

설문·웨어러블 기반 일상 혈압의 보간·추정 — 수준·위상·편차 분해와 그 정확도 한계

이규성¹ · 차상현¹ · 성봉주¹ · 박대근¹ · 김효영²

¹ 차의과학대학교 일반대학원 AI헬스케어융합학과 · ² 한국공학대학 인공지능학과 · 데이터: A DAY Inc. (DAY BP), N = 120

초록 (Abstract)

배경. 스마트워치와 설문으로 혈압(BP)을 추정하는 무가압(cuffless) 접근이 확산되고 있으나, 세션(순간) 단위 정확도의 한계와 그 원인은 충분히 규명되지 않았다.

방법. 120명에게서 7일간 하루 약 5회 측정한 참조 BP와 웨어러블·설문 변수를 사용하였다. 세션평균·이상점 정제 후 각 날을 분단위 하이브리드로 인과 보간하였다(실측 시각엔 실측 고정). 외부(KNHANES)+설문+day1 보정으로 개인 수준을, 위상 커널로 일주기를 더해 구조적 추정치 $E0 = \text{수준} + \text{위상}$ 을 만들고, 편차 $r = \text{실측} - E0$ 를 **위치 신호만으로** 모델하며(과거 BP추정변수는 제외) 단일·multi-task 모델 계열을 비교하였다. 예측 구간을 따로 두지 않고 **day2-7 전부**를 추정에 활용하였다(day1은 보정 앵커). 사람 60:20:20 분할, 평가는 실측 시점에서 MAE와 BHS 등급으로 하였다.

결과. 선형(ridge regression)·스펙트럴이 가장 잘 일반화하여 held-out(test 24명) **일평균 MAE가 SBP 2.40·DBP 2.28 mmHg(BHS A/A), 세션 SBP 5.77(B)·DBP 4.76 mmHg(B, harmonic+ridge면 DBP 4.68 A)**에 도달하였고, tree-gradient boosting은 과적합되었다. 과거 BP추정변수를 제외해도 성능은 거의 유지되었는데, $E0(\text{수준}+\text{위상})$ 가 이미 개인 level을 담기 때문이다. SHAP상 위치 기여는 ΔHR ·심박수(심혈관 각성)가 주도하며 특히 이완기에 크다.

결론. 집계형 위치 신호와 1일 보정, 구조적 $E0$ 만으로 일평균은 Grade A에 도달하나 세션 SBP는 B가 한계이며, 이는 개인 평균으로의 mean-regression과 평균화된 위상, 소멸에 가까운 residual autocorrelation의 구조적 결과다. 선행연구는 세션 Grade A에 raw PPG 파형·PTT·주기 재보정이 필요함을 일관되게 지지한다.

키워드: cuffless blood pressure, wearable, circadian phase, level-deviation decomposition, nowcasting, wearable physiology, heart-rate arousal, BHS grade, mean-regression

1. 서론

혈압(BP)은 심혈관 질환의 이환율 및 사망률을 예측하는 가장 유익하고 교정 가능한 인자 중 하나이며, 그 임상적 가치는 언제 그리고 얼마나 자주 측정하는지에 크게 좌우된다. 진료실에서의 단일 측정값은 연속적으로 변동하는 궤적상의 한 순간만을 포착하는 반면, 활동 중 측정 및 진료실 외 모니터링은 독립적인 예후 가치를 지니는 일간 및 일중 패턴을 드러낸다. 그러나 기존의 활동 중 모니터링은 수면과 일상 활동을 방해하는 간헐적 커프 가압에 의존하므로, 순응도와 기록의 시간적 밀도가 모두 제약된다.

이에 따라 커프리스 착용형 기기는 일상생활 중 부담이 적으면서도 빈번한 BP 정보를 획득하는 수단으로서 점점 더 큰 관심을 받고 있다. 생리 신호 측정과 간편한 자기보고를 결합한 스마트워치는 특히 매력적인 기반을 제공하는데, 광학 신호 및 동작 신호를 맥락 정보가 담긴 설문 응답과 결합하여 각 측정 시점에서 개인의 상태를 특징지을 수 있기 때문이다. 연속적 착용형 신호와 주기적 설문 입력을 결합한 이러한 방식은 진료실 밖 일상적 BP 평가를 위한 실용적 흐름으로 부상하였다.

이러한 추세에도 불구하고 여러 공백이 남아 있다. 대부분의 선행 연구는 이 문제를 순간 추정으로 설정하여, 단일 시점의 신호를 동시점 BP 값으로 매핑하였다. 실세계 적용에 동등하게 중요한 두 과제, 즉 희소한 커프 측정값 사이를 채우는 분 단위 BP 보간과 사전에 관측되지 않은 개인의 임의 시점 BP를 실시간으로 산출하는 분 단위 추정(nowcasting)은 하나의 프레임워크 안에서 함께 다루어지는 경우가 드물다. 또한 보다 거친 일간 요약과 대비하여 개별 측정 세션 수준에서 달성 가능한 정확도는 충분히 규명되지 않은 채로 남아 있어, 집계 특징 기반 착용형 기기의 실용적 한계가 불확실한 상태이다.

이러한 공백을 다루기 위하여 본 연구 DAY BP는 수준-위상-편차 분해를 중심으로 통합 프레임워크를 개발하였다. 수준(level) 항은 짧은 보정 기간과 외부 집단 참조값에 의해 고정되는 각 개인의 안정적 BP 설정점을 포착한다. 위상(phase) 항은 인구통계학적 군집 내에서 공유되는 구조화된 일주기 변동을

포착하며, 추정 단계에서는 이를 1일차 보정 지점(recency-by-phase 가중, length-scale $\ell = 196$ 분, confidence $\kappa = 5$)과 집단 연령×성별 cluster 일주기 곡선을 혼합하여 구성한다. 이는 개인 자신의 조밀한 동일 날 측정값을 따라가는 보간(Part 1, $\ell = 90$, $\kappa = 0.5$)의 위상과 구별된다. 편차(deviation) 항은 구조적 추정값 $E0 = \text{level} + \text{phase}$ 를 고정한 뒤, 현재의 착용형 신호만으로 설명 가능한 나머지 residual 변동을 포착한다. 이 세 성분을 분리함으로써 동일한 구조적 골격이 측정값 사이의 보간과 신규 개인에 대한 분 단위 추정을 모두 지원하며, 성능은 세션 수준과 일간 평균 수준 양쪽에서 보고된다.

본 연구의 기반이 된 데이터셋은 120명의 참가자(남성 60명, 여성 60명; 평균 연령 32.1세, 표준편차(SD) 8.1, 범위 18에서 66세; 평균 체질량지수 22.7)로 구성되었으며, 7일에 걸쳐 하루 약 다섯 세션으로 모니터링하여 8,193건의 원시 측정값을 세션 평균화한 후 약 4,128 세션을 얻었다. 참조 수축기 혈압(SBP)은 평균 113.0 mmHg(SD 14.1), 이완기 혈압(DBP)은 74.5 mmHg(SD 9.5)였으며, 급내 상관계수는 SBP 0.72, DBP 0.63으로 개인 간 변동이 지배적임을 나타내었다. 전 과정에 걸쳐 개인 수준 분할(60:20:20 = 72명/24명/24명)을 사용하였으며, 1일차는 개인의 수준을 고정하는 calibration 앵커로만 사용되어 학습이나 평가의 대상이 되지 않았고, 2일차에서 7일차 전체가 추정의 훈련, 검증, 시험에 사용되었다. 성능은 오직 측정된 시점에서만 산출되었다.

본 연구는 세 가지 기여를 한다. 첫째, 보간과 신규 개인에 대한 분 단위 추정을 단일한 수준-위상-편차 분해 아래 통합하며, 수준 항을 외부 집단 참조값으로 고정하여 개인 간 SBP 정확도를 유의하게 향상시켰다. 둘째, 편차 항에 대하여 20개의 단일 과제 및 multi-task 모델 계열을 체계적으로 벤치마킹하여, 착용형 신호만을 입력으로 하는 ridge regression과 같은 선형 및 스펙트럼 모델이 표본 외에서 가장 잘 일반화한 반면 tree 및 gradient boosting 방법은 과적합되었음을 보였다. 셋째, 집계 특징 기반 착용형 기기의 정확도 한계와 그 원인을 정량적으로 규명하여, 일간 평균은 British Hypertension Society(BHS) Grade A에 도달한 반면 세션 수준 SBP는 각 개인의 설정점에서의 mean-regression, 평균화된 위상, 그리고 소멸에 가까운 residual autocorrelation으로 인해 Grade B에 머물렀음을 입증하고, 이 한계를 문헌에서 확인된 파형, 타이밍, 재보정 요건과 연결하였다.

본 논문의 나머지는 다음과 같이 전개된다. 제2절은 코호트, 측정 프로토콜, 연구 설계를 기술하고, 제3절은 보간, 수준, 위상, 편차 성분을 상술하며, 제4절은 확립된 표준에 대비한 추정 결과와 그 정확도 한계 분석을 보고하고, 제5절은 임상적으로 준비된 커프리스 모니터링을 위한 합의, 한계, 향후 방향을 논의한다.

