

Detecting Elevated Blood Pressure from a Sensor-Embedded Toilet Seat: A Mixture-of-Experts Classifier Evaluated in Home and Clinical Settings

Hadi Wehbe, Staff Algorithm Engineer
Casana Care, Inc. · 10 September 2026

Abstract

Background. Hypertension affects about half of U.S. adults and only one in five has it controlled, largely because every established modality asks the patient to decide to measure, far less often than guidelines assume. Passive measurement, during a behavior people already perform daily, removes that decision, but is useful only if its alerts can be trusted.

Aims. We asked whether a classifier reading the Casana Smart Seat, a toilet seat that passively records electrical, mechanical, and optical cardiac signals and body weight, can identify elevated blood pressure at a false-alarm rate low enough to act on, and whether clinical performance carries over to unsupervised home use.

Methods. 1,101 recordings from 96 participants were analyzed, all from protocols Casana designed and ran under institutional review board approval with informed consent. Fourteen used a seat at home for 45 days with a supplied cuff (433 recordings); 82 contributed 668 clinic recordings with simultaneous auscultation. Three models trained at different decision boundaries were combined by a fixed gating rule, and calls aggregated by vote per participant-week at home and per participant in clinic, against the 130/80 mmHg ACC/AHA boundary.

Results. Per participant in clinic, sensitivity was 0.75 at specificity 0.94 (positive predictive value 0.92, $n = 32$). At home, sensitivity was 0.42 with no false alarms across 40 evaluable participant-weeks, though that perfect specificity rests on nine normal weeks from 14 people. Excluding near-boundary cases raised sensitivity to 0.80 in clinic and 0.56 at home.

Conclusion. The contribution is passive, repeated measurement paired with a classifier that rarely raises a false alarm: a flag is informative, a quiet result is not reassuring (negative predictive value 0.33 at home). This supports use as a passive trigger for confirmatory cuff measurement, not a substitute.

1. Introduction

Hypertension is common, consequential, and mostly uncontrolled

Hypertension, persistently high arterial blood pressure, affects an estimated 1.28 billion adults aged 30 to 79 worldwide, and is a leading cause of stroke, myocardial infarction, chronic kidney disease, and heart failure [1]. In the United States it is present in 47.7 % of adults, yet only 20.7 % of adults with hypertension have it controlled to below 130/80 mmHg, and 40.8 % are not even aware they have it [2]. Globally the picture is similar: about 46 % of adults with hypertension do not know they have the condition, and roughly one in five has it under control [1].

Those awareness figures are the reason detection, rather than treatment, is the binding constraint. Hypertension is asymptomatic until it causes damage, effective medications are inexpensive and widely available, and the risk of stroke and cardiac events rises continuously with pressure across the whole range above normal [3]. The people who go undiagnosed are not failing to respond to treatment; they are never measured often enough for anyone to notice.

The gap: every current measurement modality depends on the patient acting

The reference standards for blood pressure are accurate when applied correctly, but each one requires a person to decide to measure, and in practice that decision is made far less often than any guideline assumes. In a remote hypertension-management program that supplied free devices, education, and personalized support, a third of enrolled patients recorded no blood-pressure readings at all in a typical week [6]. Section 2.1 sets out each modality and the size of its gap in detail.

The Casana Smart Seat

The **Casana Smart Seat** is a sensor-embedded toilet seat that passively acquires cardiovascular measurements during ordinary use, requiring no wearable hardware, no user-initiated measurement, and no change in routine behavior [11]. A person sits down as they normally would; the device records for as long as they are seated and then uploads the recording, encrypted, to a cloud platform for signal processing and estimation, where longitudinal trends can be reviewed by the user and, where applicable, a care team.

Sensors integrated into the seat surface and base capture four things:

- a single-lead **electrocardiogram (ECG)**, the heart's electrical activity, sensed through skin contact with the seat;
- a **ballistocardiogram (BCG)**, the mechanical recoil of the body with each ejection of blood from the heart, sensed through the seat's force path;
- **photoplethysmography (PPG)** at the posterior thigh, an optical measurement of the pulse of blood volume arriving in the tissue;
- **body weight and its distribution**, from four load cells, one at each corner of the seat.

From these signals the platform derives heart rate, heart-rate variability, peripheral oxygen saturation, and blood pressure. Because measurement is attached to an already-habitual daily behavior, the recording cadence does not depend on motivation. In a randomized crossover trial in 49 older adults with cardiovascular disease, daily adherence to the seat was significantly higher than to a conventional cuff-and-scale telemonitoring kit over 12 weeks [10].

What this paper contributes

The contribution here is not that a classifier can separate elevated from normal blood pressure. Single-measurement classifiers already exist, and on their own they inherit the problem described above: they still require someone to take the measurement. The contribution is the combination of two properties: a measurement that repeats without anyone deciding to take it, and a decision rule deliberately operated so that it raises few false alarms.

Both halves are necessary and neither is sufficient. Passive measurement without a trustworthy decision rule produces a stream of alerts a user learns to ignore, which is the failure mode that ends engagement with a monitoring product. A precise decision rule applied to a measurement nobody takes has nothing to score.

Because adherence, rather than per-measurement accuracy, is the binding constraint on hypertension detection, a modality that removes the decision to measure is worth evaluating specifically on whether the alerts it raises can be acted upon. That is the question this paper asks, and it is why specificity and positive predictive value, rather than sensitivity, are the metrics the evaluation is built around.

This paper describes and evaluates the **mixture-of-experts** blood-pressure classifier deployed in the Smart Seat's blood-pressure pipeline. The classifier comprises multiple models, each trained independently against a different blood-pressure decision boundary. At inference, a fixed gating rule combines the expert outputs to produce a single elevated or normal classification for each recording. Recording-level classifications are then aggregated by voting to produce the final classification surfaced by the product.

The paper covers:

- what the classifier does and how it is built;
- how it performs on a clinical cohort with auscultatory reference readings taken on the seat, and on an in-home cohort of unsupervised daily use;
- how those two settings differ, what that difference reveals about moving a physiologic algorithm out of controlled conditions, and which part of the combined rule accounts for it;
- the operating characteristics a product decision should be based on, including where the classifier fails and what its output cannot be used to rule out.

2. Background and related work

2.1 Current measurement methods for blood pressure

Blood pressure is the force per unit area that circulating blood exerts on the walls of the arteries. It rises and falls with every heartbeat, peaking as the heart contracts (the **systolic** pressure) and reaching its lowest point as the heart refills (the **diastolic** pressure). A reading of "120/80 mmHg" is the systolic value over the diastolic value, in millimeters of mercury. The 2017 ACC/AHA guideline sets the categories [3]:

Table 1. Blood-pressure categories under the 2017 ACC/AHA guideline.

Category	Systolic (mmHg)		Diastolic (mmHg)
Normal	below 120	and	below 80
Elevated	120 to 129	and	below 80
Stage 1 hypertension	130 to 139	or	80 to 89
Stage 2 hypertension	140 or above	or	90 or above
Hypertensive crisis	180 or above	and/or	120 or above

The 130/80 mmHg line, where hypertension begins under that guideline, is the boundary this classifier targets. Four modalities are used to find it, one in the clinic, two at home, and one ambulatory.

Office cuff (clinic). An automated oscillometric cuff applied by a clinician is the point-in-time reference standard. The American Heart Association's measurement statement enumerates what a valid reading requires: at least five minutes of quiet seated rest, feet flat on the floor, back supported, arm supported at heart level, an appropriately sized cuff, no caffeine or exercise or smoking in the preceding 30 minutes, no talking, and an average of at least two readings [4]. Deviations are common and consequential: a systematic review of 328 studies found that individual sources of error shifted readings by between 23.6 mmHg below and 33 mmHg above true resting systolic pressure, with a cuff over clothing, an unsupported arm, a cuff that is too small, and talking during the measurement among the largest contributors [5]. Office measurement also misclassifies in both directions, because a single setting is not the population of a person's readings: across national and international registries, up to one in three patients in a given scenario has either white-coat hypertension (elevated in clinic, normal outside it) or masked hypertension (normal in clinic, elevated outside it) [8]. And adults are typically seen once or twice a year, which is not a sampling rate that can track a slowly-varying condition.

