



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원 저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리와 책임은 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)





생물학적 동등성 시험에서 결측치 처리 방법에 따른
약동학적 파라미터 및 동등성 평가 비교 연구

연세대학교 보건대학원

보건정보통계학과 보건정보통계전공

강 범 석

생물학적 동등성 시험에서 결측치 처리 방법에 따른
약동학적 파라미터 및 동등성 평가 비교 연구

지도 하 민 진 교수

이 논문을 보건학 석사학위 논문으로 제출함

2025년 6월

연세대학교 보건대학원

보건정보통계학과 보건정보통계전공

강 범석



강범석의 보건학 석사학위 논문을 인준함

심사위원 하 민 진 인

심사위원 박 소희 인

심사위원 장 현 수 인

연세대학교 보건대학원

2025년 6월

차 례

| | |
|--|----|
| 국문요약 | V |
| I. 서론 | 1 |
| 1. 연구 배경 | 1 |
| 2. 연구 목적 | 3 |
| II. 연구 배경 및 분석 체계 | 4 |
| 1. 생물학적 동등성 시험 | 4 |
| 2. 생물학적 동등성 시험에서 결측치 발생과 문제점 | 9 |
| 3. 결측치 처리 방법 | 13 |
| 4. 추가적인 결측치 시나리오 | 22 |
| 5. 자료 구성 및 데이터 특성 | 30 |
| 6. 결측치 처리 방법 및 적용 | 32 |
| 7. 분석 설계 및 주요 평가 지표 | 35 |
| III. 연구 결과..... | 37 |
| 1. PK parameter & 90% 신뢰구간 및 잔류효과 확인 | 37 |
| 2. 잔류효과 확인 | 45 |
| 3. 혈장 중 시간대별 평균 농도 그래프 | 53 |



| | |
|--------------------------------------|----|
| 4. 간헐적 누락과 후반부 절단이 혼합된 형태의 결측치 | 58 |
| 5. 시뮬레이션 | 64 |
| IV. 고찰 | 75 |
| 1. 연구 결과에 대한 고찰 | 75 |
| 2. 연구의 제한점과 의의 | 81 |
| V. 결론 | 84 |
| 참고문헌 | 87 |
| Abstract | 90 |



표 차 례

| | |
|--|----|
| 표 1. PK 파라미터 및 90% 신뢰구간에서 결측치 처리 방법 별 차이 | 44 |
| 표 2. 잔류효과에 대한 결측치 처리 방법 별 차이 (AUC_t) | 50 |
| 표 3. 잔류효과에 대한 결측치 처리 방법 별 차이 (C_{max}) | 52 |
| 표 4. 결측치 처리 방법별 약동학 파라미터 및 90% 신뢰구간 비교 | 63 |
| 표 5. 시뮬레이션을 통한 결측 메커니즘 및 결측률에 따른 POWER, TYPE 1 ERROR 비교 | 68 |

그 림 차 례

| | |
|--|----|
| 그림 1. Interrupted PK Profile | 24 |
| 그림 2. Truncated PK Profile | 24 |
| 그림 3. 시간대별 약물 농도값의 결측 분포 히트맵 | 32 |
| 그림 4. Complete Case Analysis의 혈장 중 시간대별 평균 농도 그래프 | 56 |
| 그림 5. Multiple Imputation의 혈장 중 시간대별 평균 농도 그래프 | 56 |
| 그림 6. Mean Imputation의 혈장 중 시간대별 평균 농도 그래프 | 56 |
| 그림 7. Maximum Likelihood Estimation의 혈장 중 시간대별 평균 농도 그래프 | 57 |
| 그림 8. XGBoost의 혈장 중 시간대별 평균 농도 그래프 | 57 |
| 그림 9. MI(XGBOOST)의 혈장 중 시간대별 평균 농도 그래프 | 57 |
| 그림 10. MAR 가정 하에서 Sample Size 및 CV 변화에 따른 Type I Error 및 Power 추가 분석 | 72 |
| 그림 11. MCAR 가정 하에서 Sample Size 및 CV 변화에 따른 Type I Error 및 Power 추가 분석 | 73 |
| 그림 12. MNAR 가정 하에서 Sample Size 및 CV 변화에 따른 Type I Error 및 Power 추가 분석 | 74 |

국 문 요 약

생물학적 동등성 시험에서 결측치 처리 방법에 따른

약동학적 파라미터 및 동등성 평가 비교 연구

연구목적: 본 연구는 생물학적 동등성 시험에서 결측치가 발생하는 경우, 다양한 결측치 처리 방법이 약동학 지표 및 생물학적 동등성 판정 결과에 미치는 영향을 정량적으로 비교·평가하고, 분석 조건에 따라 가장 적절한 처리 방법을 도출하는 것을 목적으로 한다.

연구방법: 2×2 교차설계 기반의 임상시험 데이터를 활용하여 분석을 수행하였으며, 결측치 처리 방법으로는 완전 케이스 분석, 평균 대체법, 다중 대체법, 최대우도추정법, XGBoost 기반 단일 대체, 그리고 XGBoost 기반 다중 대체 (MI(XGBoost)) 등 총 여섯 가지를 적용하였다. 각 방법별로 약물의 최대 농도 및 곡선하 면적에 대한 90% 신뢰구간을 산출하여 생물학적 동등성 여부를 평가하였다. 추가적으로, 결측 메커니즘과 결측률, 피험자 수, 변동계수의 다양한 조건을 조합한 시뮬레이션을 수행하여 각 방법의 검정력과 제1종 오류율 변화 양상을 분석하였다. 이와는 별도로, 절단형 프로파일과 간헐적 누락이 혼합된 결측 구조를 구성하여 각 처리 방법의 적용 가능성과 한계를 추가적으로 평가하였다.

연구결과: 다중 대체법은 임상시험 데이터 분석과 시뮬레이션 전반에서 가장 일관되고 강건한 성능을 보였으며, 최대우도추정법은 임상 데이터 분석에서 실용적인 대안으로 나타났다. 반면, 완전 케이스 분석과 평균 대체법은 결측률이 높거나 표본 수가 적은 조건에서 오류율 증가 및 검정력 저하가 두드러졌다. XGBoost는 시간 순서를 반영한 예측 정확도는 높았으나, 신뢰구간이 다른 방법에 비해 과도하게 넓게 나타나, 동등성 여부에 대한 통계적 판단이 불안정해지는 한계가 있었다. MI(XGBoost)는 임상시험 데이터에서는 일정 수준의 실용성을 보였으나, 반복 대체 간 불확실성 반영이 제한되어 통계적 추론에는 신중한 적용이 요구된다.

결론: 결측이 존재하는 생물학적 동등성 시험에서는 다중 대체법은 본 연구의 분석 조건 하에서 통계적 신뢰성과 실무 적용 가능성을 모두 충족시켜, 유력한 처리 방법 중 하나로 판단된다. 단, 분석 설계 단계에서 결측 발생 메커니즘과 데이터 구조를 충분히 고려하여 적절한 결측치 처리 전략을 수립하는 것이 필수적이다.

I. 서론

1. 연구 배경

의약품을 개발하는 과정에서 임상시험은 신약이 사람에게 안전하고 효과적인지를 검증하는 필수적인 절차이다. 특히 생물학적 동등성(Bioequivalence, BE) 시험은 기존에 허가된 오리지널 의약품과 동일한 성분의 제네릭 의약품이 체내에서 비슷하게 작용하는지를 확인하는 중요한 단계로 제네릭 의약품의 허가와 공급에 있어 과학적 근거를 제공한다. 이를 통해 환자는 기존 약물과 동등한 효과를 가지는 대안을 선택할 수 있으며, 의료비 절감 측면에서도 긍정적 영향을 기대할 수 있다.

생물학적 동등성 시험은 두 제제 간의 체내 약물 흡수 및 분포의 유사성을 정량적으로 평가하기 위해 설계되었으며, 일정한 시간 간격으로 혈액을 채취하여 약물 농도를 분석하는 방식으로 수행된다. 이 과정에서는 실험 설계의 정밀한 통제와 더불어 자료의 정확성과 일관성이 확보되어야 한다. 그러나 실제 임상시험에서는 피험자의 중도 탈락, 채혈 실패, 분석 기기 오류 등 다양한 원인으로 인해 일부 시점의 농도값이 누락되는 결측치가 발생할 수 있다. 이러한 결측은 약동학 지표의 신뢰도뿐만 아니라 시험약과 대조약 간의 동등성 평가 결과에도 영향을 미칠 수 있어, 적절한 처리가 요구된다.

이에 대해 식품의약품안전처는, 임상시험 중 발생한 결측치는 반드시 분석에서 제외할 필요는 없으며, 통계적으로 타당한 처리 방법이 시험계획서 및 통계분석계획서에 명확히 정의되어 있는 경우 분석에 포함할 수 있다고 명시하고 있다. 또한, 결측 처리 방식에 따라 분석 결과가 달라질 수 있으므로, 다양한 방법을 적용하여 비교·검토하는 것이 바람직하며, 이러한 계획은 사전에 통계분석계획서에 포함되는 것이 권장된다.

현재까지는 충분한 농도-시간 자료가 확보된 피험자만을 분석에 포함시키는 접근이 일반적으로 권장되어 왔으며, 이는 실무적으로 완전 케이스 분석을 기본 방식으로 간주하는 관행과도 일치한다. 그러나 결측이 빈번하거나 결측률이 높은 경우, 단순한 제외 방식보다 다중 대체법이나 최대우도추정법과 같은 보다 정교한 통계적 방법의 적용이 확대되고 있다. 이러한 방법을 사용할 경우, 그 근거와 적용 방식은 반드시 분석 계획서에 사전 정의되어야 하며, 명확한 통계적 정당성이 요구된다.

따라서 연구자는 결측의 발생 패턴(MCAR, MAR, MNAR 등)과 데이터 특성을 종합적으로 고려하여, 시험 설계의 타당성과 규제 수용성을 모두 충족할 수 있는 결측치 처리 전략을 수립해야 한다.



2. 연구 목적

본 연구는 생물학적 동등성 시험에서 결측치가 발생했을 때, 다양한 처리 방법이 약동학 지표(Pharmacokinetic parameters, PK parameters) 및 생물학적 동등성 평가 결과에 미치는 영향을 비교하고, 분석 조건에 따라 가장 적절한 결측치 처리 방법을 제시하는 것을 목적으로 한다.

구체적인 연구 목적은 다음과 같다.

첫째, 생물학적 동등성 시험 과정에서 나타나는 결측치의 종류와 발생 패턴을 분석한다.

둘째, 여러 결측치 처리 방법이 주요 약동학 지표의 계산 결과에 미치는 영향을 비교한다.

셋째, 각 결측치 처리 방법이 생물학적 동등성 판정 결과에 미치는 영향을 평가하며, 시험약과 대조약 간의 90% 신뢰구간이 동등성 기준 범위(80.00~125.00%)에 포함되는지를 바탕으로 동등성 충족 여부를 비교한다.

II. 연구 배경 및 분석 체계

1. 생물학적 동등성 시험

1) 생물학적 동등성 시험

생물학적 동등성 시험은 제네릭 의약품이 오리지널 의약품과 유사한 치료 효과를 나타내는지를 과학적으로 검증하는 절차로, 제네릭 의약품의 품질과 유효성을 평가하는 데 필수적인 과정이다. 제네릭 제제는 신약과 동일한 유효 성분을 포함하고 있으나, 제형, 제조 공정, 보조 성분의 차이에 따라 체내 흡수 속도나 흡수량에서 차이가 발생할 수 있다. 이에 따라 두 제형 간의 약효 유사성을 객관적으로 입증하기 위해 생물학적 동등성 시험이 수행된다.

생물학적 동등성 시험은 일반적으로 건강한 성인을 대상으로 하며, 피험자 1명이 시험약과 대조약을 모두 복용하도록 설계된다. 이는 주로 2×2 교차설계 방식으로 이루어지며, 두 제형은 무작위 순서로 투여되고, 투약 간에는 일정한 휴약기를 두어 이전 투약의 영향을 최소화한다. 이후 약물 복용 후 일정한 시간 간격으로 혈액을 채취하여, 약물이 체내에서 어떻게 흡수되고 분포되며 배출되는지를 관찰한다.

이 과정은 시험약과 대조약 각각에 대해 동일하게 반복되며, 피험자 1명에게서 시간대별 약물 농도 데이터가 두 번 수집된다. 그 결과, 수집된

데이터는 동일한 피험자 내에서 두 제형을 모두 비교할 수 있는 구조를 가지며, 시간에 따른 반복 측정값으로 구성된다. 이러한 구조는 제제 간의 체내 행동 특성을 정밀하게 비교·분석할 수 있게 해주며, 생물학적 동등성 판정의 핵심 자료로 활용된다.

두 제형 간의 동등성은 총 체내 노출량과 최대 혈중 농도에 대한 평균 비율(시험약/대조약)의 90% 신뢰구간(Confidence Interval, CI)을 기준으로 판단한다. 이들 지표는 로그 변환을 거쳐 산출되며, 평가 결과로 도출된 90% 신뢰구간이 80.00%에서 125.00% 사이의 동등성 기준 범위에 완전히 포함되는 경우, 두 제형은 통계적으로 유의한 차이가 없는 것으로 간주되어 생물학적으로 동등하다고 판단된다.

여기서 사용되는 80.00%~125.00% 범위는, 시험약과 대조약의 체내 노출 차이가 $\pm 20\%$ 이내일 경우, 두 제형이 임상적으로 유사한 약효를 나타낸다고 간주할 수 있다는 통계학적·약리학적 판단에 근거한다. 이 기준은 미국 식품의약국(FDA)의 Guidance for Industry: Statistical Approaches to Establishing Bioequivalence (2001)를 비롯하여, 유럽의약청(EMA), 세계보건기구(WHO), 일본 의약품의료기기종합기구(PMDA), 그리고 대한민국 식품의약품안전처(MFDS) 등 다수의 국제 규제기관에서 공통적으로 채택하고 있는 평가 기준이다.

생물학적 동등성 시험에서 수집되는 자료는 한 사람의 시간에 따른 약물 농도 변화를 정밀하게 반영하므로, 약물의 흡수와 대사에 대한 개인차를 고려할 수 있다는 장점이 있다. 이러한 시험을 통해 제네릭 의약품이 오리지널 의약품과 동등한 효과를 나타내는 것이 입증되면, 제네릭은 안전성과 유효성이 확보된 의약품으로 간주되어 보다 저렴한 비용으로 시장에 공급될 수 있다. 이는 환자의 의료비 부담을 줄이는 데 기여할 뿐 아니라, 보건의료 시스템 전반에서 약물 접근성을 높이고 자원 배분의 효율성을 향상시키는 데에도 중요한 역할을 한다.

생물학적 동등성 시험에서는 시험약과 대조약 간의 약동학적 유사성을 평가하기 위해 여러 핵심 약동학적 지표를 사용한다.

1. AUC_t (Area Under the Curve to time t)

특정 시간까지의 약물 총 노출량을 나타내며, 시험약과 대조약의 흡수 정도를 비교하는 핵심 지표이다.

2. AUC_{inf} (Area Under the Curve to infinity)

약물이 체내에서 완전히 제거될 때까지의 총 노출량을 의미하며, 장기적인 약물 지속성을 평가하는 데 사용된다.

3. C_{max} (Maximum Concentration)

혈중 약물 농도가 가장 높은 값을 의미하며, 약물의 흡수 속도와 강도를 평가하는 지표이다.

4. T_{max} (Time to Maximum Concentration)

최대 농도(C_{max})에 도달하기까지 걸리는 시간을 의미하며, 약물의 흡수 속도를 평가하는 기준이 된다.

5. $t_{1/2}$ (Half-life)

혈중 약물 농도가 절반으로 감소하는 데 걸리는 시간으로, 약물의 제거 속도를 평가하는 중요한 지표이다.

6. AUC_t/AUC_{inf} 비율

총 노출량(AUC_{inf}) 중 실제 측정된 노출량(AUC_t)의 비율을 나타내며, 데이터의 신뢰성을 평가하는 기준이 된다.

생물학적 동등성 시험에서는 시험약과 대조약 간의 약동학적 유사성을 평가하기 위해 AUC_t 와 C_{max} 를 주요 지표로 활용하며, 이들에 대한 기하평균비의 90% 신뢰구간을 산출하여 통계적으로 동등성을 판단한다. 해당 신뢰구간이 사전에 정의된 동등성 기준인 80.00%~125.00% 범위에 모두 포함될 경우, 시험약은 대조약과 생물학적으로 동등한 것으로 간주된다.

이러한 생물학적 동등성 여부를 판단하기 위해 다음과 같이 가설을 설정한다 (김도형 외, 2020).

귀무가설(H0): 시험약과 대조약 사이에 차이가 있으며, 동등성 마진을 벗어난다.

$$H_0 : |\mu_T - \mu_R| \geq \Delta$$

대립가설(HA): 시험약과 대조약의 차이가 동등성 마진 내에 존재한다.

$$H_A : |\mu_T - \mu_R| < \Delta$$

여기서 μ_T 는 시험약의 평균, μ_R 는 대조약의 평균을 나타내며, Δ 는 동등성 마진(80.00%~125.00%)을 의미한다. 즉, 대립가설을 지지하는 결과가 도출될 경우 시험약과 대조약은 생물학적으로 동등한 것으로 판단된다.

기하평균비율(Geometric Mean Ratio, GMR)은 각 피험자의 약동학적 지표를 로그 변환한 후, 다음과 같이 계산된다:

$$GMR = e^{\bar{x}_T - \bar{x}_R}$$

여기서는 \bar{x}_T 는 시험약의 로그 변환된 약동학적 지표 평균값, \bar{x}_R 는 대조약의 로그 변환된 약동학적 지표 평균값을 의미한다.

각 약물의 로그 변환된 평균값은 다음과 같이 계산된다.

$$T = \frac{1}{n_T} \sum_{i=1}^{n_T} \ln(X_{T,i}) , \quad R = \frac{1}{n_R} \sum_{i=1}^{n_R} \ln(X_{R,i})$$

여기서 $X_{T,i}$ 와 $X_{R,i}$ 는 각각 시험약과 대조약에 대해 피험자 i 의 AUC 또는 C_{\max} 값이며, 이를 로그 변환한 후 평균을 계산하여 로그 평균값을 구한다.

기하평균비율이 단순한 값으로만 제공되면 통계적 신뢰도가 부족할 수 있기 때문에, 90% 신뢰구간을 함께 산출하여 두 약물 간 동등성을 검토해야 한다. 신뢰구간은 아래와 같이 계산된다.

$$90\% CI = [\exp\{(\bar{X}_T - \bar{X}_R - t_{0.05, df} \cdot SE)\}, \exp\{\bar{X}_T - \bar{X}_R + t_{0.05, df} \cdot SE\}]$$

여기서 $t_{0.05, df}$ 는 자유도(df)에 따른 t분포의 95% 분위수, SE는 표준오차를 의미한다.

2. 생물학적 동등성 시험에서 결측치 발생과 문제점

생물학적 동등성 시험에서는 약물의 혈중 농도를 측정하여 시험약과 대조약의 동등성을 평가하는 과정에서 여러 가지 이유로 인해 결측치가 발생할 가능성이 크다. 이러한 결측치는 연구의 신뢰성을 저하시킬 수 있으며, 생물학적 동등성 평가 결과를 왜곡할 가능성이 있다.

생물학적 동등성 시험에서 발생하는 결측치는 크게 완전 무작위 결측(MCAR), 조건부 무작위 결측(MAR), 비무작위 결측(MNAR) 세 가지로

분류된다. 각 유형별 특성을 고려하여 결측치 처리 방법이 선택되어야 하며, 이에 대한 신중한 접근이 요구된다.

