The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.36.6.777
33(6), 777–789

Statistical analysis of estimating incubation period distribution and case fatality rate of COVID−19

Han Jeong Ki[a] · Jieun Kim[a] · Sohee Kim[a] · Juwon Park[a] · Joohaeng Lee[a] · Yang-Jin Kim[a,1]

[a]Department of statistics, Sookmyung Women's University

(Received August 24, 2020; Revised September 25, 2020; Accepted October 6, 2020)

Abstract

COVID-19 has been rapidly spread world wide since late December 2019. In this paper, our interest is to estimate distribution of incubation time defined as period between infection of virus and the onset. Due to the limit of accessibility and asymptomatic feature of COVID-19 virus, the exact infection and onset time are not always observable. For estimation of incubation time, interval censoring technique is implemented. Furthermore, a competing risk model is applied to estimate the case fatality and cure fraction. Based on the result, the mean incubation time is about 5.4 days and the fatality rate is higher for older and male patient and the cure rate is higher at younger,female and asymptomatic patient.

Keywords: case fatality rate, cure rate, incubation time, interval censoring, pandemic

1. 서론

2019년 12월 중국 우한에서 처음 발생한 COVID-19는 이후 중국 전역과 전 세계로 확산되었다. 중국 정부는 우한 의료진 15명이 확진 판정을 받았다고 발표하였으며, COVID-19의 사람 간 감염 가능성이 제기되었다. 초기에는 원인을 알 수 없는 호흡기 전염병으로만 알려졌으나, 세계보건기구(WHO)가 2020년 1월 9일 해당 폐렴의 원인이 새로운 유형의 COVID-19 (SARS-CoV-2, 국제 바이러스 분류 위원회 2월 11일 명명)라고 밝히면서 그 병원체가 확인됐다. 이후 감염 확산세가 이어지자, WHO는 1월 30일 '국제적 공중보건 비상사태 (Public Health Emergency of International Concern; PHEIC)'로 3월 11일에는 대유행병(PANDEMIC)으로 선언하였다. 우리나라에서는 1월 19일 중국을 방문하고 돌아온 중국인 여성이 최초의 감염건으로 신고된 후, 2월 대구의 종교 집단과 경북 지역의 요양 시설의 대규모의 감염과 사망 발생으로 그 위험률이 가시화되었다. 그 이후 감염 추세는 전국으로 확산되었으며 정부 당국과 국민들의 방역에 대한 경각심을 불러오게 하였으며 무증상 감염으로 인한 방역의 어려움과 감염 가능성에 대한 심리적 스트레스는 매우 높아지고 있다. 현재 감염과 치사율에 대한 활발한 연구가 진행되고 있으며 본 연구에서는 생존 분석의 다양한 기법을 적용함으로써 점염병의 특성을 이해하고자 한다. 본 논문에서 분석한 자료는 2020년 6월 31일 기준 질병관리 본부(KCDC)에서 발표한 감염 환자자료로 시각화 기법과 통계분석 기법을 적용하여 국민 보건과 관련된 특성을 규명하고자 한다 (https://www.kaggle.com/kimjihoo/coronavirus

This work was supported by Korea research grant (NRF-2017R1D1A1B03030578).

[1]Corresponding author: Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul 01910, Korea. E-mail: yjin@sookmyung.ac.kr

dataset). 특히 본 연구의 주요 관심은 바이러스의 잠복기간(incubation time)분포와 치사율 그리고 치유율을 추정하고자 한다.

첫 번째 연구 주제는 확진자의 감염 시점을 역추적하여 감염원을 알 수 있는 감염 집단을 대상으로 잠복시간의 분포를 추정한다. 바이러스 감염은 연속적인 확산을 가져오며 여기서 바이러스 전이 구간(transmission interval)은 감염 전파자(감염원, infector)와 그로 인한 감염자(infectee)의 감염 시점 간격으로 정의된다. 이 구간을 통해 그 질병이 얼마나 빨리 확산되는지를 알 수 있다 (Fine, 2003). Mettler 등 (2020)는 정확한 감염 시점을 구하는 것은 어렵기 때문에 감염원과 그로 인한 감염자들의 증상 발현 시점간의 차이를 이용할 것을 제안하며 이를 clinical onset serial interval 이라고 명하였다. 하지만 무증상 양성 환자에 대해서는 이 구간값도 직접적으로 구할 수 없게 된다. 이에 그들은 또 다른 구간인 감염원과 그로 인한 감염자의 확진 시점을 이용한 diagnostic serial interval을 제안하였다. 하지만 이 구간 역시 무증상 환자의 배제등으로 표본 편이(sample selection)를 가져올 수 있게 된다. 본 연구에서는 추적 조사를 통해 감염 발생 시점이 가능하거나 구간내로 추측할 수 있는 두 개의 코호트를 대상으로 잠복 기간의 분포를 구하고자 한다. 여기서 잠복기간은 바이러스 감염(infection)에서 증상 발현(symptom onset)까지 걸린 시간으로 정의되는데 본 연구에서는 이들 사건들의 정확한 발생 시점이 알려져 있지 않은 경우를 고려한다. 첫 번째 감염 집단은 천안 줌바 댄스 감염자료로 감염원인 댄스 강사로부터 1차 감염된 사람들 중 30명 표본 자료이며 두 번째 감염 집단은 성남 교회 감염 자료이다. 이들 자료를 이용하여 잠복 기간을 추정하기 위해선 감염자들의 감염 시점과 증상 발현 시점이 측정되어야한다. 하지만 천안 줌바 댄스 그룹에서는 정확한 감염시점을 제공하지는 않았으며 무증상 감염이 빈번하게 발생하였다 (86%). 예를 들어, 천안 그룹에서는 강사가 워크샵에서 감염된 후 종교 단체와 강습소를 통해 바이러스 전파가 이루어졌다 (Bae 등, 2020). 또한 감염자(infectee) 중 10명은 증상을 경험하였으나 나머지 감염자들은 아무런 증상을 느끼지 못하였다. 반면에 성남 종교집단 감염에 대해서는 일요일 예배를 통해 감염된 것으로 추정되기 때문에 감염자들의 정확한 감염 시점은 알 수 있었으나 감염 이후 양성 확진을 받기 전 대부분이 무증상 상태였으며 한 명만이 유증상자였다.