2. 관련 연구

광용적맥파(PPG)를 이용한 무가압(cuffless) 혈압(BP) 추정은 크게 두 가지 구별되는 특징 패러다임을 따라 연구되어 왔다. 파형 기반 접근법은 원시 PPG 맥파로부터 형태학적 기술자(descriptor)를 추출하며, 여기에는 수축기 및 이완기 정점 기하, 중복맥패임(dicrotic notch), 증대지수(augmentation index), 반사지수(reflection index), 그리고 광용적맥파의 2차 미분 특징(SDPTG)이 포함된다. 반면 집계 특징(aggregate-feature) 접근법은 일정 원도 동안의 생리 신호를 심박수, 심박변이도, 활동 강도와 같은 스칼라 기술자로 요약한다. 두 패러다임은 보존하는 정보의 입도(granularity)에서 뚜렷한 차이를 보인다. 형태학적 기술자는 동맥 순응도와 파동 반사에 의해 형성되는 박동 대 박동(beat-to-beat) 윤곽을 부호화하는 반면, 집계 기술자는 맥파 형태를 버리고 심혈관 상태의 느린 포락선(envelope)만을 보존한다. 이러한 구분은 달성 가능한 정확도에 직접적인 영향을 미치는데, 이는 개별 측정값에 대한 British Hypertension Society(BHS) Grade A 기준(절대오차의 60, 85, 95 퍼센트가 각각 5, 10, 15 mmHg 이내)이 단일 측정 수준에서 요구가 까다롭기 때문이다.

두 번째로 정보량이 큰 타이밍 채널은 맥파 전달 시간(PTT)과 맥파 도달 시간(PAT)으로부터 유도되며, 일반적으로 심전도와 같은 두 번째 동기화 신호를 필요로 한다. 이 타이밍 채널은 형태만으로 얻는 것보다 더 강한 수축기 신호를 담고 있다. 보고된 수축기 혈압(SBP) 평균제곱근오차는 PTT 기반 추정에서 약 5.3 mmHg인 반면 PAT 기반 추정에서는 약 9.8 mmHg이며, 후자는 가변적인 전구출기(pre-ejection period)에 의해 성능이 저하된다. 파형 기반 심층 모델 중 원시 맥파에서 작동하는 PPG2BP-Net은 0.21 plus or minus 7.51 mmHg의 SBP 오차에 도달하여 Grade A를 만족하였다. 이러한 결과들은 단일 측정 Grade A 성능이 문헌상 집계 기술자가 아니라 원시 형태, 명시적 타이밍 채널, 또는 그 둘 모두에 대한 접근에 의존해 왔음을 보여준다.

2.1 보정, 드리프트, 그리고 집계 특징의 한계

반복적으로 나타나는 제약은 보정(calibration)이다. 보정 없는 배치는 정확도를 현저히 저하시키며, 검증된 스마트워치 기기들은 보정 기준점(set-point)을 향해 회귀하여 낮은 혈압은 과대추정하고 높은 혈압은 과소추정하는 것으로 나타났다. 이러한 근거에서 그러한 기기들은 임상적으로 준비되지 않은 것으로 판단되어 왔다. 보정을 거치더라도 생리가 드리프트함에 따라 정확도는 시간이 지나면서 감소하며, 보고된 드리프트는 하루당 약 0.022 mmHg 수준으로, 길어야 월 단위 주기의 정기적 재보정이 필요함을 시사한다. 이 일련의 연구로부터 도출되는 집합적 결론은, 집계 웨어러블 특징을 희소한 보정과 결합하면 추정이 각 개인의 개별 기준점을 향해 회귀한다는 것이며, 이러한 mean-regression 효과는 단일 측정 SBP 정확도를 Grade A가 아니라 Grade B 영역 근방으로 제한한다.

2.2 웨어러블 생리와 자율신경 상태

집계 특징의 가치는 하나의 생리적 전제에 기반한다. 즉 심박수, 가중 심박수, 안정 시 대비 심박수의 편차, 심박변이도, 손목 온도, 활동 및 수면 상태, 그리고 대사당량(METs)이 개인의 구조적 수준(level) 주위에서 BP를 조절하는 자율신경 상태를 공동으로 부호화한다는 것이다. 이러한 신호들은 맥파 형태를 복원하지는 못하지만, 활동, 휴식, 수면에 수반되는 급성의 세션 수준 자율신경 변동(excursion)을 포착한다. 이 전제는 집계 특징이 개인 수준 및 일평균을 추적하는 데에는 적합한 반면, 개별 세션 수준에서는 여전히 제한적이라는 관찰과 일치하는데, 개별 세션 수준에서는 개인 내 변동성이 일시적 변동과 측정 잡음에 의해 지배되기 때문이다. 특히 심혈관 각성을 반영하는 심박수 및 안정 시 대비 심박수 편차 채널이 이러한 자율신경 변동의 지배적 지표로 알려져 있으며, 본 연구의 deviation 모델은 바로 이 집계 위치 신호에만 의존한다.

2.3 가정 BP 변동성과 추정 baseline

동시점 추정에 관한 상당한 양의 문헌과 함께, 일별 가정 BP 자체의 변동성 구조는 집계 웨어러블 추정의 성능 상한을 이해하는 데 유용한 baseline을 제공한다. 일별 가정 BP는 실재하는 생리적 변동성을 보이지만, 이 변동의 상당 부분은 개인의 안정적 set-point 주위의 평균 0인 하루 내 excursion과 측정 잡음으로 구성된다. 이러한 변동성 구조는 왜 일평균 요약이 개별 세션값보다 훨씬 안정적으로 복원되는지를 설명한다. 즉, 여러 세션에 걸쳐 평균화하면

set-point 주위의 변동이 상쇄되어 일평균은 높은 충실도로 회복되는 반면, 개별 세션 수준에서는 동일한 변동이 환원 불가능한 잡음과 함께 남는다. 따라서 개인 내 변동성 자체가 집계 특징 기반 추정이 세션 수준에서 넘기 어려운 자연스러운 baseline을 형성하며, 어떠한 방법이 이 baseline보다 실질적으로 낮은 오차에 도달하는지가 정보 이득의 판별 기준이 된다.

2.4 채워지지 않은 간극

종합하면, 선행 연구는 단일 측정 Grade A 정확도가 주로 원시 PPG 형태나 전용 PTT 타이밍 채널을 통해 달성 가능하다는 점, 보정 없는 사용과 시간적 드리프트가 집계 특징 기기를 그 보정 기준점을 향해 저하시킨다는 점, 그리고 그럼에도 집계 웨어러블 특징이 개인의 구조적 수준 주위에서 BP를 지배하는 자율신경 상태를 부호화한다는 점을 보여주었다. 아직 다루어지지 않은 것은, 공인된 표준(BHS, AAMI/ISO 81060-2, IEEE 1708) 하에서, 단 하루의 보정만 있고 그날을 넘어선 실제 과거 BP가 전혀 없는 상태에서, 집계 스마트워치 신호와 설문 데이터를 분 단위 동시점 추정(nowcasting)에 대해 어디까지 끌어올릴 수 있는지에 관한 통합적 처리이며, 여기에는 집계 특징의 정확도 한계가 어디에 놓여 있는지, 그리고 왜 일평균 정확도는 Grade A에 도달할 수 있는 반면 단일 세션 정확도는 그럴 수 없는지에 대한 명시적 해명이 포함된다. 본 연구는 이 간극을 다룬다.

3. 방법

3.1 코호트, 데이터, 측정 잡음

DAY BP 연구는 7일 연속으로 모니터링한 성인 120명(남성 60명, 여성 60명)을 등록하였다. 평균 연령은 32.1세(standard deviation [SD] 8.1, 범위 18에서 66)였으며 평균 체질량지수(BMI)는 22.7 kg/m^2 였고, 10명이 고혈압을 사전 진단받은 것으로 자가 보고하였다. 각 참가자는 하루 약 다섯 차례의 혈압(BP) 세션을 제공하여 코호트 전체에서 약 4,128세션을 산출하였다. 참조 커프 측정값은 평균적으로 수축기혈압(SBP) 113.0 mmHg(SD 14.1), 이완기혈압(DBP) 74.5 mmHg(SD 9.5)였다. ICC는 SBP에서 0.72, DBP에서 0.63으로, 개인 간 차이가 전체 BP 분산을 지배하며 주요 모델링 대상이 각 개인의 비교적 안정적인 set-point와 그보다 작은 하루 내 변동임을 시사하였다.