1. MCAR (완전 무작위 결측, Missing Completely at Random):

MCAR은 결측이 관측된 데이터나 피험자의 특성과 완전히 무관하게 무작위로 발생하는 경우를 의미한다. 예를 들어, 실험 기기의 일시적 오류나 데이터 입력 실수로 인해 특정 혈액 샘플이 유실된 경우가 이에 해당한다. 이러한 경우에는 결측이 분석 변수와 독립적이므로, 리스트와이즈 삭제를 통해 결측값이 포함된 사례를 제거하더라도 통계적 편향은 발생하지 않는다. 실제로 Jakobsen et al. (2017)은 MCAR 환경에서 리스트 삭제가 분석 결과에 유의미한 영향을 미치지 않는다고 보고하였다. 그러나 결측률이 높아질 경우, 분석에 사용 가능한 표본 수가 줄어들면서 검정력이 약화되고, 결과의 신뢰도가 저하될 수 있다. 특히 표본 수가 제한된 생물학적 동등성 시험에서는 이러한 영향이 더욱 치명적일 수 있으며, 이에 따라 단순 삭제보다는 다중 대체법의 적용이 바람직할 수 있다. Enders (2010)는 MCAR 환경에서 결측률이 높을 경우 다중대체법을 활용함으로써 데이터 손실을 줄이고 통계적 검정력을 유지할 수 있다고 제안하였다.

2. MAR (조건부 무작위 결측, Missing at Random):

MAR은 결측이 특정 공변량(예: 연령, 성별 등)과는 관련이 있지만, 누락된

값 자체와는 독립적인 경우를 의미한다. 예를 들어, 고령의 피험자가 채혈을 거부하여 약물 농도 데이터가 누락되는 경우, 결측은 연령과는 관련되지만 실제 혈중 농도와는 무관할 수 있다. 이와 같은 상황에서 단순 삭제를 적용할 경우, 고령자 집단의 데이터가 과소 대표되어 연구 결과에 체계적인 편향이 발생할 수 있다. 이러한 편향을 최소화하기 위해서는 단순 삭제보다 다중 대체법 또는 최대우도추정법과 같은 통계적 보정 방법이 보다 적절하다. Schafer (1997)는 MAR 조건에서 다중대체법이 데이터 손실을 최소화하면서도 추론의 타당성을 유지하는 데 효과적임을 보고하였다. 또한 Beunckens et al. (2005)은 최대우도추정법이 MAR 상황에서도 강건한 분석 성능을 보이며, 단순 삭제에 비해 우월한 선택임을 강조하였다.

생물학적 동등성 시험에서도 MAR 메커니즘은 빈번히 발생할 수 있으며, 특히 고령자 또는 특정 약물 반응 집단에서 데이터 결측이 집중될 경우, 다중대체법과 같은 방법은 분석의 대표성과 정확성을 동시에 보장할 수 있는 대안으로 작용한다.

3. MNAR (비무작위 결측, Missing Not at Random):

MNAR는 결측이 관측되지 않은 실제 값 자체와 직접적으로 관련된 경우를 의미한다. 예를 들어, 약물 농도가 지나치게 낮아 분석기기의 검출 한계(LLOQ)를 밑도는 경우나, 고농도로 인한 부작용으로 피험자가 채혈을

거부하는 경우가 이에 해당한다. 이러한 결측은 누락 패턴 자체에 정보가 포함되어 있으므로, MNAR 환경에서 단순 대체 방식은 결측 메커니즘을 반영하지 못하므로, 편향된 추정치와 왜곡된 동등성 평가로 이어질 수 있다.

저농도 환자의 데이터가 더 자주 누락될 경우, 평균 농도값은 실제보다 높게 추정되며, 이는 생물학적 동등성 판정을 왜곡시키는 원인이 될 수 있다. 따라서 MNAR 환경에서는 결측 메커니즘을 모델링한 분석이 필수적이며, 대표적으로 선택 모델과 감쇠 모델이 활용된다. Ibrahim et al. (2005)는 선택 모델이 MNAR 상황에서 분석의 타당성을 향상시킬 수 있음을 보고하였으며, Mallinckrodt et al. (2003)은 감쇠 기반 패턴-믹스처 모델이 현실적인 대안임을 제시하였다.

특히 생물학적 동등성 시험처럼 검출 한계와 피험자 이탈이 빈번한 약동학 연구에서는, 결측 발생 패턴을 사전에 분석하고 적절한 모델 기반 결측 처리법을 적용하는 것이 통계적 신뢰도 확보에 핵심적이다.

생물학적 동등성 시험에서 발생하는 결측치는 분석 결과의 신뢰도를 저하시킬 뿐만 아니라, 시험약과 대조약 간 동등성 평가 결과를 왜곡할 수 있는 주요 요인이다. 결측이 약동학적 핵심 지표에 영향을 줄 경우, 표본 수 감소에 따른 검정력 저하와 함께 오판 가능성이 높아진다. 특히 결측이 고령자, 부작용 경험자 등 특정 피험자 집단에 집중될 경우, 연구 결과에

체계적인 편향이 발생하여 잘못된 결론으로 이어질 위험이 크다.

이로 인해 실제로는 동등하지 않은 약물이 통과되거나, 동등한 약물이 부적합 판정을 받을 수 있다. 그러나 현재까지 생물학적 동등성 시험에서 발생하는 결측치에 대한 단일한 표준 처리 방법은 존재하지 않으며, 결측 메커니즘과 연구 목적에 따라 적절한 방법을 선택하는 것이 필요하다. 다양한 결측 처리 기법을 비교 적용하고 결과를 다각도로 검토하는 과정은 분석의 타당성을 확보하는 데 필수적이다.

궁극적으로, 생물학적 동등성 시험에서 결측을 효과적으로 다루기 위해서는 데이터 수집 단계에서부터 결측을 최소화하는 전략을 수립하고, 결측 발생 유형(MCAR, MAR, MNAR)을 정확히 파악한 후 그에 맞는 통계적 처리 기법을 적용해야 한다. 이를 통해 분석의 신뢰도를 높이고, 규제 수용 가능성 또한 확보할 수 있을 것이다.

3. 결측치 처리 방법

결측치 처리 방법은 연구 목적과 결측치 유형에 따라 적절히 선택되어야 하며, 잘못된 처리 방식은 분석 결과의 신뢰성을 저하시킬 수 있다. 본 절에서는 본 연구에 사용된 주요 결측치 처리 방법과 수식 기호를 정의하고,



각 방법의 원리와 특징을 설명한다.

본 연구에서는 결측치 처리 방법에 관계없이 공통적으로 사용하는 기호를 다음과 같이 정의하였다. 피험자 i 의 j 시점에서 측정된 약물 농도값은 x_{ij} 로 표기하며, 관측 여부를 나타내는 지시변수 R_{ij} 는 해당 값이 관측된 경우 1, 결측된 경우 0으로 정의하였다. 또한, 결측이 발생한 x_{ij} 를 각 처리 방법에 따라 대체하거나 예측하여 산출된 값은 \hat{x}_{ij} 로 나타내었다.

1. 평균 대체법 (Mean Imputation)

평균 대체법은 결측이 발생한 시점의 농도값을 해당 시점에서 관측된 다른 피험자들의 동일 시점 평균값으로 대체하는 방법이다. 특정 채혈 시점 j 에서 일부 피험자의 혈중 농도값이 결측된 경우, 해당 시점의 관측값 평균으로 대체된다. 이를 수식으로 나타내면 다음과 같다(Little and Rubin, 2019a).

$$\hat{x}_{ij} = \frac{\sum_{i=1}^n R_{ij} x_{ij}}{\sum_{i=1}^n R_{ij}}$$

평균 대체법은 계산이 단순하고 실무 적용이 용이하여, 결측값이 존재하더라도 전체 데이터셋의 크기를 유지할 수 있다는 장점이 있다. 그러나 모든 결측값을 동일한 값으로 대체하는 방식이기 때문에, 데이터 내 고유한 변동성과 분산 구조를 반영하지 못하고, 그 결과 전체 분산이 과소 추정되어

분석 결과에 왜곡을 초래할 수 있다. 또한, 채혈 시점 간 농도값 간의 상관관계를 약화시키거나 왜곡할 가능성도 존재한다.

평균 대체법은 결측치가 완전 무작위로 발생하는 경우(MCAR)에는 상대적으로 안전하게 사용할 수 있지만, 조건부 무작위(MAR) 또는 비무작위 결측(MNAR) 상황에서는 평균값이 특정 집단의 특성을 과대 또는 과소 반영할 가능성이 있어 분석 결과의 신뢰도를 저하시킬 수 있다. 따라서 평균 대체법은 MCAR 가정 하에서 제한적으로 사용하는 것이 바람직하다.

2) 완전 케이스 분석 (Complete Case Analysis, CCA)

완전 케이스 분석은 결측값이 포함된 관측치를 분석에서 완전히 제외하고, 결측이 없는 관측치만을 활용하여 분석을 수행하는 방법이다. 즉, 전체 데이터셋에서 하나라도 결측값이 포함된 관측치는 제거하고, 결측값이 전혀 존재하지 않는 관측치로만 구성된 새로운 데이터셋을 이용하여 분석을 수행하며, 이는 다음과 같이 표현된다 (Little and Rubin, 2019b).

$$D_{CCA} = \{x_{ij} \mid R_{ij} = 1 \quad \forall i, j\}$$

완전 케이스 분석은 결측치가 완전 무작위(MCAR)로 발생한 경우에는 분석 결과에 편향을 유발하지 않으므로, 비편향된 추정이 가능하다는 장점이 있다. 하지만 결측이 조건부 무작위(MAR)나 비무작위(MNAR)일 경우, 결측값이 특정 시점이나 특정 피험자 집단에 집중되어 발생할 수 있으며, 이러한 경우

단순히 결측 데이터를 제거하면 분석 결과에 왜곡을 초래할 수 있다. 특히 결측이 특정 집단에 집중되어 있는 경우, 해당 집단의 정보가 반영되지 않아 분석 결과의 대표성이 저하될 수 있으며, 이러한 편향은 예측 또는 추론의 정확도에 부정적인 영향을 미칠 수 있다.

또한, 완전 케이스 분석은 결측을 포함한 관측치를 제거함으로써 분석에 사용되는 표본 수가 감소하게 되며, 이는 통계적 검정력의 약화를 초래할 수 있다. 따라서 완전 케이스 분석은 결측 비율이 낮고, 결측 발생이 무작위적인 경우에만 제한적으로 사용하는 것이 바람직하다.

3) 다중 대체법 (Multiple Imputation, MI)

다중 대체법은 단일한 값이 아닌 여러 개(m)의 대체값을 생성하여 결측치를 보완한 후, 각각의 데이터셋에서 분석을 수행하고 최종적으로 통합하는 방식이다.

다중대체법에서는 결측값이 포함된 데이터에 대해 m 개의 대체 데이터셋 $x_{ij}^{(k)}$ ($k = 1, 2, \dots, m$)으로 생성하며, 각 대체 데이터셋은 관측된 값 x_{ij} 을 기반으로 결측값이 따를 수 있는 분포를 추정한 후, 해당 분포로부터 랜덤하게 시뮬레이션하여 결측값을 생성하는 방식으로 구성된다. 이 과정은 결측값의 불확실성을 통계적으로 반영하기 위한 것으로, 동일한 위치의 결측값이라 하더라도 각 대체 데이터셋마다 서로 다른 값으로 대체된다. 생성된 m 개의

대체 데이터셋 각각에서 독립적으로 분석을 수행한 후, 해당 데이터셋에서 얻어진 추정치 $x_{ij}^{(k)}$ 들을 평균하여 최종 추정치 \hat{x}_{ij} 를 산출한다. 이후 총 분산 T 는 각 대체 데이터셋 내 분산의 평균(\bar{U})과 대체 데이터셋 간 분산(B)을 통합하여 다음과 같이 계산하였다(Rubin, 1987).

$$\hat{x}_{ij} = \frac{1}{m} \sum_{k=1}^m x_{ij}^{(k)}, \quad T = \bar{U} + \left(1 + \frac{1}{m}\right)B$$

다중 대체법은 특히 결측 메커니즘이 조건부 무작위(MAR)일 때 효과적인 방법으로, 데이터의 변동성과 불확실성을 유지하면서 편향을 최소화할 수 있다. 그러나 이 방법은 대체 과정에 사용되는 모델링 가정에 대한 타당성을 요구하며, 여러 개의 대체 데이터셋을 생성하고 각각 분석 및 통합하는 과정에서 계산 비용이 크고 적용 절차가 복잡하다는 단점이 있다. 그럼에도 불구하고, 결측 비율이 높고 MAR 가정이 적절하게 충족되는 상황에서는 높은 신뢰도의 분석 결과를 제공할 수 있는 유용한 방법이다.

4) 최대우도추정법 (Maximum Likelihood Estimation, MLE)

최대우도추정법은 모집단의 모수 θ 를 추정하기 위한 대표적인 방법으로, 주어진 자료 x_1, x_2, \dots, x_n 이 특정 모수 θ 하에서 관측될 확률이 가장 크도록 하는 값을 추정하는 방식이다.

이를 위해 먼저 n 개의 독립적인 관측값이 주어졌을 때, 이들이 동시에



발생할 확률, 즉 우도함수는 다음과 같이 나타낼 수 있다(Little and Rubin, 2019c).

$$L(\theta; x_i) = \prod_{i=1}^n f(x_i | \theta)$$

이는 각 관측값 x_i 가 주어진 모수 θ 하에서 발생할 확률 $f(x_i | \theta)$ 을 모두 곱하여 $\left(\prod_{i=1}^n \right)$ 전체 데이터가 관측될 가능성을 수식적으로 표현한 것이다. 여기서 \prod 는 $i = 1$ 부터 n 까지의 값을 모두 곱하라는 의미의 곱 기호로 관측값 x_1, x_2, \dots, x_n 서로 독립적으로 발생한다고 가정할 때 각 확률값을 전부 곱하여 우도함수 $L(\theta; x_i)$ 를 구성하게 된다.

이러한 모수 추정치를 최대우도추정치라 하며, 이는 다음과 같이 정의된다.

$$\hat{\theta} = \operatorname{argmax}_{\theta} L(\theta; x_i)$$

최대우도추정법은 결측이 존재하더라도 관측된 데이터를 최대한 활용하여 분석할 수 있으며, 결측치로 인한 정보 손실을 최소화할 수 있다는 장점이 있다. 특히 혼합효과모형과 결합하여 개인 간 변동성을 반영한 분석이 가능하므로, 약동학 연구나 생물학적 동등성 시험과 같이 개별 특성이 중요한 분석에서 유용하게 사용된다.

반면, 최대우도추정법은 특정 확률 분포를 가정해야 하는 특성이 있으며, 이

가정이 부적절할 경우 추정치의 신뢰성이 저하될 수 있다. 또한 계산 과정이 복잡하고 고도의 통계적 전문성을 요구하며, 대규모 데이터셋에서는 계산 비용이 크게 증가할 수 있다.

따라서 최대우도추정법은 결측 메커니즘이 조건부 무작위(MAR)일 때 효과적인 결측치 처리 방법으로 간주되며, 복잡한 데이터 구조에서도 강력한 추정 능력을 제공하는 통계적 접근법이다.

5) XGBoost 기반 머신러닝 기법 (eXtreme Gradient Boosting)

XGBoost는 경사하강법기반의 양상을 결정 트리 모델로, 다차원적이고 비선형적인 변수 간 관계를 학습하여 결측값을 예측하는 머신러닝 기법이다. 본 연구에서는 특정 채혈 시점에서 결측이 발생한 혈중 농도값을 동일 피험자의 다른 채혈 시점 농도값 및 처리군(Treatment), 시간(Time) 등의 정보를 바탕으로 예측하는 방식으로 적용하였다(Chen and Guestrin, 2016).

결측값이 존재하는 채혈 시점에서 농도값은 다음과 같이 예측된다.

$$\hat{x}_{ij} = XGBoost(x_{i,-j})$$

여기서 $x_{i,-j}$ 는 피험자 i 의 다른 시점 농도값 및 입력 변수(시간, 처리군 등)를 의미한다.

XGBoost는 다음과 같은 목적함수를 기반으로 최적의 모델을 학습한다.

$$L(\phi) = \sum_{i=1}^n l(x_{ij}, \hat{x}_{ij}) + \sum_{k=1}^K \Omega(f_k)$$

여기서 $l(x_{ij}, \hat{x}_{ij})$ 는 실제 관측값 x_{ij} 과 예측값 \hat{x}_{ij} 간의 차이를 최소화하는 손실함수이며, $\Omega(f_k)$ 는 트리 모델의 복잡도를 제어하여 과적합을 방지하는 정규화 항이다.

XGBoost 모델은 결측이 없는 데이터를 이용해 모델을 학습한 후, 이후 결측값이 포함된 경우 학습된 모델을 활용하여 예측값을 산출한다. 이러한 방식은 전통적인 통계 기법과 달리 데이터의 분포나 결측 메커니즘(MCAR, MAR, MNAR)에 대한 가정 없이 적용할 수 있으며, 입력 변수 간의 비선형적 상호작용을 효과적으로 반영할 수 있다는 강점이 있다.

또한, 다중 공선성 문제에 영향을 받지 않으며, 일부 시점의 농도값이 결측된 경우에도 자동으로 처리되는 특성을 갖추고 있어 분석의 유연성이 높다는 장점이 있다.

6. XGBoost 기반 다중 대체법

XGBoost 기반 다중 대체(MI(XGBoost)) 방식은 다중 대체의 일반적인 구조를 따르되, 결측값을 예측하는 과정에서 전통적인 회귀 모델 대신 XGBoost 알고리즘을 활용한다는 점에서 차별화된다. 이 방식은 변수 간의 비선형 관계 및 고차 상호작용을 효과적으로 반영할 수 있어, 구조적

복잡성이 높은 데이터에 적합하다.

결측값이 발생한 시점 t 에서의 약물 농도 x_{ij} 는 피험자 i 의 다른 시점 농도값 $x_{i,-j}$, 시간 변수, 제제 구분 등으로 구성된 입력 변수 벡터 Z_{ij} 를 기반으로 다음과 같이 예측되었다:

$$\hat{x}_{ij}^{(k)} = f_{XGB}^{(k)}(Z_{ij}) \quad (k=1, \dots, m)$$

여기서 $f_{XGB}^{(k)}$ 는 XGBoost 기반 회귀 모형으로, 각 반복에서 부트스트랩 샘플링이나 하이퍼파라미터 랜덤화를 적용함으로써 모형 간 예측값의 변이를 유도하였다. 이러한 방식으로 생성된 m 개의 대체값 $\hat{x}_{ij}^{(1)}, \dots, \hat{x}_{ij}^{(m)}$ 을 활용해 개의 완전한 데이터셋을 구성하였다.

각 완전 데이터셋에 대해 동일한 분석을 수행한 후, Rubin's Rules을 적용하여 최종 예측값 \hat{x}_{ij} 와 총 분산 T 를 다음과 같이 계산하였다:

$$\hat{x}_{ij} = \frac{1}{m} \sum_{k=1}^m \hat{x}_{ij}^{(k)}, \quad T = \bar{U} + \left(1 + \frac{1}{m}\right)B$$

여기서 $\hat{x}_{ij}^{(k)}$ 는 k 번째 대체 데이터셋에서 예측된 약물 농도값, \bar{U} 는 평균 내적 분산, B 는 대체 간 분산이다. 다만, XGBoost는 본질적으로 결정론적 예측 알고리즘이므로 반복 대체 간 변이가 제한적이며, 이로 인해 B 가 과소 추정될 가능성이 존재한다. 이는 Rubin's Rules의 이론적 전제와 충돌할 수 있으며, 결과적으로 신뢰구간이 실제보다 좁게 형성될 수 있다.