이러한 불완전한 관측 자료를 이용하여 잠복기간 분포를 추정하기 위해 구간 중도 절단(interval censored data) 분석 기법을 적용하고자 한다. 구간 중도 절단 자료는 생존 분석(survival analysis)자료에서 흔히 발생되는 우중도 절단(right censored data)과는 달리 사건 발생 시간을 포함하는 두 개의 시점으로 구성된다. 대표적인 예로 HIV 감염 시점, AIDS 발현 시점등은 의료기관에서 검사를 통해 그 양성을 확인할 수 있기 때문에 정확한 감염 시점과 발현 시점 대신 마지막 음성 반응 시점과 첫 번째 양성 반응 시점으로 구성된 구간 중도 자료로 표현된다.

두 번째 연구에서는 치사율과 완치율 추정을 위해 경쟁 위험 모형의 적용을 고려해본다. 확진자들은 병원 격리 상태에서 사망 또는 완치 중 하나의 사건을 종말 사건(terminal event)으로 경험하게 된다. 따라서 두 가지 이상의 종말 사건이 발생되는 경우 일반적인 생존 분석이 아닌 경쟁위험모형(competing risk model)이 적용되어야 한다. 본 분석에서 사용하는 확진자 자료에 대해서는 확진 후 사망 시점 또는 완치 시점까지 경과된 시간을 이용한다. 여기서 여전히 격리된 양성 확진자는 위험그룹에 속하게 된다. COVID-19 확진자의 치사율과 완치율을 연령대별로 추정하고 연령과 증상유무와 연관성을 밝히고자 한다.

본 논문은 다음과 같이 구성된다. 2장에서는 COVID-19 자료에 대한 탐색적 자료 분석을 실시하며 3장에서는 본 논문에 적용된 통계적 방법론을 정리한다. 4장과 5장에서 불완전하게 측정된 감염 시점과 증상 발현 시점을 이용하여 잠복 기간 분포의 추정과 치사율을 분석한다. 6장에서는 관련된 향후 연구를 제시함으로써 논문을 마무리하고자 한다.

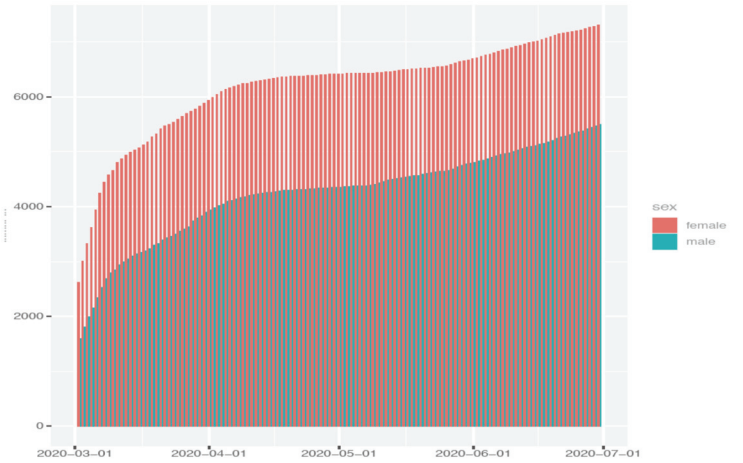

Figure 2.1. Cumulative infection numbers by gender during 2020/03/01–2020/06/01.

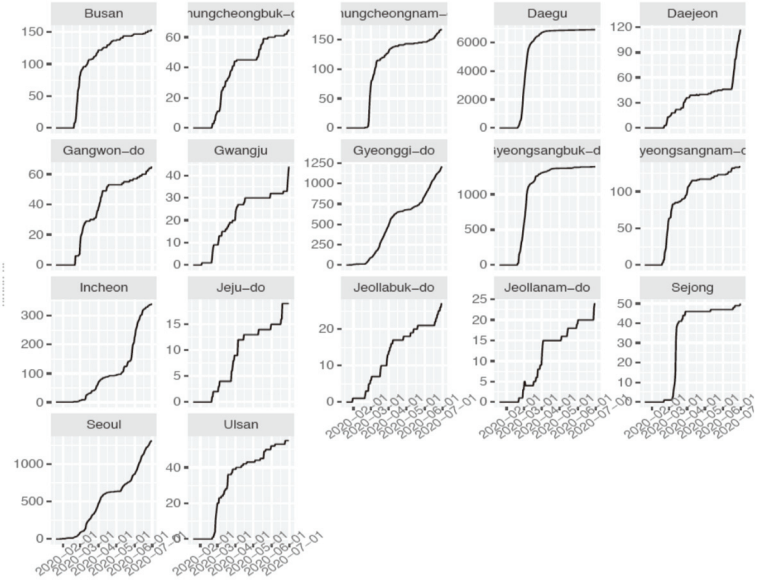

Figure 2.2. Cumulative infection number by province during 2020/03/01–2020/06/01.

