과 Zelen[9]은 이러한 시간의존형 매개변수로 발생하는 치우침의 원인을 length bias sampling에 대한 현상으로 설명하고 기존의 로그-순위 검정, 층화된 로그-순위 검정 등이 제1종의 오류가 매우 커진다는 것을 수리적으로 증명하고 시뮬레이션을 통해 그 사실을 밝혔다. 또한 치우침의 영향을 제거하기 위한 새로운 개념적인 모형과 통계적인 모형을 제안하였다.

이 연구는 건강근로자효과가 발생하는 이론적인 모형을 구축하고 그 치우침을 제거하는 새로운 방법을 제안하고자 시행되었다. 또한 새로 제안한 방법과 기존의 통계적인 방법들이 건강근로자효과가 발생하는 여러 가지 가상적인 시나리오 상에서 제1종의 오류 및 검정력이 어떻게 변화하는지 시뮬레이션을 통해 비교하므로서 새로 제안한 방법의 타당성과 활용성을 알아보고자 시행되었다.

연구대상 및 방법

1. 건강근로자효과에 대한 이론적인 틀

건강근로자효과는 혼란변수로 작용하는 초기시점에서 건강상태와 t 시점에서 시간의존형 변수인 고용상태의 영향으로 발생한다. 먼저 건강근로자고용효과는 다음과 같은 이론적인 모형으로 설명할 수 있다. 초기시점에서 상대적으로 건강한 사람은 그렇지 못한 사람에 비해 특정 위험작업에 더 많이 노출되고 또한 초기시점에서 건강상태가 나쁠수록 생존시간이 짧아진다(또는 사망의 위험이 증가한다)고 가정할 수 있다. 즉, 초기시점의 건강상태는 위험작업의 노출정도와 사망까지 생존

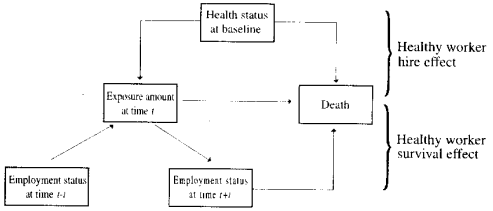

Fig. 1. Basic structure of healthy worker effect.

시간의 관련성에 순수한 혼란변수로 작용한다. 역학연구에서 초기시점의 건강상태를 정확하게 측정한다는 것은 매우 어려우므로 대신 측정오차가 동반된 건강상태를 측정하는 것이 일반적인 것으로 생각할 수 있다.

건강근로자생존효과는 다음과 같은 이론모형으로 설명할 수 있다. 먼저 (t-1) 시점에서 고용상태는 t 시점에서의 노출정도에 영향을 미치며, 또한 t 시점에서 노출정도는 (t+1) 시점에서의 고용상태에 영향을 미치고, 그리고 고용상태 변화여부는 독립적으로 생존시간에 영향을 미친다. 이 연구에서 고용상태가 변하기 전까지 개인의 위험작업에 대한 노출정도는 x_0로 일정하게 유지하다가 고용상태가 변하면 노출정도가 x_1으로 변하는 시간의존형 변수로 가정한다. 이상의 건강근로자 효과에 대한 개념적인 모형은 Fig. 1과 같다.

이상과 같은 이론모형에서 건강근로자효과를 제거하기 위하여 이 연구는 다음과 같은 방법론적인 접근을 시도하였다. 먼저 건강근로자고용효과는 초기건강상태의 혼란효과에 의해 발생하므로 이 효과를 제거하기 위해서는 초기건강상태를 기존의 통계방법을 이용하여 혼란변수로 통제하면 된다. 다음으로 고용상태의 변화는 각 근로자의 생존시간을 변화시키고 고용상태의 변화가 노출정도를 변화시키므로 단순히 고용상태가 변한 집단과 변하지 않은 집단을 구분하여 노출정도와 사망의 관계를 분석하면 length bias sampling으로 인한 치우침이 발생하므로 이를 고려하지 않은 기존의 통계방법을 적용할 수가 없다[9]. 따라서 본 연구는 각 개인의 고용상태가 변하지 않는다고 가정하였을 때와 변한 후의 개념적 생존시간들을 정의하고 또한 고용상태가 변할 때까지의 대기시간을 정의하여 이들의 경쟁관계로 고용상태의 변화가 관찰되는 것으로 모형화하여 length bias sampling으로 인한 치우침을 제거하고자 하

였다.

건강근로자생존효과를 야기하는 고용상태의 변화 여부는 다음과 같이 이론화 할 수 있다. 확률변수 T_0를 고용상태가 변하지 않는다고 가정할 때의 이론적인 생존시간으로 정의하고, W를 고용상태가 변하기까지의 대기시간으로 정의하면 시간의존형 변수인 고용상태의 변화여부 Z는 다음과 같이 정의할 수 있다.