Home cuff (home). Home monitoring is the guideline-recommended supplement to office measurement [3], and it is where the adherence gap is best quantified. In 3,390 patients enrolled in a remote hypertension-management program that provided free devices, education, and personalized support, and asked for four readings per day on at least six days a week, 32.7 % recorded no measurements in a typical week and only 34.8 % recorded the full recommended 24 to 28 [6]. That is the ceiling case: a supported, enrolled, motivated population. Home cuffs also inherit every technique-related error source from the office setting [5], without a clinician present to prevent them.

24-hour ambulatory monitoring (home and ambulatory). An ambulatory monitor inflates automatically every 15 to 30 minutes across a full day and night, and its 24-hour average is the best-validated single predictor of cardiovascular events. Its limitation is tolerability rather than accuracy. In a repeated-measures tolerability study, sleep disruption was the dominant complaint: about 67 % of participants reported the monitor woke them after they had fallen asleep and 8.6 % removed it during the night, with pain reported by 32 %, skin irritation by 37 %, and bruising by 7 %; participants who tolerated it poorly were more likely to remove it and to return too few measurements to interpret [7]. A device that wakes two thirds of its wearers is not one people repeat often, which makes ambulatory monitoring a diagnostic episode rather than a way to follow someone over time.

Consumer wearables (home and ambulatory). Wrist- and finger-worn devices that estimate blood pressure from optical or timing-based features are attractive because they demand nothing of the user beyond wearing them. No cuffless device has yet been shown to meet the accepted validation standard for automated blood-pressure devices on an unselected population, and the 2021 European Society of Hypertension guidance does not recommend them for clinical use [9]. Wear time also decays: in a cohort of older adults asked to wear a device for 14 consecutive days, 32 % (7 of 22) wore it on fewer than 4 of those days, even though those who did wear it did so for most of the day [12].

What is structurally hard to fix. All four require the person to act: to attend an appointment, to inflate a cuff, to tolerate a device overnight, to keep something charged and worn. Accuracy improvements do not touch that, because the missing measurements are not inaccurate, they are absent. A modality attached to a behavior people already perform several times a day removes the decision entirely, and that is the structural argument for measuring from a seat.

2.2 Related algorithmic approaches

The approach evaluated here is a **mixture of experts**: rather than one model asked to separate elevated from normal everywhere along the pressure range, several models are trained separately, each against a different decision boundary, and their outputs are combined at inference time. The idea of decomposing a problem across specialized networks whose outputs are then combined by a gating function goes back to Jacobs et al. [13]; what differs here is that the gate is a fixed rule rather than a learned function, so its behavior is auditable and does not need its own training data.

Placing several binary classifiers at different points along an ordered scale, and reading their pattern of responses as a graded judgment, is the standard decomposition for ordinal problems [14]. Blood pressure is naturally ordinal, and two of the three models here can be read in that frame. The third cannot, because it sits below the boundary rather than above it, and its role in the combined rule is to withhold a positive call rather than to support one.

Each model is a neural sequence encoder applied to per-heartbeat windows cut from the recording, producing a probability for the whole recording by averaging over its heartbeats. Section 4 describes the pipeline and the combination rule.

3. Data

3.1 Data collection

None of the data analyzed here was found or repurposed. Every recording was produced by a collection study that Casana designed, submitted for institutional review, and ran, in two programs: a series of supervised clinical studies, conducted at Casana's own offices and at partner clinical sites, and an at-home deployment study. Each study protocol was reviewed and approved by an institutional review board (IRB) before any data was collected, and every participant gave informed consent before enrollment. Collection spans September 2021 to June 2026.

Both programs captured the same instrument stream. At each sit, the Smart Seat records a single-lead electrocardiogram, a ballistocardiogram, a photoplethysmogram at the posterior thigh, and the load carried at each of the four corners of the seat, for as long as the participant remains seated. What differs between the programs is the reference blood-pressure measurement paired to each recording, the schedule on which sits occurred, and the covariates recorded at enrollment.

In the clinical studies, participants attended a supervised study visit and sat on a Smart Seat while a trained clinician took an auscultatory blood-pressure reading during the recording itself; age, sex, height and weight were recorded at enrollment. Most sits were taken at seated rest, and two of the studies deliberately captured sits under other conditions, including induced low and high activity, so that the corpus would cover the physiological range the device meets in use rather than resting values alone.

In the at-home deployment study, participants installed a Smart Seat in their own bathroom and used it as they normally would, with no measurement protocol attached to sitting; separately they were supplied a validated oscillometric arm cuff and asked to self-measure as often as they could manage, so that a timestamped cuff reading could be paired to the nearest sit.

3.2 Training and development data

The three models were trained on Casana's internal supervised cardiovascular corpus, which combines recordings from clinical studies (paired with auscultatory and oscillometric reference readings) with recordings from at-home deployments (paired with home-cuff readings). Each model was trained separately, on the subset of the corpus that its own decision boundary can label, with recordings whose reference pressure falls too close to that model's boundary excluded from its training data.

Training splits were built by participant, not by recording, so no participant contributes recordings to more than one split. No participant in either evaluation cohort described below appears in any training split. The evaluation cohorts are therefore participant-disjoint from training, though the clinical evaluation cohort is drawn from the same family of clinical studies that contributed to the training corpus, which means it is a held-out sample rather than an independent replication.

3.3 Evaluation data

Two cohorts were evaluated. They were chosen because they differ on the three axes that matter most for a passive at-home device: where the measurement happens, what the reference standard is, and whether the same person is seen more than once.

The in-home cohort is unsupervised daily use. Participants were given a Smart Seat to install in their own homes and use as they normally would, with no measurement protocol attached to sitting. Separately, each participant was given a validated oscillometric arm cuff and asked to self-measure their blood pressure as often as they could manage. Each cuff reading, timestamped, was paired to the nearest seat recording within a short window, and that reading is the reference standard for the cohort. Fifteen participant identifiers were allocated to the study; fourteen contributed recordings to this analysis and one contributed none. The fourteen produced 433 recordings over 45 calendar days, from 11 May to 24 June 2026.

Because these participants agreed to a high-effort self-measurement protocol, their cuff cadence should be read as a best case for home monitoring rather than a typical one. Their seat cadence, by contrast, required nothing of them:

- a median of 20.5 active days per participant (interquartile range 13 to 31.5, range 3 to 38);
- a median of 23.5 recordings per participant (interquartile range 13 to 42.5, range 3 to 77);
- a median of 6 active weeks per participant;
- within an active week, a median of 5 recordings (interquartile range 3 to 7, mean 5.6, maximum 16), across 77 participant-weeks;
- within an active day, a median of 1 recording (mean 1.5).