2. 자료 설명

본 장에서는 본격적인 통계 분석을 시작하기 전에 2020년 6월 30일까지 수집된 자료를 이용하여 시각화 방법을 통해 우리나라의 감염 현황을 파악하고자 한다. Figure 2.1은 2020년 3월 1일부터 2020년 6월 30일 약 4개월간 확진자 추세를 남녀별로 구분하여 보여준다. 여성 확진자수가 남성 확진자수보다 훨씬 많으며 감염자의 증가 추세는 일정하였다. Figure 2.2는 전국 주요 시도별 누적 감염자수를 보여준다. 대구, 경상북도, 세종시는 감염 확진 추세가 꺽이고 있는데 반해 나머지 지역에서는 계속해서 감염자수가 발생하고 있

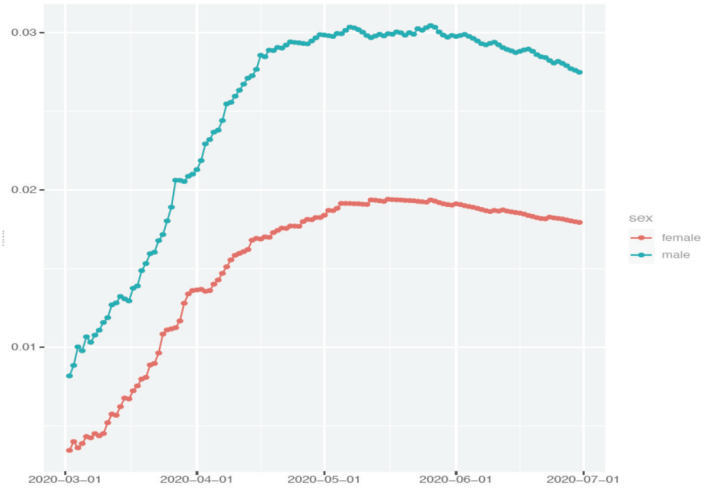

Figure 2.3. Death rate by gender during 2020/03/01–2020/06/30.

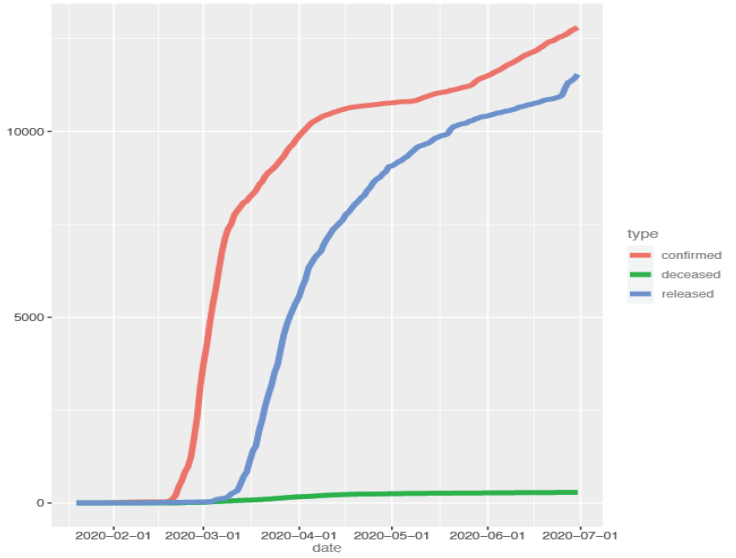

Figure 2.4. Cumulative numbers of infection, release and death during 2020/03/01–2020/06/01.

음을 알 수 있다. 특히, 대전, 경기도, 서울, 인천 지역의 확진자수는 이전 확진추세보다 더 급격히 증가함을 알 수 있다. Figure 2.3에서는 확진자 중 사망률을 보여주는데 계속 증가하다가 2020년 5월 중반부터 차츰 감소함을 알 수 있다. 여기서 남성의 사망률은 여성 사망률보다 모든 시간대에서 더 높았으며 그 격차는 시간이 경과함에 따라 더욱 커짐을 알 수 있다.

Figure 2.4에서는 4개월간 확진자수, 이들 중 완치자수와 사망자수를 보여준다. 전체 확진자수의 증가 추세는 4월 중반부터 주춤하다가 5월 중반에 다시 증가함을 알 수 있다. 이는 5월 연휴 기간의 영향으로 추측해본다.

3. 통계적 방법론

본 논문에서 사용된 두 가지 생존 분석 방법론에 대해 간단히 설명한다.

3.1. 구간 중도 절단 자료의 통계적 분석

구간 중도 절단 자료(interval censored data)는 정확한 사건 발생 시점을 알 수 없고 대신 이를 포함한 두 시점만이 제공되는 자료형태이다. 그 예로, HIV 의심 환자를 대상으로 양성 여부를 조사할 때, 환자의 바이러스 발현 시점을 정확하게 인지할 수 없기 때문에 혈액 채취를 통한 검사 결과를 이용한다. 따라서 마지막으로 음성 진단을 받은 검사 시점과 처음으로 양성 진단을 받은 검사 시점, 두 시점 사이의 어느 시점에서 발생되었음을 알 수 있다. 이러한 종류의 자료 유형은 의약학자료에서 흔히 발생된다. 하지만 많은 분석에서는 처음으로 양성진단을 받은 검사 시점 또는 두 구간 시점의 중간 시점을 사건 발생 시점으로 대체(imputation)하는데 이는 심각한 결과를 가져올 수 있다. 즉, 환자마다 구간 길이가 다를 수 있으며 환자의 상태에 따라 관측 시점이 영향을 받는 경우에 편의된 추정량을 유도하게 된다 (Sun, 2006). 본 연구에서는 감염시점이 정확히 알려져 있지 않은 경우 바이러스의 잠복 기간 분포를 추정하기 위해 구간 중도 절단 자료의 분석 방법을 적용한다.

(XL, XR)은 구간 중도 절단 시점으로 관심 있는 사건의 발생 시점 T에 대해 $XL < T < XR$의 관계를 가지게 된다. 구간 중도 절단 자료의 비모수 최대 우도 추정량(nonparametric MLE)을 구하기 위해 다음의 우도함수가 적용된다.

$$L = \prod_{i=1}^{n} [S(XL_i) - S(XR_i)]$$

여기서 $S(t) = P(T \geq t)$는 생존 함수를 의미한다. 위의 우도 함수는 겹치지 않은 시간 구간대 $\{(q_1, p_1], (q_2, p_2], \ldots, (q_m, p_m]\}$를 이용하여 다음과 같이 재표현된다.