이와같은 과정에 의해 고용상태가 변하면 위험작업에 노출되는 정도가 변하고 또한 고용상태의 변화는 독립적으로 생존시간의 변화를 야기한다. 고용상태가 변하여 변화된 생존시간을 T_1으로 정의하면 T_0와 T_1은 개념적 변수이고 실제 관찰된 생존시간은 고용상태에 따라 절단(truncation)된 형태로 관찰되므로 단지 관찰된 생존시간을 이용하여 노출정도와의 관련성을 분석하면 length bias sampling으로 야기되는 치우침이 발생한다[9].

$$Z=\begin{cases} 1\ (\text{즉, 고용상태가 변함}) & \text{만약 } T_0 > W \\ 0\ (\text{즉, 고용상태가 변하지 않음}) & \text{만약 } T_0 \le W \end{cases}$$

2. 건강근로자효과를 제거하는 새로운 통계적 방법

먼저 초기시점의 건강상태는 편의상 건강한 그룹과(S=0)과 그렇지 못한 그룹(S=1)으로 나눈다. 위에서 정의한 T_0, T_1, 그리고 W의 생존함수를 각각 S_0, S_1, 그리고 S_W라 정의하면 이들 생존시간과 t 시점에서의 노출정도는 비례위험모형을 확장한 다음과 같은 관계가 성립한다고 가정할 수 있다.

$$\begin{aligned} S_0(t)&=S_{00}(t)\exp(\beta_0 x(t)+\gamma_0 S), \\ S_1(t)&=S_{10}(t)\exp(\beta_1 x(t)+\gamma_1 S), \\ S_W(t)&=S_{W0}(t)\exp(\eta\, x(t)) \end{aligned} \qquad \text{식(1)}$$

여기서, S_{00}, S_{10}, 그리고 S_{W0}는 각각의 기저위험함수(baseline hazard function)에 대응하는 생존함수이다. 위의 모형에서 β_0는 고용상태가 변하지 않을 경우 노출정도와 사망위험의 관련성의 크기이고, β_1는 고용상태가 변한 경우의 노출정도와 사망위험의 관련성에 대한 크기로 해석할 수 있다. 그리고 η는 노출정도와 고용상태가 변할 때까지의 대기시간과의 관련성을 나타낸다.

이상의 모형에서 노출정도와 사망과의 관련성에 대한 귀무가설은 $H_0: \beta_0=\beta_1=0$와 같다. 이상의 가정과 모형에서 Nam & Zelen[9]의 방법을 확장하면 $(\beta_0,\ \gamma_0)$ 그리고 $(\beta_1,\ \gamma_1)$은 서로 분리(separable) 가능한 우도(likelihood)로 구성되며 β_0와 β_1에 대한 스코어 함수(score function)는 각각 다음과 같다.

$$U_0(\beta_0,\gamma_0)=\sum_{k=1}^{N}\int_0^{\infty}[X_k(t)-\frac{\sum_{s=1}^{N} x_s(t)(1-z_s(t))R_s(t)\exp(\beta_0 x_s(t)+\gamma_0 S_s)}{\sum_{s=1}^{N}(1-z_s(t))R_s(t)\exp(\beta_0 x_s(t)+\gamma_0 S_s)}](1-z_k(t))dN_k(t)$$

$$U_1(\beta_1,\gamma_1)=\sum_{k=1}^{N}\int_0^{\infty}[X_k(t)-\frac{\sum_{s=1}^{N} x_s(t)\, z_s(t)\ R_s(t)\exp(\beta_1 x_s(t)+\gamma_1 S_s)}{\sum_{s=1}^{N} z_s(t)\ R_s(t)\exp(\beta_1 x_s(t)+\gamma_1 S_s)}]z_k(t)dN_k(t)$$

식(2)

여기서, $N_k(t)$는 k 번째 대상의 t 시점에서 사망여부를 나타내는 셈과정(counting process)이고 $R_k(t)$는 k 번째 대상의 t 시점 바로 직전까지 위험상태 여부를 나타내는 확률과정(stochastic process)이다.

즉, 이 연구에서 제안한 방법은 다음과 같은 두 단계 검정으로 설명할 수 있다. 먼저 첫 번째 단계는 전체 자료를 이용하여 $\beta_0=0$를 검정한다. 이 경우 새롭게 제안한 스코어 함수는 콕스 비례위험 회귀모형의 스코어 함수와 비교할 때 각 시점에서의 위험집단(risk set)에 차이가 있다. 즉, 기존의 콕스 비례위험 회귀모형에서 t 시점의 위험집단은 t 시점까지 사망하지 않고 생존한 모든 관찰치가 포함되지만 새롭게 제안한 방법의 위험집단은 t 시점까지 사망하지 않고 또한 t 시점까지 고용상태가 변하지 않았던 모든 관찰치가 포함된다. 두 번째 단계는 고용상태가 변한 근로자만을 대상으로 $\beta_1=0$을 검정한다. 따라서 두 번째 단계에서 제안한 방법의 t 시점에 대한 위험집단은 t 시점까지 고용상태가 변하고 그리고 t 시점까지 사망하지 않은 관찰치이다.

이 연구에서는 Newton-Raphson 알고리즘을 이용하여 회귀계수 β_0와 β_1을 추정하였고 또한 이들 추정치들의 표준편차를 관찰정보행렬(observed information matrix)을 이용하여 추정하였다. 이들 추정치들을 이용하여 귀무가설 H_0에 대한 다음과 같은 Wald 형태의 검정통계량을 최종적으로 제안하였다. 제안한 통계량은 근사적으로 자유도가 2인 카이제곱 분포를 따른다.

$$\chi^2_2=\left(\frac{\hat{\beta}_0}{se(\hat{\beta}_0)}\right)^2+\left(\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right)^2$$

3. 시뮬레이션 모형

노출정도와 생존시간의 관련성을 검정하기 위한 다음 4가지 모형이 건강근로자효과로 인해 제1종의 오류가 5%로 유지하는지를 시뮬레이션을 통해 분석하였다. 모형은 크게 시간의존형 콕스모형 3가지와 이 연구에서 제안한 모형이다. 먼저 h(t)를 t 시점에서의 위험함수로 정의하고 $h_0(t)$를 기저위험함수로 정의한다. 또한 b_i, c_i, 그리고 d_i를 각각 노출정도, 건강상태, 그리고 고용상태에 대응되는 각 모형에서의 회귀계수로 정의한다.