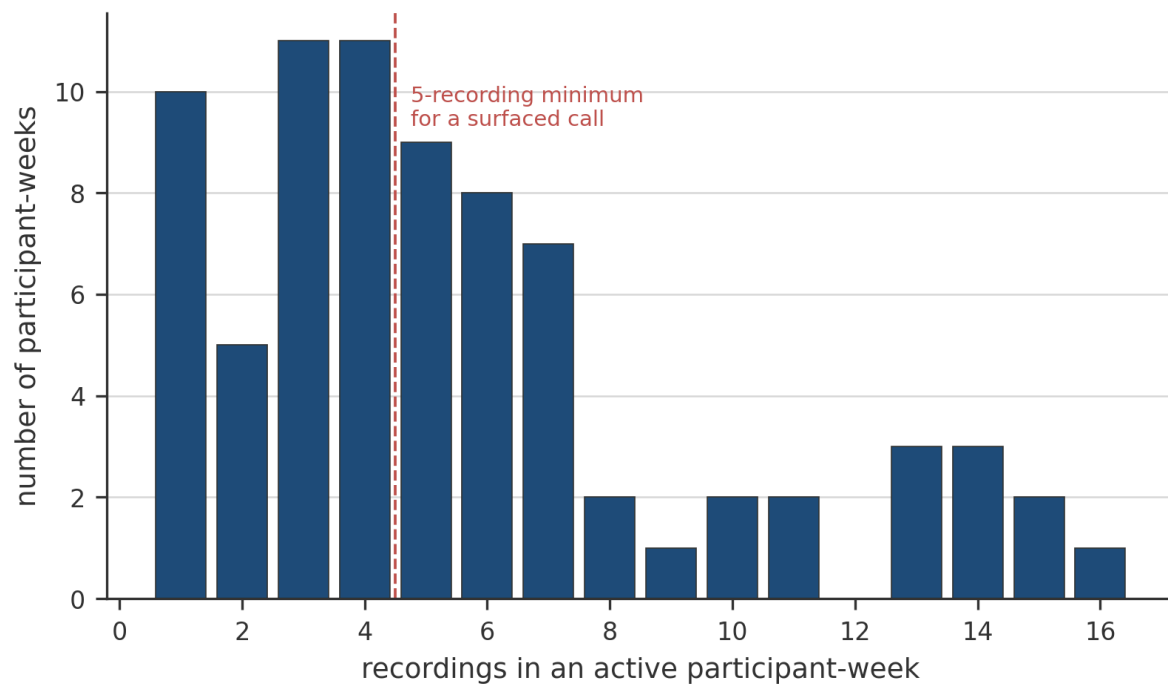


Figure 1. How often in-home participants sat on the Smart Seat, counted as recordings per active participant-week (77 participant-weeks). Median 5 recordings per week (interquartile range 3 to 7, maximum 16). The dashed line marks the minimum recording count below which no call is surfaced for the week.

The clinical cohort (internally, Holdout) is supervised measurement in a clinical setting. A trained clinician took an **auscultatory** blood-pressure reading, the classical cuff-and-stethoscope method in which the clinician listens for the Korotkoff sounds that appear and disappear as the cuff deflates, while the participant was seated on the Smart Seat. That reading, taken during the seat recording rather than before or after it, is the reference standard for this cohort. Eighty-two participants contributed 668 recordings, drawn from nine clinical studies conducted between September 2021 and October 2025. Participants contributed a median of 4 recordings each (range 1 to 36), so this cohort supports repeat measurement within a visit but not across days.

Table 2. The two evaluation cohorts.

	In-home cohort	Clinical cohort
Setting	Participants' own homes, unsupervised	Clinical study visit, supervised
Reference standard	Self-measured oscillometric cuff, paired to the nearest recording	Auscultation by a trained clinician, taken during the recording
Participants contributing recordings	14	82
Recordings	433	668
Recordings per participant, median (range)	23.5 (3 to 77)	4 (1 to 36)
Period covered	45 days, May to June 2026	September 2021 to October 2025
Repeat visits	Yes, a median of 6 weeks per participant	No, one visit per participant

Table 3. Baseline characteristics of the clinical cohort (n = 82 participants)

Characteristic	Distribution
Age (years)	mean 54.2 (SD 15.1), median 55, range 23 to 82
Sex	58 female, 24 male
Height (cm)	mean 167.8 (SD 10.5), range 148 to 193
Weight (kg)	mean 79.9 (SD 19.9), range 48 to 162
Body mass index (kg/m ²)	mean 28.2 (SD 5.8), range 19.4 to 55.1
Missing values	none, on any characteristic above

The two cohorts also differ in the blood pressures they contain, which matters for reading the results.

Table 4. Reference blood pressure in each cohort, across all analyzed recordings.

	In-home cohort (n = 433)	Clinical cohort (n = 668)
Systolic (mmHg), mean (SD)	128.3 (17.4)	127.4 (25.1)
Systolic, median (IQR)	129 (121 to 138)	122 (106 to 147)
Systolic, range	75 to 178	82 to 204
Diastolic (mmHg), mean (SD)	81.0 (13.4)	77.8 (13.5)
Diastolic, median (IQR)	81 (74 to 86)	73 (68 to 90)
Diastolic, range	47 to 131	50 to 120
Truly elevated / truly normal	315 / 118	344 / 324
Within 5 mmHg of the boundary	159 (36.7 %)	13 (1.9 %)

The means are nearly identical and the distributions are not. The in-home cohort is packed against the classification boundary, with more than a third of its recordings within 5 mmHg of it and 73 % of them counted as elevated, while the clinical cohort spreads across a much wider range and splits almost evenly.

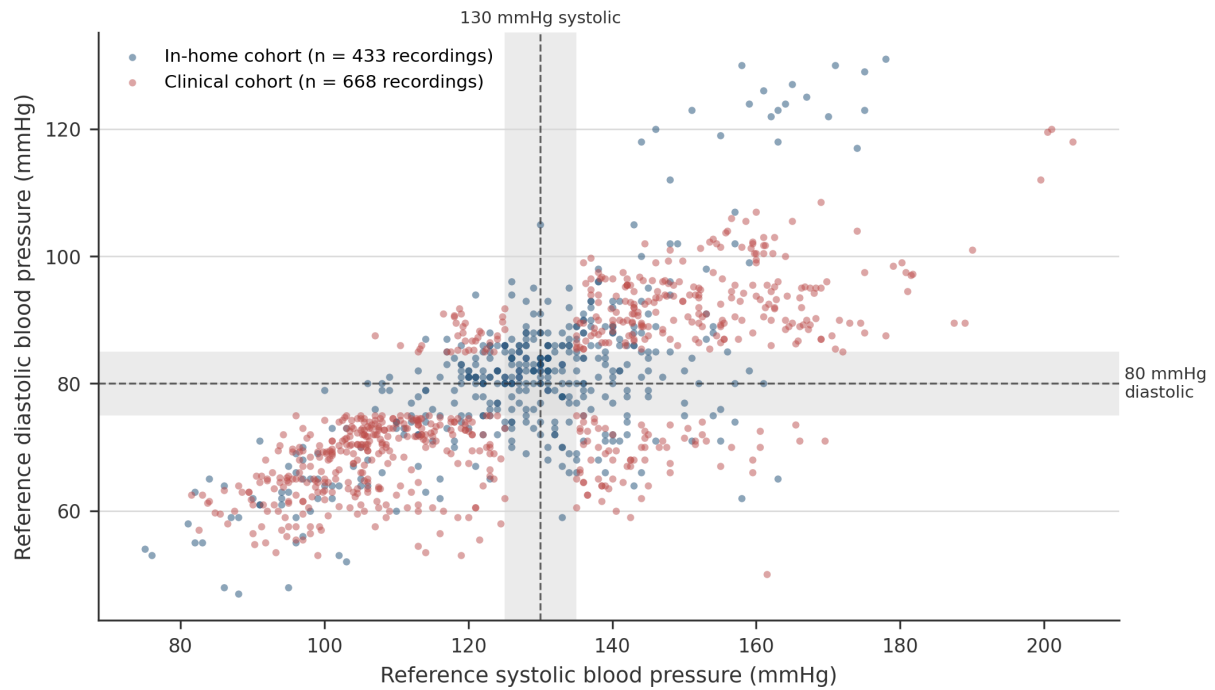


Figure 2. Reference blood pressure for every analysed recording. Dashed lines mark the 130 over 80 mmHg boundary that separates elevated from normal; the shaded strips are the 5 mmHg ambiguous band excluded under the second, looser truth definition. The in-home cohort clusters tightly on the boundary (159 of 433 recordings fall inside the band), whereas the clinical cohort spreads across a much wider range (13 of 668 inside the band).

3.4 Data validation

All 1,101 recordings were processed successfully. There were no recordings rejected for insufficient signal quality, so the denominators reported in Section 5 are the full cohorts rather than a surviving subset.