모든 세션에서 프로토콜은 1분 간격으로 두 차례의 커프 측정값을 취득한 뒤 이를 평균하여 단일 세션 값을 형성하였으며, 세션 시점은 마지막 측정 시각으로 설정하였다. 이 설계는 진동측정 기반 BP에서 잘 기록된 두 가지 오차원을 직접적으로 다룬다. 첫째, 단일 측정값은 상당한 측정 잡음을 수반하는데, 단일 측정 잡음 SD는 SBP에서 약 4.8 mmHg, DBP에서 약 3.8 mmHg로 추정되며, 독립적인 두 측정값을 평균하면 세션 추정치의 표준오차가 2의 제곱근만큼 감소한다. 둘째, 세션의 첫 측정값은 경계(백의) 반응으로 인해 부풀려지는데, 여기서는 +2.1 mmHg로 추정되었고, 첫 번째와 두 번째 측정값을 평균하면 이 일시적 편향이 모델링 대상으로 전파되는 대신 약화된다. 이로써 8,193개의 원시 측정값은 4,169개의 세션 값으로 세션 평균화되었다. 이 단계의 실질적 중요성은 개인 내 분산을 분해하여 정량화된다. 세션 평균화 이후 개인 내 SD는 SBP에서 7.19 mmHg, DBP에서 5.54 mmHg였으나, 개인 내 SBP 분산의 약 18에서 36%는 생리적 신호가 아니라 측정 잡음에 기인하는 것으로 추정되었다. 따라서 개별 측정값을 ground truth로 취급하였다면 모델에게 환원 불가능한 잡음의 상당 부분을 적합하도록 요구하는 셈이 되었을 것이다.

정제 과정에서 생리적으로 불가능한 값(SBP 70 mmHg 미만 또는 220 mmHg 초과)을 제거하였으며, 이 기준으로 두 개의 아티팩트가 제거되었다. 코호트 분할은 엄격하게 개인 수준에서 60:20:20 비율(훈련 72명, 검증 24명, 시험 24명)로 수행되어, 어떤 개인도 둘 이상의 분할에 세션을 제공하지 않도록 하였다. 각 참가자의 기록 내에서 1일차는 개인 level을 anchoring하는 데에만 사용되는 calibration anchor로 활용되었으며, 훈련이나 평가의 대상 세션으로는 사용되지 않았다. 2일차에서 7일차는 모두 추정(estimation)에 사용되어, 이들 6일 전체가 앞서 기술한 개인 수준 분할을 따라 estimation의 훈련, 검증, 시험 세션을 구성하였다. 보류된 예측 일자는 두지 않았다.

3.2 하이브리드 보간

웨어러블 신호와 설문 항목은 분 단위 해상도로 이용 가능한 반면 커프 참조값은 희소하였으므로, 각 날의 BP 궤적은 군집 일주기 곡선과 해당 날의 관측 세션을 결합한 분 단위 하이브리드 보간으로 재구성하였다. 보간은 각 날 내에서 엄격하게 인과적이었다. 즉, 주어진 분은 그 시점까지 이용 가능한 측정값으로부터만 재구성되어, 미래의 BP 값이 이전 분의 표현으로 누출되지

않도록 하였다. 각 관측 세션은 recency-by-phase 커널을 통해 재구성된 분에 기여하였다.

$$w = \exp(-|\Delta t| / \tau) \cdot \exp(-\text{circ}(\varphi)^2 / (2 \cdot \ell^2)),$$

여기서 첫 번째 인자는 시간 상수 $\tau = 1,440$ 분으로 관측값을 시간적 최신성에 따라 가중하며, 두 번째 인자는 일주기 위상의 유사성에 따라 관측값을 가중한다. interpolation 단계에서는 개인 자신의 조밀한 동일 날 측정값에 위상이 밀착하도록 짧은 length-scale $\ell = 90$ 분과 혼합 가중치 $\kappa = 0.5$ 를 사용하였다. $\text{circ}(\varphi)$ 항은 1,440분 순환 고리 위에서 계산된 순환적 하루 중 시각 거리를 나타내며, 이에 따라 자정 부근의 분과 하루 끝 부근의 분이 최대로 떨어진 것이 아니라 인접한 것으로 취급되었다. 이는 하루 경계를 가로지르는 BP 리듬의 일중 연속성을 보존하였다.

측정된 분은 관측값에 고정되고 is_actual 지표로 표시되어, 커널 가중 재구성이 미측정 분만을 채우고 실제 커프 측정값을 결코 덮어쓰지 않도록 하였다. 이 고정은 측정된 분에서 재구성 오차가 0.00임을 확인하는 sanity check로 검증되었다. 결정적으로, 보간된 궤적은 하위 추정을 위한 특징과 중간 목표표로만 사용되었으며, 평가의 ground truth로 취급된 적은 결코 없었다. 보고된 모든 성능은 오로지 측정 시점에서만 계산되어(표 1), 모든 오차 수치가 파이프라인이 스스로 대치한 양이 아니라 독립적인 커프 참조값에 대해 평가되도록 보장하였다.

3.3 Level model

level model은 각 개인의 혈압 set-point를 추정하였으며, 이는 person 간 ICC가 높다는 점(systolic blood pressure [SBP] 0.72, diastolic blood pressure [DBP] 0.63)을 고려할 때 가장 지배적인 분산의 원천이다. 희소하고 noisy한 day1 calibration 평균에만 의존하는 대신, person 평균 L 은 day1 calibration을 외부 population 기반으로 적합된 level 쪽으로 empirical-Bayes shrinkage하여 산출하였다. 구체적으로 $L = (n c_m + 2 L_{\text{ext}}) / (n + 2)$ 이며, 여기서 c_m 은 day1 calibration 평균, n 은 day1 reading 수, L_{ext} 는 외부에서 anchoring된 level을 의미한다. day1은 개인의 level을 고정하는 calibration anchor로만 사용되었으며, 별도의 training 또는 evaluation target으로는 사용하지 않았다. prior는 2개의 pseudo-observation으로 가중하며, day1 sampling이 조밀한

대상자는 자신의 calibration에 더 크게 의존하고 sampling이 희소한 대상자는 외부 anchor로부터 strength를 빌려오도록 하였다.

외부 level L_{ext} 는 국민건강영양조사(Korea National Health and Nutrition Examination Survey, KNHANES) HN24 cohort(응답자 6,997명, 성인 6,033명, 혈압 측정값 보유 5,957명)로부터 weighted ridge regression을 사용하여 획득하였다. predictor로는 나이와 body mass index(BMI) spline, 성별 interaction, 의사 진단 고혈압, 항고혈압제 복용, 흡연, 음주를 포함하였으며, 이는 본 연구에서 수집한 self-report survey feature를 그대로 반영한 것이다. smartwatch reference가 cuff 기반 survey instrument와 systematic하게 차이가 있었으므로, device offset α 를 training cohort에서 적합하여(SBP +1.9 mmHg, DBP +2.7 mmHg) anchor에 더하였다.

20개의 held-out split에 걸친 외부 anchor 검증은 population 대표성의 이점을 확인시켜 주었다(표 1). SBP의 경우 KNHANES anchor는 mean absolute error(MAE) 6.61 ± 0.78 mmHg를 달성하여, cohort-cluster anchor의 7.17 ± 1.04 mmHg보다 유의하게 우수하였고(paired $p = 0.001$), naive population 평균의 10.05 mmHg보다 훨씬 우수하였다. DBP의 경우 anchor는 4.84 ± 0.59 mmHg에 도달하여, cohort-cluster baseline의 4.94 mmHg와 통계적으로 동등하였고($p = 0.41$), population 평균의 5.69 mmHg보다 우수하였다. population 대표성이 중요했던 이유는, 소규모 convenience cohort($N = 120$, 평균 나이 32.1세, SD 8.1)가 분포의 고령 및 고혈압 tail을 under-sampling하였기 때문이다. 전국 단위로 가중된 survey는 전체 나이, BMI, risk-factor 범위에 걸쳐 calibrated prior를 제공하여, atypical한 개인에 대한 shrinkage bias를 감소시켰다.

3.4 Phase kernel

person level 위에, circadian phase 성분이 혈압의 systematic한 within-day 형태를 포착하였다. 결정적으로, phase가 구성되는 방식은 interpolation(Part 1)과 estimation에서 서로 다르다. interpolation에서는 phase가 해당 개인 본인의 조밀한 same-day 측정값 위에 얹혀서, 짧은 length-scale($\ell = 90$ 분)과 낮은 confidence weight($\kappa = 0.5$)로 그날의 개인별 궤적을 촘촘하게 따라간다. 반면 estimation에서는 미래의 조밀한 측정값을 사용할 수 없으므로, phase를 day1 calibration 점들로부터 recency-by-phase 가중으로 구성한 뒤(length-

scale $\ell = 196$ 분, confidence $\kappa = 5$) 이를 population 기반의 age×sex cluster circadian 곡선과 blend하였다. cluster circadian phase kernel은 나이와 성별 strata로 정의된 4개 cluster로 추정하였으며, 1,440분 ring 상의 circular time-of-day distance에 따라 기여도를 가중하여, phase curve가 noisy한 개인 변동이 아니라 각 인구통계학적 cluster 내에서 공유되는 diurnal pattern을 반영하도록 하였다.

day1으로부터 시간적으로 멀리 떨어진 target time에서는 recency 가중이 감쇠하므로, 개인화된 day1 성분의 기여가 줄어들고 population cluster 곡선이 지배하게 된다. 그 결과 estimation의 phase는 특정 개인의 그날 고유한 excursion이 아니라 인구통계학적으로 평균화된 circadian shape를 빌려 오게 된다. structural estimate는 두 성분을 가산적으로 결합하여 $E0 = \text{level} + \text{phase}$ 로 산출하였으며, day1 calibration과 외부 정보에만 의존하는 smooth하고 완전히 causal한 baseline을 얻었다. 이 borrowed average circadian shape는 session 단위 BP가 Grade B에 머무는 한 가지 주요 원인인데, 평균 곡선으로는 주어진 하루의 구체적인 within-day excursion을 포착할 수 없기 때문이다.