- 모형 1 : 시간의존형 노출정도와 시간비의존형 건강상태를 독립변수로 한 콕스모형 :
  $h(t) = h_0(t)\exp(b_1x(t)+c_1S)$

- 모형 2 : 시간의존형 노출정도, 시간비의존형 고용상태(연구 종료시점에서의 상태), 그리고 시간비의존형 건강상태를 독립변수로 한 콕스모형 :
  $h(t) = h_0(t)\exp(b_2x(t)+c_2S+d_2Z)$

- 모형 3 : 시간의존형 노출정도와 시간의존형 고용상태 그리고 시간비의존형 건강상태를 독립변수로 한 콕스모형 :
  $h(t) = h_0(t)\exp(b_3x(t)+c_3S+d_3Z(t))$

- 모형 4 : 시간의존형 노출정도와 시간비의존형 건강상태를 독립변수로 하여 본 연구에서 새롭게 제안한 모형

이상의 4가지 모형의 차이점은 다음과 같다. 먼저, 모형 1은 고용상태의 변화를 공변수로서 모형에 포함하지 않았고 모형 2는 고용상태의 변화를 공변수로서 모형에 포함하였으나 시간비의존형 변수로 간주하였다. 그리고 모형 3은 고용상태의 변화를 시간비의존형 변수로 간주한 콕스모형이다. 그리고 모형 4는 식(1)의 모형과 식(2)의 스코아 함수를 통하여 검정하는 본 연구에서 새롭게 제안한 방법이며 고용상태의 변화를 두 단계로 나누어 분석하므로서 그 효과를 통제하는 방법이다.

가. 건강근로자고용효과에 대한 모수

초기시점의 건강상태에 대한 자료는 건강한 집단과 그렇지 않는 집단으로 동일한 표본수로 이분화 하여 참값의 자료를 발생하였다. 건강한 근로자의 초기 노출정도는 일양분포(uniform distribution) U(2/3, 1), 그렇지 않는 근로자는 U(1/3, 1)로서 건강한 근로자의 노출정도의 평균이 높게 자료를 발생하였다. 한편 직업에 고용될 때의 건강상태를 측정오차가 없이 정확히 측정한다는 것은 실제 역학연구에서 어려우므로 민감도와 특이도가 각각 80%인 대리변수를 사용하므로서 분류오류의 영향도 아울러 조사하였다. 즉, 건강한 집단내에서의 각 근로자가 건강하다고 분류될 확률이 80%, 건강하지 않는 집단내에서의 근로자가 건강하지 않다고 분류될 확률이 80%가 되도록 이항분포를 이용하여 참값의 건강상태를 수정하고 이를 건강상태에 대한 대리변수로 사용하였다.

나. 건강근로자생존효과에 대한 모수

건강근로자생존효과는 초기 노출정도와 고용상태가 변하기까지의 대기시간 W, 고용상태의 변화로 인한 평균노출정도의 변화, 그리고 고용상태의 변화로 인한 평균생존시간의 변화를 모수화하여 자료를 발생하였다.

노출정도와 고용상태의 변화는 식 (1)의 η값을 0.5로 하여 노출이 많을수록 고용상태가 빨리 변화도록 하였다. 만약 k 번째 근로자의 초기 노출정도가 x_{0k}인 경우에 w시점에서 고용상태가 변하면 w시점 이후의 노출정도는 U(0, x_{1k})에서 자료를 발생하여 노출정도가 감소하는 것으로 하였다. 위험작업의 노출이 없고 고용상태가 변하지 않는다고 가정하였을 때 건강한 근로자와 그렇지 못한 근로자가 사망할 때까지의 평균생존시간을 m_{0h}와 m_{0u}로 모수화 하였다. 또한 위험작업의 노출이 없고 초기 시점에서부터 고용상태가 변한다고 가정하였을 때 건강한 근로자와 그렇지 못한 근로자가 사망할 때까지의 평균생존시간을 m_{1h}와 m_{1u}로 모수화 하였다.

다. 전체 자료발생 및 검정

건강근로자효과를 야기하는 모수들 중 본 연구에서는 (m_{0h}, m_{0u}, m_{1h}, m_{1u})와 (β_0, β_1)을 변화하고 나머지 값들은 고정하였다. 만약 $m_{0h} \neq m_{0u}$, $m_{1h} \neq m_{1u}$이면 건강근로자고용효과가 있고, $m_{0h} \neq m_{1h}$, $m_{0u} \neq m_{1u}$이면 건강근로자생존효과가 있다는 것을 의미한다. 그리고, $m_{0h} \neq m_{0u}$,

$m_{1h} \neq m_{1u}$, $m_{0h} \neq m_{1h}$, $m_{0u} \neq m_{1u}$이면 건강근로자고용효과와 생존효과가 모두 있으며, $m_{0h}=m_{0u}=m_{1h}=m_{1u}$인 경우는 건강근로자효과가 없음을 의미한다.