4. Methods

A reader needs to be able to picture three things before the results mean anything: the quantity being measured, the standard it is compared against, and the algorithm itself. Blood pressure and its categories are set out in Section 2.1; the reference standards are described per cohort in Section 3.3 and their limitations in Section 4.4; the signals the algorithm reads are shown in Figure 3 and the pipeline in Figure 4.

4.1 Hardware and preprocessing

This classifier uses all four sensing modalities in the Smart Seat.

Sensor	What it measures
Single-lead ECG	The heart's electrical activity, through skin contact with the seat
Ballistocardiogram	The body's mechanical recoil as the heart ejects blood, through the seat's force path
Photoplethysmography, posterior thigh	The optical pulse of blood volume arriving in the tissue
Load cells, one per corner	The weight carried at each corner of the seat

Figure 3 shows six seconds of the three cardiovascular channels from a single recording. The same heartbeat is visible in all three: as a sharp electrical spike in the ECG, as a mechanical disturbance in the ballistocardiogram a fraction of a second later, and as a slower rise in the photoplethysmogram as blood reaches the thigh. The timing relationships between those three events, and the shape of the pulse in each, are what carry information about pressure.

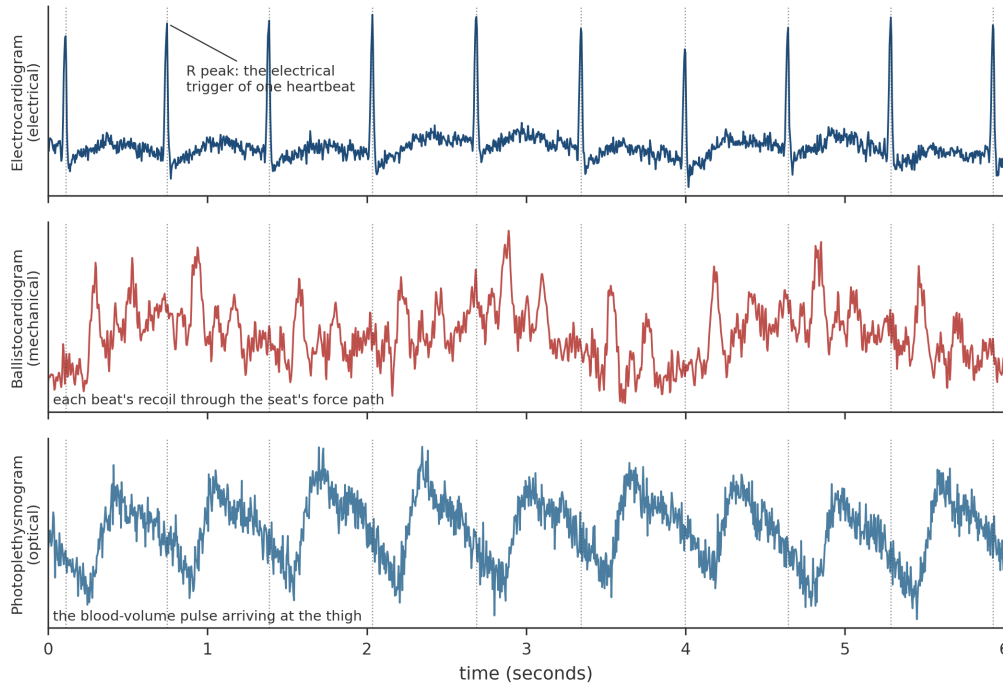


Figure 3. 6 seconds of the three cardiovascular channels the Smart Seat records simultaneously during one ordinary sit, from a single recording in the clinical cohort. Dotted lines mark successive heartbeats, located from the electrocardiogram. The same beat appears in all three channels: as an electrical trigger, as the body's mechanical recoil, and as the blood-volume pulse reaching the thigh. Amplitudes are normalized per channel, so the panels show timing and shape rather than absolute magnitude.

Each recording is split into fixed windows centred on the detected heartbeats; each window is checked against per-channel signal-quality criteria and discarded if it fails; and the surviving windows are packaged as the model input. A recording that yields too few usable windows produces no call.

4.2 Algorithm

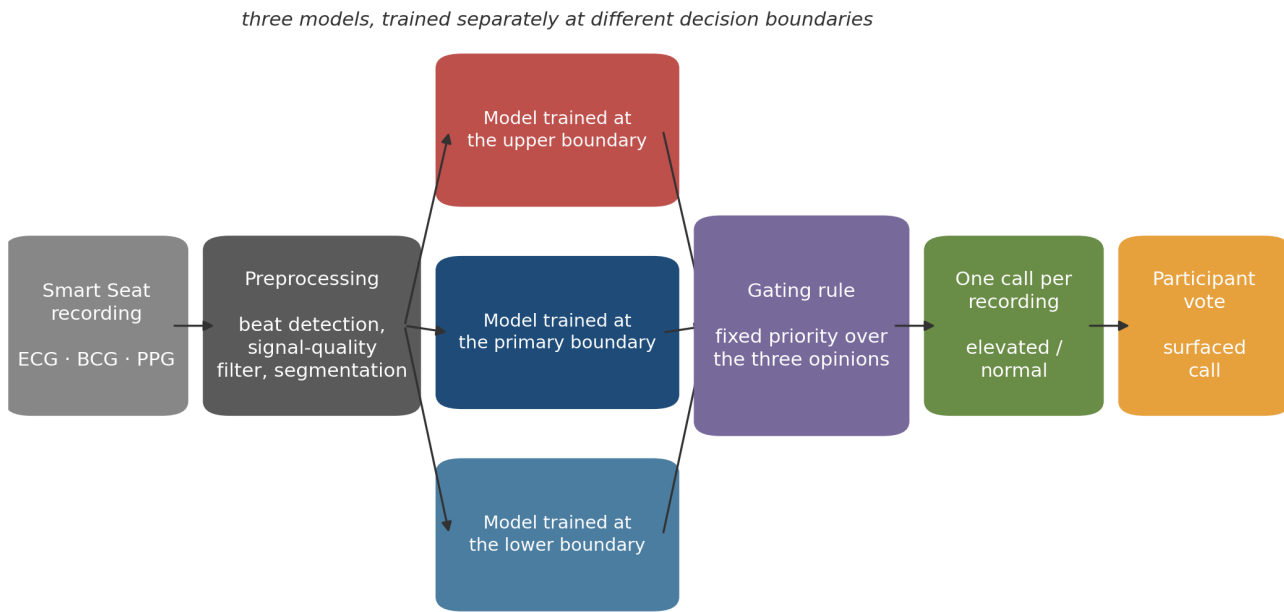


Figure 4. How a recording becomes a surfaced call. Shared preprocessing turns one sit into a set of per-heartbeat segments; three models, each trained separately against a different blood-pressure decision boundary, score the same segments; a fixed gating rule resolves their three opinions into one elevated or normal call for the recording; and a vote across the participant's recordings produces the call the product surfaces.

Three models score the same preprocessed windows. Each was trained separately, against a different blood-pressure decision boundary: one at the clinical boundary the product reports against, one placed higher in the range, and one placed below the normal range entirely. Each produces a single probability for the recording by averaging its per-window outputs.

A fixed gating rule then resolves those three probabilities into one call. The rule is a priority order rather than a weighted average, and it reads as follows: if the model trained at the upper boundary is confident, the recording is called elevated and no other opinion is consulted; otherwise, if the model trained at the lower boundary is confident, the recording is called normal; otherwise, the model trained at the primary boundary decides, calling the recording elevated if it is confident and normal if it is not.

The two outer models are each trained on a problem that is easier to separate than the primary one, because their boundaries sit well away from the bulk of the population, so each can be trusted at a lower level of confidence than the primary model requires. The primary model is held to a deliberately demanding level of confidence, so that it contributes a positive call only when it is sure and otherwise leaves the recording at the default.