$$L = \prod_{i=1}^{n} \sum_{j=1}^{m} a_{ij} f_j , \tag{3.1}$$

여기서 $a_{ij} = I((q_j, p_j] \subset (XL_i, XR_i))$로 정의되며 $\sum_{j=1}^{m} f_j = 1$을 만족하는 $f_j = E\big(T \in (q_j, p_j]\big),\ j = 1, \ldots, m$를 구하기 위해 Turnbull이 제시한 자기 일치(self-consistency)알고리즘이 적용된다 (Turnbull, 1974). 이 추정량은 R 패키지 `interval`에 속해 있는 `icfit`함수 또는 SAS의 `iclifetest` 프로시죠 (`PROC iclifetest`)를 이용하여 구할 수 있다. 하지만 이러한 비모수 추정량은 사건 발생 확률 또는 평균 계산에 있어 모수적 방법을 이용한 분포 추정량과 비교하여 한계를 가지게 된다. 따라서 본 분석에서는 평균 잠복 기간을 계산하기 위해 구간 중도 절단 자료에 대해 모수분포 $F(\theta)$를 적용하였다. 예를 들어, 와이블 분포는 두 가지 모수 $\theta = (\gamma, \alpha)$를 가지며 확률 밀도 함수와 생존 함수는 각각 다음과 같이 정의된다.

$$f(t; \theta) = \frac{\gamma}{\alpha}\left(\frac{t}{\alpha}\right)^{\gamma - 1} \exp\left(-\left(\frac{t}{\alpha}\right)^{\gamma}\right),\ \ t > 0$$

$$S(t; \theta) = \exp\left(-\left(\frac{t}{\alpha}\right)^{\gamma}\right),\ \ t > 0$$

여기서 평균, 분산 그리고 중위수는 각각 $E(T) = \alpha\Gamma(1 + 1/\gamma)$, $\mathrm{Var}(T) = \alpha^2\{\Gamma(1 + 2/\gamma) - \Gamma^2(1 + 1/\gamma)\}$,

$T_{med} = \alpha(\log 2)^{1/\gamma}$가 된다. 본 논문에서 와이블 분포외에도, 지수분포, 로그 정규 분포, 그리고 로그 로지스틱 분포를 적용한다.

3.2. 경쟁 위험 모형

경쟁 위험 모형(competing risk model)은 일반적인 생존 분석자료와 달리 두 개 이상의 종말 사건(terminal event)을 가지는 사건사 자료(event history data)로 이해할 수 있다. 예를 들어, 백혈병 환자에게 골수 이식 후 원인 A에 의한 생존시간이 관심이 있다고 하자. 이때 다른 원인 B에 의해 환자가 사망하는 경우, 그 사망은 원인 A와 완전히 독립임을 가정할 수 없다 (Kim, 2016).

본 연구에서 COVID 양성확진자는 치료를 받는 과정에서 완치 또는 사망이라는 두 가지 종말 사건(terminal event)을 가진다. 따라서 주요 관심이 사망 사건의 분포일 때, 완치 사건과 아직 치료를 받고 있는 확진자들은 중도 절단 사건으로 간주된다. 하지만 완치 사건은 중도 절단과 달리 환자의 건강 상태가 호전되어 더 이상 위험 그룹에 속하지 않는 위험탈퇴(risk free)그룹으로 인식될 수 있다. 반면에 연구 기간동안 중도 절단되어 위험그룹에서 빠진 경우는 그의 완치 또는 사망 여부를 알 수 없어 완전한 위험탈퇴 그룹으로 간주할 수 없다. 따라서 완치그룹과 일반적인 중도 절단 사건을 동일한 성질로 간주할 수 없으며 완치 사건을 또 다른 종말 사건으로 간주하여 분석하는 경쟁 위험 모형을 적용하고자 한다.

생존 분석에서 적용되는 가장 일반적인 자료 형태인 우중도 절단자료는 $T = \min(T^*, C), \delta = I(T^* \le C)$로 T^*는 사망 시간 또는 관심있는 사건 발생시간을 의미하며 C는 중도 절단 시간을 표시한다.

경쟁 위험 모형하에서는 m개의 원인을 가진 종말 사건이 있다고 가정한다. 우리의 경우에는 두 가지 종말 사건 (사망과 완치)을 가지게 된다. 따라서 분석 자료는 $(T = \min(T^*, C), J \times \delta)$으로 표시되며 여기서 T는 중도 절단을 포함하여 여러 가지 종말 사건 중 가장 먼저 관측되는 사건 발생 시간을 나타내며 $J\{= 1,2,\dots,m\}$는 사건의 원인을 표시한다. 만약 중도 절단될 경우, $\delta = 0$이 된다. 경쟁 위험 모형에서는 원인별 위험 함수(cause specific hazard function)를 다음과 같이 정의한다.

$$\alpha_j(t) = \lim_{h \to 0} \frac{\Pr(t < T \le t + h,\ J = j | T \ge t)}{h}.$$

위 위험 함수는 다른 모든 사건 원인이 가능할 때, 원인 j에 의한 사건이 발생될 위험률을 의미하며 공변량의 효과를 고려할 경우 $\alpha_j(t|x)$로 확장될 수 있다. 여기서 원인별 누적발생 함수(cause-specific cumulative incidence function; CIF)는

$$F_j(t) = P(T \le t, J = j) = \int_0^t \alpha_j(s) S(s-) ds = \int_0^t \alpha_j(s) \exp\left\{ - \int_0^s \sum_{k=1}^m \alpha_k(u) du \right\} ds$$

으로 t시점까지 원인 j 사건이 발생될 확률을 의미한다. 여기서 $\lim_{t \to \infty} F_j(t) = \Pr(J = j) = F_j$이며, 중도 절단이 없는 경우에는 $\sum_{j=1}^m F_j = 1$를 만족한다. 따라서 각 F_j를 부분포(subdistribution)라고도 한다. CIF는 다음과 같이 추정되며

$$\hat{F}_j(t) = \sum_{k: t_k \le t} \hat{\alpha}_j(t_k) \hat{S}(t_k)\,, \tag{3.2}$$

여기서

$$\hat{\alpha}_j(t_k) = \frac{d_{jk}}{n_k}, \quad \hat{S}(t_k) = \prod_{l=1}^{k} \left(1 - \frac{d_l}{n_l}\right)$$