이상의 모든 생존시간은 형상모수(shape parameter)가 2인 Weibull 분포에서 자료를 발생하였다. 또한 생존시간과 독립적으로 중도절단시간에 대한 자료를 발생하여 생존시간보다 중도절단시간이 작으면 중도절단된 것으로 간주하여 최종적으로 관찰되는 생존시간 자료를 발생하였다.

위의 4가지 모형에서 노출정도와 사망의 관계에 대한 각 검정들의 제1종의 오류가 5%로 잘 유지되는지의 여부는 다음과 같은 절차를 통해 조사하였다. 먼저 각 모형에서 관심있는 귀무가설은 다음과 같다.

$$H_{0i} : b_i = 0 (i=1, 2, 3), \quad H_{04} : \beta_{1h}=\beta_{1u}=0$$

노출정도가 사망과 관계가 없는 경우, 발생한 자료를 이용하여 각각의 가설에 대한 유의수준 5%인 검정을 시행하여 가설이 기각되는지의 여부를 조사하였다. 이러한 시뮬레이션 과정을 500번 반복하여 총 기각되는 횟수를 500으로 나누어 각 모형에서의 제1종의 오류에 대한 확률을 추정하였다. 한번의 시뮬레이션 과정에서 각각의 귀무가설이 기각되는 확률이 이항분포 B(1, 0.05)를 따르므로 이를 독립적으로 500번 반복한 자료에서 추정한 제1종의 오류는 근사적으로 정규분포 $N(0.05, \frac{0.05\times0.95}{500})$를 따른다. 따라서 각각의 모형에서 추정한 제1종의 오류가 (0.031, 0.069) 범위 내에 있으면 제1종의 오류가 5%로 잘 유지된다고 할 수 있다.

한편 제1종의 오류가 잘 유지되는 모형내에서 어떤 모형이 더 좋은지는 노출정도와 사망의 관계가 있는 경우에 발생한 시뮬레이션 자료에서 동일한 방법으로 추정한 모형의 검정력이 가장 높은 방법이 가장 좋은 검정방법으로 판단하였다. 전체적인 시뮬레이션 과정은 Fig. 2와 같고, MATLAB을 이용하여 프로그래밍 하였다.

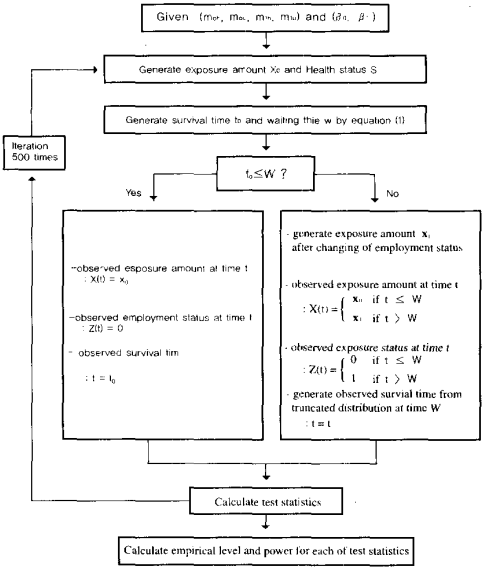

Fig. 2. Procedure of simulation in the case of no censoring.

연구 성적

1. 제1종의 오류에 대한 결과

건강근로자생존효과와 고용효과가 모두 없는 경우($m_{0h}=m_{1h}=m_{0u}=m_{1u}$), 모형 2를 제외한 나머지 모형들은 제1종의 오류가 5%로 유지되었으나 모형 2는 제1종의 오류에 심각한 문제가 있다. 이러한 이유는 η값을 0.5로 하여 노출이 많을수록 고용상태가 빨리 변하기 때문에 고용상태가 변한 집단과 그렇지 않는 집단으로 구분하여 분석하면 length bias sampling의 문제가 여전히 존재하기 때문이다.

건강근로자고용효과는 존재하나 생존효과가 없는 경우($m_{0h}=m_{1h}\neq m_{0u}=m_{1u}$), 이 연구에서 사용한 4가지 모형 모두 건강상태를 민감도와 특이도가 각각 80%인 대리변수로 분석하면 제1종의 오류가 매우 커짐을 알 수 있다. 그러나 측정오차가 없는 참값의 건강상태를 모형에 포함하면 모형 2를 제외한 나머지 방법들은 제1종의 오류가 5%로 잘 유지됨을 알 수 있다. 또한 중도절단된 생존자료가 있는 경우에도 비슷한 결과를 보였다.

건강근로자생존효과는 존재하나 고용효과가 없는 경우($m_{0h}=m_{0u}\neq m_{1h}=m_{1u}$), 모형 1과 모형 2는 제1종의 오류가 매우 커지나, 모형 3과 모형 4는 제1종의 오류가 5%로 유지됨을 알 수 있다.