The rule is deliberately asymmetric. The outer models can each override the primary one, but in opposite directions: the upper model can only add positive calls and the lower model can only remove them. The default, when no model is confident, is to call the recording normal.

Each recording's call then feeds an aggregation that produces the call the product surfaces. Recordings are grouped into the unit the product reports on, the group is skipped entirely unless it contains a minimum number of recordings, and the group is called elevated when at least a set fraction of its recordings were called elevated.

4.3 Configuration

Table 5. Configuration governing how the reported results should be interpreted.

Parameter	Value	Rationale
Classification boundary	130/80 mmHg	The point at which hypertension begins under the 2017 ACC/AHA guideline [3]; the number reported is not interpretable without it
Ambiguous band, second evaluation	5 mmHg either side	Absorbs part of the reference standard's own measurement error, so that a near-boundary miss is not scored the same as a clear one
Aggregation unit	Participant-week for the in-home cohort, participant for the clinical cohort	Matches how the product reports, and reflects that clinical participants attend one visit
Minimum recordings per unit	5	Below this, no call is surfaced at all
Vote fraction to call a unit elevated	30 % of its recordings	Sets how much within-unit agreement a surfaced positive call requires

4.4 Analysis methods

The reference standards, and their limits. The clinical cohort's reference is auscultation performed by a trained clinician during the Smart Seat recording. This is the classical method underlying the guideline thresholds themselves, and it is the stronger of the two references here because it is simultaneous with the recording and performed by a trained observer. It is not error-free: observer terminal-digit preference, cuff sizing, and arm position all shift auscultatory readings, and the same systematic review cited in Section 2.1 quantifies those effects [5]. The in-home cohort's reference is an oscillometric reading the participant took themselves, paired to the nearest Smart Seat recording rather than taken during it. It therefore carries both the technique errors of unsupervised home measurement and a timing gap, and blood pressure genuinely varies within a day. The in-home reference is the weaker standard, and comparisons should be made within a cohort rather than across the two.

Truth definitions. The primary definition is strict: a recording is *elevated* if its reference systolic pressure is at or above 130 mmHg or its reference diastolic pressure is at or above 80 mmHg, and *normal* otherwise. Every analyzed recording gets a label under this definition.

A second, looser definition repeats every analysis with near-boundary cases removed. A recording is *elevated* if reference systolic is at or above 135 mmHg or reference diastolic is at or above 85 mmHg; *normal* if reference systolic is below 125 mmHg and reference diastolic is below 75 mmHg; and *ambiguous*, excluded from scoring, in between. The purpose is to ask what performance looks like when the reference standard's own error cannot be the explanation for a disagreement: a participant whose true reading is 131/81 mmHg could be scored either way by a cuff with a few mmHg of error, whereas one at 168/102 mmHg could not. Reporting both is more honest than picking one, because the strict definition penalizes near-boundary noise and the loose definition quietly discards more than a third of the in-home cohort.

Aggregation. A *participant-week* is one participant's recordings within a single calendar week, so a participant contributing over six weeks yields up to six participant-weeks and the 40 scored in-home units come from fourteen people rather than forty. For the aggregated analyses, recordings are grouped by participant-week (in-home) or by participant (clinical), groups with fewer than 5 recordings are dropped, the group's predicted label is elevated when at least 30 % of its recordings were called elevated, and the group's true label is the truth definition applied to the mean of its reference systolic and diastolic readings. The vote fraction is deliberately below half: a participant whose pressure reads elevated in a substantial minority of a week's sits is worth flagging, so the rule is built to trigger on that minority rather than to require consensus.

Metrics. Writing TP, FP, TN and FN for the counts of true positives, false positives, true negatives and false negatives, where positive means elevated:

- **Sensitivity** (also called recall) is $TP / (TP + FN)$: of the units whose reference blood pressure was elevated, the fraction called elevated. This is the population-detection property: how many of the people who do have elevated blood pressure would be flagged.
- **Specificity** is $TN / (TN + FP)$: of the units whose reference was normal, the fraction called normal. This is the user-trust property: how rarely someone with normal blood pressure is alarmed.
- **Positive predictive value** (PPV, also called precision) is $TP / (TP + FP)$: of the units called elevated, the fraction that really were. Unlike the first two, it depends on how common elevated blood pressure is in the cohort, so it cannot be compared across cohorts with different prevalence.
- **Negative predictive value** (NPV) is $TN / (TN + FN)$: of the units called normal, the fraction that really were. This is what a quiet result is worth.

4.5 Evaluation procedure

Three analyses were run, in this order. Section 5 reports them one for one, in the same order.

1. **Recording-level agreement under the strict truth definition**, per cohort. Every analyzed recording is scored against its own paired reference reading. This is the finest granularity available and the one that isolates the classifier from the aggregation.
2. **Agreement at the aggregation level the product reports on**, under the strict truth definition: participant-weeks for the in-home cohort, participants for the clinical cohort, each requiring at least 5 recordings and a 30 % vote.
3. **Both of the above repeated under the loose truth definition**, with recordings and units within 5 mmHg of the boundary excluded, to separate near-boundary disagreement from real disagreement.

5. Results

5.1 Per recording, the classifier is specific but insensitive, and much weaker at home than in clinic

Table 6. Recording-level performance under the strict 130/80 mmHg truth definition.

	In-home cohort	Clinical cohort
Recordings scored	433	668
Truly elevated / truly normal	315 / 118	344 / 324
Sensitivity	0.257	0.631
Specificity	0.847	0.910
Positive predictive value	0.818	0.882
Negative predictive value	0.299	0.699
True positives / false negatives	81 / 234	217 / 127
True negatives / false positives	100 / 18	295 / 29

In the clinical cohort the classifier identified 217 of 344 truly elevated recordings and correctly passed 295 of 324 truly normal ones. In the in-home cohort it identified 81 of 315 truly elevated recordings, missing 234, while correctly passing 100 of 118 truly normal ones. Sensitivity is 0.374 lower in the home setting than in the clinic; specificity is 0.063 lower.

Negative predictive value differs even more sharply than sensitivity: 0.699 in the clinical cohort against 0.299 in the in-home cohort. Of the 334 in-home recordings the classifier called normal, 234 were not.