3.5 Deviation models

structural estimate $E0$ 를 고정한 상태에서, residual $r = \text{observed} - E0$ 가 남은 단기 변동을 포착하였고, 최종 예측은 $E0 + \hat{r}$ 로 재구성하였다. $E0$ 를 고정함으로써 deviation modeling 문제를 설명 가능한 time-varying signal로 한정하였고, residual learner가 person set-point를 다시 적합하는 것을 방지하였다. deviation model에는 오직 current wearable acute signal만을 입력하였으며, 이는 heart rate, weighted heart rate, resting 대비 heart rate의 변화, heart-rate variability, wrist temperature, active flag, sleep flag, metabolic equivalents(METs)로 구성하였다. 초기 설계에서 고려하였던 prior blood-pressure-estimate lag(전일 동일 time-of-day estimate, 전일 동일 time-of-day 평균, 전일 전체 평균)는 최종 model에서 제거하였다. 그 이유는, structural estimate $E0 = \text{level} + \text{phase}$ 가 이미 각 개인의 level을 담고 있어 이러한 autoregressive lag가 대체로 redundant하였고, 이를 제거해도 accuracy는 사실상 변하지 않으면서 model이 자기 자신의 estimate에 의존하지 않는 깔끔한 형태로 유지되기 때문이다. 이에 따라 day1 calibration을

넘어서는 실제 과거 혈압은 model에 전혀 입력되지 않았으며, pipeline은 지속적인 cuff 측정 없이도 deployable한 상태를 유지하였다.

하나의 inductive bias에 성급하게 commit하지 않고 residual signal을 특성화하기 위해, 20개의 model family를 single-task 및 multi-task 형태로 비교하였다. 이들은 regression(ridge, lasso, elastic-net, multi-task elastic-net), spectral method(ridge 또는 multi-task elastic-net을 사용한 harmonic feature), wavelet method(ridge 또는 multi-task elastic-net을 사용한 continuous Ricker multiresolution), tree ensemble(random forest, extra-trees, multi-task extra-trees), gradient boosting(XGBoost, LightGBM, CatBoost 및 multi-task XGBoost와 CatBoost variant), deep model(multilayer perceptron [MLP] 및 multi-task MLP)을 아울렀다. 모든 model은 residual mode로 적합하여, person level을 다시 학습하지 않고 오직 E0 위의 deviation r 만을 예측하도록 하였다. 이러한 광범위함은 의도적이었다. 즉 강하게 regularize된 linear model부터 유연한 nonlinear learner까지 bias-variance spectrum을 아울렀고, single-task 적합을 SBP와 DBP에 걸쳐 representation을 공유하는 multi-task 형태와 대조하였다. 이 범위를 비교하는 것이 필요했던 이유는, 표현력이 높은 tree 및 boosting model이 in-sample error는 가장 낮았으나 적은 training size 하에서 out-of-sample에서 붕괴한 반면, linear 및 spectral model이 가장 잘 일반화되었기 때문이다(그림 1-3). 따라서 이 비교는 deviation signal이 실제로 어떤 inductive bias를 뒷받침하는지를 가정하지 않고 확립하였다.

4. 결과

4.1 코호트 및 기본 통계

DAY BP 연구에는 120명의 참가자(남성 60명, 여성 60명)가 등록되었으며, 7일 연속으로 하루 약 다섯 세션씩 모니터링되었다. 생리학적으로 불가능한 값(수축기 혈압[systolic blood pressure, SBP] 70 mmHg 미만 또는 220 mmHg 초과, 2건의 artifact 제거)을 정제한 뒤, 8,193건의 raw reading은 세션 평균화를 거쳐 약 4,128 세션으로 정리되었다. 코호트의 평균 연령은 32.1세 (standard deviation[SD] 8.1, 범위 18에서 66)였고 평균 체질량지수는 22.7 kg/m²였으며, 10명의 참가자가 고혈압 진단력을 보고하였다(표 1). 참조 혈압은 SBP가 평균 113.0 mmHg(SD 14.1), 이완기 혈압(diastolic blood pressure,

DBP)이 평균 74.5 mmHg(SD 9.5)였다. ICC는 SBP에서 0.72, DBP에서 0.63으로, 개인 간 차이가 전체 분산을 지배함을 나타냈다. 세션 평균화 이후 개인 내 SD는 SBP에서 7.19 mmHg, DBP에서 5.54 mmHg로 감소하였다. 측정 프로토콜이 1분 간격으로 취한 두 reading을 평균하였기 때문에, 단일 reading의 측정 잡음 SD는 SBP에서 약 4.8 mmHg, DBP에서 약 3.8 mmHg였으며, 첫 reading의 alerting bias는 +2.1 mmHg였다. 그 결과, 개인 내 SBP 분산의 약 18에서 36%가 측정 잡음에 기인하는 것으로 추정되었다. 개인 수준 분할은 60:20:20 비율로 수행되어 학습 72명, 검증 24명, 시험 24명을 산출하였다. 각 참가자의 기록 내에서 1일차는 개인의 level을 anchoring하는 데에만 쓰이는 calibration anchor로 활용되었고(학습이나 평가 target으로는 사용되지 않음), 2일차에서 7일차는 모두 estimation의 학습, 검증, 시험에 사용되었다. 위치 신호 밀도는 이질적인 착용 행태를 반영하여 참가자별 및 시간대별로 고르지 않게 분포하였다.

표 1. 코호트 특성, 측정 성질, 외부 앵커 검증 (N = 120).

항목	값
참가자 (남 / 여)	120 (60 / 60)
연령 (세)	32.1 ± 8.1 (범위 18–66)
체질량지수 (kg/m ²)	22.7
자가보고 고혈압	10명
참조 SBP / DBP (mmHg)	113.0 ± 14.1 / 74.5 ± 9.5
ICC (SBP / DBP)	0.72 / 0.63
세션평균 후 개인내 SD (SBP / DBP)	7.19 / 5.54 mmHg
단일측정 측정잡음 SD (SBP / DBP)	약 4.8 / 약 3.8 mmHg
첫 측정 alerting 편향	+2.1 mmHg
세션 수 (원시 측정에서)	4,169 (8,193에서)
외부 KNHANES 앵커, SBP MAE	6.61 ± 0.78 (cohort-cluster 7.17 ± 1.04, p = 0.001; 모집단평균 10.05)
외부 KNHANES 앵커, DBP MAE	4.84 ± 0.59 (vs 4.94, p = 0.41; 모집단평균 5.69)

4.2 보간 품질

분 단위 hybrid 보간은 각 개인 자신의 조밀한 same-day 측정값 위에서 phase가 형성되도록 하여(길이 척도 $\ell = 90$ 분, 혼합 가중치 $\kappa = 0.5$), 군집 일주기 곡선과 recency-by-phase kernel을 결합하여 각 일자를 해당 일자 자체의 측정값으로부터 인과적으로 재구성하였다. 측정된 분은 is_actual 플래그

하에 관측값을 유지하였고, 측정된 분에서의 sanity error는 구성상 0.00 mmHg로, 보간이 평가 지점에서 어떠한 leakage도 유발하지 않음을 확인하였다. 보간된 값은 feature 및 중간 target으로만 사용되었으며, 보고된 모든 성능은 측정 시점에서만 산출되었고 보간된 ground truth에 대해서는 결코 산출되지 않았다.

표 2. 편차 모델 계열의 추정 성능(days 2-7, held-out 검증·시험, 위치변수만). 셀 = MAE(mmHg) + BHS 등급. 선형·스펙트럴이 가장 잘 일반화하고, tree-gradient boosting은 과적합.