또한 건강근로자생존효과와 고용효과가 모두 있는 경우 ($m_{0h}\neq m_{0u}$, $m_{1h}\neq m_{1u}$, $m_{0h}\neq m_{1h}$, $m_{0u}\neq m_{1u}$), 모형 3과 모형 4는 참값의 건강상태를 분석에 이용하면 제1종의 오류가 5%로 유지하나 모형 1과 모형 2는 제1종의 오

Table 1. Empirical 5% level of tests (number of replication is 500)

Occurrence [&]		Parameter [†]	% of	% of	Model 1 [†]		Model 2		Model 3		Model 4	
HHE	HSE	(m_{0h}, m_{0u}, m_{1h}, m_{1u})	response	censoring	Surrogate	True [¶]	Surrogate	True	Surrogate	True	Surrogate	True
No	No	(2.0, 2.0, 2.0, 2.0)	59.1	0	.042	.044	.986	.992	.054	.060	.040	.064
Yes	No	(2.0, 1.5, 2.0, 1.5)	52.0	0	.122	.046	1.000	.994	.188	.052	.256	.052
No	Yes	(2.0, 2.0, 1.5, 1.5)	59.1	0	.854	.872	1.000	1.00	.056	.060	.050	.064
Yes	Yes	(2.0, 1.5, 1.5, 1.0)	52.0	0	.998	.962	1.000	1.00	.210	.056	.276	.050
No	No	(2.0, 2.0, 2.0, 2.0)	32.9	58.3	.060	.058	.712	.760	.048	.054	.052	.058
Yes	No	(2.0, 1.5, 2.0, 1.5)	29.9	53.9	.120	.056	.902	.748	.156	.054	.162	.060
No	Yes	(2.0, 2.0, 1.5, 1.5)	32.9	54.9	.510	.534	.972	.984	.050	.054	.046	.054
Yes	Yes	(2.0, 1.5, 1.5, 1.0)	29.9	50.5	.884	.732	.998	.998	.174	.048	.158	.060

[†] Model 1 : Cox' s model which includes health status and time dependent exposure variable as covariates
Model 2 : Model 1 + time fixed employment status
Model 3 : Model 1 + time dependent employment status
Model 4 : proposed model which includes health status and time dependent exposure variable as covariates

[†] m_{0h}(m_{0u}) : mean baseline survival time of healthy (unhealthy) person without experiencing a change of employment status
m_{1h}(m_{1u}) : mean baseline survival time of healthy (unhealthy) person with experiencing a change of employment status

[&] HHE : healthy worker hire effect ; HSE : healthy worker survival effect

[¶] Proxy (true) variable of health status is used in the model

Table 2. Power[¶] of Cox's time dependent covariate model and proposed model by varying (β_0, β_1)

Parameter (β_0, β_1)	No censored			Censored			
	% of response	Time dependent covariate model [†]	Proposed model [†]	% of response	% of censoring	Time dependent covariate model [†]	Proposed model [†]
(0.4, 0.0)	44.8	.114	.158	26.8	47.6	0.108	0.106
(0.8, 0.0)	37.7	.336	.414	23.4	44.4	0.246	0.292
(1.2, 0.0)	31.0	.616	.760	20.1	40.9	0.550	0.592
(1.5, 0.0)	26.5	.828	.928	17.6	38.3	0.746	0.834
(0.0, 0.4)	52.2	.212	.202	30.1	49.8	0.132	0.120
(0.0, 0.8)	52.2	.596	.654	30.1	49.2	0.322	0.338
(0.0, 1.2)	52.2	.922	.960	30.1	48.6	0.566	0.662
(0.0. 1.5)	52.2	.988	.998	30.1	48.2	0.750	0.846
(0.4, 0.4)	44.8	.392	.282	26.8	46.9	0.236	0.190
(0.8, 0.8)	37.7	.840	.736	23.4	43.3	0.632	0.492
(1.2, 1.2)	31.0	.986	.952	20.1	39.6	0.914	0.820

[¶] : Mean baseline survival time (m_{0h}, m_{0u}, m_{1h}, m_{1u}) is (2.0, 1.5, 1.5, 1.0)
[†] : include health status, time dependent employment status and time dependent exposure variable as covariates
[†] : include health status and time dependent exposure variable as covariates

류가 매우 커짐을 알 수 있다. 이러한 결과는 중도절단된 생존자료가 있는 경우에도 비슷하였다 (Table 1).

2. 검정력에 대한 결과

모형 1과 모형 2는 건강근로자생존효과를 전혀 통제하지 못하므로 모형 3과 모형 4의 검정력만을 비교하였다(Table 2). 또한 초기시점의 건강상태를 대리변수로 분석하면 건강근로자고용효과에 대한 잔여 혼란효과로 인해 제1종의 오류가 조절되지 않는다. 따라서 검정력 분석에서는 참값의 건강상태를 모형에 포함하고, 그리고 건강근로자고용효과와 생존효과가 모두 있는 경우, 모형 3과 모형 4를 비교하였다.

고용상태가 변하지 않는다고 가정하였을 때 노출정도와 사망위험의 관련성 β_0와 고용상태가 변한 후의 노출정도와 사망위험의 관련성 β_1의 크기를 변화하면서 두 모형의 검정력을 비교하였다. 전반적으로 두 방법의 검정력은 비슷하였으나 $\beta_1=0$인 경우, β_0가 커질수록 (또한 $\beta_0=0$인 경우, β_1이 커질수록) 이 연구에서 제안한

방법의 검정력이 시간의존형 콕스모형 보다 더 높았다. 그러나 β_0와 β_1이 같이 변하는 경우에는 시간의존형 콕스모형의 검정력이 본 연구에서 제안한 모형의 검정력보다 약간 더 높았다. 이와같은 경향은 중도절단된 자료가 있는 경우에도 동일하였다.