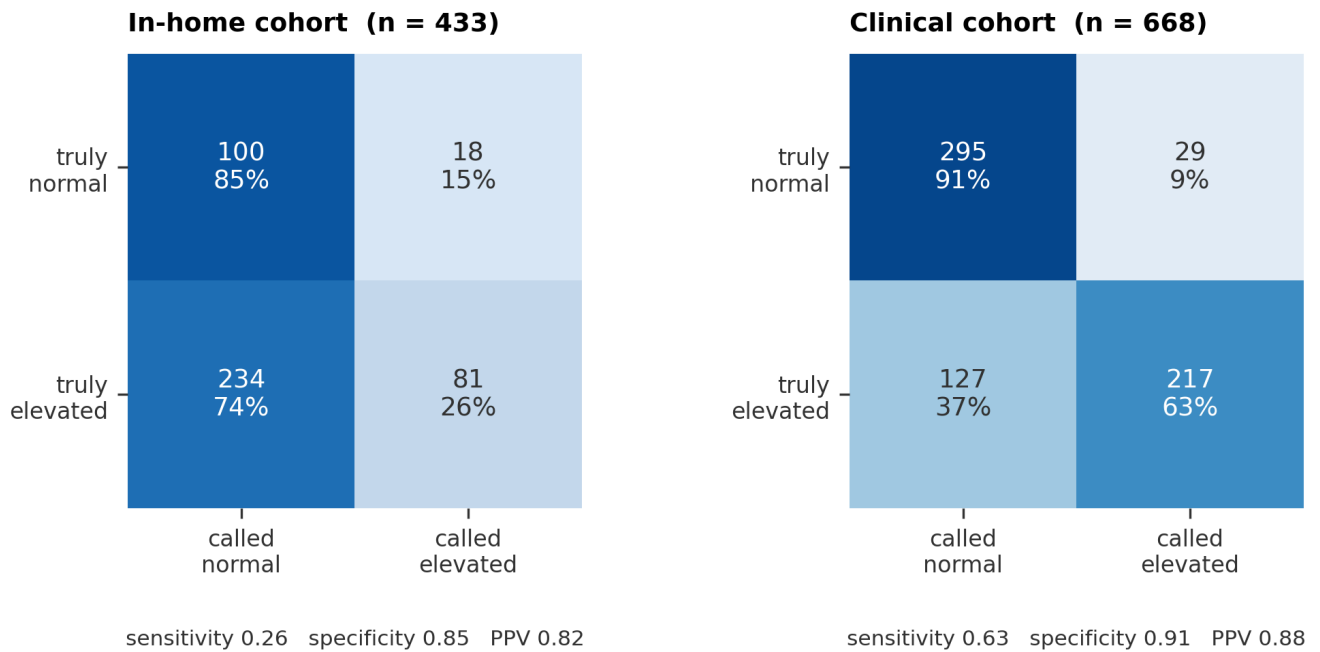


Figure 5. Recording-level agreement with the reference standard under the strict 130 over 80 mmHg truth definition. Each cell gives the recording count and, below it, the percentage of that row: the top row splits truly normal recordings, the bottom row truly elevated ones, so the bottom-right percentage is sensitivity and the top-left is specificity. The classifier misses most elevated recordings in the in-home cohort and about a third of them in the clinical cohort.

5.2 Aggregating recordings raises sensitivity in both cohorts and eliminates in-home false alarms

Table 7. Performance at the level the product surfaces a call, under the strict truth definition. Units require at least 5 recordings and a 30 % elevated vote.

	In-home cohort	Clinical cohort
Unit	Participant-week	Participant
Units formed	77	82
Units with enough recordings to score	40	32
Truly elevated / truly normal	31 / 9	16 / 16
Sensitivity	0.419	0.750
Specificity	1.000	0.938
Positive predictive value	1.000	0.923
Negative predictive value	0.333	0.789
True positives / false negatives	13 / 18	12 / 4
True negatives / false positives	9 / 0	15 / 1

Aggregation raised sensitivity in both cohorts without costing specificity in either. In the clinical cohort sensitivity rose from 0.631 per recording to 0.750 per participant while specificity rose from 0.910 to 0.938. In the in-home cohort sensitivity rose from 0.257 per recording to 0.419 per participant-week, and every false

positive disappeared: all 9 truly normal participant-weeks were called normal, and all 13 positive calls were correct.

The in-home specificity of 1.000 must be read together with the denominator that produced it. It rests on 9 truly normal participant-weeks, contributed by a subset of the 14 participants, out of 40 evaluable participant-weeks in total. At that size a single additional false positive would move specificity to 0.900 and positive predictive value to 0.929, so the interval around these two numbers is wide and the point estimate of 1.000 should not be read as a demonstration that the classifier never produces a false alarm at home. Section 7 sets out the consequences of the sample size for every in-home number in this paper.

The gain comes from the vote fraction rather than from averaging. Because a unit is called elevated when 30 % of its recordings were, a participant-week needs only a minority of elevated sits to be flagged, which recovers weeks where the classifier caught the elevation intermittently. It also means the aggregated numbers are not a smoothed version of the per-recording ones: they describe a different, deliberately more sensitive decision rule applied on top of them.

Requiring at least 5 recordings is a substantial exclusion. It removed 37 of 77 in-home participant-weeks (48.1 %) and 50 of 82 clinical participants (61.0 %), the latter because clinical participants contributed a median of 4 recordings from a single visit. The aggregated clinical result therefore describes 32 of the cohort's 82 participants.

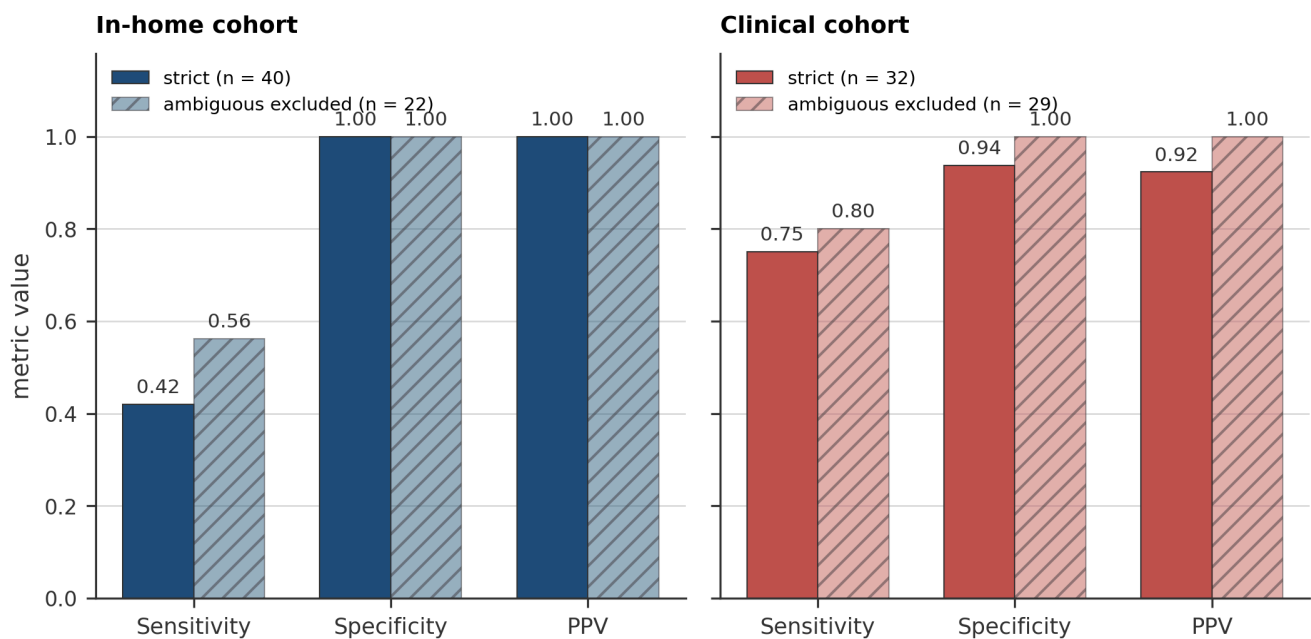


Figure 6. Performance at the level the product surfaces a call: participant-weeks for the in-home cohort and participants for the single-visit clinical cohort, each requiring at least 5 recordings and a 30% elevated vote. Solid bars use the strict 130 over 80 mmHg truth definition; hatched bars exclude units whose averaged reference falls within 5 mmHg of the boundary. Specificity is 1.00 in the in-home cohort (all 9 truly normal weeks called normal) but sensitivity is 0.42; the clinical cohort reaches sensitivity 0.75 at specificity 0.94.

5.3 Excluding near-boundary cases barely moves the per-recording result and lifts the aggregated one substantially

Table 8. Both levels repeated with cases within 5 mmHg of the 130/80 mmHg boundary excluded. Change is relative to the strict result for the same cohort and level.

	In-home cohort	Clinical cohort
Per recording		
Recordings scored (excluded)	274 (159)	655 (13)
Sensitivity	0.275 (+0.018)	0.631 (0.000)
Specificity	0.919 (+0.072)	0.920 (+0.010)
Positive predictive value	0.902 (+0.084)	0.897 (+0.015)
Aggregated		
Units scored (excluded)	22 (18)	29 (3)
Sensitivity	0.562 (+0.143)	0.800 (+0.050)
Specificity	1.000 (0.000)	1.000 (+0.062)
Positive predictive value	1.000 (0.000)	1.000 (+0.077)

The exclusion is highly asymmetric between cohorts: it removes 159 of 433 in-home recordings (36.7 %) and 13 of 668 clinical recordings (1.9 %).

