Full-day, in home validation of infant body position measurements from inertial sensors

John M. Franchak[1], Maximilian Tang[1], Hailey Rousey[1], & Chuan Luo[1]

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

Correspondence concerning this article should be addressed to John M. Franchak, UC Riverside Department of Psychology, 900 University Avenue, Riverside, CA 92521. E-mail: franchak@ucr.edu

Abstract

Abstract

Keywords: body position, motor development, everyday experiences, sitting, machine learning

Word count: X

Full-day, in home validation of infant body position measurements from inertial sensors

From moment to moment, infants' movements facilitate and constrain how they can interact with their surroundings. Changes in *body position*—whether infants are supine on their backs, prone on their bellies, sitting, upright, or held by a caregiver—have immediate consequences for vision, object exploration, and social interaction. When sitting and upright, infants have a better view of faces and distant objects compared to their view while in a prone position (Franchak et al., 2018; Kretch et al., 2014; Luo & Franchak, 2020). Infants struggle to manipulate objects while supine and prone, but sitting affords object exploration (Soska & Adolph, 2014). Upright walking changes how infants and caregivers interact compared with crawling in a prone position; while walking infants move farther away, share toys in different ways, and hear different language from caregivers (Chen et al., 2022; Karasik et al., 2011, 2014; West & Iverson, 2021). As infants grow older and acquire new abilities, such as independent sitting and walking, the spend more time sitting and upright and less time held, supine, and prone (Adolph & Tamis-LeMonda, 2014; Franchak et al., 2018; Franchak, 2019; Thurman & Corbetta, 2017). Thus, characterizing individual differences in the day-to-day accumulation of body position experiences informs developmental theory by revealing differential opportunities for learning (Franchak, 2020).

In this paper, we present an inertial sensing method to classify infant body position from moment-to-moment across an entire day, and validate its accuracy using over 100 hours of video recorded across 35 in-home data collection sessions. Our method takes inspiration from a more mature technology: Long-form audio recordings of infants' language experiences using wearable audio recorders. We begin by describing the impact of long-form audio recordings on research in developmental psychology, and identify the key features that should be replicated in long-form recordings of motor behavior. Next, we review the current state-of-the-art in measuring infant motor behavior—video and survey data—and their limitations in capturing real-time, full-day behavior. Finally, we discuss

the advantages of using inertial sensing to classify motor behavior. Despite promising past results in brief, supervised sessions (Airaksinen et al., 2022, 2020; Franchak et al., 2021), the current investigation takes a needed step forward by testing accuracy over long, unsupervised recordings. Although in the current investigation we focus on classification of body position, this approach can be extended to categorize other aspects of movement, such as locomotion (Airaksinen et al., 2020) and infant-caregiver contact (Yao et al., 2019).

Inspiration from Long-Form Audio Methods

The LENA® recorder is a commercial device that is worn by infants in a custom shirt pocket that has sufficient battery life and storage to record for an entire day. Closed-source LENA® algorithms analyze the audio recordings to provide automatic counts of useful metrics, such as the number of words spoken by adults in the vicinity of the participant. Other long-form audio methods rely on custom-built recorders (Wass et al., 2022), apply alternative classification algorithms to LENA® data (Micheletti et al., 2022; Räsänen et al., 2020), or manually transcribe audio recorded by LENA® devices to improve accuracy or identify behaviors beyond the built-in categories (Bergelson, Casillas, et al., 2019; Mendoza & Fausey, 2021).