그럼에도 불구하고, MI(XGBoost) 방식은 데이터의 구조적 특성(예: 시계열성, 변수 간 상호작용 등)을 충실히 반영할 수 있다는 점에서 보조 분석이나 민감도 분석의 수단으로 활용 가능하다. 특히 선형 모델 기반 대체가 부적절한 상황에서, 머신러닝 기반 예측을 접목한 실험적 대안으로서 유효성이 있다.

4. 추가적인 결측치 시나리오

1) 간헐적 누락과 후반부 절단이 혼합된 형태의 결측치

실제 임상시험에서는 단일 형태의 결측보다 다양한 유형의 결측이 복합적으로 발생하는 사례가 흔하다. 특히 피험자의 자발적 중단, 일시적 이탈, 채혈 지연, 시료 분실 등 복합적인 원인으로 인해 시간대별 약물 농도 데이터가 부분적으로 누락되거나 측정이 조기 종료되는 경우가 빈번히 보고된다. 이러한 결측 구조는 일반적으로 Interrupted PK Profile(간헐적 누락, 그림 1)과 Truncated PK Profile(후반부 절단, 그림 2)이 혼합된 양상으로 나타나며, 기존의 결측 분류(MCAR, MAR, MNAR)만으로는 포괄하기 어려운 혼합형 결측 구조에 해당한다.

이와 같은 결측은 약동학 곡선의 상승기와 하강기 모두에 영향을 미칠 수 있으며, 특히 약동학 파라미터 지표의 정확한 추정에 심각한 영향을 줄 수 있다. 곡선의 중간 구간에서 결측이 발생하면 보간이 필요하고, 후반부 시점이

누락되면 외삽 구간이 길어져 추정의 불확실성이 이중으로 확대된다. 이로 인해 분석 결과의 분산 증가는 물론, 생물학적 동등성 평가의 핵심 지표인 기하평균비(GMR) 및 신뢰구간 폭 변화로 이어질 수 있다.

또한 혼합형 결측은 단순히 결측의 존재 유무를 넘어서, 결측의 위치, 시계열적 패턴의 연속성, 그리고 약동학적으로 중요한 구간과의 중첩 여부에 따라 분석에 미치는 영향이 상이하다. 따라서 이러한 시나리오에서는 단순 대체나 삭제 방식보다는, 결측 발생 위치와 패턴을 반영할 수 있는 정교한 모델 기반 접근법(MI, MLE, XGBoost 등)이 보다 적절하다. 특히 다중 대체법과 최대우도추정법은 곡선 전반의 패턴을 보존하면서 결측을 처리할 수 있어, 복합 결측 시나리오에 강건한 방법론적 대안이 될 수 있다.

결론적으로, 후반부 절단된 프로파일과 간헐적으로 누락된 프로파일이 혼합된 결측 구조는 실제 임상시험에서 빈번히 관찰되는 현실적인 결측 형태로, 이를 명확히 정의하고 정량적으로 평가할 수 있는 기준 마련이 필요하다. 생물학적 동등성 분석 시에도 단순한 결측률만 고려할 것이 아니라, 결측의 위치, 지속 시간, 시계열적 특성 등을 함께 고려하여 분석 설계를 정교화하는 노력이 요구된다.

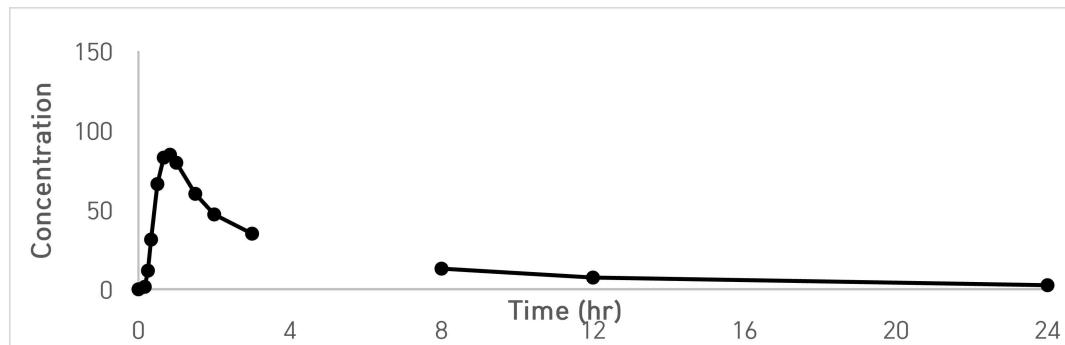


그림1. Interrupted PK Profile

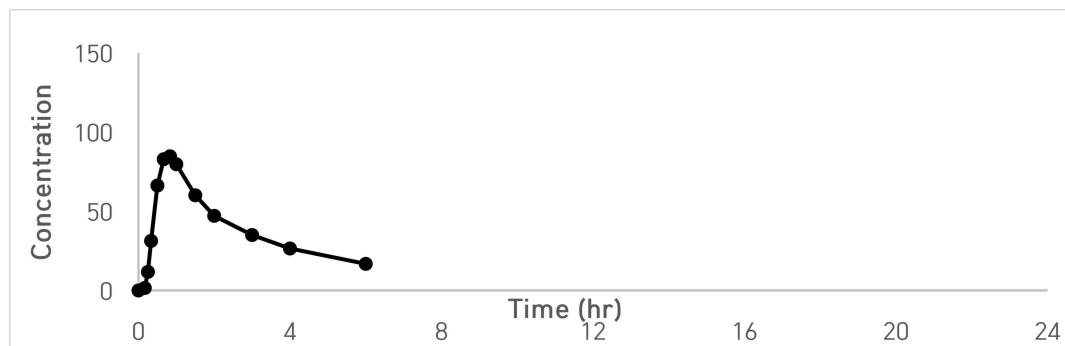


그림2. Truncated PK Profile

2) 시뮬레이션

본 연구는 결측치 메커니즘 및 결측률이 약동학적 분석 결과에 미치는 영향을 평가하기 위하여 시뮬레이션 기반 접근을 활용하였다. 시뮬레이션은 실제 임상시험 환경에서 발생할 수 있는 다양한 결측 구조와 오차 조건을 통제된 상황에서 반복적으로 구현할 수 있다는 점에서, 분석 전략의 타당성을



검증하고 민감도를 비교 평가하는 데 매우 유용한 도구이다. 실제 임상시험에서는 피험자의 중도 탈락, 시료 누락, 분석기기 오류 등 다양한 요인으로 인해 결측이 빈번히 발생하며, 약동학적 파라미터나 잔차 오차 구조 또한 피험자마다 상이하게 나타나는 경우가 많다. 이러한 현실적 제약 하에서는 단일 분석 결과에 대한 해석이 신뢰성을 담보하기 어려우며, 분석 방법 간 성능 차이를 정량적으로 비교할 필요가 있다.

시뮬레이션에 사용된 혈중 농도 데이터는 단일 구획 모델과 1차 흡수 및 제거 과정을 기반으로 생성하였다. 약물 농도는 다음의 약동학 수식을 통해 계산하였다(Gabrielsson et al., 2016; Rowland et al., 2011; Shargel et al., 2012).

$$C(t) = \frac{F \times D \times k_a}{V \times (k_a - k_e)} \times (e^{-k_e \times t} - e^{-k_a \times t})$$

여기서, $C(t)$ 는 투여 후 t 시간의 혈중 농도, F 는 생체이용률, D 는 투여 용량, k_a 는 흡수 속도 상수, k_e 는 제거 속도 상수, V 는 분포용적, t 는 투여 후 경과 시간을 의미한다.

관찰 농도는 이론 농도에 측정 오차를 반영하여 생성되었으며, 오차 구조는 약물 농도에 비례하는 곱셈 오차 모형을 기반으로 구성하였다. 고농도 구간에서는 상대적으로 큰 오차를, 저농도 구간에서는 최소 오차를 보장하기 위하여

표준편차의 하한값을 2 ng/mL로 고정하였다. 관찰값은 다음 수식으로 정의된다.

$$C_{obs}(t) = \max(0, C(t) + \varepsilon), \quad \varepsilon \sim N(0, \sigma^2), \quad \sigma = \max(C(t) \cdot CV, 2)$$

관측 농도 수식 $C_{obs}(t) = \max(0, C(t) + \varepsilon)$ 는 시뮬레이션을 통한 약물 농도 생성 시 현실적인 측정 오차 및 음수 농도 방지를 동시에 반영하기 위한 구조이다. 먼저, $C(t) + \varepsilon$ 는 이론적으로 계산된 농도 $C(t)$ 평균이 0이고 분산이 σ^2 인 정규분포를 따르는 잔차 오차 ε 를 더한 값으로, 실제 측정 과정에서의 생리학적 변동성과 기기 오차를 모사한다. 이때, 표준편차 σ 는 $C(t) \cdot CV$ 와 2 중 큰 값을 택하여 저농도 구간에서도 일정 수준 이상의 오차를 유지하게 하며, 이는 과소분산 방지 및 LLOQ 기반의 현실적 분석 조건을 고려한 설정이다. 마지막으로, 전체 항에 대해 $\max(0, \cdot)$ 를 적용하여 음수로 계산되는 농도값은 모두 0으로 절단함으로써, 비물리적인 결과를 배제하고 실제 임상시험의 분석 조건에 가까운 데이터를 생성하도록 하였다.

이 수식은 Gabrielsson and Weiner (2016), Rowland and Tozer (2011) 등에서 제시된 기본적인 오차 모델을 기반으로 하되, 저농도 영역에서의 과소분산을 방지하기 위해 최소 표준편차 개념을 반영한 사용자 정의 방식이다. 이와 같이 정규분포 기반의 잔차 오차에 대해 표준편차 하한을 설정하는 접근은 약동학 시뮬레이션에서 빈번히 사용되는 현실적 오차 처리

방법 중 하나이다.

시뮬레이션은 총 24명의 가상 피험자를 대상으로 1000회 수행되었으며, 각 피험자는 TR 또는 RT 순서로 무작위 배정되었다. 약물 투여는 공복 상태에서 단회 경구 투여로 가정하였고, 혈중 농도 측정은 0시간 포함 총 16개 시점으로 설정하였다. 약동학 파라미터는 피험자 간 개체차를 반영하기 위해 k_a 는 평균 1.1(표준편차 0.2), k_e 는 평균 0.187(표준편차 0.015)의 정규분포에서 무작위로 추출하였다. 분포용적 V는 12.5로 고정하였으며, CV는 30%로 설정하여 잔차 오차를 반영하였다. 이러한 설정은 성인 인체의 평균적인 약물동태 특성과 중간 수준의 생체내 변동성을 기반으로 한다.

결측 메커니즘은 실제 임상시험에서 발생 가능한 다양한 결측 원인을 반영하여 MCAR, MAR, MNAR의 세 가지로 구분하여 적용하였다. MCAR는 분석기기 오류 또는 시료 분실 등 피험자의 특성과 무관하게 무작위로 결측값이 발생하는 구조로, 전체 데이터 중 일정 비율(5%, 10%, 20%)을 임의로 결측 처리하여 구현하였다. MAR은 k_a 가 낮은 피험자를 중심으로 탈락 가능성이 높다고 가정하고, 이들을 대상으로 시간대별 탈락 확률 분포를 적용하여 중도 시점 이후의 농도값을 결측 처리하였다. MNAR은 T8 시점의 관측값이 낮은 피험자를 대상으로 이후 시간대의 모든 농도를 결측 처리함으로써, 실제 관측값 자체가 결측 발생에 영향을 미치는 구조를

모사하였다.

결측치 처리 후, 각 피험자의 AUC_t 및 C_{max} 는 각각 trapezoidal rule과 최대 농도를 통해 산출되었으며, 로그 변환 후 선형 혼합효과모형을 적용하여 기하평균비 및 90% 신뢰구간을 계산하였다. 분석 결과가 동등성 판정 기준(80.00~125.00%) 내에 포함될 경우 동등으로, 포함되지 않을 경우 비동등으로 판단하였다.

분석 성능 지표로는 Type I Error와 Power를 활용하였다. Type I Error는 실제로 두 제제가 동등함에도 불구하고 동등성 기준을 벗어나 비동등으로 잘못 판단된 비율을 의미하며, 본 연구에서는 $F = 1.0$ 조건에서 시뮬레이션을 반복 수행하고 동등 판정을 받지 못한 비율로 계산하였다(Yu et al., 2018).

$$C_{Ref}(t) = \frac{1.0 \times \text{Dose} \times k_a}{V(k_a - k_e)} (e^{-k_e t} - e^{-k_a t}), \quad C_{Test}(t) = \frac{1.0 \times \text{Dose} \times k_a}{V(k_a - k_e)} (e^{-k_e t} - e^{-k_a t})$$

반면 Power는 실제로 두 제제가 비동등함에도 불구하고 정확히 이를 감지한 비율, 즉 통계적 검정력이 있는지를 평가하는 지표로, $F = 0.8$ 조건에서 동등성 기준을 벗어난 비율을 통해 산출하였다(Zhang et al., 2023).

$$C_{Ref}(t) = \frac{1.0 \times \text{Dose} \times k_a}{V(k_a - k_e)} (e^{-k_e t} - e^{-k_a t}), \quad C_{Test}(t) = \frac{0.8 \times \text{Dose} \times k_a}{V(k_a - k_e)} (e^{-k_e t} - e^{-k_a t})$$

또한 각 결측 메커니즘(MCAR, MAR, MNAR), 결측률(5%, 10%, 20%),



생체내 변동계수($CV = 30\%, 40\%, 50\%$) 및 피험자 수($n = 24, 48, 72$)의 조합에 따라 시뮬레이션을 총 100회 반복 수행하였으며, 분석 방법(Complete Case Analysis, Mean Imputation, Multiple Imputation, Maximum Likelihood Estimation)별 성능을 비교하였다. 시뮬레이션 결과는 heatmap 형태로 시각화하여 각 처리 방법의 민감도 및 보수성 차이를 직관적으로 비교하였다.

본 연구의 시뮬레이션 설계 및 결측치 처리 방법은 미국 식품의약국(FDA, 2001) 가이드라인과 기존 약동학 문헌에서 널리 채택된 표준 방법론에 기반하여 수행되었으며, 실제 임상시험 설계 및 분석 전략 수립 시 고려 가능한 현실적 구조를 반영하였다.



5) 자료구성 및 데이터 특성

본 연구는 무작위배정, 공개, 2×2 교차 설계의 생물학적 동등성 시험 자료를 기반으로 하였다. 해당 시험은 공복 상태에서 단회 경구 투여 후 약동학적 특성을 평가하는 방식으로 진행되었으며, 총 48명의 건강한 성인이 참여하였다. 피험자는 두 군 중 하나에 무작위로 배정되었고, 각 군은 두 개의 시험기로 구성된 교차 설계에 따라 시험약과 대조약을 각각 투여받았다. 두 시험기 사이에는 잔류 약물의 영향을 방지하기 위한 휴약기가 설정되었다.

시험은 공복 상태에서 단회 경구 투여되었으며, 투약 후 일정 시간 간격으로 혈액을 채취하여 혈중 약물 농도를 측정하였다. 약동학적 분석을 위해 각 시험기마다 총 16회의 혈액 채취가 이루어졌고, 채혈 시점은 투약 전(0시간)을 포함하여 사전에 정의된 시간 간격에 따라 설정되었다.

본 연구에서 발생한 결측치는 다음과 같은 미씽 메커니즘을 포함할 가능성이 있다:

1 . 완전 무작위 결측 (MCAR, Missing Completely at Random)

실험 기기 오류, 샘플 보관 문제 등의 기술적 원인으로 인해 무작위로 데이터가 누락되는 경우이다.

2 . 조건부 무작위 결측 (MAR, Missing at Random)

피험자가 임상시험용 의약품의 투여 중단을 요청하거나, 시험 참여 동의를 철회하는 경우에 해당할 가능성이 있다. 또한 특정 인구 집단(예: 부작용 경험자, 특정 연령대)에서 결측이 집중될 가능성이 있으며, 중도 탈락자 중 피험자가 자발적으로 시험을 중단한 경우 MAR에 해당할 가능성이 높다.

3. 비무작위 결측 (MNAR, Missing Not at Random)

시험 책임자(또는 위임을 받은 시험 담당 의사)가 피험자의 안전을 고려하여 시험을 중단한 경우에 해당한다. 이는 주로 심각한 이상반응을 경험한 피험자가 포함될 가능성이 크며, 생물학적 동등성 평가에 중요한 영향을 미칠 수 있다. 중도 탈락자 중 의료진이 피험자의 안전을 위해 시험을 중지한 경우 MNAR일 가능성이 높다.

연구 데이터에서 피험자의 중도 탈락 사유를 분석한 결과, 일부는 자발적으로 동의를 철회한 경우로 MAR로 간주될 수 있다. 반면, 일부는 안전성 문제로 인해 연구진에 의해 중도 중단되었으며, 이는 MNAR 특성을 가질 가능성이 있다. 이러한 결측 메커니즘을 고려하여 적절한 결측치 처리 방법을 적용하고, 분석 결과에 미치는 영향을 비교·평가하는 것이 중요하다.

6) 결측치 처리 방법 및 적용

분석에 사용된 피험자들의 시간대별 약물 농도값의 결측 여부를 시각화한 헛맵의 예시이다(그림 3). 세로축은 시험대상자 번호를, 가로축은 채혈 시간을 나타내며, 색이 채워진 영역은 관측값이 존재하는 구간을, 공백은 결측된 구간을 의미한다.

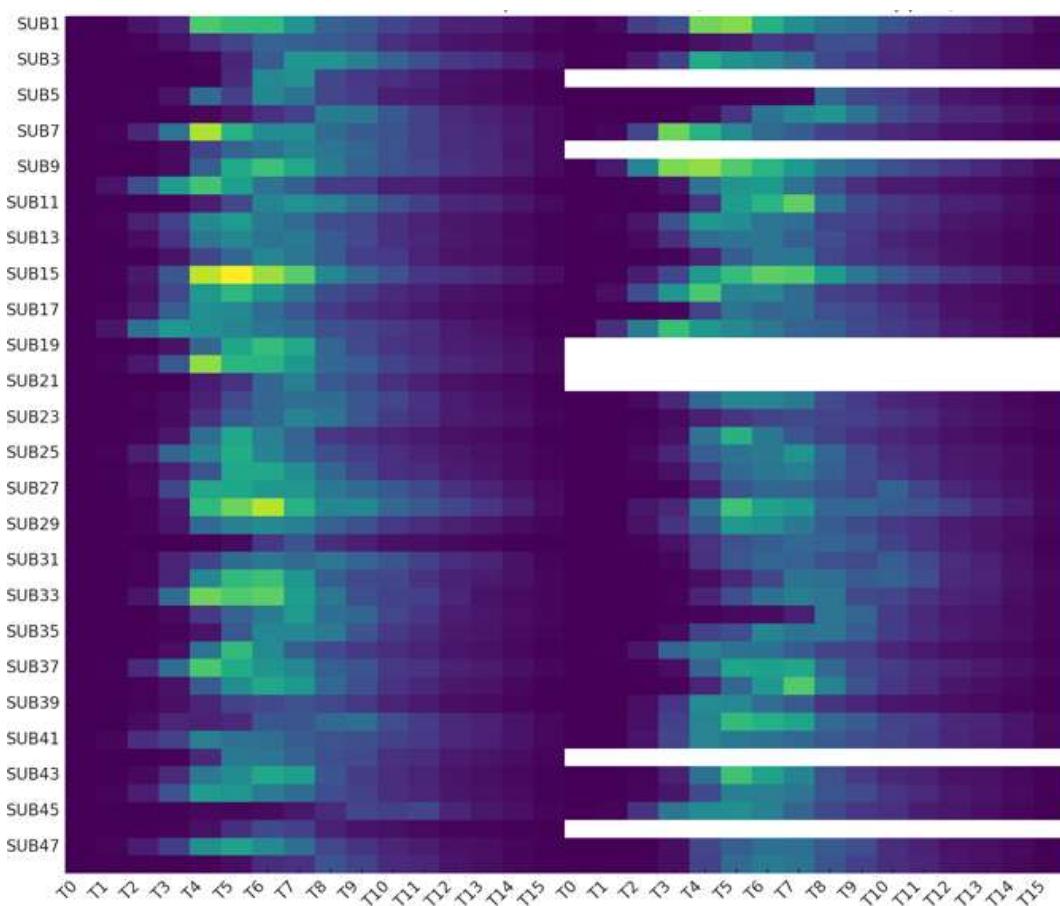


그림3. 시간대별 약물 농도값의 결측 분포 헛맵

완전 케이스 분석(CCA)는 결측치가 발생한 7명의 피험자를 분석에서 완전히 제외하고, 결측이 없는 41명의 데이터만을 활용하여 분석을 수행하였다. 결측값을 포함한 데이터는 고려하지 않고 제거함으로써 분석의 일관성을 유지하는 방식이다.