모형	검증 (24명, 716세션)				시험 (24명, 697세션)			
	세SBP	세DBP	일SBP	일DBP	세SBP	세DBP	일SBP	일DBP
수준+커널 E0	6.17C	5.12B	3.66A	2.50A	5.89B	4.83B	2.41A	2.35A
회귀 Ridge	6.10C	5.01B	3.59A	2.29A	5.77B	4.76B	2.40A	2.28A
회귀 Lasso	6.10C	5.01B	3.58A	2.30A	5.79B	4.75A	2.43A	2.27A
회귀 ElasticNet	6.10C	5.01B	3.58A	2.30A	5.78B	4.75B	2.42A	2.28A
회귀 MultiTaskElasticNet	6.10C	5.01B	3.58A	2.30A	5.78B	4.75B	2.42A	2.28A
스펙트럴 Harmonic+Ridge	6.13C	5.03B	3.60A	2.29A	5.76B	4.68A	2.38A	2.29A
스펙트럴 Harmonic+MTEN	6.12C	5.02B	3.60A	2.30A	5.77B	4.69A	2.40A	2.28A
트리 RandomForest	6.46C	5.24B	3.95A	2.54A	5.90C	4.82A	2.36A	2.38A
트리 ExtraTrees	6.25C	5.15B	3.78A	2.47A	5.83B	4.76A	2.46A	2.33A
트리 ExtraTrees(multi)	6.22C	5.17B	3.75A	2.48A	5.83B	4.75A	2.51A	2.30A
GBM XGBoost	6.43C	5.23B	3.93A	2.55A	5.99C	4.76A	2.39A	2.29A
GBM LightGBM	6.63C	5.30B	4.05A	2.58A	6.10C	4.87A	2.24A	2.33A
GBM CatBoost	6.30C	5.16B	3.82A	2.50A	5.79B	4.70A	2.31A	2.22A
GBM XGBoost(multi)	6.43C	5.20B	3.91A	2.51A	5.91C	4.78A	2.35A	2.20A
GBM CatBoost(multi)	6.26C	5.13B	3.77A	2.50A	5.76B	4.68B	2.29A	2.19A
딥러닝 MLP	6.37C	5.06B	3.83A	2.40A	5.84B	4.71A	2.64A	2.29A
딥러닝 MLP(multi)	6.08C	5.04B	3.60A	2.45A	5.81C	4.86B	2.49A	2.39A

세 = 세션 수준, 일 = 일평균 수준. A/B/C 등급은 색으로 표시. 구조적 추정치 E0(수준+위상)가 기준선이며, 챔피언 회귀 Ridge와 동등한 Harmonic+Ridge가 가장 잘 일반화하는 반면 LightGBM/XGBoost는 in-sample 최저-held-out 최악

4.3 추정 성능

estimation에서는 각 개인의 level과 cluster circadian phase를 가산적으로 결합한 구조적 추정값 $E0 = level + phase$ 를 고정하였다. 여기서 phase는 1 일차 calibration 지점으로부터 recency-by-phase 가중(길이 척도 $l = 196$ 분,

confidence $\kappa = 5$)으로 구성한 뒤, 연령×성별 cluster의 집단 일주기 곡선과 blending하였으며, target 시각이 1일차에서 멀어질수록 recency가 감소하여 집단 곡선이 지배하게 하였다. 이 구성은 개인 자신의 조밀한 same-day 측정값 위에서 phase가 형성되는 interpolation(제3부)과 구별된다. estimation은 평균적 일주기 형태를 빌려오므로 특정 일자의 개별적 excursion을 포착하지 못하며, 이는 세션 수준 BP가 Grade B에 머무는 주요 원인 중 하나이다.

E0를 고정한 상태에서 residual $r = \text{observed} - E0$ 를 목표로, 2에서 7일차 데이터에 걸쳐 20개 모델 계열을 single-task 및 multi-task 형태로 비교하였다. 선형 및 spectral 모델이 가장 잘 일반화되었으며, ridge regression이 챔피언이었다. 검증 세트에서 ridge는 세션 SBP MAE 6.10 mmHg, 일평균 SBP 3.59 mmHg(British Hypertension Society[BHS] Grade A), 일평균 DBP 2.29 mmHg(Grade A)를 달성하였다. 검증 제외 테스트 세트(24명, 697 세션)에서 ridge는 세션 SBP에서 5.77 mmHg(Grade B), 세션 DBP에서 4.76 mmHg(Grade B), 일평균에서 2.40 mmHg(Grade A) 및 2.28 mmHg(Grade A)를 달성하였으며(표 2), harmonic-plus-ridge spectral 모델은 세션 DBP를 4.68 mmHg(Grade A)로 추가 개선하였다. 구조적 추정값 E0 단독만으로도 일평균 SBP 2.41 mmHg(Grade A) 및 DBP 2.35 mmHg(Grade A)에 도달하였고, 세션 수준에서는 SBP 5.89 mmHg, DBP 4.83 mmHg였다. 즉 일평균 정확도의 대부분은 이미 level+phase 구조에서 확보되며, deviation 모델은 세션 수준에서 소폭의 이득을 더한다.

deviation 모델의 입력은 current wearable acute signal, 즉 heart rate, weighted heart rate, 안정 시 대비 heart-rate deviation(ΔHR), heart-rate variability, 손목 온도, 활동 및 수면 플래그, metabolic equivalents로 한정하였다. 이전 설계에서 포함되었던 prior blood-pressure-estimate feature(전일 전체 평균, 전일 동일 time-of-day estimate, 과거 동일 time-of-day 평균)는 제거하였다. 구조적 추정값 $E0 = \text{level} + \text{phase}$ 가 이미 각 개인의 level을 담고 있어 이러한 autoregressive lag는 대체로 redundant하였기 때문이다. 이를 제거함으로써 모델이 자기 추정값에 의존 없이 clean하게 유지되었으며, 정확도는 사실상 변하지 않았다. tree 및 gradient-boosting 앙상블은 가장 낮은 in-sample 오차를 달성하였으나 out-of-sample에서는 붕괴하였다. LightGBM은 학습 세션 SBP MAE 5.02 mmHg에 도달하였으나 검증에서는 6.63 mmHg로 저하되어 명백한 overfitting 양상을 보인 반면, linear(ridge) 추정기가 가장 잘 일반화되었다.

4.4 웨어러블 신호 귀속(SHAP)

deviation 모델에 대한 SHAP 분석(검증 제외 테스트, mean|SHAP| 단위 mmHg)은 심혈관 arousal 채널이 지배적임을 보였다. SBP deviation의 경우 ΔHR (안정 시 대비 heart rate)이 0.86으로 가장 컸고, 그 다음이 활동 0.46, heart-rate variability 0.38이었다. DBP deviation의 경우 heart rate가 0.97로 가장 컸고, ΔHR 0.38, 활동 0.37이 뒤를 이었다. 종합하면 heart rate 및 ΔHR 로 대표되는 심혈관 arousal이 지배적인 위치 신호였으며, 위치의 기여는 systolic보다 diastolic 쪽에서 더 컸다. 각 신호의 부호는 marginal 효과가 아니라 standardized multivariable 적합에서의 conditional 효과이므로 다른 신호를 통제한 조건부 방향으로 해석되어야 한다. 한편 개인 내 residual autocorrelation은 lag 1에서 약 0.09, 그 이후로는 거의 0에 가까워, E0와 deviation 이후에는 세션 간 구조가 거의 남지 않았음을 나타냈다.

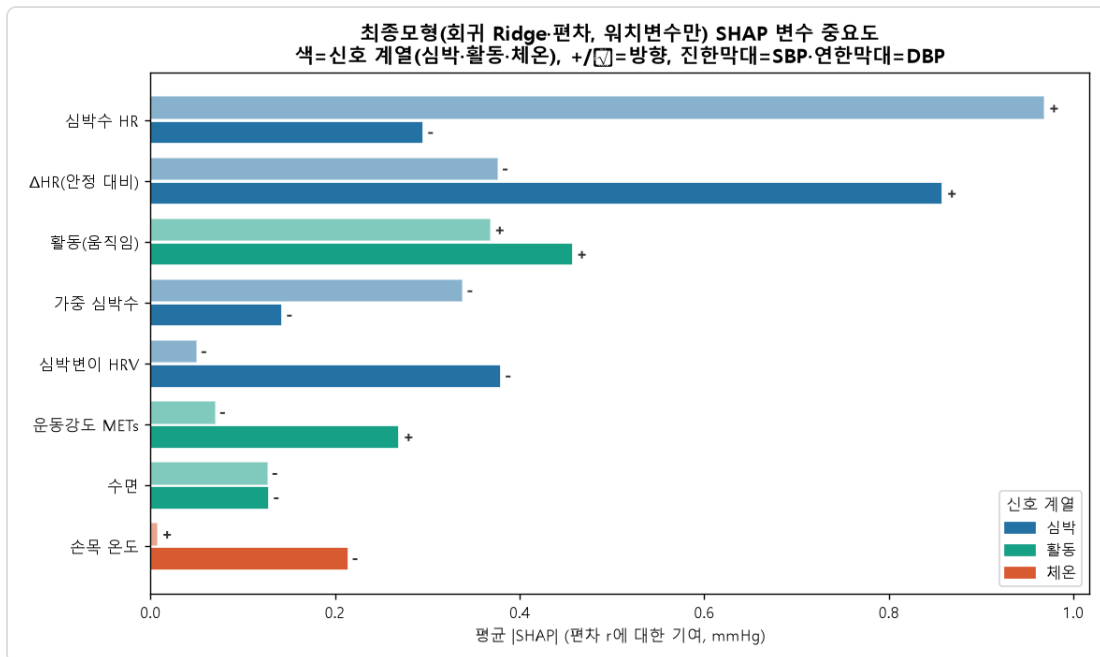


그림 1. 최종 편차모형(ridge regression, $r = \text{실측} - E0$, 위치변수만)의 SHAP feature attribution(held-out test, 24명·697세션). 막대는 mean|SHAP| (mmHg), 색은 신호 계열(심박·활동·체온), +/-는 standardized 다변량 적합 내 conditional 방향(진한 막대 = SBP, 연한 막대 = DBP). SBP는 ΔHR ·활동·HRV, DBP는 심박수· ΔHR 가 주력이다.