Long-form audio recordings have had a transformational impact on language development research by allowing researchers to characterize opportunities for learning in daily life. Measuring the amount of speech heard by infants in the home (Weisleder & Fernald, 2013) or in a daycare setting (Perry et al., 2018) revealed individual differences input that predict later vocabulary. Full-day language recording synchronized with other data sources allows researchers to identify how auditory input and vocal production interact with other processes. Beyond individual differences in aggregated data, long-form recordings can be used to determine the temporal schedule of experiences. For example, infants' daily experiences hearing music are clustered in time, with "bursty" episodes of hearing music separated by relatively long periods during which music is absent (Mendoza

& Fausey, 2022). Synchronizing audio recordings with other data sources extends researchers' ability to characterize daily experiences. Linking LENA® speech measurements to repeated, time-stamped text-message surveys about infant device placement revealed that infants heard less caregiver speech during moments that they were restrained in devices such as swings, exersaucers, and car seats (Malachowski et al., 2023). A custom-built wearable ECG and audio recorder allowed Wass et al. (2022) to discover that infant arousal increases the likelihood of infant vocalization across the day.

We identified five key features of long-form audio methods that should be replicated in analogous studies of motor behavior. First, wearable audio recorders are *mobile*. Measurement is not limited to a particular room because the recording device travels with the participant. Data are recorded to onboard device memory, so participants do not need to be in range of a receiver. Second, wearable audio recording is *unobtrusive*. Participants' reactivity to observation, such as from a video camera, may influence behavior. For example, caregivers spoke more frequently to infants during a video-recorded portion of a home recording compared with audio-only segments captured by a LENA® device (Bergelson, Amatuni, et al., 2019). Third and fourth, recordings capture *real-time data* over a *full day*. The ability to record real-time data is vital for making inferences about processes that happen on the timescale of minutes or even seconds within in an individual as opposed to comparisons of aggregated data between infants. Synchronizing real-time data to other data streams helps to reveal sources of variability within an individual (e.g., Wass et al., 2022). Full-day recordings are essential for capturing experiences across the heterogeniety of daily routines that moderate behavior (e.g., play, feeding, errands) (Kadooka et al., 2021, April; Tamis-LeMonda et al., 2018). "Burstiness" of behavior means that long recordings are needed to capture clusters of events amid long periods in which they may be absent (Barbaro & Fausey, 2022; Warlaumont et al., 2021). Fifth, *automatic classification* means that the approach can scale to analyze large numbers of participants over long recordings without the bottleneck of manual annotation/transcription.

Automatic classification can only replace human annotation if it is sufficiently accurate and unbiased. An independent assessment of the LENA® algorithms found mixed results about the accuracy of different outcomes. For example, correlations between human transcribed counts of adult words and child vocalizations against LENA®'s automatic counts were strong, $r = .698$ and $r = .649$, respectively (Cristia et al., 2020). However, poor agreement was found for other metrics, such as the number of "conversational turns" between the child and communicative partners and the identification of male speakers.

Thus, for some use cases (and for some metrics), long form audio recordings provide a mobile, unobtrusive way to automatically score real-time data over a full day. In the remainder of the paper, we turn to the question of how to replicate these qualities in long form recordings of motor behavior.

Limitations of Video and Survey Methods

Video and survey methods are the current state-of-the-art in assessing motor behavior. Although, each method has advantages and disadvantages for characterizing infants' everyday motor experiences that complements the other, neither method on its own can provide comparable data to the long form audio recordings reviewed in the previous section.

Video observation is the most common way of measuring infant motor behavior in home recordings. Most often, an experimenter with a handheld camera follows infants from room to room to ensure that their movements are visible throughout the recording session (Chen et al., 2022; Herzberg et al., 2021; Karasik et al., 2011). The primary advantage of video recording is that it captures real-time behavior. Infants' body position, locomotion, and reaching comprise events that occur on the timescale of seconds, so standard video recording is adequate to score gross motor behavior. However, requiring an experimenter to operate a camera is obtrusive, whereas relying on a stationary camera means that infants

will be absent from view as they move from place to place. Moreover, video observation cannot easily scale to long durations or large numbers of participants. Logistically, an experimenter cannot follow behind infants to record their behavior from morning to night (and were they to do so, they would likely alter infants' and caregivers' behavior). Typical video recording sessions last 45-120 minutes (Chen et al., 2022; Herzberg et al., 2021; Karasik et al., 2011), short of capturing the variety of activities across the full daily routine. Even if full day videos were available, the lack of suitable automatic classification tools means that the human cost of annotation makes it difficult for video methods to scale to testing large numbers of participants. Our annotation of body position takes approximately 2-5 hours to complete for every hour of video (depending on how often infants switch positions), meaning that a full "waking day" of approximately 11 hours for a 12-month-old (Galland et al., 2012) could take 22-55 hours of labor to full annotate.