로 추정된다. t_k는 모든 원인에 의한 사건 발생 시점을 합쳐서 크기 순서로 나열하였을 때 k번째 시점을 의미하며, d_{jk}는 t_k시점에서의 원인 j에 의한 사건 수, n_k는 t_k시점에서의 위험그룹 수, $d_l = \sum_{j=1}^{m} d_{jl}$은 t_l시점에서 발생된 총 사건수를 각각 의미한다. 만약, $\hat{S}_j(t)$이 원인 j에 의한 사건만을 실패로 간주하고 다른 원인에 의한 사건 발생을 중도 절단으로 간주하여 구한 Kaplan-Meier 추정량이라고 할 때, $\hat{F}_j(t)$와 $1 - \hat{S}_j(t)$는 다음의 관계를 가진다.

$$1 - \hat{S}_j(t) \geq \hat{F}_j(t) .$$

즉, 원인별 사건 발생을 중도 절단으로 간주할 경우 사건의 누적 발생률은 과대 추정될 수 있음을 알 수 있다. 발생확률에 대한 공변량의 효과를 추정하기 위해 다음의 원인 별 부분포 위험 함수(subdistribution hazard function) (Fine과 Gray, 1999)를 적용한다.

$$\lambda_j(t; x) = \lim_{h \to 0} \frac{1}{h} \Pr\{t \leq T < t + h, J = j | T \geq t \cup (T \leq t \cap J \neq j), x\} = \frac{dF_j(t, x)/dt}{1 - F_j(t, x)} . \tag{3.3}$$

위 함수의 가장 큰 특징은 위험 함수에서 경쟁 사건을 경험한 개체를 적절한 가중치를 이용하여 위험 그룹에 포함시키는 것이며 위에서 설명한 원인별 위험함수와 다른 의미를 가진다. 관련된 자료 분석을 위해 R 패키지의 cmprsk를 사용하였다.

4. 사례를 이용한 바이러스 잠복기간 분포 추정

잠복기간은 바이러스 감염 사건과 징후 발생(symptom onset)까지 경과된 시간을 의미한다. 하지만 이 두 사건의 정확한 발생 시점을 구하는 것은 매우 어렵다. 본 연구에서는 두 가지 감염그룹 자료를 이용하고자 하며 그 감염원이 정확하게 알려진 경우의 1차 감염만을 고려한다. 첫 번째 코호트는 천안 줌바 댄스강습 관련 감염으로 댄스 강사로부터 감염된 30명의 증상 발현 시간 또는 확진시점을 이용하고자 한다. 또 다른 코호트는 성남 교회집단 감염으로 여기서 한명의 감염원으로부터 일요일 예배 참석을 통해 감염된 49명의 자료를 이용하고자 한다.

천안 코호트의 감염원(infector)인 줌바댄스 강사는 2월 15일 워크샵에서 감염된 후 2월 20일에 증상이 발현되었으며 2월 26일에 양성 확진을 받게 된다. 여기서 확진을 받기 전 댄스 강습과 종교 시설 방문등의 활동을 통해 감염 전파가 이루어졌을 것이라 추측한다. 강사로부터 감염된 양성확진자 중 30명의 샘플을 이용한다. 하지만 그들의 정확한 감염 시점은 알 수 없으며 양성 확진 전 22명이 증상 발현을 보였다. 관련 증상 발현 시점과 양성 확진 판정 시점의 자료의 일부가 Table 4.1에 주어져 있으며 Figure 4.1는 감염원과 두 명의 감염자의 전이 경로 도식화를 보여준다. 본 분석에서 사용할 두 번째 감염 그룹인 성남 종교 집단 자료에서는 감염 시점은 알려져 있으나 49명 중 한명만이 증상 발현을 보이게 된다.

잠복기간은 $X = T_2 - T_1$으로 정의되며 T_1은 감염시점이고 T_2는 증상 발현 시점을 의미한다. 감염시점과 증상발현 시점의 기지(known)여부에 따라, 네 가지 경우에 대해 잠복기간이 정의된다. 예를 들어, 댄스 강사의 감염 시점과 증상 발현 시점은 2월 15일과 2월 20로 알려져 있다. 이 경우, 잠복 기간은 5 ($= T_2 - T_1$)일이 된다. 하지만 그에 의한 감염된 사람들의 정확한 감염시점은 우리가 분석한 자료에서 알 수 없으며

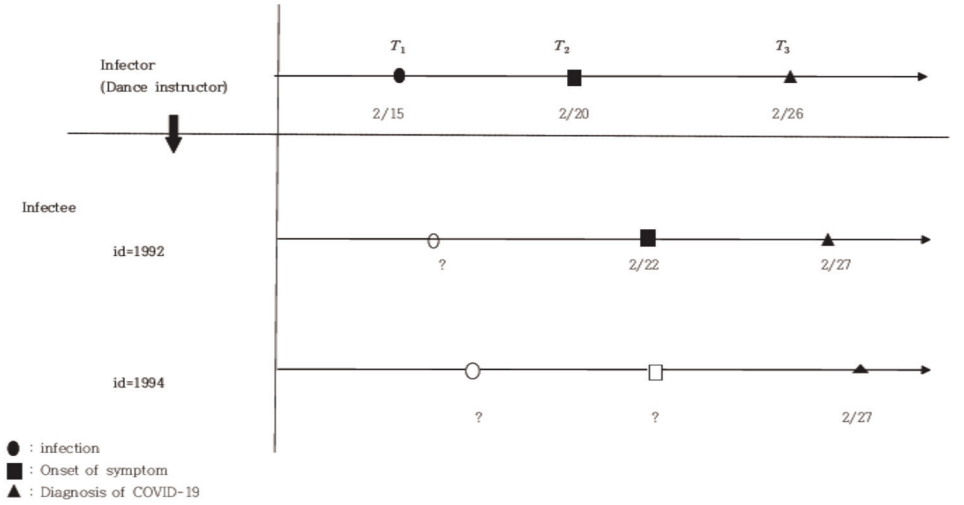

Figure 4.1. Two examples of transition from an infector(Dance instructor) in Cheonan Jumba dance.