평균 대체법(Mean Imputation)는 결측이 발생한 7명의 피험자에 대해 동일한 시간대에서 관측된 다른 피험자들의 평균값으로 결측치를 대체하였다. 즉, 특정 시간 지점에서 약물 농도가 누락된 경우, 같은 시간대에 측정된 나머지 피험자들의 약물 농도의 평균값을 계산하여 해당 피험자의 결측값을 보완하였다. 이를 통해 데이터셋의 완전성을 유지하면서도 분석을 수행할 수 있도록 하였다.

다중 대체법(MI)은 결측이 발생한 7명의 피험자에 대해 단일 값이 아닌 5개의 대체값(MI1~MI5)을 생성하여 결측을 보완하였다. 이는 확률적 모델을 기반으로 동일한 시간대에서 약물 농도의 분포를 고려하여 잠재적인 대체값을 생성하는 방식으로, 각 결측값에 대해 5개의 독립적인 값이 생성되었다. 이후 이 다섯 개의 데이터셋을 각각 분석한 후 최종적으로 통합된 결과를 도출하였다.

최대 우도 추정(MLE)는 결측값을 직접 대체하지 않고, 전체 관측 데이터를 기반으로 가능성 함수(likelihood)를 구성하여 결측값을 내재적으로 보정하는

방식이다. 본 연구에서는 결측이 발생한 7명을 포함한 전체 48명의 데이터를 대상으로, SAS의 PROC MIXED 프로시저를 이용해 혼합효과모형을 구축하였다. 제제(treatment), 시기(period), 순서(sequence)를 고정효과로, 피험자 간 차이를 무작위효과로 설정하였다. 분석은 결측값을 별도로 대체하지 않고 모델 기반으로 직접 추정하여 수행되었으며, 모든 피험자를 포함한 상태에서 약동학적 지표 및 생물학적 동등성 평가가 가능하도록 하였다.

XGBoost 회귀모형은 결측이 발생한 7명의 피험자에 대해 누락된 약물 농도를 예측하여 보완하는 데 사용되었다. 시간, 제제구분, 인접 시간대의 약물 농도 등을 입력 변수로 설정하고, 결측이 없는 데이터를 학습용으로 사용하여 모델을 훈련하였다. 이후 학습된 모델을 통해 각 피험자의 결측 시점 약물 농도를 예측하고, 해당 값으로 대체하여 전체 데이터셋을 완성하였다. 이를 바탕으로 약동학적 지표와 생물학적 동등성 평가를 수행하였다.

MI(XGBoost)는 XGBoost 회귀모형을 활용하여 결측이 발생한 7명의 피험자에 대해 누락된 약물 농도를 예측하고, 이를 기반으로 5개의 완전한 데이터셋(MI1~MI5)을 구성하였다. 예측에는 시간, 제제구분, 인접 시간대의 약물 농도 등을 입력 변수로 활용하였으며, 결측이 없는 피험자의 데이터를

학습용으로 사용하여 모델을 훈련하였다. 이후 동일한 예측 모형을 통해 생성된 예측값을 반복적으로 적용하여 5개의 대체 데이터셋을 만들고, 각각의 데이터에 대해 분석을 수행한 뒤 결과를 통합할 수 있도록 하였다.

7) 분석 설계 및 주요 평가 지표

본 연구는 1상 생물학적 동등성 시험에서 발생하는 결측치 처리 방법이 주요 약동학적 지표 및 생물학적 동등성 평가 결과에 미치는 영향을 분석하고, 최적의 결측치 처리 방법을 도출하는 것을 목표로 한다. 생물학적 동등성 시험에서 결측치는 분석 결과의 신뢰성과 정확성을 저하시킬 수 있으므로, 이를 적절히 처리하는 것이 필수적이다. 이를 위해 본 연구에서는 AUC_t , AUC_{inf} , C_{max} , T_{max} , $t_{1/2}$ 등의 주요 약동학적 지표를 대상으로 완전 케이스 분석, 평균 대체법, 다중 대체법, 최대우도추정법, XGBoost 기반 머신러닝 기법, 그리고 이를 활용한 다중 대체 방식(MI(XGBoost)) 등 여섯 가지 결측치 처리 방법을 적용하여 결과를 비교·분석하였다. 또한, 90% 신뢰구간이 80.00%~125.00% 범위 내에 포함되는지를 기준으로 생물학적 동등성 판정 결과의 변화를 확인하였으며, 잔류효과, 순서 효과, 시기 효과를 검정하여 결측치 처리 방법이 실험 설계 요인에 미치는 영향을 분석하였다. 이를 위해 혼합효과 모델(Mixed Effects Model)을 활용하여 고정효과(Fixed Effects)와 무작위효과(Random Effects)를 고려하고, 분산분석(ANOVA)을



통해 각 처리 방법에 따른 데이터 변동성을 평가하였다. 본 연구를 통해 생물학적 동등성 시험에서 결측치를 보다 효과적으로 처리할 수 있는 방법을 제안하고, 분석 결과의 신뢰성과 검정력을 높이는 데 기여하고자 한다.

III. 연구 결과

1. PK 파라미터 및 90% 신뢰구간

본 연구에서는 생물학적 동등성 시험에서 발생한 결측치를 처리하는 방법에 따라 주요 약동학적 지표의 추정값과 생물학적 동등성 평가의 핵심 지표인 기하평균비율(GMR) 및 이에 대한 90% 신뢰구간이 어떻게 달라지는지를 정량적으로 분석하였다. 비교 대상이 된 결측치 처리 방법은 완전 케이스 분석(CCA), 평균 대체법(Mean), 다중 대체법(MI), 최대우도추정법(MLE), 머신러닝 기반의 XGBoost, 그리고 이를 활용한 다중 대체 방식인 MI(XGBoost)까지 총 여섯 가지였다. 생물학적 동등성 평가는 AUC_t 및 C_{max} 의 기하평균비율과 90% 신뢰구간을 통해 수행되었으며, 통상적인 동등성 범위(80.00%~125.00%) 내 포함 여부로 판정하였다.

AUC_t 기준으로, CCA 방법은 GMR이 101.6%, 90% 신뢰구간은 97.83~105.57%로 가장 좁은 신뢰구간을 보였다. 이는 결측값이 포함된 피험자를 제외하여 표본 수가 줄고 전체 분산이 축소된 결과로 해석된다. MI 방법은 GMR이 102.7%, 신뢰구간은 97.52~108.24%로, 결측으로 인한 불확실성을 반영하면서도 기준 범위 내에서 안정적인 추정값을 제공하였다.

Mean 방법은 GMR이 103.8%, 신뢰구간은 99.72~108.08%로 평균 대체에 따른 분산 과소 추정으로 신뢰구간이 실제보다 좁게 형성된 경향이 있었다. MLE 방법은 GMR이 106.6%, 신뢰구간은 100.63~112.89%로 개별 피험자 반응을 모형에 반영한 결과로 비교적 넓은 신뢰구간을 보였다. XGBoost 방법은 GMR이 103.6%, 신뢰구간은 91.83~116.89%로 여섯 방법 중 가장 넓은 신뢰구간을 보였으며, 하한값이 기준 범위 하단에 근접하여 생물학적 동등성 판정에 있어 보수적 해석이 필요하다. MI(XGBoost) 방법은 GMR이 104.2%, 신뢰구간은 100.06~108.44%로 나타났으며, 반복 예측을 통한 대체값 간 분산이 억제되어 신뢰구간이 비교적 좁게 형성되었다.

C_{\max} 에서도 유사한 경향이 확인되었다. CCA 방법은 GMR이 103.3%, 신뢰구간은 93.94~113.60%로 AUC_t 와 유사하게 좁은 구간을 보였다. MI는 GMR이 106.1%, 신뢰구간은 96.51~116.73%였고, Mean은 GMR이 105.8%, 신뢰구간은 96.73~115.67%로 두 방법 모두 기준 범위 내에서 안정적인 추정값을 보였다. MLE는 GMR이 110.5%, 신뢰구간은 99.63~122.56%로 가장 높은 중앙값을 보였으며, XGBoost는 GMR이 104.3%, 신뢰구간은 90.60~120.06%로 가장 넓은 구간과 기준 하단 근접 하한을 나타냈다. MI(XGBoost)는 GMR이 106.5%, 신뢰구간은 97.08~116.46%로 나타났으며, 기준 범위 내에서 비교적 안정적인 결과를 보였다.

각 방법별 AUC_t 및 C_{max} 의 절대 평균값(Mean ± SD)은 표 1에 제시하였다. AUC_t 에서는 시험약의 평균값이 가장 높은 방법은 MLE(379.5 ± 149.5)였고, 그 뒤를 MI(369.7 ± 135.2), MEAN(367.0 ± 134.7), MI(XGBoost)(367.0 ± 134.8)가 비슷한 수준으로 따랐다. CCA는 358.0 ± 139.2였으며, XGBoost는 가장 낮은 350.1 ± 150.8을 기록하였다. 대조약에서는 MI(357.7 ± 127.8)가 가장 높았다. MLE는 352.8 ± 128.5, MEAN은 352.8 ± 126.0으로 동일한 평균값을 보였다. XGBoost는 336.1 ± 141.0으로 가장 낮았고, CCA(352.0 ± 132.6)와 MI(XGBoost)(351.7 ± 126.2)는 중간 수준이었다. 전반적으로 시험약이 대조약보다 AUC_t 평균값이 높았으며, MLE는 제제 간 차이가 가장 컼고, MEAN과 MI(XGBoost)는 차이가 작았다.

C_{max} 에서는 시험약 중 MLE(114.0 ± 38.07)가 가장 높았고, MI(111.5 ± 35.32), CCA(110.5 ± 36.12), MI(XGBoost)(110.0 ± 35.50), MEAN(109.5 ± 35.67), XGBoost(105.5 ± 41.09) 순이었다. 대조약의 경우 CCA(106.8 ± 36.22), MI(105.6 ± 35.41)가 비교적 높았고, XGBoost(100.3 ± 39.64)가 가장 낮았다. MLE(103.4 ± 35.94), MEAN(103.9 ± 35.39), MI(XGBoost)(103.9 ± 35.46)는 중간 수준이었다. 시험약은 모든 방법에서 대조약보다 높은 C_{max} 값을 보였으며, MLE에서 제제 간 차이가 가장 컸다.

AUC_{inf}는 AUC_t와 유사한 경향을 보였다. 시험약에서는 MLE(411.7 ± 172.2)가 가장 높았고, MI(400.5 ± 157.0), MEAN(397.0 ± 156.4), MI(XGBoost)(396.7 ± 156.5), CCA(385.9 ± 160.4), XGBoost(378.8 ± 172.9) 순이었다. 대조약은 MI(384.6 ± 146.5)가 가장 높았고, MEAN(378.9 ± 144.4), MLE(379.6 ± 147.1), MI(XGBoost)(377.5 ± 144.7), CCA(377.2 ± 150.6), XGBoost(360.9 ± 159.8) 순이었다. 전반적으로 AUC_{inf}에서도 시험약이 대조약보다 높은 값을 보였으며, 각 방법별 상대적 순위는 AUC_t와 거의 일치하였다.

T_{max}는 MEAN, MLE, MI 방법에서 시험약과 대조약 모두 0.83시간(0.33~2.00)으로 동일하게 나타났다. 반면, CCA 방법에서는 시험약이 0.83시간, 대조약이 0.80시간으로 소폭의 차이를 보였다. 또한 XGBoost 및 MI(XGBoost) 방법에서는 각각 시험약-대조약 간 0.86-0.82시간, 0.84-0.82시간으로 차이를 보여, 머신러닝 기반 예측 방법에서 미세한 편차가 발생했음을 시사한다.

t_{1/2}는 전반적으로 6.6~6.9시간 범위 내에 있었으며, AUC_t/AUC_{inf} 비율은 모든 방법에서 평균 약 94%로 나타나 적절한 샘플링이 이루어졌음을 시사한다. 이들 지표는 동등성 평가에 직접 사용되지는 않지만, 결과의 신뢰성을 보완하는 참고자료로 활용될 수 있다.

요약하면, CCA는 가장 좁은 신뢰구간을 보였으나, 결측값이 있는 피험자를 완전히 제외함으로써 표본 손실이 발생하고, 이에 따른 검정력 저하 가능성이 존재한다. Mean은 결측값을 단순 평균으로 대체함으로써 분산이 과소 추정되어 신뢰구간이 인위적으로 좁아질 수 있으며, 분석 결과에 왜곡을 초래할 우려가 있다. MI는 결측에 따른 불확실성을 다중 대체 방식으로 반영하면서도 기준 범위 내에서 안정적인 추정 결과를 제공하였다. MI(XGBoost) 또한 평균값과 GMR이 MI와 유사한 수준으로 나타났으나, 머신러닝 기반 예측 특성상 반복 대체 간 분산이 억제되어 신뢰구간이 상대적으로 좁게 형성되는 경향이 있었다. MLE는 결측치를 별도로 대체하지 않고 전체 자료를 기반으로 분석이 가능하여 데이터 손실 없이 분석 효율성이 높았으며, 개인 간 변이를 반영함으로써 보다 유연한 추정이 가능했다. 다만, 일부 지표에서 GMR이 비교적 높게 산출되는 경향이 있어 해석 시 보수적 접근이 필요하다. XGBoost는 데이터의 복잡한 패턴을 반영하는 데 강점을 보였지만, 신뢰구간이 가장 넓고 하한이 생물학적 동등성 기준에 근접하여 해석에 주의가 요구된다.

결론적으로, 본 연구는 동일한 임상시험 자료에 서로 다른 결측치 처리 방법을 적용할 경우 약동학적 파라미터의 추정값 및 생물학적 동등성 평가 결과에 유의한 차이가 발생할 수 있음을 실증적으로 제시하였다. 특히 MI와 MLE는 결측이 존재하는 조건에서도 통계적 신뢰성과 검정력을 균형 있게



확보할 수 있는 방법으로 평가되며, 실제 임상시험 분석에서 적용 가능성과 실용성이 높은 대안으로 판단된다. 반면, Mean과 CCA는 적용이 간편하다는 장점에도 불구하고 결과 해석의 왜곡 가능성이 존재하므로, 결측률이 높은 상황에서는 신중한 적용이 필요하다. XGBoost는 분석적 해석이나 보조 지표로는 활용 가치가 있으나, 규제기관 제출을 위한 생물학적 동등성 평가에서는 보수적이고 제한적인 활용이 요구된다. MI(XGBoost)는 머신러닝 기반 예측의 장점을 결합한 실험적 접근으로, 반복 대체 기반의 이론적 전제에는 한계가 있으나, 패턴 반영의 장점을 살린 시각적 검토나 보완적 분석 수단으로 고려할 수 있다.

표 1. PK 파라미터 및 90% 신뢰구간에서 결측치 처리 방법 별 차이

| PK Parameter (Mean±SD) | METHOD | 시험약 | 대조약 |
|-----------------------------|------------------|------------------|------------------|
| AUC_t (hr · ng/mL) | CCA | 358.0 ± 139.2 | 352.0 ± 132.6 |
| | MI | 369.7 ± 135.2 | 357.7 ± 127.8 |
| | MEAN | 367.0 ± 134.7 | 352.8 ± 126.0 |
| | MLE | 379.5 ± 149.5 | 352.8 ± 128.5 |
| | XGBOOST | 350.1 ± 150.8 | 336.1 ± 141.0 |
| AUC_{inf} (hr · ng/mL) | MI(XGBOOST) | 367.0 ± 134.8 | 351.7 ± 126.2 |
| | CCA | 385.9 ± 160.4 | 377.2 ± 150.6 |
| | MI | 400.5 ± 157.0 | 384.6 ± 146.5 |
| | MEAN | 397.0 ± 156.4 | 378.9 ± 144.4 |
| | MLE | 411.7 ± 172.2 | 379.6 ± 147.1 |
| C_{max} (ng/mL) | XGBOOST | 378.8 ± 172.9 | 360.9 ± 159.8 |
| | MI(XGBOOST) | 396.7 ± 156.5 | 377.5 ± 144.7 |
| | CCA | 110.5 ± 36.12 | 106.8 ± 36.22 |
| | MI | 111.5 ± 35.32 | 105.6 ± 35.41 |
| | MEAN | 109.5 ± 35.67 | 103.9 ± 35.39 |
| T_{max} (hr) | MLE | 114.0 ± 38.07 | 103.4 ± 35.94 |
| | XGBOOST | 105.5 ± 41.09 | 100.3 ± 39.64 |
| | MI(XGBOOST) | 110.0 ± 35.50 | 103.9 ± 35.46 |
| | CCA | 0.83 (0.33~2.00) | 0.80 (0.33~2.00) |
| | MI | 0.83 (0.33~2.00) | 0.83 (0.33~2.00) |
| MEAN | 0.83 (0.33~2.00) | 0.83 (0.33~2.00) | |
| | MLE | 0.83 (0.33~2.00) | 0.83 (0.33~2.00) |

| | XGBOOST | 0.86 (0.33~2.00) | 0.82 (0.33~2.00) | | |
|---|-------------|------------------|------------------|-------|--------|
| | MI(XGBOOST) | 0.84 (0.33~2.00) | 0.82 (0.33~2.00) | | |
| $t_{1/2}$ (hr) | CCA | 6.73 ± 1.21 | 6.52 ± 1.21 | | |
| | MI | 6.83 ± 1.50 | 6.61 ± 1.22 | | |
| | MEAN | 6.84 ± 1.50 | 6.60 ± 1.21 | | |
| | MLE | 6.88 ± 1.51 | 6.64 ± 1.28 | | |
| | XGBOOST | 6.84 ± 1.50 | 6.60 ± 1.21 | | |
| | MI(XGBOOST) | 6.80 ± 1.50 | 6.58 ± 1.22 | | |
| AUC _t /AUC _{inf} (%) | CCA | 93.7 ± 3.3 | 94.1 ± 2.8 | | |
| | MI | 93.4 ± 3.4 | 93.8 ± 3.0 | | |
| | MEAN | 93.4 ± 3.4 | 93.9 ± 3.0 | | |
| | MLE | 93.2 ± 3.5 | 93.7 ± 3.2 | | |
| | XGBOOST | 93.5 ± 3.4 | 94.0 ± 3.0 | | |
| | MI(XGBOOST) | 93.5 ± 3.4 | 93.9 ± 3.0 | | |
| | | AUC _t | C _{max} | | |
| | | LOWER | UPPER | LOWER | UPPER |
| 기하평균비율 (90% 신뢰구간) | CCA | 97.83 | 105.57 | 93.94 | 113.60 |
| | MI | 97.52 | 108.24 | 96.51 | 116.73 |
| | MEAN | 99.72 | 108.08 | 96.73 | 115.67 |
| | MLE | 100.63 | 112.89 | 99.63 | 122.56 |
| | XGBOOST | 91.83 | 116.89 | 90.60 | 120.06 |
| | MI(XGBOOST) | 100.06 | 108.44 | 97.08 | 116.46 |

2. 잔류효과 확인

본 연구에서는 2×2 교차설계로 실시된 생물학적 동등성 시험 자료를 바탕으로, 결측치 처리 방법에 따라 잔류효과 분석 결과가 어떻게 달라지는지를 평가하였다. 잔류효과 검정은 약물 복용 순서, 시험 시기, 그리고 시험약과 대조약 간 차이를 각각 고정효과로 포함한 혼합효과모형을 이용하여 수행하였으며, 분석 지표로는 AUC_t 와 C_{max} 가 사용되었다.