In contrast, survey methods such as daily diaries/inventories or ecological momentary assessment (EMA) are mobile, unobtrusive, can be applied across an entire day, and do not need laborious annotation. Diary studies provide caregivers with logs or structured interviews to estimate from memory how much time infants spend in particular activities (Karasik et al., 2022; Majnemer & Barr, 2005). Ecological momentary assessment uses text-message or app-based notifications to prompt caregivers to make repeated estimates of infants' behavior over the course of a day (Franchak, 2019; Kadooka et al., 2021, April). And although the responses are valuable in aggregate, survey methods lack the real-time temporal resolution to describe moment-to-moment changes in behavior. For diaries and interviews, limits on caregivers' memory mean that they will report what was most frequent, but cannot remember events that happen on the scale of seconds and minutes. At best, EMA surveys prompt caregivers to make hourly observations; increasing the number of surveys per day would be too burdensome for the respondent. Thus, despite being a useful tool for estimating broad developmental changes and individual differences in infants' motor experiences, survey methods are not suited for capturing within-participant

temporal dynamics.

Promise of Inertial Sensing Methods

Measuring infant movement with inertial movement units (IMUs) is a promising avenue for long form recordings of motor behavior in the home (Barbaro, 2019; Bruijns et al., 2020; Cliff et al., 2009; Lobo et al., 2019). Lightweight sensors (10-30 g) can be embedded in garments to make recordings fully *mobile*, and they are *unobtrusive* because they do not require a researcher's presence nor do they record sensitive audio/video that might influence participants' behavior. Many commercially-available and inexpensive IMUs have $>$ 12 hour battery life with onboard storage to record *real-time*, *full-day* motion data at a sampling rate that is higher than typical video (e.g., 50-100 Hz). Moreover, past work has successfully recorded the rate of leg kicks (e.g., Deng et al., 2019) and activity intensity (Schneller et al., 2017) in infants and children across multiple days.

The open question is whether *automatic classification* is sufficiently accurate to measure movement categories that are relevant to developmental and clinical research, and whether measurement validity is acceptable over long recording periods. IMUs typically contain accelerometers that measure linear acceleration paired with gyroscopes that measure angular acceleration. Unlike motion tracking systems that might be used in a lab, IMUs do not provide data about the position of the body in space. Thus, data processing algorithms are needed to classify the raw sensor data (i.e., linear and angular acceleration timeseries) into meaningful categories (e.g., supine, prone, sitting, upright). The difficulty of the classification task depends on the categories of interest. More basic aspects of movement, such as overall activity intensity, can be identified by taking the magnitude of acceleration (irrespective of direction) and applying thresholds or cut-points to define when a movement has occurred at a particular intensity (Armstrong et al., 2019; e.g., Hager et al., 2017).

Categorizing body position—supine on the back, prone on the belly, sitting, upright, or held off the ground by a caregiver—is too complex for cut-point definitions to be accurate. First, the magnitude of movement can vary greatly *within* a body position. An upright infant can be standing still or can be walking briskly across the room. A prone infant and be stationary in "tummy time", or they can crawl in myriad ways (Adolph et al., 1998). Moreover, the configuration of the arms, legs, and torso within a body position can vary greatly in everyday contexts. Infants can sit on the floor in a tripod position with support from an arm, in a "V" position with legs fully extended, in a "W" position with knees bent. Sitting on a caregivers' lap without the need to maintain balance means that the legs can dangle and the torso can lean in different directions. Sitting in a high chair or car seat reduces the magnitude of postural sway within sitting, and creates even more possibilities for how the arms and legs may move relative to the torso. Indeed, creating an all-encompassing set of rules for how to annotate sitting from video is no trivial task because of the various ways that sitting can occur in daily life. Finally, caregivers frequently pick up and transport infants, creating motion signals that need to be differentiated from independent activity (Kwon et al., 2019; Patel et al., 2019). Thus, modern approaches to human activity recognition have used machine learning to classify activity categories based on features derived from IMU data in adults (Arif & Kattan, 2015; Preece et al., 2009), children (Nam & Park, 2013; Ren et al., 2016; Stewart et al., 2018), and infants (Airaksinen et al., 2020; Franchak et al., 2021; Yao et al., 2019). Synchronized video with ground-truth human annotations creates training data for a machine learning algorithm, such as a random forest model, that can be later used to predict categories from IMU data that was not annotated. Crucially, this allows automatic classification to scale to full day recording by relying on a relatively smaller set of video annotation.