고찰 및 결론

건강근로자효과는 특정 위험작업에 노출되는 근로자들의 사망률(또는 질병발생률)이 일반인구집단의 사망률보다 낮은 경우의 현상으로 정의하고 있으며[10], 그 원인을 설명하고자 현재까지 많은 연구들이 시도되었다. Choi[11]는 이 부분에 권위가 있는 9명의 전문가로부터 건강근로자효과에 대한 정의와 그 원인 등을 문의하였다. 건강근로자효과를 어떻게 정의할 것인가에 대해 이들 전문가들의 의견은 거의 일치한 반면, 그 원인에 대한 의견으로 10가지의 가능한 선택 치우침(selection bias), 2가지의 가능한 정보 치우침(information bias), 그리고 2가지의 가능한 혼란변수로 인한 치우침(confounding bias) 등으로 상당히 다양하였다. 그러나 이들 대부분의 전문가들은 충분치 못한 추적관찰, 건강한 근로자의 선택, 그리고 건강한 근로자가 계속적으로 그 직업에 남아있는 세 가지를 건강근로자효과에 대한 원인으로 생각하였다[10]. 여기서 충분치 못한 추적관찰은 건강한 근로자가 계속적으로 그 직업에 남아 있고 또한 질병에 새롭게 이환된 사람들이 그 직업을 그만두는 경향으로 설명할 수 있으므로 결국 나머지 두 가지 원인으로 설명할 수 있다.

Gilbert[10]는 1944년 부터 1978년까지 Hanford 핵발전소에서 근무하는 백인 남성들의 코호트 자료를 이용하여 건강근로자효과를 6가지 가능한 혼란변수로 야기되는 현상으로 설명하고자 하였으며 이 중 본 연구에서 중요하게 간주하는 고용상태의 변화가 있는 근로자의 사망률이 높았음을 보고하였다. Gamble 등[12]은 유기용제에 노출되었던 퇴직근로자의 사망률이 높았고 특히 65세 이전에 퇴직한 근로자의 사망률이 높음을 보고하였다. 또한 기존의 많은 연구에서도 고용상태가 변한 근로자의 사망률이 높음을 보고하였다[13, 14]. 한편 Steenland[3]는 근로자의 인년을 활동중인 기간(active period)과 그렇지 않은 비활동중인 기간(inactive period)으로 나누어 사망률을 추정하였으며 활동중인 기간에서의 사망률은 일반인구집단보다 낮았으나 비활동중인 기간의 사망률은 높음을 보고하여 고용상태의 변화를 건강근로자효과에 대한 주요한 매개변수로 강조하였다.

이상의 연구결과들은 이 연구에서 제안한 모형의 타당성과 관련성이 있다. 즉, 고용상태의 변화가 생존시간을 변화시키므로 단순히 고용상태가 변하지 않은 집단과 변한 집단을 구분하여 노출정도와 사망의 관계를 분석하면 length bias sampling으로 인한 치우침이 발생하므로(9), 본 연구는 고용상태가 변하지 않는다고 가정하였을 때의 개념적 생존시간과 고용상태가 변한 후의 개념적 생존시간을 따로 정의하여 노출정도와의 관계를 보고자 하였다. 특히 고용상태의 변화 여부를 고용상태가 변하지 않는다고 가정하였을 때의 개념적 생존시간과 고용상태가 변하기까지 대기시간의 경쟁관계로 발생하는 현상으로 모형화하고 length bias sampling으로 인한 치우침을 제거하는 새로운 방법을 제안하였다는 데 그 의의가 있다고 할 수 있다.

본 연구의 시뮬레이션 결과 초기시점의 건강상태를 시간비의존형 공변수로, 그리고 t 시점에서의 고용상태를 시간의존형 공변수로 한 콕스모형은 노출정도와 생존시간의 관계에 대한 제1종의 오류에 큰 문제가 없었다. 또한 이 연구에서 새롭게 제안한 방법도 제1종의 오류에 문제가 없었다. 그러나 두 방법에서 검정력을 비교해 보면, 고용상태가 변한 후의 노출정도와 사망위험간에 관련성이 없을 때(즉, $\beta_1=0$), β_0의 변화에 따라 이 연구에서 새롭게 제안한 방법의 검정력이 시간의존형 콕스모형보다 더 높았다. 이러한 이유는 두 모형에서 정의하고 있는 위험집단의 정의에 차이가 있고 특히 본 연구에서 제안하고 있는 모형의 위험집단이 length bias sampling으로 인한 치우침을 효과적으로 고려하기 때문으로 생각된다. 또한 고용상태가 변하기 전의 노출정도와 사망위험간에 관련성이 없는 경우에도(즉, $\beta_0=0$), β_1의 변화에 따라 이 연구에서 제안한 방법의 검정력이 높았다. 그러나 β_0와 β_1이 모두 0이 아닐 때는 시간의존형 콕스모형이 이 연구에서 제안한 방법보다 검정력이 높았다. 이러한 이유는 두 모수가 동시에 0이 아니므로 고용상태가 변한 집단과 변하지 않은 집단의 관찰된 생존시간의 차이가 적어지므로 length bias sampling으로 인한 효과가 작아진다. 따라서 본 연구에서 제안한 방법은 스코아함수에서 전체표본을 고용상태 변화여부에

따라 나누어 분석하므로 표본수의 상대적인 감소로 이러한 검정력의 감소가 발생하였을 것으로 생각되며 이 부분에 대한 추후 연구가 필요할 것으로 생각된다.