Table 4.1. Symptom onset and confirmed date of infectee in Cheonan Jumba dance

Id	Infection case	Symptom onset date	Confirmed date
1981	gym facility in Cheonan	2020-02-24	2020-02-25
1982	gym facility in Cheonan	2020-02-24	2020-02-25
1984	gym facility in Cheonan	2020-02-23	2020-02-26
1988	gym facility in Cheonan	2020-02-23	2020-02-26
1989	gym facility in Cheonan	2020-02-25	2020-02-27
1990	gym facility in Cheonan	2020-02-22	2020-02-27
1991	gym facility in Cheonan	2020-02-24	2020-02-27
1992	gym facility in Cheonan	2020-02-22	2020-02-27
1994	gym facility in Cheonan		2020-02-27
1995	gym facility in Cheonan		2020-02-27
1996	gym facility in Cheonan	2020-02-26	2020-02-27
1998	gym facility in Cheonan	2020-02-24	2020-02-27
2001	gym facility in Cheonan	2020-02-22	2020-02-28

대신 그들의 증상 발현 시점 또는 양성 진단 시점만을 알게 된다. Table 4.1의 확진자 id가 1981–1992는 증상 발현시점(T_2)을 알고 있다. 이 경우, 그의 감염 시점은 강사가 워크샵으로부터 돌아온 2월 16일(TL)부터 증상 발현 시점 하루 전($TR = T_2 - 1$)이라고 간주한다. 따라서 그의 잠복 기간은 다음과 같이 구간 중도 절단 형태를 가지게 된다 ($T_2 - TR \le X \le T_2 - TL$). 만약 증상 발현이 없었다면 (id = 1994, 1995, 1996), 감염 시점과 증상 발현 시점 모두가 알려져 있지 않다는 것이다. 예를 들어, id = 1994의 감염 시점은 $TL \le T_1 \le TR$으로 TL은 2월 16일, TR은 2월 21일로 정하였다. 여기서 2월 21일은 증상 발현이 있는 유증상자들을 대상으로 구한 증상 발현 시점의 최소값의 하루 전 시점($TR = \min(T_2) - 1$)이다. 또한 id = 1994의 증상 발현 시점은 $T_2 < T_3$의 정보를 이용하여 잠복 시점은 $(T_3 - TR, T_3 - TL)$의 구간을 가지게 된다. 정리하면 감염 시점이 알려져 있지 않은 경우(천안자료의 모든 경우가 해당됨), 감염의 왼쪽 구간 값(TL)은 2월 16일이었고 오른쪽 구간 값(TR)은 증상 발현 여부에 따라 증상 발현 시점이 알려진 경우는 $TR = T_2 - 1$

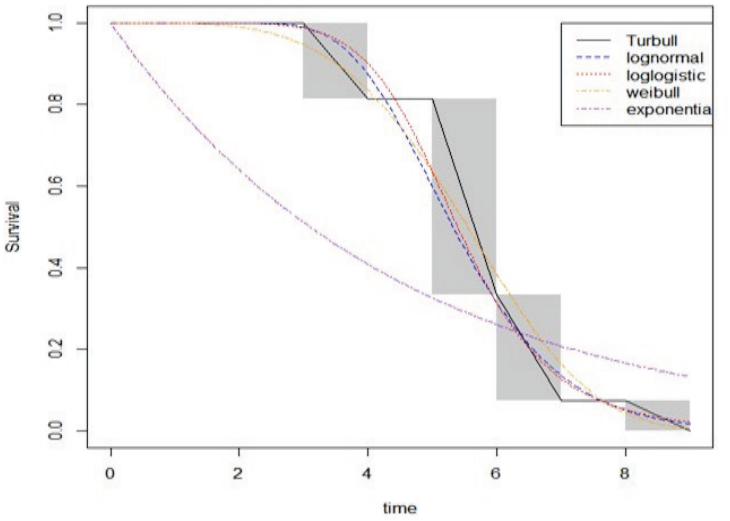

Figure 4.2. NPMLE and four PMLEs for distribution of incubation time.

Table 4.2. Estimation of mean incubation time

Distribution	Mean(day)	95 % CI	AIC	BIC
Exponential	4.468	(3.491, 5.720)	129.510	136.253
Weibull	5.501	(5.042, 6.031)	61.169	65.908
Log Logistic	5.394	(4.702, 6.049)	60.047	64.786
Log Normal	5.492	(4.931, 6.657)	60.521	65.259

로, 증상 발현이 없는 경우는 $TR = \min(T_2) - 1$을 적용하였다. 자료분석에서 적용된 자료는 1차 감염으로 같은 감염원을 가지고 있으며, 감염은 첫 번째 증상 발현 시점전에 발생함을 가정한다. 따라서, 증상 발현 시점이 알려져 있지 않은 경우에는, 확진 시점과의 관계 ($T_2 < T_3$)를 이용하여 잠복기간은 $X \in (T_3 - TR, T_3 - TL) = (XL, XR)$의 구간으로 정의하였다.

비슷한 방법으로 두 번째 코호트인 성남 교회 감염자들의 잠복 기간을 정의해보자. 이 그룹에서 49명의 감염 시점은 3월 8일 교회 예배 때로 알려져 있다. 하지만 48명 중 47명이 무증상 감염으로 증상 발현 시점은 알려져 있지 않다. 따라서 잠복시간은 다음의 구간 중도 시간 $[1, T_3 - T_1) = [XL, XR]$이 적용된다.