Three prior investigations have used machine learning to categorize infant body position from IMUs towards the goal of collecting full-day data. Airaksinen et al. (2020) tested 4- to 8-month-olds in a laboratory visit, and found 95% accuracy in distinguishing

between body position categories that crawling infants could perform on the floor (excluding times that infants were held by caregivers). Using a wider age range of 6-18 months, Franchak et al. (2021) found 98% accuracy (*kappa* = 95%) in a laboratory validation study in categorizing body position that included infants who could both crawl and walk and also included a category for caregiver holding. Most recently, Airaksinen et al. (2022) conducted a validation study of body position classification in either a home or clinic testing in 4- to 19-month-olds, refining their previous method to detect moments that infants were carried by caregivers. Classification accuracy did not vary between lab and home settings, and was generally high (95%, *kappa* = .93). Although all three studies yielded promising classification accuracy, accuracy was assessed in brief (15-60 minute) sessions supervised by a researcher, leaving the open question of how well body position classification will scale to testing across an entire day of natural home life.

Goals of the Current Study

Accordingly, the overarching goal of the current study is to test the feasibility and validity of long-form body position recording in the home during unsupervised, everyday behavior. Supervised recordings from past work (Airaksinen et al., 2022, 2020; Franchak et al., 2021), whether in the home or in the lab, let researchers set up the situation to encourage or restrict certain behaviors. Usually, caregivers are asked to play with the infant. However, across a real day, non-play activities (e.g., eating lunch in a high chair) create challenging situations for applying automated classification of body position. Will models trained on video-recorded observations at the beginning of the day generalize to predict behavior at a later time? Assessing the validity of temporally *distal* periods is a crucial step to establish whether automatic classification can be used to measure body position across a day.

In the current study, we report the feasibility and validity of body position classification over the full day in the home based on 35 testing sessions from 22 infants

aged 4-14 months. Participants received a custom pair of infant leggings embedded with 4 IMUs (one on each ankle and one on each hip) and a video camera to collect ground truth data about infant body position. A *proximal comparison* period began when participants received the equipment and completed a guided phone call during which caregivers were asked to elicit different body positions based on prompts from the experimenter. Although not directly supervised, this period was most similar to previous recordings because it occurred during a convenient time for the infant and caregiver to play while they received instructions from the experimenter. The **first goal** of the current study was to determine the accuracy of body position classification during the proximal comparison period using this novel, semi-supervised procedure in participants' homes. Past work found varying better performance using "individual models"—models that were trained on one participant's data to predict their later behavior—compared with a "group model" that aggregated data from all infants to create a single body position classifier (Franchak et al., 2021), so we compared both modeling approaches in the current investigation.

A second, *distal comparison* period followed the proximal comparison period and captured approximately 90 minutes of home behavior that was completely unsupervised. Caregivers and infants could (and did) do whatever they wished. Because this recording happened a considerable amount of time after the initial setup and instructions from the experimenter, accuracy could decline if caregivers or infants moved the garment or sensors. Moreover, increasing variation in everyday activities during the distal comparison creates a greater challenge for classification, testing whether machine learning models can generalize to novel test cases. Thus, the **second goal** of our study was to assess accuracy during the distal comparison.