AUC_t 기준으로 볼 때, 대부분의 결측치 처리 방법에서는 세 고정효과 모두에서 통계적으로 유의한 결과가 관찰되지 않았다(표 2).

CCA 방법에서는 sequence($p = 0.9736$), period($p = 0.5225$), treatment($p = 0.4792$)로 모든 고정효과에서 유의한 차이가 나타나지 않았다.

MI 방법에서는 sequence($p = 0.6087$), period($p = 0.4984$), treatment($p = 0.4953$)로 모든 고정효과에서 유의한 차이가 나타나지 않았다.

Mean 방법에서는 sequence($p = 0.6629$), period($p = 0.5647$), treatment($p = 0.1249$)로 모든 고정효과에서 유의한 차이가 나타나지 않았다.

MLE 방법에서는 sequence($p = 0.8528$), period($p = 0.3034$),

treatment($p = 0.0689$)로 모든 고정효과에서 유의한 차이가 나타나지 않았다.

XGBoost 방법에서는 sequence($p = 0.6908$), treatment($p = 0.6244$)는 유의하지 않았으나, period($p = 0.0342$) 효과는 통계적으로 유의하게 나타났다. 이는 머신러닝 기반 예측 모델이 시기 간 미세한 차이에 민감하게 반응할 수 있음을 시사하며, 데이터의 시계열적 패턴에 민감한 XGBoost 알고리즘 특성이 반영된 결과로 해석될 수 있다.

MI(XGBoost) 방법에서는 sequence($p = 0.6867$), period($p = 0.6574$), treatment($p = 0.0946$)로 모든 고정효과에서 유의한 차이가 나타나지 않았다.

C_{\max} 를 기준으로 한 분석에서는 모든 결측치 처리 방법에서 sequence, period, treatment 효과가 통계적으로 유의하지 않았다(표 3).

CCA 방법에서는 sequence($p = 0.7195$), period($p = 0.2390$), treatment($p = 0.5674$)로 모든 고정효과에서 유의한 차이가 나타나지 않았다.

MI 방법에서는 sequence($p = 0.5457$), period($p = 0.1618$), treatment($p = 0.3339$)로 모든 고정효과에서 유의한 차이가 나타나지 않았다.

Mean 방법에서는 sequence($p = 0.4985$), period($p = 0.3074$), treatment($p = 0.2972$)로 모든 고정효과에서 유의한 차이가 나타나지 않았다.

MLE 방법에서는 sequence($p = 0.8670$), period($p = 0.1817$), treatment($p = 0.1125$)로 모든 고정효과에서 유의한 차이가 나타나지 않았다.

XGBoost 방법에서는 sequence($p = 0.5164$), period($p = 0.2713$), treatment($p = 0.6183$)로 모든 고정효과에서 유의한 차이가 나타나지 않았다.

MI(XGBoost) 방법에서는 sequence($p = 0.5360$), period($p = 0.2722$), treatment($p = 0.2624$)로 모든 고정효과에서 유의한 차이가 나타나지 않았다.

이와 같이, 결측치 처리 방법에 따라 AUC_t 기준에서는 특정 효과(XGBoost의 period effect)에서 통계적으로 유의한 결과가 관찰된 반면, C_{max} 는 전반적으로 결측치 처리 방법의 영향을 받지 않고 안정적으로 나타났다.

이러한 결과는 AUC_t 가 약물의 체내 노출량을 누적적으로 반영하는 지표로서, 결측치 처리 방식에 따라 보다 민감하게 반응할 수 있음을



시사한다. 특히 시기효과와 제제효과는 결측 보완 과정에서 약물 농도 곡선의 후반부가 영향을 받을 수 있어, 결측이 후기 시점에 집중된 경우 분석 결과에 왜곡이 발생할 수 있다. 반면 C_{max} 는 단일 시점의 최고 농도에 기반하므로, 결측이 T_{max} 이후에 집중된 경우에는 분석에 미치는 영향이 상대적으로 작다. 따라서 생물학적 동등성 시험에서는 결측치 처리 방식에 따른 자표별 민감도 차이를 이해하고, 분석 결과를 해석할 때 이러한 맥락을 충분히 고려하는 것이 중요하다.

표 2. 잔류효과에 대한 결측치 처리 방법 별 차이(AUC_t)

| Effect | METHOD | Den DF | F-Value | P-value |
|-----------------------|-----------|--------|---------|---------|
| CCA | Sequence | 39 | 0.00 | 0.9736 |
| | Period | 39 | 0.42 | 0.5225 |
| | Treatment | 39 | 0.51 | 0.4792 |
| MI | Sequence | 45 | 0.31 | 0.6087 |
| | Period | 45 | 0.83 | 0.4984 |
| | Treatment | 45 | 1.09 | 0.4953 |
| AUC _t MEAN | Sequence | 45 | 0.19 | 0.6629 |
| | Period | 45 | 0.34 | 0.5647 |
| | Treatment | 45 | 2.44 | 0.1249 |
| MLE | Sequence | 45 | 0.03 | 0.8528 |
| | Period | 45 | 1.08 | 0.3034 |
| | Treatment | 45 | 3.47 | 0.0689 |
| XGBOOST | Sequence | 45 | 0.16 | 0.6908 |
| | Period | 45 | 4.76 | 0.0342 |
| | Treatment | 45 | 0.24 | 0.6244 |



| | | | | |
|-----------------|-----------|----|------|--------|
| MI (XGBOOST) | Sequence | 45 | 0.16 | 0.6867 |
| | Period | 45 | 0.20 | 0.6574 |
| | Treatment | 45 | 2.91 | 0.0946 |

표 3. 잔류효과에 대한 결측치 처리 방법 별 차이 (C_{max})

| Effect | METHOD | | Den DF | F-Value | P-value |
|-----------|-----------|--|--------|---------|---------|
| CCA | Sequence | | 39 | 0.13 | 0.7195 |
| | Period | | 39 | 1.43 | 0.2390 |
| | Treatment | | 39 | 0.33 | 0.5674 |
| MI | Sequence | | 45 | 0.47 | 0.5457 |
| | Period | | 45 | 2.28 | 0.1618 |
| | Treatment | | 45 | 1.20 | 0.3339 |
| C_{max} | Sequence | | 45 | 0.47 | 0.4985 |
| | Period | | 45 | 1.07 | 0.3074 |
| | Treatment | | 45 | 1.11 | 0.2972 |
| MLE | Sequence | | 45 | 0.03 | 0.8670 |
| | Period | | 45 | 1.84 | 0.1817 |
| | Treatment | | 45 | 2.62 | 0.1125 |
| XGBOOST | Sequence | | 45 | 0.43 | 0.5164 |
| | Period | | 45 | 1.24 | 0.2713 |
| | Treatment | | 45 | 0.25 | 0.6183 |



| | | | | |
|-----------------|-----------|----|------|--------|
| MI (XGBOOST) | Sequence | 45 | 0.39 | 0.5360 |
| | Period | 45 | 1.24 | 0.2722 |
| | Treatment | 45 | 0.29 | 0.2624 |



3. 혈장 중 시간대별 평균 농도 그래프

혈장 중 시간대별 평균 농도 그래프는 약물의 흡수, 분포 및 제거 과정을 시각적으로 확인할 수 있는 가장 직관적인 도구로서, 생물학적 동등성 시험에서는 약동학적 지표들의 수치적 해석뿐 아니라 약물의 시간에 따른 동태를 시각적으로 비교하는 데 필수적이다.

먼저 CCA는 결측치가 있는 피험자를 완전히 제외하고 총 42명의 피험자만을 대상으로 평균 곡선을 산출하였다(그림 4). 그 결과, 초기 흡수 구간에서는 약물 농도가 빠르게 상승한 후 T_{max} 를 지나 완만하게 감소하는 전형적인 농도-시간 곡선 형태를 보였으며, 표본 수가 적어 전체적인 변동성이 작고 곡선이 부드럽게 표현되는 경향이 관찰되었다. 평균 대체법은 각 시간대의 평균값으로 결측을 단순 보완함으로써 시간에 따른 농도 변화의 분산이 인위적으로 감소하였고, 이에 따라 곡선이 상대적으로 평탄하게 나타났다. 이는 실제 생체 내 동태를 충분히 반영하지 못한다는 점에서 해석에 주의가 필요하다.

MI는 모든 피험자의 결측치를 다중 대체를 통해 보완하고, 각각의 대체 결과를 통합하여 평균 곡선을 생성하였다(그림 5). 곡선의 전체적인 형태는 CCA와 유사하게 나타났으나, T_{max} 이후 구간에서는 보다 완만한 하강 곡선과 넓은 신뢰 구간을 보였다. 이는 결측값에 대한 불확실성을 충실히 반영한



결과로 해석되며, 시간 경과에 따른 체내 농도 변화에 대한 보다 현실적인 예측이 가능하다는 점에서 의미가 있다.

Mean은 각 시간대의 평균값으로 결측을 보완하여 평균 곡선을 생성하였다(그림 6). 시간에 따른 농도 변화의 분산이 인위적으로 감소하면서, 곡선이 상대적으로 평坦하게 나타나는 경향을 보였고, 이는 실제 생체 내 동태를 충분히 반영하지 못할 수 있어 해석에 주의가 필요하다.

MLE는 혼합효과모형을 기반으로 개별 피험자의 반응 차이를 추정하여 분석한 결과, 평균 곡선이 실측치 기반의 평균 농도 곡선과 가장 근접한 형태로 나타났다(그림 7). 이는 개별 수준의 정보를 최대한 활용함으로써 곡선의 정밀도와 예측 가능성성이 높아졌음을 의미하며, 분석 결과의 신뢰성을 높이는 데 기여할 수 있다.

XGBoost는 머신러닝 기반의 예측 기법으로, 채혈 시점, 혈중 농도, 투약 정보 등 입력 특성 간의 복잡한 비선형 상호작용을 반영하여 평균 곡선을 도출하였다(그림 8). 대부분의 구간에서 임상시험 데이터와 유사한 패턴을 유지하였으나, T_{max} 이후 곡선의 후반부에서는 tail effect가 발생하였다. 이는 알고리즘이 이상치나 비정형적 패턴에 과도하게 민감하게 반응했을 가능성을 시사하며, 높은 예측 민감도를 가지는 동시에 이상값에 대한 반응성이 커 해석에는 신중을 기할 필요가 있다.



XGBoost 기반 MI는 머신러닝 회귀모형을 통해 결측 값을 예측하여 반복 대체한 후, 이 결과를 통합하여 평균 곡선을 도출하였다(그림 9). 임상시험 데이터와 유사한 전반적 패턴을 유지하였으나, T_{\max} 이후 구간에서는 tail effect가 관찰되었다. 이는 알고리즘이 이상값이나 비정형적 패턴에 민감하게 반응했을 가능성을 시사한다. 특히 반복 대체 간 예측값의 다양성이 제한되기 때문에, 평균 곡선은 부드럽고 신뢰구간은 상대적으로 좁게 나타날 수 있으며, 이는 Rubin's Rule 기반의 통계적 추정과는 괴리가 있음을 시사한다. 따라서 해당 접근은 시각적 비교를 위한 보조 도구로 해석되어야 하며, 이론적 MI의 전제 충족 여부와는 구별되어야 한다.

이처럼 각 방법별로 생성된 시간대별 평균 농도 그래프는 결측치 처리 방식의 차이가 분석 결과에 미치는 영향을 직관적으로 보여준다. 따라서 생물학적 동등성 시험 설계 및 분석 시, 결측치를 단순히 대체하는 것을 넘어 곡선 형태 및 신뢰구간의 특성을 함께 고려한 시각적 검토가 중요한 평가 기준으로 활용되어야 할 것이다.

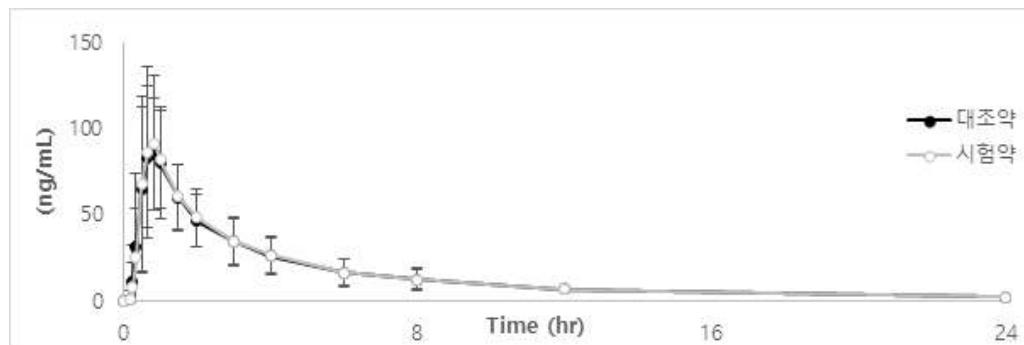


그림 4. CCA에 대한 혈장 중 시간대별 평균 농도 그래프

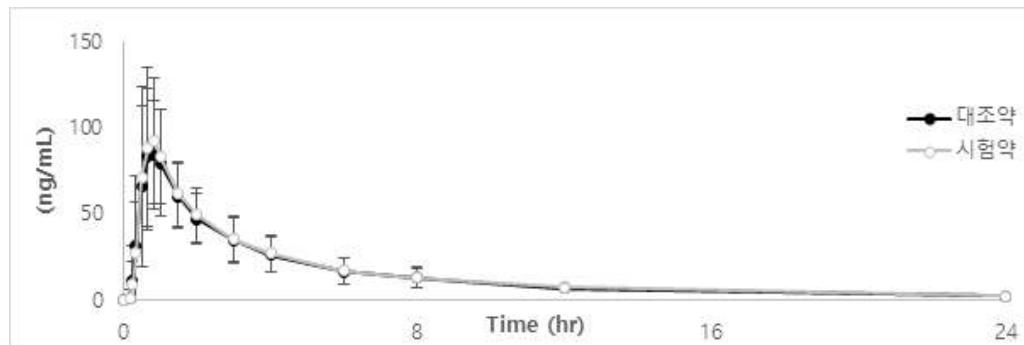


그림 5. Multiple Imputation에 대한 혈장 중 시간대별 평균 농도 그래프

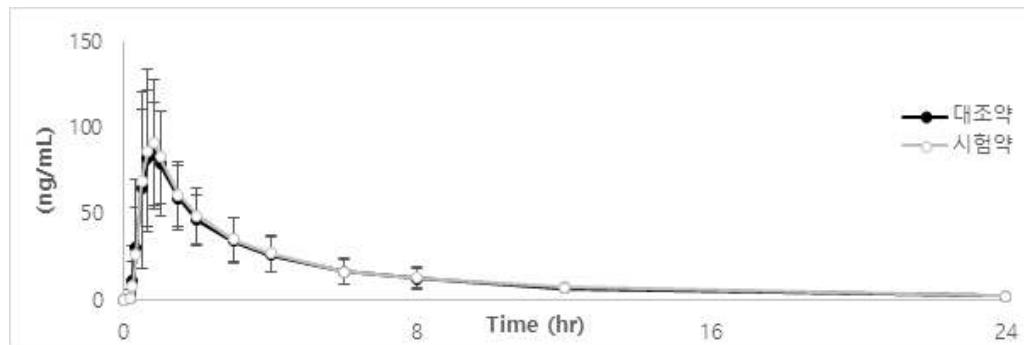


그림 6. Mean Imputation에 대한 혈장 중 시간대별 평균 농도 그래프

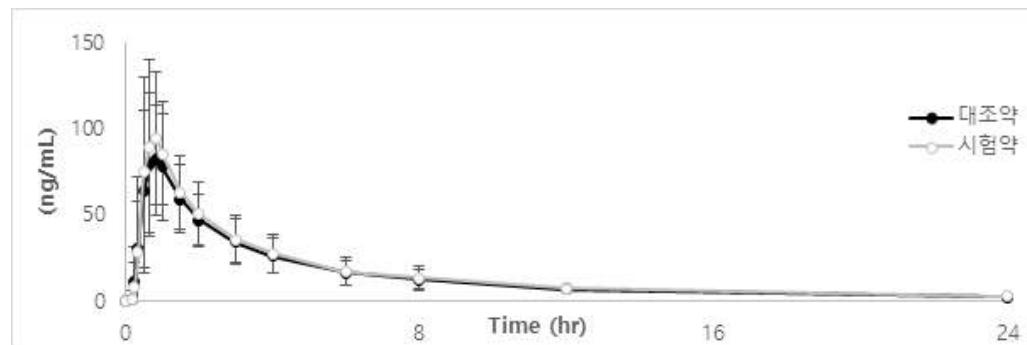


그림 7. MLE에 대한 혈장 중 시간대별 평균 농도 그래프

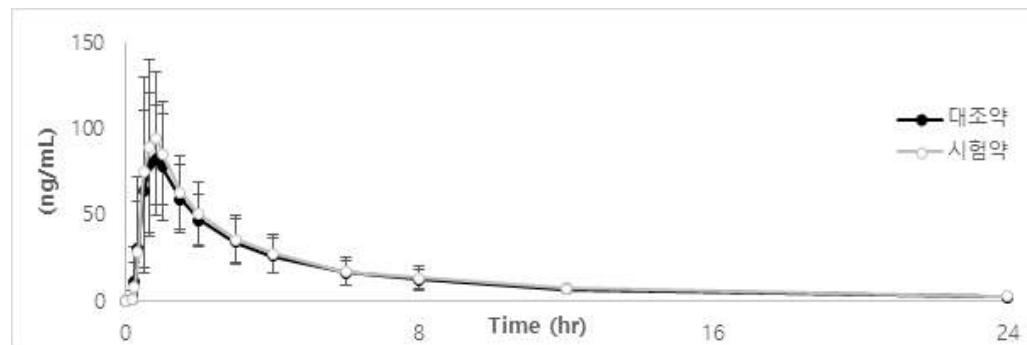


그림 8. XGBOOST에 대한 혈장 중 시간대별 평균 농도 그래프

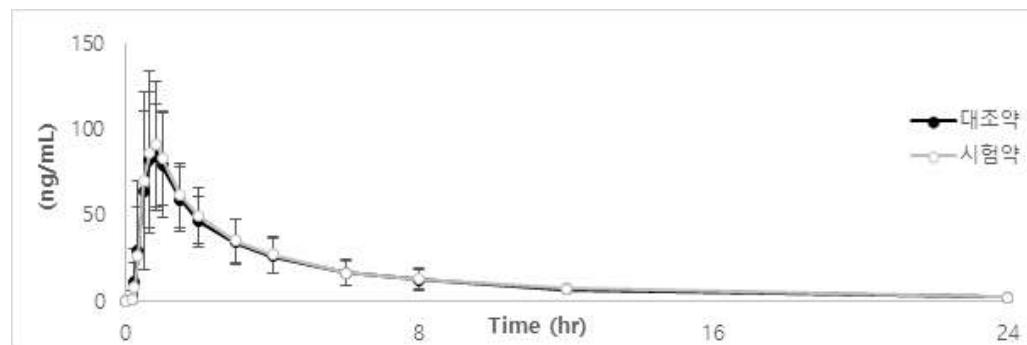


그림 9. MI(XGBOOST)에 대한 혈장 중 시간대별 평균 농도 그래프

4. 간헐적 누락과 후반부 절단이 혼합된 형태의 결측치

AUC_t 기준으로 가장 좁은 90% 신뢰구간을 보인 방법은 CCA였다. GMR은 102.6%, 신뢰구간은 98.82%~106.43%로 결측이 있는 피험자를 제외함으로써 분석 대상 수가 줄고, 이로 인해 전체 분산이 축소되어 신뢰구간 폭이 감소한 것으로 해석된다.