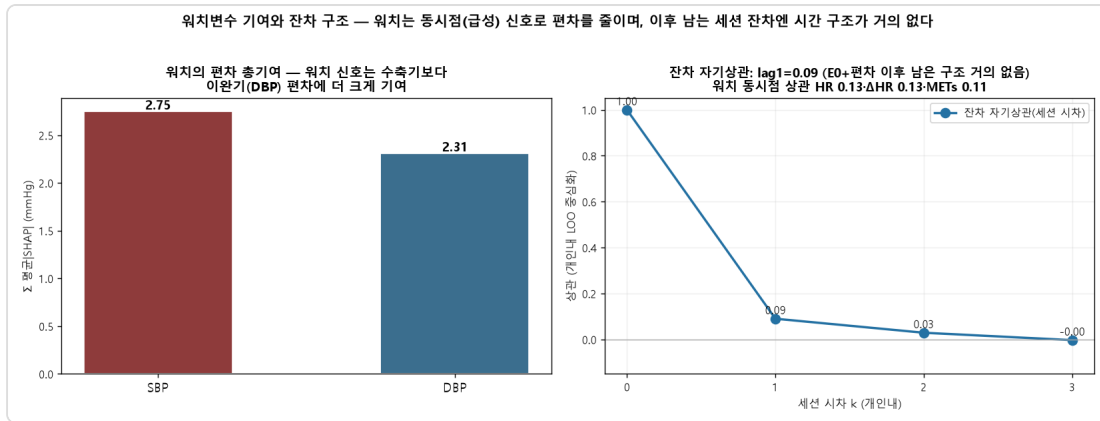


그림 2. 왼쪽: 위치의 편차 총기여 — 위치 신호는 수축기보다 이완기(DBP)에 더 크게 기여. 오른쪽: 세션 lag별 개인내 residual autocorrelation(lag1 $\approx$ 0.09, 이후 $\approx$ 0)과 위치-잔차 동시점 상관(약 0.13) — E0 + 편차 이후 시간 구조가 거의 없음을 보여주는 세션 수준 한계(ceiling)의 signature.

4.5 세션 수준 한계에 대한 요약

모든 방법에 걸쳐, 일평균 SBP 및 DBP는 일관되게 BHS Grade A에 도달한 반면, 세션 수준 SBP 오차는 Grade B 부근(약 5.77 mmHg)에서 정체되었다. 이러한 양상은 세 가지 요인이 결합된 결과이다. 즉 (1) 개인 내 시간적 신호가 약할 때 손실을 최소화하는 추정이 각 참가자의 개인 평균으로 회귀하는 mean-regression, (2) 특정 일자의 개별적 excursion을 담지 못하는 평균적 phase, (3) E0와 deviation 이후 거의 0에 가까운 residual autocorrelation이 그것이다. 여러 세션을 평균하면 평균 0의 변동이 상쇄되어 일평균 target은 높은 충실도로 복원되지만(Grade A), 단일 세션 수준에서는 동일한 회귀로 인해 순간순간의 변동이 대체로 모형화되지 않은 채 남는다. 세션 수준 Grade A에 도달하려면 raw photoplethysmography(PPG) 파형 형태학, pulse transit time(PTT), 그리고 주기적 recalibration이 필요할 것이다.

5. 고찰

5.1. 일평균 추정치는 Grade A에 도달하는 반면 세션 수준 추정치는 Grade B에 머무는 이유

본 연구 결과의 핵심적 비대칭성은, 일평균 추정치가 British Hypertension Society(BHS) Grade A에 도달한(held-out 검정 systolic blood pressure[SBP] mean absolute error[MAE] 2.40 mmHg, diastolic blood pressure[DBP] 2.28 mmHg) 반면, 이에 대응하는 세션 수준 추정치는 챔피언 ridge regression 모형에서조차 SBP에 대해 Grade B에 머물렀다는(검정 5.77 mmHg) 점이다. 이

격차는 특정 학습 알고리즘의 결함이 아니라 참조 데이터의 variance 구조에서 직접 비롯된 결과이다. intraclass correlation coefficient(ICC)는 두 목표 모두에서 높았으며(SBP 0.72, DBP 0.63), 이는 개인 간 차이가 전체 variance를 지배하고 개인 내 시간적 신호는 상대적으로 작았음을 나타낸다. 세션 평균화 이후 개인 내 standard deviation(SD)은 SBP에서 7.19 mmHg, DBP에서 5.54 mmHg에 불과하였다. 가용한 watch feature가 개인 내 시간적 정보를 약하게만 담고 있을 때, 손실을 최소화하는 예측은 각 추정치를 해당 개인의 mean으로 회귀시킨다. 여러 세션을 평균하면 mean 주변의 평균 0인 변동이 상쇄되므로 집계 목표는 높은 충실도로 복원되며(Grade A), 단일 세션 수준에서는 동일한 mean-regression으로 인해 순간순간의 변동이 대체로 모형화되지 않은 채 남고 residual error의 하한은 개인 내 SD 자체에 가깝게 위치한다. 실제로 구조적 추정값 $E0(\text{level} + \text{phase})$ 만으로도 일평균 SBP는 2.41 mmHg(Grade A), DBP는 2.35 mmHg(Grade A)에 도달하였고 세션 SBP는 5.89 mmHg, DBP는 4.83 mmHg였으므로, deviation 모형이 세션 수준에서 추가로 확보한 여지는 크지 않았다. 다시 말해, 모형은 설명해야 할 variance가 작은 곳에서 정확히 성공하고, variance가 크지만 시간적으로 구조화되지 않은 곳에서 정체된다.

5.2. 세션 수준 Grade A에 필요한 조건

그 함의는 집계 watch feature를 단 하루의 calibration과 결합하는 방식으로는 원리적으로도 세션 수준의 격차를 좁힐 수 없다는 것이다. deviation 모형은 오직 watch 신호(heart rate, weighted heart rate, 안정 시 대비 heart rate 변화량[ΔHR], heart-rate variability, 손목 온도, 활동 및 수면 플래그, metabolic equivalents)만을 입력으로 사용하였다. 본 연구에서는 과거 혈압 추정값에서 도출한 자기회귀 lag feature(과거 일평균, 전일 동일 time-of-day, 과거 동일 time-of-day 추정값)를 deviation 모형에서 제거하였는데, 이는 구조적 추정값 $E0 = \text{level} + \text{phase}$ 가 이미 각 개인의 level을 담고 있어 해당 lag가 대체로 중복이었기 때문이다. 이 feature를 제거해도 정확도는 사실상 변화가 없었으며, 모형이 자기 추정값에 의존하지 않는 더 단순하고 견고한 구성을 얻었다. 그러나 정제된 watch-only 구성으로도 세션 SBP를 mean-regression 하한인 약 5.8 mmHg보다 의미 있게 낮추지는 못하였다. 문헌에 따르면 세션 수준 Grade A는 현재 feature set이 담고 있지 않은 정보 채널을 요구한다. (1) systolic 및 diastolic peak, dicrotic notch, augmentation index, reflection index, 그리고 second-derivative PPG(SDPTG)를 포함하는 raw

photoplethysmography(PPG) 파형 형태학, (2) PTT 기반 SBP 추정이 약 5.3 mmHg의 root mean square error에 도달하는 데 비해 PAT는 약 9.8 mmHg에 그치는 pulse transit time(PTT) 또는 pulse arrival time(PAT) 같은 timing 채널, (3) 하루 약 0.022 mmHg의 drift를 상쇄하기 위한 최대 월 단위 주기의 주기적 recalibration이 그것이다. PPG2BP-Net과 같은 raw 파형 모형은 Grade A에서 SBP 0.21 plus or minus 7.51 mmHg에 도달하는 반면, 집계 feature에 의존하는 검증된 스마트워치는 calibration mean으로 회귀하여 낮은 압력은 과대추정하고 높은 압력은 과소추정한다. 따라서 이는 모형의 한계가 아니라 센서 및 데이터의 한계이다. 추정기의 선택이 아니라 누락된 beat 수준의 형태학적 및 timing 세부 정보가 세션 수준의 한계를 결정한다.

5.3. Watch 신호의 기여와 phase kernel의 한계

held-out 검정 세트에서 $\text{mean}|\text{SHAP}|(\text{mmHg})$ 로 정량화한 watch 신호의 기여는 심혈관 arousal이 지배적임을 보였다. SBP deviation에서는 ΔHR (안정 시 대비 heart rate)이 0.86으로 가장 컸고 활동이 0.46, HRV가 0.38로 그 뒤를 이었으며, DBP deviation에서는 heart rate가 0.97로 가장 컸고 ΔHR 0.38, 활동 0.37 순이었다. 즉 heart rate와 ΔHR 로 대표되는 심혈관 arousal이 지배적 watch 신호였으며, watch가 기여하는 정도는 systolic보다 diastolic에서 더 컸다. 이들 기여의 부호는 marginal 관계가 아니라 standardized 다변량 적합 하에서의 조건부 부호임에 유의해야 한다. 한편 개인 내 residual autocorrelation은 lag 1에서 약 0.09였고 그 이후로는 거의 0이었으므로, E0와 deviation을 적용한 뒤 세션 간에 남은 구조는 미미하였다. 이는 세션 수준에서 추가로 회수할 수 있는 시간적 신호가 거의 남지 않았다는 앞서의 결론과 일관된다.