After this second video recording period, we asked caregivers to keep infants wearing IMUs for the rest of the day until their regular bedtime, creating the **first real-time, full-day dataset of infant body position**. Interpreting such data required caregivers to

log when infants napped, when they removed the sensor garment for diaper changes or other reasons, and when infants went to bed at the end of the day. The **third goal** of the study was to examine the quality of the full-day data. Could body position be successfully recorded over the desired period? Finally, the novel full-day dataset affords us a unique chance to ask whether the estimated time infants spent in different body positions align with past results using video and survey methods. Thus, the **fourth goal** of the current study was to determine whether full-day body position measurements conformed to expectations about age differences in body position. Based on past results (Franchak, 2019), infants should spend increasingly more time sitting and upright but less time supine over the age range tested (4 to 14 months).

Methods

Participants and Design

Infants were recruited in one of two age groups: *Younger* infants were between 4 and 7 months and *Older* infants were between 11 and 14 months. There were 9 infants in the 4-7 month group (X female) and 14 in the 11-14 month group (X female). Families were recruited through social media advertisements and from community events in Southern California. The ethnicity for infants was reported as X Hispanix/Latinx and X for not Hispanix/Latinx. Race was reported as X Caucasian, X Asian, X ~~~ and X others. Parents were compensated $30 in cash for each visit they completed. The BLINDED Institutional Review Board reviewed and approved all procedures associated with the study. All participants gave their informed consent to participate.

Most participants were tested in a single session ($n = 17$), but 6 participants contributed between 2-4 sessions as part of an ongoing longitudinal study. Only 1 session was excluded due to a technical error—one of the four IMU sensors failed to record, resulting in an unusable set of data for classification. Across the two age groups, we report data on a total of 34 sessions, with 14 sessions from younger infants and 20 sessions from

older infants. Across sessions, younger infants' age ranged from 3.84 to 7.16 ($M = 4.99$) and older infants' age ranged from 10.74 to 14.23 ($M = 11.75$).

Apparatus

Four inertial movement units (IMUs) were used to record infant movement across the day (MC10 Biostamp). A custom garment was made to hold the IMUs: Internal pockets were sewn into a snug-fitting pair of infant leggings so that IMUs would stay close to the body and so that infants could not pull out the sensors. On each side of the body there was a pocket over the hip and a pocket just above the ankle. Each sensor recorded accelerometer and gyroscope data at 62.5 Hz throughout the day, with sufficient battery and on-device storage to record for approximately 12 hours. Infants also wore a LENA® recorder throughout the day in the front pocket of a LENA® shirt, located near the infant's chest.

Video recordings were captured using an action camera on a miniature tripod (Insta360 ONE R) that caregivers placed in the room that the infant was in. The proprietary "Boosted Battery Base" attached to the action camera to allow for a total of 3 hours of recording. However, as described below, this created two video files temporally separated by a gap of approximately 40-45 s. Caregivers also received a log sheet to record times that infants napped as well as times that the sensor garment was removed from the infant (e.g., baths, diaper changes).

Procedure

Figure 1) shows an exemplar timeline of the entire procedure and recording periods for a single participant. On the day of the visit, a researcher arrived at the participant's home in the morning between and prepared all the equipment at the doorstep. In order to synchronize all three recording devices, the researcher began by first turning on the video camera and the LENA audio device (the IMUs were already configured and placed at

arrival because they required a proprietary sensor dock to begin recording in the laboratory). To create an easily recognizable synchronization point between the video recording and IMU data, the researcher dropped or struck the leggings (containing the IMUs) on a surface in view of the camera, as in Franchak et al. (2021). All the equipment—once recording and with synchronization information recorded—was placed inside a large bucket and left outside the family's front door.

The researcher then called the caregiver on the phone and walked them through a set of procedures needed to properly set up the equipment and record video for classifier training and testing. At the start of this "guided call", the caregiver was instructed to place the camera in an area that captured the majority of the room. Next, they were asked to put the pair of leggings and shirt on their infant, with the researcher providing guidance about how to correctly orient the garments.