한편 방법론적인 측면이 아니라 노출정도와 사망의 관계에 대한 실제 역학연구에서는 고용상태가 변하지 않는다고 가정하였을 때 생존시간과 노출정도의 관계를 분석하는 것이 의미가 있을수 있으며, Robins[6, 7]의 통계적인 모형도 이러한 관점에서 모형화되었다. 따라서 $H_0:\beta_0=0$ ($\beta_1=0$)에 대한 대립가설 $H_1:\beta_0\neq 0$ ($\beta_1=0$)의 검정이 일반적일 것으로 생각되며 이러한 경우에 더 높은 검정력을 보인 이 연구에서 제안한 방법이 기존의 콕스모형보다 건강근로자 생존효과를 제거하는데 효과적일 것으로 생각된다.

한편 시간의존형 변수인 고용상태를 통제하지 않는 경우 건강근로자생존효과로 인해 제1종의 오류는 매우 크게 증가하고, 특히 모형 2에서와 같이 고용상태를 시간비의존형 변수로 통제하면 이 문제는 더욱 심각하게 된다. 또한 건강근로자고용효과를 야기하는 초기시점의 건강상태가 측정오차를 동반하고 만약 측정오차가 커지면 제1종의 오류에 심각한 문제가 발생한다는 것을 이 연구의 시뮬레이션 결과는 보여주고 있다.

건강근로자생존효과를 제거하기 위하여 그 원인이 되는 고용상태를 어떻게 분석하느냐에 따라 이제까지 사용된 방법론을 크게 두가지 형태로 나눌수가 있다. 첫 번째는 고용상태를 혼란변수로 간주하여 건강근로자생존효과를 제거하는 방법이다. 이 방법으로는 고용된 시점부터의 생존시간이 충분히 긴 근로자만을 대상으로 고용상태를 층화하여 분석하는 방법[15], 노출변수에 시차(lag)를 두어 분석하는 방법[10], 현재의 고용상태를 혼란변수로 간주하고 회귀모형에 공변수로 추가하여 통제하는 방법[16] 등이 있다. 이상의 방법들은 이론적인 모형의 검토를 통해 제안된 것이 아니라 경험적 자료의 증거로부터 제안된 관념적인 방법들로 생각할 수 있다. 그러나 고용상태를 시간비의존형 변수로 분석하거나(모형 2), 층화하여 분석하는 경우 노출정도와 사망의 관계에 대한 제1종의 오류에 심각한 문제가 발생한다는 본 연구의 시뮬레이션 결과를 볼 때 위의 방법들은 정확한 방법론에 근거를 둔 이론적인 검정과정을 거칠 필요가 있다고 생각된다.

두 번째는 고용상태를 혼란변수와 시간의존형 매개변수의 역할을 동시에 하는 변수로 간주하고 건강근로자생존효과를 G-null test, G-estimation 방법, 그리고 이를 응용한 SNFTM(structural nested failure time model) 모형을 이용하여 제거하는 방법이 있다[5-8]. 현재까지 Robins의 모형이 건강근로자생존효과를 이론적으로 가장 잘 설명하는 것으로 생각되고 있다. 그러나 이 방법은 자료의 형태에 따라 그 적용방법이 통일되지 못하고, 계산과정이 복잡하다는 단점이 있으며 또한 시뮬레이션을 통해 여러 가지 건강근로자효과가 발생할 수 있는 상황에서의 모형에 대한 타당성이 검토된 적이 없었다. 향후 본 연구에서 새롭게 제안한 방법과 Robins의 방법을 비교하는 연구가 필요하다고 할 수 있다.

지난 2000년 일년 동안 Journal of Occupational and Environmental Medicine(JOEM)을 검토한 결과, 코호트 연구를 이용하여 위험인자의 노출정도와 사망(또는 질병발생)의 관련성을 살펴본 연구는 총 8편이었다. 이 8편의 논문 중 위험인자 노출정도에 따른 사망발생 정도를 비교하기 위한 방법으로 5편은 표준화 사망비 또는 표준화 발생비를 구하였고, 1편은 로지스틱 회귀분석을 이용하였으며 나머지 2편은 콕스 비례위험모형을 이용하였다. 표준화 사망비 및 표준화 발생비나 회귀분석을 이용한 논문 5편 중 3편에서 연구의 한계점으로 건강근로자효과를 지적하는 정도였고, 콕스의 비례위험모형을 이용한 2편의 논문에서도 고용상태 변화를 모형에 고려하지 않았다[12, 17-23]. 즉 2000년 일년 동안 코호트 연구로 노출정도와 질병발생 및 사망과의 관계를 연구한 논문들의 대부분이 건강근로자생존효과를 야기하는 고용상태의 변화여부를 고려하지 않았음을 볼 때 이 부분에 대한 내용이 계속적으로 강조될 필요성이 있다고 생각되며 이러한 의미에서 이 연구가 가지는 방법론적인 의의는 매우 크다고 할 수 있다. 아울러 근로자가 직업에 고용될 초기시점의 건강상태를 정확하게 평가하는 것도 매우 중요한 문제이며 이에 대한 연구도 계속적으로 이루어져야 할 것으로 생각된다.