이렇게 정의된 구간 중도 절단된 잠복 기간의 분포를 추정하기 위해 (3.1)의 우도 함수를 적용하여 구한 Turnbull's 비모수 최대 우도 추정량(NPMLE)이 Figure 4.2(Turnbull)에 표시되어있다. 이 추정량의 계단 폭의 크기에 근거하여 바이러스 잠복기간의 확률이 가장 큰 시간대는 5–6일이었다. Table 4.2는 네 가지 모수 분포(지수분포, 와이블 분포, 로그 로지스틱 분포, 로그 정규분포)를 이용하여 구한 평균, 평균의 95% 신뢰구간, AIC, BIC를 보여준다. 본 분석에서는 R 패키지의 survival의 survfit 함수를 적용하였다. AIC와 BIC 결과를 통해 로그 로지스틱 분포의 적합도가 가장 좋음을 알 수 있다. 또한 Figure 4.2에서 로그 로지스틱 분포의 생존 함수 추정량이 비모수 최대 우도 추정량에 가장 가까이 그려지는 것을 볼 수 있다. 이 때 구한 평균 잠복 기간은 5.39일이고 95% 신뢰구간은 (4.70, 6.05)일이었다. 여기서 신뢰구간을 구하기 위해 붓스트랩 퍼센타일(percentile)방법을 적용하였다 (Efron과 Tibshirani, 1993).

Table 5.1. Case fatality rate and cure rate%

Age	Case fatality rate	Cure rate
10's	-	54.70%(54.54–54.98)
20's	-	66.07%(65.97–66.17)
30's		57.05%(65.98–57.12)
40's		66.53%(66.46–66.59)
50's	1.01%(1.00–1.01)	57.11%(57.03–57.19)
60's	3.00%(2.99–3.01)	57.11%(57.03–57.19)
70's	10.83%(10.77–10.89)	52.29%(52.12–52.45)
≥80	14.03%(13.97–14.10)	40.67%(40.53–40.81)

Table 5.2. Regression model for CFR and CR

Cov	CFR		CR: Cure rate	
	β (se)	*p*-value	*β (se)*	*p*-value
AGE(1;AGE ≥ 50)(0; AGE < 50)	3.44(0.432)	<0.0001	–0.443(0.060)	<0.001
Symptom	–1.347(0.713)	0.059	–0.235(0.080)	0.003
Gender(Male = 1)	0.959(0.256)	0.002	–0.131(0.051)	0.010

5. 경쟁 위험 모형을 이용한 치사율과 완치율 추정

본 절에서는 COVID-19 자료에서 양성 확진 후 의료 기관에 격리 치료를 받는 감염자를 대상으로 치사율(case fatality rate)를 추정하고자 한다 (Jewell 등, 2007). 여기서 모든 감염자는 이론적으로 사망할 수 있는 위험 상태에 놓여있으며 따라서 그들 모두는 위험군으로 간주되게 된다. 하지만 우리 나라 감염 환자군의 통계 결과를 보면 2020년 8월 10일 시점에서 감염자 14,660명 중에서 13,729명이 완치되었으며 305명이 사망하였다. 본 연구에서는 이러한 치사율과 완치율을 연령 별로 구해보고자 한다.

이를 위해 앞에서 설명한 두 가지 종말 사건을 가지는 경쟁 위험 모형을 적용할 수 있다. 2020년 6월 30일까지 수집된 환자들의 표본인 5,165명의 자료를 이용한다. 이들 중에 78명(1.5%)이 사망하였고 2,158명이 치료 중이며 2,929명이 완치(56.70%)되었다. 본 분석에서는 시간별 확진자들의 서로 다른 확진 시점과 사망시점 그리고 완치 시점을 반영하기 위해 CIF 식 (3.2)를 이용하여 다음의 치사율과 치유율 추정량을 적용한다.

$$\hat{F}_1^k(t^*) = \sum_{t^* \le t_j} \frac{d_{j1}}{n_j} \hat{S}(t_j -), \quad \hat{F}_2^k(t^*) = \sum_{t^* \le t_j} \frac{d_{j2}}{n_j} \hat{S}(t_j -),$$

여기서 k는 Age(연령대)별로 계산됨을 의미하며 $t_1 < t_2 < \cdots < t_m$은 모든 유형의 사건들의 발생 시간을 시간 순으로 정렬했을 때, t^*를 관측치들의 최대값이고 d_{j1}은 t_j시점에서 사망자 수를 d_{j2}는 t_j시점에서 완치자수를 각각 의미한다. n_j는 t_j시점에서의 위험개체수를 $\hat{S}(t_j)$는 추정된 생존 함수이다.

Table 5.1은 연령대별 치사율과 치유율을 보여준다. 주어진 결과에 근거하면 나이가 증가할수록 치사율이 증가함을 알 수 있다. 하지만 80대 이상을 제외하고는 서로 비슷한 치유율을 보여준다.

Table 5.2에서는 치사율과 치유율에 대한 나이와 증상 유무의 연관성 결과를 보여준다. 특히 Table 5.1에서 50대 미만의 사망 건수가 없기때문에 나이를 연속형으로 간주하는 것보다 50세 전후로 나누어 적용하였

다. 경쟁 위험 모형에 대한 회귀 모형을 위해 식 (3.3)의 부분포 비례위험모형(subdistribution proportional hazard model)을 적용하였다. 치사율에 대해서는 나이변수는 예상대로 매우 유의적이었으며 증상 유무는 유의하지 않았다. 특히, 50세 이상의 치사 위험도는 50세 미만의 $\exp(3.447) = 31.40$ 배이었다. 치유율은 50세 미만 감염자는 50세 이상 치유율의 $\exp(0.443) = 1.56$ 배이며 증상이 없을 때 치유 가능성이 더 높음을 알 수 있다.