The two levels respond to it very differently. Per recording, discarding more than a third of the in-home cohort moved sensitivity only from 0.257 to 0.275, and clinical recording-level sensitivity did not move at all (all 13 excluded clinical recordings were truly normal under the strict definition). Aggregated, the same exclusion moved in-home sensitivity from 0.419 to 0.562 and clinical sensitivity from 0.750 to 0.800, and lifted clinical specificity and positive predictive value to 1.000.

That contrast is worth stating plainly, because the aggregated loose numbers are the most favourable in this paper and the least well supported. The in-home aggregated loose result rests on 22 participant-weeks, of which 6 are truly normal, after excluding 18 of the 40 eligible weeks as ambiguous; the clinical one rests on 29 participants, of which 14 are truly normal. Under the loose definition the in-home cohort still retains 200 truly elevated recordings at the per-recording level, and the classifier missed 145 of them.

6. Discussion

Recap. We set out to determine whether a classifier reading the Smart Seat's existing sensors can identify elevated blood pressure specifically enough to be worth surfacing to a user, and whether clinical-setting performance carries over to unsupervised home use. Three models trained at different decision boundaries, combined by a fixed gating rule, reached a sensitivity of 0.75 at a specificity of 0.94 per participant in a clinical cohort with auscultatory reference, and a sensitivity of 0.42 at a specificity of 1.00 per participant-week in an in-

home cohort with self-measured cuff reference. Read against the two properties Section 1 set out, the results show that the classifier does produce a high enough specificity to surface to users. The classifier remains highly specific at home, but finds a smaller share of elevated cases there than under controlled conditions in-clinic.

The home-to-clinic gap is a finding, not only a caveat. The largest single effect in this evaluation is the 0.374 drop in per-recording sensitivity between the clinical and in-home cohorts, and it deserves treatment as a subject of study rather than as a deduction from the headline numbers. It is a direct measurement of what happens when the same trained models, at the same fixed operating point, are moved from supervised measurement to unsupervised use in people's own homes.

The mechanism is visible in the score distribution rather than in the calls. The primary-boundary model produced substantially more positive calls in the clinical cohort than in the in-home cohort. Rather than making confident predictions that turned out to be wrong, the model became systematically less confident: its scores compressed below the operating threshold, so recordings that would have been called elevated in clinic fell instead to the default negative call. A model that is confidently wrong and a model that is uncertain fail in different ways and call for different remedies, and this one was uncertain.

Four explanations are consistent with that pattern, and this study separates them only partially.

- **Input distribution shift.** Clinical recordings were collected under supervision using defined protocols, and 189 of the 668 were obtained under conditions other than seated rest, including 66 under deliberately induced low or high activity. In-home recordings were taken under naturally occurring conditions, spanning variation in posture, hydration, time of day, and recency of activity that no protocol constrained. The clinical corpus also spanned four hardware generations while the in-home deployment used a single one.
- **Real-world physiology and behavior.** A person sitting at home is not performing a measurement, and the physiological state the seat samples is correspondingly less standardized than a seated-rest protocol produces. Posture on the seat, and how soon the sit follows exertion, are unconstrained in a way they never are in clinic.
- **Reference measurement noise.** The in-home reference is self-measured and paired to the nearest recording rather than taken during it, so some share of the apparent misses are disagreements with a noisy label rather than with the participant's true pressure. Section 5.3 bounds this contribution: removing every near-boundary case, more than a third of the cohort, recovered only 0.018 of per-recording sensitivity, so reference error near the boundary is not the main mechanism.
- **Calibration rather than discrimination.** Because the failure takes the form of compressed confidence rather than confident error, recalibrating the operating thresholds, or the score distribution itself, for the home setting may recover sensitivity without retraining the models. That is a testable prediction this evaluation does not test, because the thresholds were fixed before either cohort was scored and deliberately not re-tuned per cohort.

The reading we take from this is that a physiologic algorithm validated under clinical supervision should not be assumed to transfer to unsupervised home use at the same operating point, and that the transfer is something to measure rather than infer. The encouraging part is that compressed confidence is the more tractable of the two failure modes: it points at calibration and at training data drawn from the deployment setting, rather than at a model that has learned the wrong thing.

How much of the gap is boundary crowding. The obvious hypothesis is that the in-home cohort is simply crowded against the boundary: 36.7 % of its recordings sit within 5 mmHg of it, against 1.9 % in the clinical cohort, so more of its recordings are genuinely hard to call. Section 5.3 shows this accounts for part of the gap

and not all of it. At the recording level, removing every near-boundary case, more than a third of the cohort, moved in-home sensitivity only from 0.257 to 0.275, leaving 145 missed elevated recordings that are not marginal. At the aggregated level the same exclusion did more work, lifting in-home sensitivity from 0.419 to 0.562, because averaging a week's readings pulls borderline weeks across the line in a way it cannot do for a single recording. Boundary crowding therefore explains a meaningful share of the aggregated gap and very little of the per-recording one.

Comparison with the alternatives. The competing modalities in Section 2.1 fail on adherence, and the seat's cadence is where it wins. In-home participants produced a median of 5 recordings per active week (interquartile range 3 to 7) with no measurement effort at all, in a cohort where a median participant was active on 20.5 days. Set that against the best-supported home-cuff cohort available: of 3,390 patients given free devices, education and personal support, 32.7 % recorded nothing in a typical week [6]. The randomized crossover comparison of the seat against a conventional telemonitoring kit found significantly higher daily adherence for the seat [10]. Ambulatory monitoring, the most informative modality per study, wakes about two thirds of its wearers overnight [7], which is why it is used once rather than continuously.

But adherence only converts into detection if the classifier detects. A sensitivity of 0.42 per participant-week means the classifier identified 42 % of elevated participant-weeks. For a person who is not measuring at home, those weeks would carry no blood-pressure information at all, and a flag that prompts a confirmatory cuff reading is the difference between noticing and not. That argument holds at population scale rather than in this cohort, whose participants were also using a supplied cuff. The honest reading is that the Smart Seat solves the harder of the two problems, getting the measurement at all, and that the classifier is currently the binding constraint in the home setting rather than the sensing.

The clinical proposition is asymmetric, and the asymmetry is the point. A positive call and a negative call from this classifier carry very different weight, and saying which is which is more useful than any single summary number. A positive call is actionable: in the in-home cohort every positive call was correct (13 of 13 participant-weeks) and in the clinical cohort 12 of 13 were, so a user told that their week looks elevated has been told something worth following up, and the cost of following up is one cuff reading. A negative call is not reassuring: in-home negative predictive value is 0.333, so roughly two of every three quiet participant-weeks belonged to someone whose reference blood pressure was in fact elevated.

That asymmetry places the classifier precisely. It is not a replacement for cuff-based blood-pressure measurement, and it is not a comprehensive hypertension detector. It is a passive trigger for confirmatory measurement, and judged as one the operating characteristics are the right ones, because a trigger is judged on whether acting on it is usually warranted rather than on whether it fires every time it could. Judged as a measurement, or as a screening test meant to rule out disease, the same numbers would be disqualifying. Two consequences follow for how a result may be presented: a quiet result must never be surfaced as evidence that blood pressure was normal, and any user-facing wording implying that no flag means no problem would be actively misleading in this cohort.

The 13 of 13 figure carries the same caveat attached to it in Section 5.2. It rests on a small number of participant-weeks from 14 people, and it establishes that false alarms were rare in this sample, not that they are rare in general.

The longitudinal question is where the scientific upside lies. Every result in this paper is cross-sectional: a call compared against a reference reading taken at approximately the same time. That design cannot establish that the classifier tracks change within a person, and Section 7 states as much. It is also the narrowest use of what a passive modality produces. The property that distinguishes measurement from a seat is not its accuracy

at any single sit, it is that sits recur without being scheduled, in this cohort a median of 5 recordings in an active week from participants who did nothing to produce them. A modality with that cadence can ask a question episodic measurement cannot: whether a change in an individual's cardiovascular state becomes visible in the aggregate of many imprecise passive measurements before it becomes visible in the occasional precise one.