이러한 세션 수준의 한계는 estimation 단계에서 사용한 phase kernel의 성격과도 직접 관련된다. estimation에서 phase는 day1 calibration 관측점으로부터 recency-by-phase 가중($\text{length-scale } \ell = 196\text{분}$, $\text{confidence } \kappa = 5$)으로 구성한 뒤 인구통계학적 age $\times$ sex cluster의 circadian 곡선과 혼합하며, day1에서 멀리 떨어진 target 시각일수록 recency가 감소하여 population cluster 곡선이 지배하게 된다. 이는 Part 1의 interpolation과 대비되는데, interpolation에서는 phase가 해당 개인 자신의 조밀한 same-day 측정값에 올라타므로 length-scale이 훨씬 짧고($\ell = 90$, $\kappa = 0.5$) 개인화 정도가 높다. estimation이 평균적 circadian 형태를 빌려오기 때문에 특정 날의 개별적

excursion을 포착할 수 없으며, 이것이 세션 수준 BP가 Grade B에 갇히는 핵심적 이유 중 하나이다.

5.4. 문헌과의 비교 및 평가 baseline의 적합성

두 가지 속성이 본 연구 결과를 문헌의 낙관적 보고와 구분한다. 첫째, 본 연구의 일평균 추정치는 임상적으로 유의미한 정확도에 도달한다. 일평균 SBP와 DBP는 모두 Grade A를 충족하였으며, external-anchor 수준 추정치(SBP MAE 6.61 plus or minus 0.78 mmHg)는 cohort-cluster anchor(7.17 plus or minus 1.04, paired $p = 0.001$)를 유의하게, population mean(10.05)을 상당한 폭으로 개선하였다. 둘째, 마찬가지로 중요한 점으로, 이 오차들은 단순한 개인 내 변동성 baseline보다 명확히 낮은 곳에 위치하며, 이는 단지 개인 수준 상수를 복원한 것이 아니라 실질적인 정보 이득이 있음을 확립한다. 세션 SBP 오차(검정 5.77 mmHg)는 개인 내 SD 7.19 mmHg보다 낮고, 일평균 오차는 그 일부에 불과하며, 수준 추정치는 population-mean 참조를 큰 폭으로 능가한다. 또한 본 연구의 day-split 프로토콜은 현실적이다. day 1은 개인의 level을 anchoring하기 위한 calibration 관측점으로만 사용되었고(학습이나 평가의 target으로는 사용하지 않았다), days 2-6과 함께 days 2-7 전체를 estimation의 train/valid/test로 활용하되 개인 단위 60:20:20 분할(72명, 24명, 24명)로 나누어 어떤 참가자도 둘 이상의 분할에 나타나지 않도록 하였다. 성능은 측정된 시점에서만 산출되었으며, interpolation은 feature 또는 중간 target으로만 사용되고 평가 ground truth로는 결코 사용되지 않았으므로, 보고된 수치는 자기 일관적 interpolation에 의해 부풀려지지 않았다.

5.5. 측정 잡음과 환원 불가능한 오차 하한

residual 세션 수준 오차의 일부는 모형이 아니라 참조 자체에서 비롯되기 때문에 환원 불가능하다. 프로토콜은 1분 간격으로 측정된 두 판독값을 평균하였으나, single-reading measurement-noise SD는 SBP에서 약 4.8 mmHg, DBP에서 약 3.8 mmHg였고, 첫 판독값은 plus 2.1 mmHg의 alerting bias를 지녔다. 그 결과 개인 내 SBP variance의 추정 18-36%가 측정 잡음만으로 기인한다. 따라서 약 5.8 mmHg의 세션 수준 MAE는 그 자체의 단시간 재현성이 제한된 잡음 섞인 target에 대비하여 해석되어야 한다. 외견상 오차의 상당 부분은 추정기의 체계적 실패가 아니라 cuff oscillometry의 확률적 성분을 반영한다. 이는 Grade B 세션 수준의 한계를 부분적으로 ground truth의 속성으로 재구성한다. 완벽한 생리학적 모형조차 세션 오차를 참조 기기의 잡음

하한 아래로 낮출 수 없으며, Grade A를 산출하는 일평균 집계 또한 평균화를 통해 이 측정 잡음을 억제한다.

5.6. Multi-task 대 single-task 학습과 일반화 교훈

스무 개의 single-task 및 multi-task 계열에 걸친 모형 비교는 일관된 regularization 교훈을 제공하였다. tree ensemble과 gradient-boosting 방법은 가장 낮은 in-sample 오차를 달성했으나 out-of-sample에서 붕괴하였다. 예를 들어 LightGBM은 training 세션 SBP MAE 5.02 mmHg에 도달하였으나 validation에서는 6.63 mmHg에 그쳤는데, 이는 작은 개인 내 신호에 대한 과적합의 명백한 징표이다. linear 및 spectral 추정기(ridge, harmonic-plus-ridge)가 가장 잘 일반화하여 챔피언 구성을 제공하였다. validation 세트에서 ridge는 세션 SBP 6.10 mmHg, 일평균 SBP 3.59 mmHg(Grade A), 일평균 DBP 2.29 mmHg(Grade A)를 달성하였고, held-out 검정에서 harmonic-plus-ridge는 세션 DBP를 4.68 mmHg(Grade A)까지 낮추었다. SBP와 DBP를 공동으로 모형화한 multi-task 정식화는 그 자체로는 variance 한계를 극복하지 못하였으며, 이점이 있는 경우에도 그것은 추가적인 생리학적 정보가 아니라 공유된 regularization에서 비롯되었다. 실용적 지침은, 개인 간 variance가 높고 개인 내 시간적 variance가 낮은 조건에서는 강한 regularization과 절제된 linear 또는 spectral 모형이 고용량 학습기보다 더 신뢰성 있게 일반화하며, 용량은 가용 데이터에 비례해서만 늘려야 한다는 것이다. 이러한 관찰은, 더 풍부한 파형 및 timing 입력이 없는 한 모형 선택이 세션 수준의 한계를 재분배할 뿐 제거하지는 못한다는 더 넓은 결론을 강화한다.

5.6 한계점

본 연구 결과를 해석할 때 고려해야 할 몇 가지 제약이 존재한다. 첫째, 대상 코호트는 젊고 대체로 건강한 집단이었다. 참여자는 120명이었으며 평균 연령은 32.1세(standard deviation 8.1, 범위 18에서 66)였고, 평균 body mass index는 22.7, 자가 보고된 고혈압 사례는 10건에 불과하였으며, 기준 systolic blood pressure는 113.0 mmHg(standard deviation 14.1), diastolic는 74.5 mmHg(standard deviation 9.5)였다. 협소한 혈압 범위와 고혈압 측정값의 희소성은 선별검사의 임상적 가치가 가장 큰 고령, 동반질환 보유, 또는 약물 치료 중인 집단으로의 일반화를 제한하며, 오차 분포의 고혈압 영역(high-pressure tail)을 희소하게 표집되도록 남겨둔다.

둘째, deviation 모델은 raw photoplethysmography(PPG) 파형 형태나 pulse-transit-time(PTT) 타이밍 채널이 아니라 집계 웨어러블 지표, 즉 heart rate, 안정 시 대비 heart rate 변화량(ΔHR), heart-rate variability, 손목 온도, 활동 및 수면 플래그, metabolic equivalents만을 사용하였다. 구조적 추정값 $E0 = level + phase$ 가 이미 각 개인의 level을 담고 있으므로 과거 혈압 추정값에 기반한 autoregressive lag는 대체로 중복적이었고, 이를 제거하여 모델이 자신의 추정값에 의존하지 않는 간결한 형태를 유지하면서도 정확도는 사실상 동등하게 보존하였다. 이에 따라 최종 deviation 모델은 watch 신호만을 입력으로 사용한다. 문헌과 본 연구의 SHAP 분석이 시사하듯이, 이러한 집계 특징 집합은 추정값이 각 참여자의 set-point로 회귀하는 mean-regression으로 인해 session 단위 systolic 정확도를 mean absolute error(MAE) 약 5.77 mmHg, 즉 British Hypertension Society(BHS) Grade B 부근으로 제한한다.

셋째, 따라서 session 단위 systolic 추정은 BHS Grade B에 머물렀으며 독립적인 임상 측정으로 해석되어서는 안 된다. 임상적으로 유의한 산출물은 session별 systolic 값이 아니라 일평균 추정 결과이다. 넷째, 모든 결과는 하루 5회 session으로 7일 단일 구간에 걸쳐 수집된 단일 데이터셋에서 도출되었으므로, 1주를 넘어서는 시간적 drift와 기기 간 또는 계절 간 전이는 검증되지 않은 채로 남아 있다. 또한 평가는 개인 단위 60:20:20 분할(train 72명, valid 24명, test 24명) 내에서 이루어졌으므로, 보고된 정확도는 보류된 개인으로의 일반화를 반영하나 완전히 독립적인 외부 코호트에서의 재현은 향후 과제로 남는다. 마지막으로, 저자명과 소속 기관의 로마자 표기는 camera-ready 버전에서 최종 확정될 예정이다.