MI 방법의 GMR은 103.7%, 신뢰구간은 99.65%~107.92%로 나타났으며, 결측으로 인한 불확실성을 반영한 현실적인 추정 결과로 볼 수 있다. Mean은 GMR 103.9%, 신뢰구간 99.72%~108.05%로, 평균값 대체 특성상 분산이 과소 추정되어 신뢰구간이 실제보다 좁게 나타나는 경향이 있었다.

MLE의 GMR은 102.5%, 신뢰구간은 97.59%~107.58%였다. 이는 혼합효과모형 기반으로 개별 피험자 반응을 반영하여 결측에 대한 불확실성을 유연하게 추정한 결과로 해석된다. XGBoost는 GMR 104.6%, 신뢰구간 97.63%~111.88%로 다섯 가지 방법 중 가장 넓었으며, 신뢰구간 폭이 커서 추정값에 대한 불확실성이 상대적으로 크다는 점에서 해석 시 주의가 필요하다.

C_{max} 기준에서도 유사한 경향이 관찰되었다. CCA의 GMR은 104.5%, 신뢰구간은 94.79%~114.08%로 AUC_t와 마찬가지로 가장 좁은 구간을 형성하였다. MI는 GMR 107.1%, 신뢰구간 97.59%~116.87%, Mean은 GMR

107.1%, 신뢰구간 96.25%~116.29%로 두 방법 모두 높은 평균값을 보였으며, 신뢰구간 폭은 유사하였다. MLE는 GMR 110.6%, 신뢰구간 100.33%~122.07%로 가장 높은 중앙값과 가장 넓은 신뢰구간을 보였고, XGBoost는 GMR 109.5%, 신뢰구간 99.55%~120.59%로 기준 경계에 근접하였다.

AUC_t의 경우, 시험약은 MLE(369.8 ± 136.7)에서 가장 높은 평균값을 보였고, 그 뒤를 MEAN(366.9 ± 134.7), MI(366.6 ± 134.7), CCA(362.6 ± 142.2), XGBoost(356.1 ± 141.8) 순으로 따랐다. 대조약 또한 MLE(361.9 ± 135.9)가 가장 높았으며, MI(352.9 ± 126.1), MEAN(352.2 ± 124.6), CCA(353.2 ± 135.3), XGBoost(339.3 ± 130.9) 순으로 나타났다. 전반적으로 모든 방법에서 시험약의 AUC_t 평균값이 대조약보다 높았으며, 특히 MLE에서는 두 제제 간 평균값 차이가 가장 작았다.

C_{max}의 경우, 시험약은 MLE(114.9 ± 38.61)에서 가장 높은 값을 보였으며, MI(109.4 ± 35.70), MEAN(109.4 ± 35.69), CCA(110.0 ± 36.71), XGBoost(108.4 ± 36.61)가 유사한 수준으로 나타났다. 대조약에서는 CCA(104.8 ± 33.93)가 가장 높았고, MLE(103.6 ± 34.53), MI(103.0 ± 34.87), MEAN(102.3 ± 33.71), XGBoost(99.7 ± 36.01)

순으로 확인되었다. C_{max} 역시 시험약이 대조약보다 높은 경향을 보였으며, 방법에 따라 제제 간 차이 폭에 차이가 존재하였다.

AUC_{inf}는 AUC_t와 유사한 상대적 분포를 보였다. 시험약의 경우, MLE(399.8 ± 158.6)가 가장 높았으며, MI(396.5 ± 156.4), MEAN(397.0 ± 156.4), XGBoost(384.9 ± 164.2), CCA(391.5 ± 164.6) 순이었다. 대조약에서도 MLE(390.2 ± 157.2)가 가장 높았고, XGBoost(364.3 ± 149.8)가 가장 낮았다. 전반적으로 시험약이 대조약보다 AUC_{inf} 평균값이 높게 나타났으며, 방법 간 순위는 AUC_t와 유사한 양상을 보였다.

T_{max} 는 시험약에서 0.84~0.85시간, 대조약에서 0.81~0.86시간의 중앙값을 나타냈고, 모든 방법에서 범위는 0.33~2.00시간으로 동일하였다. 방법 간 뚜렷한 차이는 관찰되지 않았다.

$t_{1/2}$ 는 평균 6.5~6.8시간 범위 내에서 나타나, 제제 및 방법 간 유의한 차이가 없는 것으로 확인되었다.

AUC_t/AUC_{inf} 비율은 모든 방법에서 평균 약 93% 수준을 유지하였으며, 이는 혈중 농도-시간 곡선이 충분히 포획되었음을 의미하며 적절한 샘플링이 이루어졌음을 시사한다.

이러한 약동학적 보조 지표들은 생물학적 동등성 판정의 직접적인 기준은

아니지만, 자료의 신뢰성과 분석의 정량적 해석을 보완하는 참고 지표로서 활용 가치가 있다.

요약하면, CCA는 가장 좁은 신뢰구간을 보였지만 표본 손실에 따른 검정력 저하 우려가 있으며, Mean은 분산 과소 추정으로 신뢰구간이 실제보다 좁게 나타날 수 있다. MI는 결측에 대한 불확실성을 반영하면서도 안정적인 추정 결과를 제공하였고, MLE는 데이터 손실 없이 분석이 가능하지만 GMR이 높게 나타나 해석에 주의가 요구된다. XGBoost는 결측 패턴 반영에는 유리하나, 신뢰구간이 가장 넓고 기준선에 근접하는 특성상 생물학적 동등성 평가에 있어 보수적인 해석이 필요하다.

결론적으로, 동일한 원자료에 서로 다른 결측치 처리 방법을 적용할 경우 생물학적 동등성 분석 결과에 유의한 차이가 발생할 수 있음이 수치적으로 입증되었다. 특히 MI와 MLE는 결측이 존재하는 상황에서도 분석 신뢰성과 검정력을 균형 있게 확보할 수 있는 방법으로 판단되며, 실제 임상시험 분석에 있어 적용 가능성이 높다. 반면, Mean과 CCA는 간편하다는 장점에도 불구하고 생물학적 동등성 판단 결과를 왜곡할 가능성성이 있으므로, 결측률이 높은 경우 신중한 적용이 요구된다. XGBoost는 분석적 시사점을 제공할 수 있으나, 규제기관 제출을 위한 주요 분석에는 보조적 또는 탐색적 방법으로 제한적으로 고려하는 것이 바람직하다.



표 4. 결측치 처리 방법별 약동학 파라미터 및 90% 신뢰구간 비교

| PK Parameter (Mean±SD) | METHOD | 시험약 | 대조약 |
|---------------------------------|---------|------------------|------------------|
| AUC _t (hr · ng/mL) | CCA | 362.6 ± 142.2 | 353.2 ± 135.3 |
| | MI | 366.6 ± 134.7 | 352.9 ± 126.1 |
| | MEAN | 366.9 ± 134.7 | 352.2 ± 124.6 |
| | MLE | 369.8 ± 136.7 | 361.9 ± 135.9 |
| | XGBOOST | 356.1 ± 141.8 | 339.3 ± 130.9 |
| AUC _{inf} (hr · ng/mL) | CCA | 391.5 ± 164.6 | 379.3 ± 155.3 |
| | MI | 396.5 ± 156.4 | 379.0 ± 144.5 |
| | MEAN | 397.0 ± 156.4 | 378.3 ± 143.0 |
| | MLE | 399.8 ± 158.6 | 390.2 ± 157.2 |
| | XGBOOST | 384.9 ± 164.2 | 364.3 ± 149.8 |
| C _{max} (ng/mL) | CCA | 110.0 ± 36.71 | 104.8 ± 33.93 |
| | MI | 109.4 ± 35.70 | 103.0 ± 34.87 |
| | MEAN | 109.4 ± 35.69 | 102.3 ± 33.71 |
| | MLE | 114.9 ± 38.61 | 103.6 ± 34.53 |
| | XGBOOST | 108.4 ± 36.61 | 99.7 ± 36.01 |
| T _{max} (hr) | CCA | 0.85 (0.33~2.00) | 0.81 (0.33~2.00) |
| | MI | 0.84 (0.33~2.00) | 0.82 (0.33~2.00) |
| | MEAN | 0.84 (0.33~2.00) | 0.82 (0.33~2.00) |
| | MLE | 0.84 (0.33~2.00) | 0.82 (0.33~2.00) |
| | XGBOOST | 0.85 (0.33~2.00) | 0.86 (0.33~2.00) |
| t _{1/2} (hr) | CCA | 6.73 ± 1.59 | 6.51 ± 1.30 |
| | MI | 6.83 ± 1.50 | 6.60 ± 1.21 |
| | MEAN | 6.84 ± 1.50 | 6.60 ± 1.22 |
| | MLE | 6.80 ± 1.53 | 6.67 ± 1.27 |



| | XGBOOST | 6.76 ± 1.54 | 6.54 ± 1.25 | | |
|--|---------|------------------|------------------|--------|--------|
| AUC _t /AUC _{inf} (%) | CCA | 93.7 ± 3.5 | 94.0 ± 3.2 | | |
| | MI | 93.4 ± 3.4 | 93.9 ± 3.0 | | |
| | MEAN | 93.4 ± 3.4 | 93.9 ± 3.0 | | |
| | MLE | 93.5 ± 3.5 | 93.7 ± 3.2 | | |
| | XGBOOST | 93.6 ± 3.5 | 94.0 ± 3.1 | | |
| | | AUC _t | C _{max} | | |
| | | LOWER | UPPER | LOWER | UPPER |
| 기하평균비율 (90% 신뢰구간) | CCA | 98.82 | 106.43 | 94.79 | 114.08 |
| | MI | 99.65 | 107.92 | 97.59 | 116.87 |
| | MEAN | 99.79 | 108.05 | 98.25 | 116.29 |
| | MLE | 97.59 | 107.58 | 100.33 | 122.07 |
| | XGBOOST | 97.63 | 111.88 | 99.55 | 120.59 |

5. 시뮬레이션

본 연구에서는 결측 메커니즘(MCAR, MAR, MNAR)과 결측률(5%, 10%, 20%) 조건에 따라 네 가지 결측치 처리 방법(완전 케이스 분석(CCA), 평균 대체법(Mean Imputation), 다중 대체법(MI), 최대우도추정법(MLE))의 성능을 비교하였다. 분석 지표는 Power와 Type I Error를 기준으로 평가하였으며, 시뮬레이션 반복 간 결과의 신뢰도 및 일관성을 확인하기 위해 각 지표의 표준편차도 함께 제시하였다.

전반적으로 MI는 모든 결측 메커니즘과 결측률 조건에서 가장 안정적이고 일관된 성능을 보였다. Power는 98.2%~99.6% 수준으로 유지되었고, Type I Error는 최대 6.2% 이내로 통제되었다. MAR 및 MNAR 조건에서 결측률이 20%로 증가하더라도 Power는 각각 99.2%, 98.7%로 유지되었으며, Type I Error도 각각 4.0%, 3.9% 수준으로 나타났다. Power의 표준편차는 6.3%~13.3%, Type I Error의 표준편차는 12.2%~24.1% 범위로, 반복 간의 변동성도 비교적 작았다.

MLE는 MAR 및 MNAR 조건에서는 MI와 유사한 수준의 Power를 보였고, 일정 수준의 Type I Error 통제도 가능했지만, MCAR 조건에서는 Type I Error가 비정상적으로 상승하는 경향을 보였다. MAR 조건에서는 Power가 98.2%~99.0%, Type I Error는 1.5%~3.8%로 비교적 안정적이었고, MNAR

조건에서도 Power는 98.5%~99.5%, Type I Error는 2.3%~8.2%로 유지되었다. 그러나 MCAR 조건에서는 결측률이 20%인 경우 Power는 98.7%였으나, Type I Error는 평균 17.4%, 표준편차는 37.9%로 매우 크게 증가하였다. 이는 MCAR 조건 하에서는 분포 추정의 불안정성으로 인해 신뢰구간이 좁아지고, 실제로 동등한 경우에도 유의한 결과로 해석될 가능성이 상대적으로 높아짐을 시사한다.

MEAN은 계산이 간단하고 전반적으로 낮은 오류율을 보였지만, 분산을 과소 추정하여 실제보다 신뢰구간이 좁게 형성되는 경향이 있어 Type I Error 해석에 주의가 필요하다. MAR과 MNAR 조건에서는 Power가 97.0%~99.7%로 높게 유지되었고, Type I Error는 대부분 0.9%~1.7% 수준으로 안정적이었다. MCAR 조건에서도 Power는 비교적 높게 유지되었으며(98.8%~99.7%), Type I Error는 최대 6.2% 수준이었다. Power의 표준편차는 5.5%~13.7%, Type I Error의 표준편차는 9.5%~12.2% 범위였다. 다만, 무작위 결측 상황에서는 데이터의 분포가 왜곡될 수 있는 점을 고려해야 한다.

CCA는 결측률이 낮은 경우 비교적 높은 Power와 낮은 Type I Error를 나타냈으나, 결측률이 증가할수록 분석 대상자 수가 급감하고 통계적 신뢰성이 현저히 저하되는 현상이 나타났다. 특히 MCAR 조건에서는 결측이

전체 피험자에 무작위로 고르게 확산되는 구조적 특성으로 인해, 대부분의 피험자가 하나 이상의 시간대에서 결측값을 갖게 되었고, 이로 인해 CCA는 분석에 포함할 수 있는 피험자 수가 급격히 줄어들었다. 실제로 MCAR 10% 조건에서 Power는 97.5%로 유지되었지만, Type I Error는 평균 63.9%, 표준편차 48.1%로 과도하게 상승하였으며, 20% 결측률 조건에서는 유효 표본 부족으로 분석 자체가 불가능한 사례도 발생하였다. 이는 단순한 결측률 수준보다도 결측의 분포 방식이 분석 결과에 더 큰 영향을 미친다는 점을 시사하며, CCA의 분석 구조가 이러한 상황과 근본적으로 부적합함을 의미한다. CCA는 피험자 단위로 결측 유무를 판단하고, 하나의 시간점이라도 결측이 존재할 경우 해당 피험자를 전체 분석에서 제외하기 때문에, 결측률이 낮더라도 피험자 수준에서의 탈락률은 기하급수적으로 증가할 수 있다. 이러한 구조적 특성은 분석 대상자 수의 급감과 함께 분산 추정의 왜곡을 유발하며, 그 결과로 Type I Error가 과대 추정되어 분석의 신뢰성을 심각하게 훼손하게 된다.

요약하면, MI는 본 연구의 시뮬레이션 조건 하에서 전반적으로 우수한 통계적 성능을 보였으며, 실무 적용 시 유력한 선택지로 고려될 수 있다. MLE는 MAR 및 MNAR 조건에서는 적절한 방법이 될 수 있으나, MCAR 조건에서는 Type I Error 통제에 주의가 필요하다. MEAN은 단순한 방식임에도 불구하고 전반적으로 양호한 성능을 보였으나, 무작위 결측



조건에서는 분포 왜곡 가능성에 유의해야 한다. CCA는 결측률이 높아질수록 분석 대상 수가 급감하며 Type I Error가 비정상적으로 증가할 수 있으므로, 실무 적용 시 신중한 고려가 필요하다. 따라서 실제 분석에서는 결측 메커니즘과 결측률 수준을 사전에 검토한 후, MI와 같은 모델 기반의 대체 방법을 우선적으로 적용하는 것이 바람직하다.

표 5. 시뮬레이션을 통한 결측 메커니즘 및 결측률에 따른 POWER, TYPE 1 ERROR 비교

| Mechanism | Rate | Method | Power | Type 1 Error |
|-----------|------|--------|-----------|--------------|
| MAR | 5% | CCA | 98.8±10.9 | 1.7±13.0 |
| | | MEAN | 98.1±13.7 | 0.9±9.5 |
| | | MI | 99.4±7.7 | 2.6±15.9 |
| | | MLE | 98.9±10.2 | 1.5±12.2 |
| | 10% | CCA | 99.1±9.5 | 1.9±13.7 |
| | | MEAN | 98.7±11.4 | 1.1±10.4 |
| | | MI | 99.3±8.3 | 1.8±13.3 |
| | | MLE | 99.0±9.9 | 2.5±15.6 |
| | 20% | CCA | 99.2±8.9 | 2.7±16.2 |
| | | MEAN | 98.5±12.2 | 1.5±12.2 |
| | | MI | 99.2±8.9 | 4.0±19.6 |
| | | MLE | 98.2±13.4 | 3.8±19.1 |
| MNAR | 5% | CCA | 99.3±8.4 | 0.7±8.3 |
| | | MEAN | 98.9±10.5 | 0.9±9.5 |
| | | MI | 99.6±6.3 | 1.5±12.2 |
| | | MLE | 98.6±11.6 | 2.3±15.0 |
| | 10% | CCA | 98.9±10.5 | 1.7±13.0 |
| | | MEAN | 99.1±9.5 | 1.0±10.0 |
| | | MI | 99.4±7.7 | 2.4±15.3 |
| | | MLE | 99.5±6.9 | 3.3±18.0 |
| | 20% | CCA | 98.8±10.9 | 2.7±16.2 |
| | | MEAN | 98.9±10.4 | 1.7±13.0 |
| | | MI | 98.7±11.3 | 3.9±19.4 |
| | | MLE | 98.5±12.1 | 8.2±27.5 |
| MCAR | 5% | CCA | 97.8±14.7 | 15.6±36.3 |
| | | MEAN | 99.7±5.5 | 1.2±10.9 |
| | | MI | 99.5±7.1 | 2.6±15.9 |
| | | MLE | 98.7±11.2 | 2.4±15.4 |
| | 10% | CCA | 97.5±15.7 | 63.9±48.1* |
| | | MEAN | 98.8±10.9 | 1.4±11.8 |
| | | MI | 99.6±6.3 | 4.4±20.5 |
| | | MLE | 98.5±12.2 | 5.2±22.2 |
| | 20% | CCA | — | — |
| | | MEAN | 99.5±7.1 | 1.4±11.8 |
| | | MI | 98.2±13.3 | 6.2±24.1 |
| | | MLE | 98.7±11.2 | 17.4±37.9 |

* CCA는 결측 피험자를 모두 제외하므로, 분석 대상 수 감소 및 Type I Error 과대 추정 가능성 있음.