참고문헌

1. Arrighi HM, Hertz-Picciotto I. The evolving concept of the healthy worker survival effect. *Epidemiology* 1994; 5(2): 189-96.
2. Arrighi HM, Hertz-Picciotto I. Controlling the healthy worker survivor effect: an example of

arsenic exposure and respiratory cancer. *Occup Environ Med* 1996; 53: 455-62.

3. Steenland K, Stayner L. The importance of employment status in occupational cohort mortality studies. *Epidemiology* 1991; 2(6): 418-23.
4. Checkoway H, Pearce N, Crawford-Brown DJ. Research methods in occupational epidemiology : Monographs in epidemiology and biostatistics. New York: Oxford University Press, 1989.
5. Robins J. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chron Dis* 1987; 40: 139S-161S.
6. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period-Application to control of the healthy worker survival effect. *Mathematical Modeling* 1986; 7: 1393-512.
7. Robins JM, Blevins D, Ritter G, Wulfsohn M. G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the of AIDS patients. *Epidemiology* 1992; 3: 319-36.
8. Robins JM. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* 1992; 79: 321-34.
9. Nam CM, Zelen M. Comparing the survival of two groups with an intermediate clinical event. *Lifetime Data Analysis* 2001; 7: 5-19.
10. Gilbert ES. Some confounding factors in the study of mortality and occupational exposures. *Am J Epidemiol* 1982; 116(1): 177-88.
11. Choi BCK. Definition, sources, magnitude, effect modifiers, and strategies of reduction of the healthy worker effect. *J Occup Med* 1992; 34(10): 979-88.
12. Gamble JF, Lewis RJ, Jorgensen G. Mortality among three refinery/petrochemical plant cohorts. II. retirees. *JOEM* 2000; 42(7): 730-6.
13. Delzell E, Monson RR. Mortality among rubber worker. IV. General mortality patterns. *J Occup Med* 1981; 23: 850-6.
14. Vinni K, Hakama M. Healthy worker effect in the total Finnish population. *Br J Ind Med* 1980; 37: 180-4.
15. Fox AJ, Collier PF. Low mortality rates in industrial cohort studies due to selection for work and survival in the industry. *Br J Prev Soc Med* 1976; 30: 225-30.
16. Gilbert E, Marks S. An analysis of the mortality of workers in a nuclear facility. *Radiat Res* 1979; 79: 122-48.
17. Chan CK, Leung CC, Tam CM, et al. Lung cancer mortality among a cohort of men in a silicotic register. *JOEM* 2000; 42(1): 69-75.
18. Cocco P, Rice CH, Chen JQ, et al. Non-malignant respiratory disease and lung cancer among chinese workers exposed to silica. *JOEM* 2000; 42(6): 639-44.
19. Dalager NA, Kang HK, Mahan CM. Cancer mortality among the highest exposed US atmospheric nuclear test participants. *JOEM* 2000; 42(9): 798-805.
20. Danielsen TE, Langard S, Andersen A. Incidence of cancer among welders and other shipyard workers with information on previous work history. *JOEM* 2000; 42(1): 101-9.
21. Lewis RJ, Gamble JF, Jorgensen G. Mortality among three refinery/petrochemical plant cohorts. I. 1970 to 1982 active/terminated workers. *JOEM* 2000; 42(7): 721-9.
22. Nakanishi N, Okamoto M, Nakamura K, et al. Cigarette smoking and risk for hearing impairment: A longitudinal study in Japanese male office workers. *JOEM* 2000; 42(11): 1045-9.
23. Sathiakumar N, Dezell E. An updated mortality study of workers at a dye and resin manufacturing plant. *JOEM* 2000; 42(7): 762-71.

=Abstract=

A study on Statistical Method for Controlling the Effect of Intermediate Events - Application to the Control of the Healthy Worker Effect

Chung Mo Nam[1], Jinheum Kim[2], Dae Ryong Kang[3],
Yeon-Soon Ahn[4], Hoo-Yeon Lee[1], Dae Hee Lee[1]

Department of Preventive Medicine and Public Health, Yonsei University[1],
Department of Applied Statistics, University of Suwon[2],
Graduate School of Health Science and Management, Yonsei University[3],
Occupational Safety and Health Research Institute, Korea Occupational Safety and Health Corporation[4]

Purpose : The healthy worker effect is an important issue in occupational epidemiology. This study was conducted to propose a new method to test the relation between exposure and mortality in the presence of the healthy worker effect.

Methods : In this study, the healthy worker hire effect was assumed to operate as a confounding variable of health status at the beginning of employment and healthy worker survival effect as a confounding and intermediate variable of employment status. In addition, the proposed method reflects the length bias sampling caused by changing of an employment status. Simulation studies were also carried out to compare the proposed method with Cox' s time dependent covariates models .

Results : The theoretical development of the healthy worker survival effect is based on the result that an observation with change of an employment status requires that the survival time without intermediate event exceeds the waiting time for the intermediate event. According to our simulation studies, both the proposed method and Cox' s time dependent covariates model which includes the change of employment status as time dependent covariates seem to be satisfactory at 5% significance level. However, Cox' s time dependent covariates models without or with the change of employment status as time fixed covariate are unsatisfactory. The proposed test is superior in power to tests based on Cox' s model.

Conclusions : The healthy worker effect may not be controlled by classical Cox' s proportional hazards models. The proposed method performed well in the presence of healthy worker effect in terms of level and power.

Key Words: Healthy worker effect, Intermediate, Length bias sampling, Simulation, Level