That is a different claim from the one evaluated here, and it carries the larger clinical consequence. A within-person trend does not need absolute accuracy to be useful, provided it is stable enough that its own shifts are informative, and the per-recording noise that limits a single point estimate averages down across a week of sits in a way it cannot within one reading. Testing it needs a prospective home study designed around the question: enough participants to characterize demographics, enough months per participant for trends to exist, and a reference schedule dense enough to establish when a real change occurred. That study is the natural successor to this one.

Future direction.

- **Enroll a larger and demographically characterized in-home cohort.** Fourteen contributing participants cannot bound generalization across age, sex, or body composition, and the aggregated in-home result rests on 40 participant-weeks drawn from those same fourteen people.
- **Test whether per-setting recalibration closes the home-to-clinic gap.** The failure mode identified above is compressed confidence rather than confident error, which predicts that thresholds or score calibration fitted to the home setting could recover sensitivity without retraining. The prediction is inexpensive to test and was deliberately not tested here, because the thresholds were fixed before scoring.
- **Follow participants long enough to test trends rather than points,** along the lines set out above, which is the study the seat's cadence uniquely enables.

7. Limitations

- **The in-home aggregated result is 40 participant-weeks from 14 people, not 40 independent observations.** Weeks from the same participant share that participant's physiology, seat, and measurement habits, so the effective sample size is far closer to 14 than to 40, and the confidence intervals implied by the headline numbers are much wider than the denominators suggest. The perfect in-home specificity in particular rests on just 9 truly normal participant-weeks contributed by a subset of 14 people; a single additional false positive would move positive predictive value from 1.00 to 0.93. The clinical aggregated result is similarly thin, resting on 32 participants of whom 16 were truly elevated.
- **The two reference standards are not interchangeable.** The clinical reference is clinician auscultation taken during the recording. The in-home reference is a self-measured oscillometric reading paired to the nearest recording in time, carrying both unsupervised technique error and a timing gap during which blood pressure genuinely moves. Part of the in-home performance gap is attributable to reference noise, and this study cannot say how much.
- **The datasets used cannot establish that the classifier tracks change within a person over time.** Every analysis compares a call against a contemporaneous reference. Cross-sectional agreement is not longitudinal tracking, and the longitudinal claim is the one a trend-monitoring product ultimately rests on.

8. Conclusion

The result that matters is not either performance number on its own, it is what the two of them are together. A measurement that recurs without the user deciding to take it, combined with a decision rule operated so that it rarely raises a false alarm, produces something neither property delivers alone: an alert a person has reason to act on, arriving in weeks when they would otherwise have had no blood-pressure information at all.

For a person using a Smart Seat at home, that means a flag is worth taking seriously. Every in-home flag in this evaluation was correct, and a flag was raised in about four of every ten weeks in which their blood pressure was genuinely elevated. In the clinical cohort the classifier identified approximately three quarters of visits with genuinely elevated pressure, and 12 of 13 positive classifications were correct. It also means the reverse inference is unavailable: a week with no flag was more likely than not an elevated week anyway, so silence carries no reassurance.

That asymmetry, rather than any single metric, is what the evidence supports. The classifier is best understood as a passive trigger for confirmatory measurement with a blood-pressure cuff, where a high positive predictive value is the property that matters and the cost of an unnecessary confirmation is one cuff reading. It is not a blood-pressure measurement and not a substitute for one, and the absence of a flag must not be surfaced as evidence that blood pressure was normal. The in-home figures rest on 40 participant-weeks from 14 participants, so they establish the shape of this result rather than its precise value. The study that would extend it is a larger prospective home deployment asking whether frequent passive measurement reveals change within a person earlier than episodic measurement does.

Declaration of AI assistance

This paper was prepared with substantial assistance from an AI coding agent (Claude, Anthropic), directed and reviewed by the authors. What was and was not automated is set out here so a reader can weigh it.

What the AI did. It wrote the analysis code that produced the results reported here, generated the figures, retrieved and checked the literature cited, and drafted the prose of every section.

What it did not do. It did not generate any result. Every number in this paper comes from executing the evaluated classifier and the packaged analysis code over the recorded data, and the classifier under evaluation is production software that predates this paper and was not written for it. No statistic, and no citation, was produced from a language model's recall.

Safeguards applied. The gating rule described in Section 4.2 was re-derived from the three per-recording model probabilities and reconciled against the classification the deployed software recorded, on all 1,101 recordings, with no disagreements. Every quantitative claim in the prose was checked against the computed statistics so that no number could drift from its source during editing. Every reference was retrieved from the publisher or index record rather than recalled, and the figures quoted from each were read from the source. The evaluation runs end to end from the packaged recordings, and its recording-level results reproduce a previously vetted analysis of the same classifier exactly.

Human accountability. The authors defined the evaluation scope, the cohorts, and the operating point, reviewed the full draft, corrected it, and are responsible for its claims. Domain review caught errors no automated check in the pipeline could detect, including a miscount of the in-home cohort's enrollment and a reference cited to the wrong pagination, and those corrections are incorporated here.

References

1. World Health Organization. *Hypertension: key facts*. 2023. <https://www.who.int/news-room/fact-sheets/detail/hypertension> (accessed 7 August 2026).
 2. National Center for Health Statistics. *Hypertension Prevalence, Awareness, Treatment, and Control Among Adults Age 18 and Older: United States, August 2021 to August 2023*. NCHS Data Brief No. 511. Centers for Disease Control and Prevention, October 2024. <https://www.cdc.gov/nchs/products/databriefs/db511.htm> (accessed 7 August 2026).
 3. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. *Hypertension*. 2018; 71(6):e13-e115. doi:10.1161/HYP.0000000000000065
 4. Muntner P, Shimbo D, Carey RM, et al. Measurement of blood pressure in humans: a scientific statement from the American Heart Association. *Hypertension*. 2019; 73(5):e35-e66.
 5. Kallioinen N, Hill A, Horswill MS, Ward HE, Watson MO. Sources of inaccuracy in the measurement of adult patients' resting blood pressure in clinical settings: a systematic review. *Journal of Hypertension*. 2017; 35(3):421-441.
 6. Unlu O, et al. Patient engagement with home blood pressure monitoring. *JAMA Cardiology*. 2025. doi:10.1001/jamacardio.2025.5196
 7. Viera AJ, Lingley K, Hinderliter AL. Tolerability of the Oscar 2 ambulatory blood pressure monitor among research participants: a cross-sectional repeated measures study. *BMC Medical Research Methodology*. 2011; 11:59.
 8. Gorostidi M, Vinyoles E, Banegas JR, de la Sierra A. Prevalence of white-coat and masked hypertension in national and international registries. *Hypertension Research*. 2015; 38(1):1-7.
 9. Mukkamala R, Yavarimanesh M, Natarajan K, Hahn JO, Kyriakoulis KG, Avolio AP, Stergiou GS. Evaluation of the accuracy of cuffless blood pressure measurement devices: challenges and proposals. *Hypertension*. 2021; 78(5):1161-1167.
 10. VandeWeerd C, Sappington E, Poon L, Okeahialam-Iheanyi C, Roberts M. A comparison of traditional remote monitoring (RPM) devices versus a novel, toilet seat-based monitoring device. *Innovation in Aging*. 2022; 6(Suppl 1):584. doi:10.1093/geroni/igac059.2192
 11. Casana Care, Inc. *Smart Seat: product overview*. <https://casanacare.com/> (accessed 7 August 2026).
 12. Ding H, Ho K, Searls E, et al. Assessment of wearable device adherence for monitoring physical activity in older adults: pilot cohort study. *JMIR Aging*. 2024; 7:e60209.
 13. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. *Neural Computation*. 1991; 3(1):79-87.
 14. Frank E, Hall M. A simple approach to ordinal classification. In: De Raedt L, Flach P, eds. *Machine Learning: ECML 2001*. Springer; 2001:145-156.
-