그림 10~12은 결측 메커니즘(MCAR, MAR, MNAR), 결측률(5%, 10%, 20%), 변동계수(CV), 샘플 수(n)의 변화에 따른 네 가지 결측치 처리 방법(CCA, Mean, MI, MLE)의 Power 및 Type I Error Rate 변화를 Heatmap으로 시각화한 결과이다. 각 방법은 분석 설계 요인에 따라 서로 다른 민감도와 신뢰도를 보이며, 결측치 처리 방식의 선택이 결과 해석에 중요한 영향을 미친다는 점을 시사한다.

Power Heatmap을 보면, MI는 거의 모든 결측 메커니즘과 조건에서 검정력(Power)을 안정적으로 유지하였다. 샘플 수가 48명 이상이면 결측률과 CV 수준에 관계없이 Power가 99% 이상으로 유지되었으며, 특히 MAR 및 MNAR 조건에서 가장 강건한 성능을 보였다. 이는 MI가 결측 정보를 복수의 분산 기반 대체로 통계적 정밀도를 보완하며 분석의 신뢰성을 높이는 방식이기 때문으로 해석된다.

Mean은 결측률이 10% 이하이거나 샘플 수가 충분한 조건($n \geq 48$)에서는 양호한 Power를 보였으나, CV가 50%로 높은 경우나 결측률이 20%에 달할 경우, Power가 90% 이하로 급격히 저하되는 양상이 관찰되었다. 이는 평균 대체가 분산을 과소 추정하여 신뢰구간을 비정상적으로 좁게 만들기 때문이며, 샘플 수 증가 시 그 왜곡이 다소 완화되는 경향도 확인되었다.

CCA는 결측률이 10% 이상일 경우, 결측으로 인해 분석 대상자 수가 현저히 감소하여 통계적 검정에 필요한 유효 표본 확보가 어려워지는 문제가 나타났다. 특히

MCAR 조건에서는 결측률이 10% 또는 20%인 경우, Sample Size와 무관하게 Power 계산 자체가 불가능한 경우가 다수 발생하였다. 이는 결측값이 존재하는 피험자를 전부 제외하는 방식 특성상, 유효 표본 수가 급감하며 분산 추정 자체가 불가능해지는 구조적 한계에 기인한다.

MLE는 MAR 및 MNAR 조건에서는 MI와 유사하게 높은 Power를 유지하였다. 특히 샘플 수가 작거나 CV가 높은 조건에서도 Power가 99% 수준으로 안정적으로 유지되는 양상을 보였다. 이는 MLE가 관측값 기반의 가능도를 최대화하는 방식으로 결측치를 모델 내재적으로 보정하며, 동등성 판정 기준을 충족시키는 데 필요한 추정 효율을 유지할 수 있었기 때문이다.

Type I Error Heatmap에서는 MI가 가장 안정적인 오류율 통제를 보였다. 모든 메커니즘에서 0.0%~1.0% 수준으로 유지되었으며, 이는 MI가 결측으로 인한 분산 왜곡을 최소화하면서도 동등성 기준(80.00~125.00%)을 적절히 유지했음을 시사한다. Mean 방식 역시 대부분의 조건에서 2% 이내에서 안정적이었지만, 일부 MCAR 조건($n = 24$, $CV = 40\sim50\%$)에서는 오류율이 6%까지 상승하는 경향이 확인되었다.

CCA는 Type I Error 면에서도 가장 불안정한 결과를 나타냈다. MCAR 조건에서 $n = 48$, $CV = 30\%$ 일 때 Type I Error가 27%까지 상승하였고, MNAR 조건에서는 $n = 24$, $CV = 50\%$ 조건에서 21.4% 수준까지 도달하였다. 이는 피험자 제거 방식으로 인해 분석 대상의 수가 급감하고, 군 간 불균형이 초래되어 동등성 판정 기준을 왜곡하는 데 따른 결과이다.

MLE는 MAR 및 MNAR 조건에서는 Type I Error를 2% 수준에서 안정적으로 유지하였으나, MCAR 조건에서만 특이하게 오류율이 급상승하였다. 특히 $n = 24$, CV = 40~50%, 결측률 20% 조건에서 Type I Error가 93~97%에 달하는 등, 극단적인 결과를 보였다. 이는 MCAR의 완전 무작위 속성으로 인해, 결측 발생이 예측 가능한 구조와 무관하게 이루어지므로, MLE는 관측된 데이터에 기반한 분포 추정 과정에서 제한된 정보만을 활용하게 된다. 이로 인해 MLE는 표준 오차를 과소 추정하거나 신뢰구간을 비정상적으로 좁게 산출하는 경향이 있으며, 이는 실질적인 차이가 없는 경우에도 통계적으로 유의한 결과로 오판하게 만들어 Type I Error의 과대 발생으로 이어질 수 있다.

Heatmap 기반 결과를 종합하면, 결측 메커니즘(MCAR, MAR, MNAR), 결측률, 변동계수(CV), 샘플 수(N)의 조건 변화에 따라 결측치 처리 방법의 성능이 유의하게 달라지며, 특히 Power와 Type I Error 간의 균형이 분석 결과의 신뢰성에 핵심적임을 확인할 수 있다. 실제 임상시험 환경에서 MAR이 가장 일반적인 메커니즘으로 간주되며, 이 조건에서 MI와 MLE는 높은 Power와 낮은 오류율을 안정적으로 보였으므로, 실무 적용에 적합한 방법으로 평가된다. 반면, CCA는 결측률이 높아지면 분석 자체가 불가능해질 수 있으며, Mean은 단순 대체의 한계로 인해 해석 시 주의가 필요하다. 특히 MCAR 조건에서의 MLE 적용은 통계적 왜곡의 가능성성이 크므로, 이러한 메커니즘이 의심되는 경우에는 MLE 사용을 신중히 고려해야 한다.

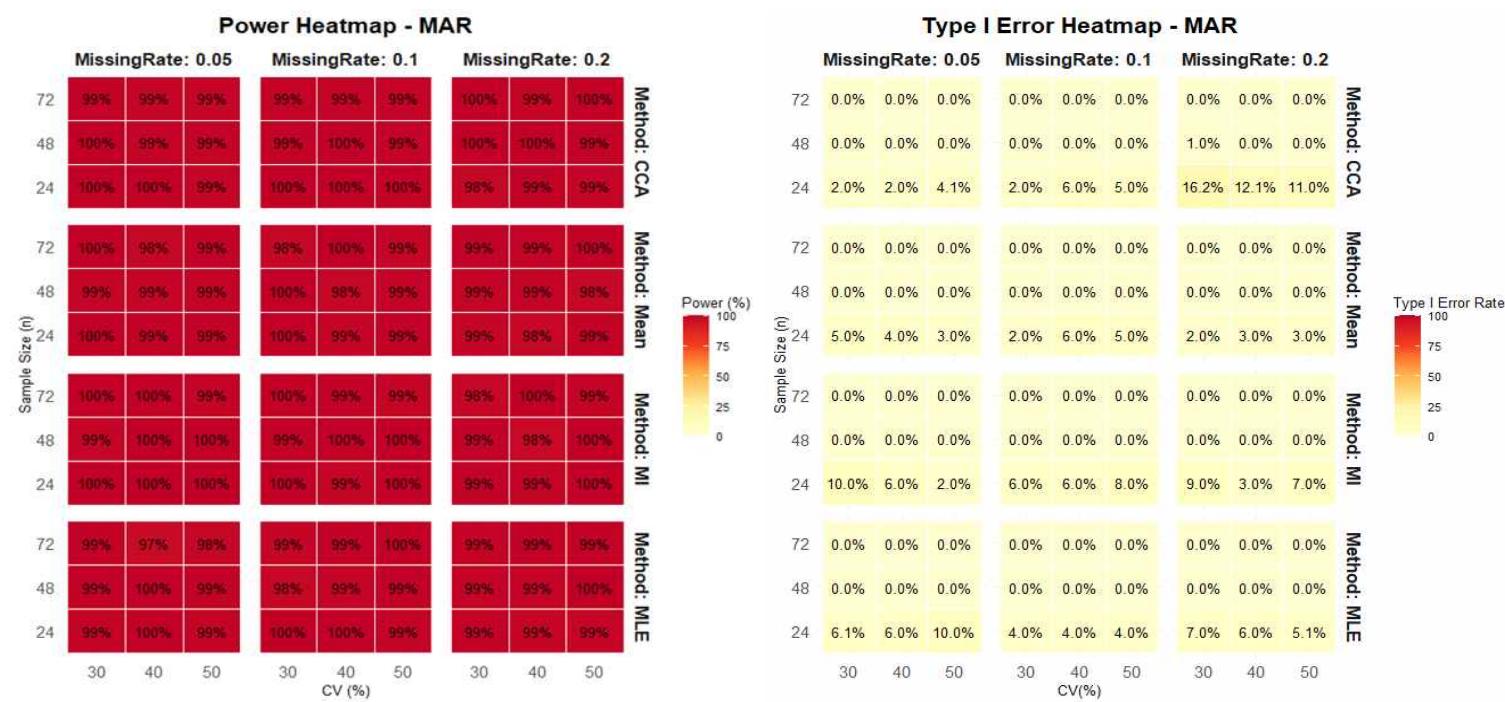


그림 10. MAR 가정 하에서 Sample Size 및 CV 변화에 따른 Type I Error 및 Power 추가 분석

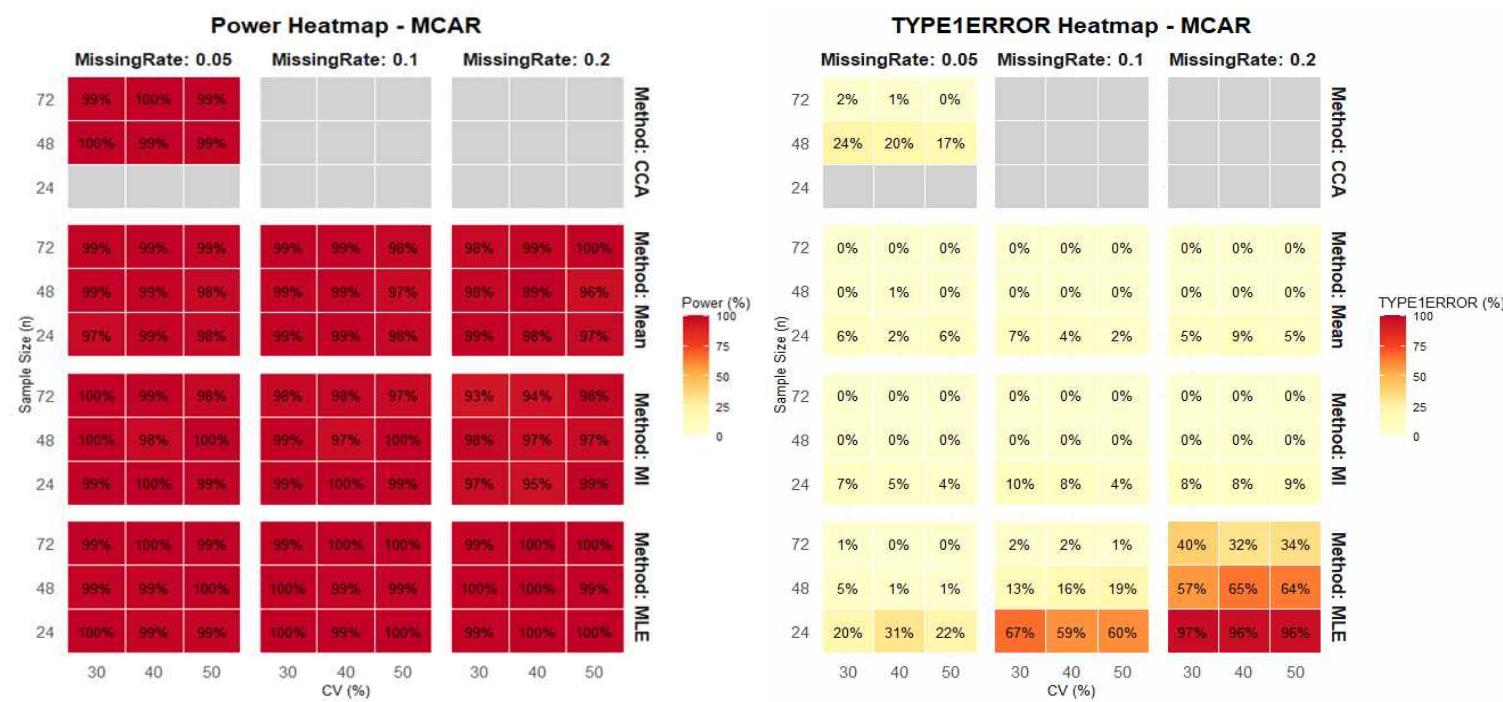


그림 11. MCAR 가정 하에서 Sample Size 및 CV 변화에 따른 Type I Error 및 Power 추가 분석

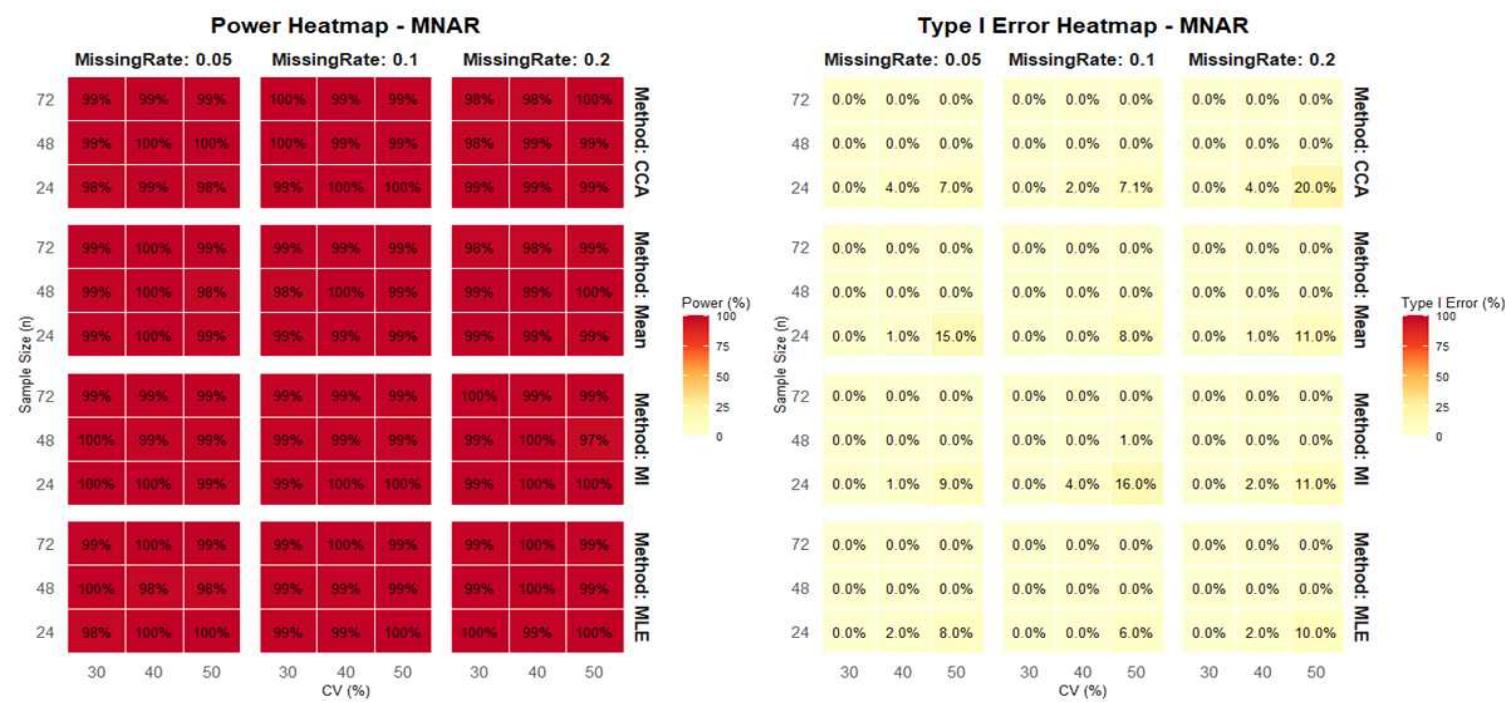


그림12. MNAR 가정 하에서 Sample Size 및 CV 변화에 따른 Type I Error 및 Power 추가 분석

IV. 고찰

1) 연구 결과에 대한 고찰

본 연구는 생물학적 동등성 시험에서 결측치 처리 방식이 약동학 지표 및 동등성 판정 결과에 미치는 영향을 정량적으로 비교·평가하였다. 이를 위해 임상시험 데이터를 기반으로 완전 케이스 분석(CCA), 평균 대체법(Mean), 다중 대체법(MI), 최대우도추정법(MLE), 그리고 XGBoost 예측모형을 적용하여 분석 결과를 비교하였으며, 결측 메커니즘(MCAR, MAR, MNAR) 및 결측률(5%, 10%, 20%)에 따른 시뮬레이션 기반 성능 평가 및 Heat map을 병행하였다.

임상시험 데이터 분석에서는 MI와 MLE가 주요 약동학 지표의 분포를 안정적으로 보존하면서도 통계적 일관성과 신뢰구간 기반의 분석 성능에서 강점을 보였다. MI는 Rubin의 규칙을 기반으로 결측으로 인한 불확실성을 통합 추정에 반영할 수 있었고, MLE는 반복적 우도 기반 추정을 통해 관측치를 최대한 활용함으로써 고차원 분석에서도 적절한 성능을 유지하였다. 반면 CCA는 피험자 수의 급감으로 인해 통계적 검정력이 저하되고, 신뢰구간 폭이 과도하게 축소되는 경향을 보여 동등성 오판 가능성이 높았다. MEAN은 계산이 간단하고 분석 가능성이 높으나, 시간대별 평균값을 단순 대체함으로써 개별 간의

변동성을 반영하지 못하며, 특히 C_{max} 와 같은 민감한 지표에서는 해석의 타당성이 떨어졌다. XGBoost는 시간, 제제, 인접 시점 농도 등 다차원 정보를 활용하여 결측 시점의 약물 농도를 비교적 효과적으로 예측하였으며, 시계열 패턴 복원 측면에서는 일정 수준의 성능을 보였다. 그러나 곡선 후반부에서 과장된 예측이 반복되고, 예측값 기반 신뢰구간이 동등성 경계값 인근에 위치하며 분석 결과의 불안정성이 자주 발생하였다. 예측 과정의 불확실성을 정량화 할 수 없고, 모델 해석의 투명성이 부족하다는 점에서 기존 규제기관이 요구하는 통계적 기준을 충족하지 못하였다. 본 연구에서는 이에 따라 XGBoost를 보조적 방법으로 한정하였다. MI(XGBoost)는 임상시험 데이터에서 약물 농도의 시간 경과에 따른 패턴을 비교적 정확히 반영하였고, AUC 및 C_{max} 와 같은 주요 약동학 지표에 대해서도 MI 및 MLE와 유사한 수준의 추정 결과를 보였다. 이는 변수 간 상호작용, 비선형 구조, 결측이 특정 시점에 국한된 데이터 특성으로 인해 예측 정확도가 높게 나타난 것으로 해석된다. 예측값의 분포는 원자료와 유사한 형태를 유지하였으나, 반복 대체 간 분산이 충분히 확보되지 않아 신뢰구간이 과도하게 축소되거나 동등성 경계값 인근에 위치하는 경우가 있었다. 특히 C_{max} 에서는 예측값이 과장되거나 왜곡되는 사례도 반복적으로 나타났다. 본 연구에서는 이러한 한계를 고려하여 MI(XGBoost)를 결측 보완을 위한 보조적 수단으로 한정하였으며, 결과 해석 시 주의가 필요하다고 판단하였다.