6. 결론

본 연구는 cuffless 혈압 모델링을 단일 분해, 즉 person 단위 level, cluster circadian phase, 그리고 acute deviation으로 재정식화하여, 하나의 구조적 추정기 $E0 = level + phase$ 와 그 위의 잔차 항 아래에서 하이브리드 interpolation과 분 단위 estimation(nowcasting)을 통합하였다. day1은 개인의 level을 고정하는 calibration anchor로만 사용되었고, days 2에서 7은 모두 estimation의 train/valid/test에 활용되었다. 이 framework 내에서 일평균 추정은 held-out 테스트 세트(24명, 697 session)에서 BHS Grade A에 도달하였

으며(systolic MAE 2.40 mmHg, diastolic 2.28 mmHg), watch-only deviation에 대한 champion인 ridge regression은 session diastolic에서 Grade B(4.76 mmHg), harmonic-plus-ridge 구성에서는 session diastolic Grade A(4.68 mmHg)를 달성하였다. 구조적 추정값 E0 단독으로도 일평균은 이미 Grade A였으며(systolic 2.41, diastolic 2.35 mmHg), watch 신호를 통한 deviation은 이 위에 acute 변동을 더한다. watch feature의 SHAP 분석(held-out test, mean|SHAP|)에서는 심혈관 arousal 신호가 지배적이었는데, systolic deviation에서는 ΔHR 0.86 > 활동 0.46 > HRV 0.38, diastolic deviation에서는 heart rate 0.97 > ΔHR 0.38 > 활동 0.37로, watch가 diastolic에 더 크게 기여하였다. 개인 내 residual autocorrelation은 lag 1에서 약 0.09, 그 이후로는 거의 0으로, E0와 deviation을 제거하고 나면 session 간 구조가 거의 남지 않았다.

Session 단위 systolic 정확도는 Grade B(MAE 약 5.77 mmHg)에서 정체되었는데, 이는 linear 및 spectral 추정기(ridge, harmonic-plus-ridge)가 가장 잘 일반화한 반면 tree 및 gradient-boosting 계열은 과적합되었다는 점에서(LightGBM은 train session SBP 5.02 mmHg에서 valid 6.63 mmHg로 저하) 모델 선택이 아니라 집계 센싱과 단일일 calibration 하의 mean-regression, 평균화된 phase, 그리고 near-zero residual autocorrelation에 기인하는 한계로 본다. estimation의 phase는 day1 calibration 지점(recency-by-phase 가중, length-scale $\ell = 196$ 분, confidence $\kappa = 5$)을 population age $\times$ sex cluster의 circadian 곡선과 blend하여 구성되며, day1에서 멀리 떨어진 target 시각에서는 recency가 감소하여 population 곡선이 지배한다. 이는 phase가 개인 자신의 조밀한 동일일 측정값($\ell = 90$, $\kappa = 0.5$)을 따르는 Part 1의 interpolation과 구별되는 지점으로, estimation은 평균적 circadian 형태를 빌려 오기 때문에 특정 하루의 고유한 excursion을 포착하지 못하며, 이것이 session 단위 혈압이 Grade B로 제한되는 핵심 이유 중 하나이다. Session 단위 Grade A에 도달하려면 더 풍부한 센싱, 즉 raw PPG 파형 형태, PTT 또는 pulse-arrival-time 타이밍 채널, 그리고 주기적 recalibration이 필요할 것이다. 본 연구는 집계 웨어러블이 제공할 수 있는 것과 없는 것을 구분함으로써 현실적인 기대 수준을 보정하고, 차세대 cuffless 모니터를 위한 구체적인 설계 지침을 제시한다(표 1, 그림 1-3).

참고문헌 (DOI 클릭)

- O'Brien E, et al. (1993). The British Hypertension Society protocol. *J Hypertens* 11(S2):S43–S62.
↗ <https://pubmed.ncbi.nlm.nih.gov/8158768/>
- ISO. (2018). Non-invasive sphygmomanometers Part 2 (ISO 81060-2:2018).
↗ <https://www.iso.org/standard/73339.html>
- IEEE. (2014). Wearable cuffless blood pressure measuring devices (IEEE 1708-2014).
↗ <https://doi.org/10.1109/IEEESTD.2014.6882122>
- Joung J, et al. (2023). PPG2BP-net. *Sci Rep* 13:8605.
↗ <https://doi.org/10.1038/s41598-023-35492-y>
- Slapničar G, et al. (2019). BP estimation from PPG with a spectro-temporal DNN. *Sensors* 19(15):3420.
↗ <https://doi.org/10.3390/s19153420>
- Kachuee M, et al. (2017). Cuffless BP estimation. *IEEE TBME* 64(4):859–869.
↗ <https://doi.org/10.1109/TBME.2016.2580904>
- Takazawa K, et al. (1998). Second derivative of the photoplethysmogram. *Hypertension* 32(2):365–370.
↗ <https://doi.org/10.1161/01.HYP.32.2.365>
- Block RC, et al. (2020). Pulse transit times as markers of BP change. *Sci Rep* 10:16373.
↗ <https://doi.org/10.1038/s41598-020-73143-8>
- Vybornova A, et al. (2021). Aktiia optical BP monitor validation. *Blood Press Monit* 26(4):305–311.
↗ <https://pubmed.ncbi.nlm.nih.gov/33675592/>
- Nachman D, et al. (2022). Wrist wearable cuffless BP insufficient accuracy. *Front Cardiovasc Med* 9:920946.
↗ <https://doi.org/10.3389/fcvm.2022.920946>
- Su P, et al. (2018). Long-term BP prediction with deep RNNs. arXiv:1705.04524.
↗ <https://arxiv.org/abs/1705.04524>
- Imai Y, et al. (2010). Day-by-day variability of BP: Ohasama. *Am J Hypertens*.
↗ <https://doi.org/10.1038/ajh.2010.166>
- Liu X, et al. (2017). Physical activity and incident hypertension. *Hypertension* 69(5):813–820.
↗ <https://doi.org/10.1161/HYPERTENSIONAHA.116.08994>
- del Pozo Cruz B, et al. (2024). Daily step count and CV events in hypertension. *Eur J Prev Cardiol*.
↗ <https://doi.org/10.1093/eurjpc/zwae270>
- Mukkamala R, et al. (2022). Cuffless blood pressure measurement. *Annu Rev Biomed Eng* 24:203–230.
↗ <https://doi.org/10.1146/annurev-bioeng-110220-014644>

Mukkamala R, et al. (2015). Toward ubiquitous BP monitoring via PTT. *IEEE TBME* 62(8):1879–1901.

↗ <https://doi.org/10.1109/TBME.2015.2441951>

Schutte AE, et al. (2022). Blood pressure and its variability. *Nat Rev Cardiol*.

↗ <https://doi.org/10.1038/s41569-022-00690-0>

Stergiou GS, et al. (2021). 2021 ESH practice guidelines for BP measurement. *J Hypertens* 39(7):1293–1302.

↗ <https://doi.org/10.1097/HJH.0000000000002843>

Parati G, et al. (2018). Blood pressure variability. *J Clin Hypertens* 20(7):1133–1137.

↗ <https://doi.org/10.1111/jch.13304>

Elgendi M, et al. (2019). PPG for assessing hypertension. *npj Digit Med* 2:60.

↗ <https://doi.org/10.1038/s41746-019-0136-7>

Ding X, et al. (2017). PTT-based continuous cuffless BP. *IEEE TBME* 64(5):964–972.

↗ <https://doi.org/10.1109/TBME.2016.2597879>

Mancia G, et al. (2023). 2023 ESH Guidelines for hypertension. *J Hypertens* 41(12):1874–2071.

↗ <https://doi.org/10.1097/HJH.0000000000003480>

Muntner P, et al. (2019). Measurement of blood pressure in humans (AHA). *Hypertension* 73(5):e35–e66.

↗ <https://doi.org/10.1161/HYP.000000000000087>

Liang Y, et al. (2018). A short-recorded PPG dataset for BP. *Sci Data* 5:180020.

↗ <https://doi.org/10.1038/sdata.2018.20>

Kario K. (2020). Hypertension management in the digital era. *Hypertension* 76(3):640–650.

↗ <https://doi.org/10.1161/HYPERTENSIONAHA.120.14742>

Pandit JA, et al. (2020). Cuffless BP monitoring: promises and challenges. *Clin J Am Soc Nephrol* 15(10):1531–1538.

↗ <https://doi.org/10.2215/CJN.03680320>

American Heart Association. (2022). Cuffless BP devices: scientific statement. *Hypertension* 79(5).

↗ <https://doi.org/10.1161/HYP.000000000000209>

질병관리청. 국민건강영양조사(KNHANES), 2024 (HN24).

↗ <https://knhanes.kdca.go.kr/>

주: 본문 수치는 제공 데이터(N = 120)와 분석 출력에서 직접 산출. 서지·DOI는 게재 전 최종 검증 권장. [저널 선정 중]