6. 결론

본 논문에서는 2020년 8월 현재에도 전세계적으로 확산되고 있는 코로나 바이러스의 국내 감염 현황과 관련된 특성을 추정하기 위해 생존 분석기법을 적용하였다. 잠복 기간의 분포를 추정하기 위해 불완전하게 측정된 감염시점과 증상발현 시점에 대해서 구간 중도 절단 개념을 적용하였다. 감염원이 알려진 두 개의 양성 확진 그룹을 대상으로 여러 가지 모수 분포를 적용하여 로그 로지스틱 분포가 최적의 분포로 선택되었다. 이 분포에 근거하여 구한 바이러스 평균 잠복 기간은 5.4일 (95% CI: (4.70-6.05)일)로 추정되었다. 이는 Ki 등 (2020)가 추정한 3.9일보다는 조금 긴 시간이었다. 이러한 차이점은 본 추정량은 무증상자 자료를 포함한 것으로 구간 중도 절단 자료에 의한 그 불확실성이 반영된 결과라 판단된다. 두 번째 연구 주제는 치사율과 치유율을 추정하기 위해 두가지 종말 사건을 경험할 수 있는 가능성을 고려하여 경쟁 위험 모형을 적용하였다. 치사율은 남성이 여성보다 더 높았으며 노령일때 아주 높았다. 치유율은 여성일수록, 젊을수록 그리고 무증상 환자일 때 더 높았다. 현재 관련 연구를 보면 치사율은 호흡기, 심장 질환 관련 유병자와 관련이 있음이 알려져 있다. 본 연구에서 사용한 자료는 이러한 정보가 포함되어 있지 않아 그 영향력을 분석하지 못한 점이 아쉬웠다. 이와 함께 모형의 적합도 검정을 위해 좀 더 많은 연구가 진행되어야 할 것이다. 또한 지역별 감염율과 치사율 치유율의 확산을 연구하기 위해 교통량등 이동정보가 필요할 것이다. 향후 연구로는 다상태 모형(multi state model)의 적용을 통해 바이러스 감염 확산 그리고 소멸에 이르는 과정을 모형화하고 이와 관련된 요인을 규명고자 한다 (Cook과 Lawless, 2019). 또한 시공간 모형(spatial temporal model)을 적용함으로써 공간의 이동을 통해 역동적인 모형을 구할 수 있을 것이다 (Diggle과 Giorgi, 2019).

감사의 글

본 연구에서 사용한 자료는 Kaggle에서 다운로드 받은 자료입니다. 자료 사용을 허락해주신 김지훈님께 감사드립니다.

References

Bae, S., Kim, H., Jung, T. Y., *et al.* (2020) Epidemiological characteristics of COVID-19 outbreak at fitness centers in Cheonan Korea, *Journal of Korean Medical Science*, **35**, 1–9.

Cook, R. J. and Lawless, J. F. (2019). *Multistate Models for the Analysis of Life History Data*, CRC Press.

Diggle, P. J. and Giorgi, E. (2019). *Model-Based Geostatistics for Global Public Health*, CRC Press.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk, *Journal of the American Statistical Association*, **94**, 496–509.

Fine, P. E. (2003). The interval between successive cases of an infectious disease, *American Journal of Epidemiology*, **158**, 1039–1047.

Jewell, N. P., Lei, X., Ghani, A., Donnelly, C. A., Leung, G. M., Ho, L. M., Cowling, B., and Hedley, A. J.

(2007). Non-parametric estimation of the case fatality ratio with competing risks data: An application to severe acute respiratory syndrome (SARS), *Statistics in Medicine,* **26,** 1982–1998.

Ki, M. Task force for 2019-nCOV (2020). Epidemiologic characteristics of early cases with 2019 novel coronavirus(2019-nCoV) disease in Korea, *Epidemiology and Health*, **42,** 1–18.

Kim, Y. J. (2016). *Survival Analysis,* Free-academy.

Mettler, S. K., Kim, J., and Maathuis, M. H. (2020). Diagnostic serial interval as a novel indicator for contact tracing effectiveness exemplified with the SARS-CoV-2/COVID-19 outbreak in South Korea.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer New York.

Turnbull, B. W. (1974). Nonparametric estimation of survivorship function with doubly censored data, *Journal of the American Statistical Association*, **69,** 169–173.

COVID-19 바이러스 잠복 시간 분포 추정과 치사율 추정을 위한 생존 분석의 적용

기한정[a] · 김지은[a] · 김소희[a] · 박주원[a] · 이주행[a] · 김양진[a,1]

[a]숙명여자대학교 통계학과

(2020년 8월 24일 접수, 2020년 9월 25일 수정, 2020년 10월 6일 채택)

요 약

COVID-19는 지난 2019년 12월부터 중국에서 발생하여 전세계적으로 확산된 대유행병이 되었다. 본 연구에서는 한국 질병 관리 본부에서 공개한 오픈 자료를 이용하였으며 시각화 기법을 통해 확진자의 남녀별 지역별 추세를 조사하였다. 또한 평균 바이러스 잠복기간을 추정하기 위해 감염원이 알려진 두 감염 그룹의 증상 발현 시점과 양성 확진 시점을 활용하였다. 하지만 양성 확진자 중 86%가 무증상으로 정확한 증상 발현시점을 알 수 없었다. 또한 주어진 자료에서는 감염시점도 알려져 있지 않아 감염시점과 증상 발현 시점차로 정의되는 잠복기간은 정확하게 측정하기가 어렵다. 이에 생존 분석의 한 기법인 구간 중도 절단을 적용하여 잠복기간의 분포를 추정하였다. 여러가지 모수 분포를 적용한 결과 최적의 분포하에서 평균 잠복 기간은 5.4일 (95% 신뢰구간 (4.70,6.01)일)이었다. 본 분석에서는 확진자 표본을 이용하여 치사율과 치유율을 구하기 위해 경쟁 위험 모형을 적용하였다. 분석 결과 50대이상의 치사 위험률은 50대미만 그룹의 30배이상이며 남성 양성 확진자가 사망할 확률이 더 높았다. 또한 여성이고 나이가 젊고 무증상일 때 치유될 가능성이 더 높았다.

주요용어: COVID19, 대유행병, 잠복기간, 구간 중도 절단, 경쟁 위험 모형, 치사율, 치유율

This work was supported by Korea research grant (NRF-2017R1D1A1B03030578).

[1] 교신저자: (04310) 서울시 용산구 청파로47길 100, 숙명여자대학교 통계학과. E-mail: yjin@sookmyung.ac.kr