시뮬레이션 결과에서도 MI는 모든 결측 메커니즘과 결측률을 조건에서 가장 일관되고 안정적인 성능을 보였다. Type I Error는 대부분 5% 이내로 유지되었고, Power는 전 구간에서 98.2%~99.6% 수준으로 유지되었다. 반면 MLE는 MAR 및 MNAR 조건에서는 비교적 안정적이었으나, MCAR 조건에서는 오류율이 급등하였다. 특히 피험자 수가 24명, 변동계수가 50%, 결측률이 20%인 조건에서 Type I Error가 96%에 달하였으며, 분산을 과소 추정하고 신뢰 구간을 비정상적으로 좁히는 구조적 한계가 주요 원인으로 확인되었다.

CCA는 결측률이 10% 이상이고 변동계수가 40% 이상이며 피험자 수가 24명인 조건에서 Power 계산이 이루어지지 않는 경우가 다수 발생하였다. 이로 인해 분석 자체가 불가능한 사례가 반복되었으며, MCAR 10% 조건에서는 Type I Error가 63.9%까지 상승하여 피험자 단위 결측 제거 방식의 실무적 한계를 명확히 드러냈다. MEAN은 피험자 수가 48명 이상인 조건에서는 안정적인 Power를 보였으며, 모든 MCAR 조건에서 Type I Error를 0~2% 수준으로 통제하였다. 그러나 평균 대체 구조 특성상 개별 변동성을 반영하지 못하며, 고변동 · 고결측 조건에서는 통계적 왜곡 가능성이 있다.

Heatmap 기반의 고차원 조건 분석에서도 MI는 모든 조건에서 가장 안정적인 성능을 보였다. Power와 Type I Error는 각각 98.2~99.6%, 0.9~6.2% 범위 내에서 유지되었다. 반면 MLE는 MCAR 조건에서 Type I Error가 최대

96%까지 상승하였고, CCA는 분석 결과 자체가 결측되는 경우가 많아 실질적인 적용에 한계가 있었다. MEAN은 비교적 높은 Power와 낮은 오류율을 보였지만, 구조적 단순화로 인해 분산이 과소 평가될 위험이 있다.

또한, XGBoost 기반의 예측값을 활용한 다중 대체(MI)도 실험적으로 적용하였다. 이 방법은 결측이 발생한 시간점의 약물 농도를 XGBoost 회귀모형으로 예측하고, 동일한 데이터를 반복 대체하여 복수의 완전 데이터를 구성하는 방식이다. 그러나 이 방식은 MI의 핵심 원칙인 반복 대체 간의 불확실성 반영을 충족하지 못한다. XGBoost는 결정론적 예측 모델이기 때문에 동일한 입력에 대해 항상 동일한 예측값을 반환하며, 반복 대체 간 차이가 거의 발생하지 않는다. 그 결과 Rubin's Rules 기반 분산 추정이 왜곡되며, 신뢰구간이 비정상적으로 축소되거나 과도하게 확대될 수 있다.

시뮬레이션 결과에서도 이러한 구조적 한계는 명확히 드러났다. CV 30% 이상, missing rate 10% 이상, 소표본($n = 24$) 조건에서는 Type I Error가 75~94%까지 상승하는 등, 통계적 유의수준(5%)을 명백히 벗어나는 오류가 반복적으로 관찰되었다. 이 경향은 MCAR, MAR, MNAR 모든 결측 메커니즘에서 일관되게 나타났으며, CV가 증가할수록, 결측률이 높아질수록 오류율은 극단적으로 치솟았다. 특히 MAR 조건에서는 CV 30%, missing rate 10% 상황에서 조차 Type I Error가 80%를 초과하였고, MNAR 조건에서는 같은 조

건에서 86.7%까지 상승하였다. 이는 대체 간 분산이 사실상 0에 수렴함에 따라 통계적 검정이 정상 작동하지 않고 있음을 의미한다.

반면, 표본 수가 증가($n = 72$)하면 Type I Error는 비교적 안정적으로 수렴하는 것처럼 보인다. 예를 들어 CV 30%, missing rate 10% 조건에서 MCAR에서는 3.3%, MAR은 14.3%, MNAR은 3.3%로 일부 수치는 허용 가능한 수준 내에 존재했다. 그러나 이는 반복 대체 간 분산이 축소된 결과로, 실제 불확실성이 반영된 것이 아니다. 즉, 수치만 보면 정상처럼 보일 수 있지만 이는 XGBoost가 내부적으로 분산을 제거한 채 예측을 반복했기 때문에 발생한 착시 효과에 불과하다. 신뢰구간 산출 구조 자체가 왜곡된 상태이므로, 통계적으로는 심각한 문제로 간주되어야 한다.

임상시험 데이터에서는 시뮬레이션과 상반된 양상이 관찰되었다. 변수 간 복잡한 상호작용과 비선형 구조가 존재하고, 결측이 특정 시간점에만 국한되어 있으며, 나머지 변수에는 예측에 필요한 정보가 충분히 남아 있었기 때문에 XGBoost의 예측 정확도가 높게 유지되었다. 예측값도 원래 분포와 유사하게 나타나며 신뢰구간의 왜곡이 심하지 않았고, MI와 유사한 통계적 결과가 도출되었다.

따라서 XGBoost 기반의 예측값을 활용한 다중 대체(MI)는 결측값 처리의 보조적 도구로는 활용 가능하지만, 통계적 추론을 위한 정규 다중 대체 방법

의 대안으로 사용하기에는 이론적·실무적 제약이 존재하며, 불확실성 반영을 위한 추가적인 조치 없이는 적용에 유의가 필요하다.

실제 임상시험에서는 결측치가 다양한 임상적 상황에서 발생하며, 그 메커니즘은 분석 방법의 선택에 결정적 영향을 미친다. MAR은 채혈 지연, 이상반응, 순응도 저하 등 예측 가능한 변수와 연관된 결측으로, 실무에서 가장 일반적으로 가정되는 메커니즘이다. 규제기관(FDA, EMA 등)에서도 MAR 메커니즘 하에서의 결측치 발생이 통계적으로 정당화될 경우, MI나 MLE 등의 분석 방법을 허용하고 있으며, 본 연구 결과에서도 이들 방법은 MAR 조건에서 높은 Power와 안정적인 오류 통제를 보였다. MCAR은 결측이 완전히 무작위로 발생하는 이상적인 메커니즘으로, 실제 임상시험에서는 드물게 나타난다. 본 연구의 시뮬레이션에서도 MCAR 조건 하에서 분산이 과소 추정되고, 이로 인해 신뢰구간이 비정상적으로 좁아지며 Type I Error가 과도하게 상승하는 현상이 반복적으로 관찰되었다. 한편 MNAR은 결측 발생이 관측되지 않은 변수에 의존하는 경우로 정의되며, 실제 임상에서도 발생 가능성성이 존재하지만, 그 존재 여부를 데이터만으로 확인하기 어렵고, 복잡한 구조적 모델링이 요구되기 때문에 실무에서는 대부분 MAR로 가정하여 분석이 이루어진다.

결론적으로, MI는 실제 임상시험에서 가장 현실적인 조건인 MAR 하에서 높은 적용 가능성과 통계적 안정성을 보여 생물학적 동등성 평가에 있어 우선적



으로 고려할 수 있는 분석 방법으로 판단된다. MLE 역시 적절히 설계된 경우 강력한 대안이 될 수 있으나, 결측 메커니즘이 명확하지 않거나 무작위성이 큰 경우에는 신중한 적용이 필요하다. 각 방법의 장단점과 적용 조건을 명확히 인식하고, 분석 목적과 데이터 특성에 부합하는 전략적 선택이 요구된다.

2) 연구의 제한점과 의의

본 연구의 제한점은 다음과 같다.

첫째, 본 연구의 시뮬레이션은 단일 구획 모델을 기반으로 하여 흡수 및 제거가 일정한 일차속도론을 따른다는 가정 하에 수행되었으며, 이는 실제 약물의 복잡한 체내 동태를 충분히 반영하지 못할 수 있다.

둘째, MI 및 XGBoost 적용 시 반복 횟수, 내부 파라미터 설정, 예측 변수 구성 등에 따라 결과가 달라질 수 있으며, 이는 분석자의 선택에 따라 결과의 재현성과 해석의 일관성에 영향을 줄 가능성이 있다..

셋째, 결측 구조는 후반부 절단 및 간헐적 누락이 혼합된 형태로 제한하여 구성하였기 때문에, 실제 임상시험에서 빈번하게 발생하는 중도 이탈, 이상반응 기반 탈락, 불완전한 순응도 등에 따른 다양한 결측 양상을 충분히 반영하지 못하였다.

이러한 제한점에도 불구하고, 본 연구는 다음과 같은 의의를 지닌다.



첫째, 임상시험 데이터를 기반으로 현실적인 결측 조건을 반영하여 시뮬레이션을 설계하고, MCAR, MAR, MNAR 등 다양한 결측 메커니즘과 결측률 변화에 따른 분석 결과를 정량적으로 비교함으로써 생물학적 동등성 연구에서 결측 처리 전략 수립에 참고할 수 있는 근거를 제시하였다.

둘째, 기존 연구에서 상대적으로 제한적으로 다루어진 결측치 처리 방법 간의 성능 차이를 체계적으로 분석하였고, 특히 MI와 MLE의 통계적 장점과 실무 적용 가능성을 실증적으로 제시하였다.

셋째, Power 및 Type I Error에 대한 민감도 분석 결과를 Heatmap으로 시각화하여 결측률, 변동계수, 표본 수 등 설계 요인에 따른 분석 방법의 성능 차이를 직관적으로 제시하였고, 이는 실제 생물학적 동등성 평가 전략 수립에 있어 실질적인 판단 기준을 제공할 수 있다.

넷째, MI(XGBoost)의 반복 대체 간 불확실성 반영이 제한된다는 점을 고려할 때, 향후 연구에서는 MI 알고리즘 내에서 랜덤 서브샘플링(random subsampling)과 예측 평균 일치법(Predictive Mean Matching, PMM)을 결합한 이중 무작위화 기법을 적용함으로써 대체 간 변이를 보다 효과적으로 반영할 수 있을 것이다. 또한 MI-LMM, MI-XGBoost, MI-BART 등 다양한 베이스 모델 기반의 다중 대체 방법을 비교·분석함으로써, 머신러닝 기반 MI 기법의 통계적 타당성과 실무 적용 가능성에 대한 보다 심층적인



검토가 가능할 것으로 기대된다. 특히 BART는 베이지안 추론 과정에서 사후분포(posterior distribution)로부터의 표본 추출이 용이하다는 특성으로 인해, 베이지안 MI 방법론의 유력한 대안으로 활용될 수 있다.

V. 결론

본 연구에서 비교한 각 결측치 처리 방법은 결측 메커니즘과 분석 조건에 따라 상이한 성능 특성과 한계를 보였다.

완전 케이스 분석(CCA)은 구조가 단순하고 구현이 용이하다는 장점이 있으나, 결측률이 높거나 분석 대상자 수가 적은 경우 통계적 검정력이 급격히 저하되고 Type I Error가 과도하게 상승하는 한계를 나타냈다.

평균 대체법(Mean Imputation)은 결측률이 낮고 샘플 수가 충분한 조건에서는 일정 수준의 분석 성능을 보였으나, 분산 과소 추정으로 인해 신뢰구간이 인위적으로 좁아지는 경향이 있었다. 대부분의 조건에서 Type I Error는 유의수준 이내로 유지되었지만, 고결측률 및 고변동 조건에서는 오류율이 증가하는 등, 제한적인 상황에서 오류율 증가 가능성이 확인되었다.

다중 대체법(MI)은 결측 메커니즘이나 결측률, 변동계수, 샘플 수 등 다양한 분석 조건 변화에도 가장 강건한 성능을 보였으며, 임상시험 데이터 분석과 시뮬레이션 모두에서 Type I Error와 Power를 안정적으로 유지하였다.

최대우도추정법(MLE)은 MAR 및 MNAR 조건에서는 MI에 준하는 수준의 정확도를 보여 유효한 대안이 될 수 있었으며, 특히 모델 기반 접근이 가능한

경우 높은 적용 가능성을 보였다. 다만 MCAR 조건에서는 신뢰구간 과소 추정으로 인해 오류율이 급격히 증가하는 현상이 확인되어, 결측 구조에 대한 사전 검토 없이 사용하는 것은 바람직하지 않다.

XGBoost는 시계열 패턴 복원 측면에서 예측 정확도는 높았으나, 신뢰구간 폭이 다른 방법에 비해 과도하게 넓었고, 80.00~125.00%의 생물학적 동등성 기준 경계에 민감하게 반응하여 결과 해석이 보수적으로 나타날 수 있었다. 또한 기계학습 기반 모델 특유의 해석 불투명성은 분석 결과의 신뢰성에 영향을 미칠 수 있다.

XGBoost 기반 MI는 임상시험 데이터에서는 주요 약동학 지표에 대해 비교적 정확한 예측 성능을 보였으나, 시뮬레이션에서는 반복 대체 간 분산 확보가 미흡하여 신뢰구간이 왜곡되고 Type I Error가 급증하는 등 통계적 추론의 불안정성이 확인되었다. 이에 따라 해당 기법은 보조적인 분석 수단으로 제한적으로 활용될 수 있다.

이러한 결과를 종합하면, 생물학적 동등성 시험에서 결측치가 존재하는 경우, 다양한 분석 조건 하에서 일관되게 안정적인 성능을 보인 다중 대체법(MI)이 가장 적절한 결측치 처리 방법으로 판단된다. 최대우도추정법(MLE)은 MAR 및 MNAR 조건에서는 MI에 상응하는 수준의 분석 성능을 보여 신뢰할 수 있는 대안이 될 수 있으나, MCAR 조건에서는



오류율이 급격히 증가하는 경향을 보여 적용에 주의가 요구된다. XGBoost 기반 MI는 임상시험 데이터에서는 일정 수준의 예측 정확도를 보였으나, 반복 대체 간 분산 확보가 충분하지 않아 통계적 추론에는 한계가 있으며, 보조적 수단으로의 활용에 그쳐야 한다. MEAN 대체법은 분석 환경에 따른 성능 편차가 크고 통계적 왜곡 가능성이 높아 실무 적용 시 주의가 필요하다. XGBoost 단일 대체법은 시간 순서를 반영한 예측 성능은 높았으나, 신뢰구간이 과도하게 넓게 나타나 통계적 판단에 불안정성을 야기할 수 있어 제한적 활용이 권장된다. 반면, CCA는 실무에서 기본적으로 사용되는 접근법이지만, 결측 메커니즘에 민감하여 데이터 손실 및 편향 발생 가능성에 유의해야 한다.

참고문헌

권인선. 2×2 교차설계법에서 결측치가 있는 경우 생물학적 동등성 검증 [석사학위논문]. 서울: 가톨릭대학교; 2010.

박상규. 2×2 교차설계법에서 모집단 생물학적 동등성 검정 방법 비교 [석사학위논문]. 대전: 충남대학교; 2008.

박상규. 결측치가 존재하는 2×2 교차실험에서 생물학적 동등성 연구 [석사학위논문]. 대전: 충남대학교; 2009.

조보배. 2×2 교차계획법에서 결측치가 있을 때의 결측치 처리 방법 비교에 관한 연구 [석사학위논문]. 대전: 충남대학교; 2010.

최익성. 임상시험의 결측치 처리 방법에 관한 비교 연구 [석사학위논문]. 서울: 서울대학교; 2011.

최지연. 결측치가 있는 2×2 교차설계에서 모집단 생동성의 통계적 평가 [석사학위논문]. 대전: 충남대학교; 2012.

Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD Conference; 2016 Aug; San Francisco (CA). New York: ACM; 2016. p. 785–94.

Gabrielsson J, Weiner D. Pharmacokinetic and pharmacodynamic data analysis: concepts and applications. 5th ed. Boca Raton (FL): CRC Press; 2016.

Gong Y. Alternative model-based data analysis approach to demonstrate bioequivalence [dissertation]. Rockville (MD): U.S. Food and Drug Administration; 2020.

Lachin JM. Fallacies of last observation carried forward analyses. *Clin Trials*. 2016;13(2):161–8.

Lee J, Kim H, Park S, Lee H, Choi M, Jung Y, et al. Model-based approach for designing an efficient bioequivalence study for highly variable drugs. *Pharmaceutics*. 2021;13(11):1101.

Little RJA, Rubin DB. Statistical analysis with missing data. 3rd ed. Hoboken (NJ): Wiley; 2019.

Philipp E, Bertrand J, Mentré F, Loingeville F, Dette H, Zhao L, et al. Model-based bioequivalence approach for sparse pharmacokinetic bioequivalence studies. *Stat Med*. 2024;43(3):e10088.

Rowland M, Tozer TN. Clinical pharmacokinetics and pharmacodynamics: concepts and applications. 4th ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2011.

Rubin DB. Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons; 1987.

Shargel L, Wu-Pong S, Yu ABC. Applied biopharmaceutics & pharmacokinetics.

7th ed. New York: McGraw-Hill Education; 2012.

U.S. Food and Drug Administration (FDA). Guidance for industry: statistical approaches to establishing bioequivalence. Rockville (MD): FDA; 2001.

Yu LX, Jiang W, Zhang X, Lionberger RA, Makhlof FT, Schuirmann DJ, Lesko LJ. Between-batch pharmacokinetic variability inflates type I error in bioequivalence studies. Clin Pharmacol Ther. 2018;104(3):525–30.
doi:10.1002/cpt.535.

Zhang Y, Liu X, Zhai S, Ma L, Yu H. A Bayesian approach to pilot–pivotal trials for bioequivalence assessment. BMC Med Res Methodol. 2023;23(1):123.
doi:10.1186/s12874-023-02120-2.

ABSTRACT

Comparison of Missing Data Handling Methods in Estimating Pharmacokinetic Parameters and Evaluating Bioequivalence

Beomseok Kang

Department of Health Informatics & Biostatistics

The Graduate school of Public Health

Yonsei University, Seoul, Korea

(Directed by Professor Minjin Ha, Ph.D.)

Study Background

In bioequivalence (BE) studies, missing data frequently occur due to factors such as subject dropout or sample processing failure, which may compromise the reliability of pharmacokinetic (PK) parameter estimation and BE determination. Despite growing interest, few studies have systematically compared the effects of various missing data handling

techniques in BE trials.

Study Subjects and Methods

This study analyzed clinical data from a 2×2 crossover BE trial to examine the impact of six missing data handling methods—Complete Case Analysis, Mean Imputation, Multiple Imputation, Maximum Likelihood Estimation, XGBoost-based Single Imputation, and MI(XGBoost). The 90% confidence intervals of PK parameters (C_{\max} , AUC_t) were computed for each method, and BE conclusions were compared. Additionally, simulation studies were conducted under varying sample sizes, coefficients of variation, missing data mechanisms (MCAR, MAR, MNAR), and missing rates to evaluate the Type I error and power for each method.

Study Results

Multiple Imputation (MI) consistently yielded robust performance across real and simulated data. MLE was also effective under MAR assumptions. In contrast, CCA and Mean Imputation led to inflated Type I error rates and reduced power, especially under high missingness. XGBoost-based methods showed high prediction accuracy but suffered from overly wide

confidence intervals, affecting BE decision stability.

Conclusion

Among the tested methods, MI demonstrated the most balanced trade-off between statistical validity and practical feasibility. However, careful consideration of the missing data mechanism and data structure is crucial during the analysis planning stage. This study offers practical recommendations for method selection in regulatory BE evaluations.

Keywords: Bioequivalence, Missing Data, Pharmacokinetics, Multiple Imputation (MI), Maximum Likelihood Estimation (MLE), XGBoost, Simulation