Afterwards, the researcher asked the caregiver to complete a number of guided activities with their infant. Within view of the camera, the caregiver was asked to place their infant in several different positions: lying supine, lying prone, sitting on the floor, standing upright, held by the caregiver while the caregiver walked back and forth, crawling, walking, and sitting in a restrained seat (e.g., high chair). Depending on the infants' age and motor skill level, the positions could be done independently or were completed with assistance from the caregiver. The researcher kept time to ensure at least 1 minute of behavior for each activity. Once completed, the caregiver was then instructed to play with their infant for 10 minutes within view of the camera to collect additional training data with the infant in positions that would be typical of play.

Afterwards, they were to go about their day as usual with the infant wearing the sensor garment until their bedtime, only taking off the sensor for naps, baths, and diaper changes. The caregiver logged the times the sensors were removed (blank areas in the timeline in Figure 1)) or the child took a nap (gray areas in the timeline in Figure 1)) so

that those times could be excluded from analysis. The following day a researcher picked up the equipment, verified the paperwork was signed, and compensated the participant.

Because the camera only had the battery life to record for ~3 hours (divided into two 90-minute video files), this divided the day into different periods for analysis. As seen in the bottom of Figure 1), the *video period* comprised the first three hours of recording starting from the researcher's arrival when they turned on the camera. The first 90-minute video file, termed the *proximal comparison*, contained the activities during the guided call followed by a period of infants and caregivers resuming their normal activities. Because this video contained the synchronization point, the data in this period had temporal synchrony between IMU and video data that contained errors of no more than 30-60 ms (1-2 video frames). When the first video file ended, a second video file was recorded, termed the *distal comparison*. This video recorded the next 90 minutes of natural activity. However, because there was a variable gap of ~40 s between the two videos, temporal synchrony in the distal comparison video was approximate containing offsets of ~ 5 s in either direction.

Body Position Annotation

The proximal and distal comparison videos for each participant were annotated by trained human coders to identify infant *body position* into one of 5 mutually-exclusive categories: supine, prone, sitting, upright, or held by caregiver. All coding was done using Datavyu software (datavyu.org).

Supine was coded when the infant was lying on their back or was reclined up to a 45 degrees angle. Supine was also coded in the rare cases when the infant was laying on their side. Prone was coded when the infant was lying on their stomach, was on all fours in a downward dog position, or was crawling. We scored sitting to include any form of the following seated positions: 1) infants sat with their buttocks on a surface, such as on the floor or a caregiver's lap, 2) infant was in a kneeling-sit position, in which their knees were

on the ground with their legs tucked underneath the buttocks, and 3) infant was in a seating device, such as a high chair, that kept the torso oriented perpendicular to the ground (a reclined position, such as in a young infant's car seat, would be counted as supine). Upright was coded when the infant was standing on the ground with two feet or walking (regardless of whether infants' balance was assisted by a caregiver or with their hands holding onto something for support). When an infant was carried by a caregiver, held was coded. However, when the caregiver was sitting with the infant in their lap the infant's body position was coded as if the caregiver was a surface (e.g., if the infant was sitting on the caregiver's lap this was coded as sitting). Times during the video when the infant was out of view were excluded. Periods when the leggings were being adjusted or taken off the infant were also excluded, as were transitions between body positions.

A primary coder completed annotation for the full length of the video, while an independent reliability coder completed annotation for the first thirty minutes of each video. Interrater reliability was based on the proportion of video frames that the two coders chose the same body position code. Overall agreement averaged 90.5% across video files, ranging from 68.4%-100%. Cohen's *kappa* averaged 85.5% across video files, ranging from 31.0%-100%.

Body Position Classification

The same machine learning classification process was used as in prior work (Franchak et al., 2021). Using the synchronization point, human-coded body annotations from video were linked to the corresponding times in the IMU timeseries data. A single, merged dataset was created with synchronized accelerometer signals (in three orientations: X, Y, and Z) and gyroscope signals (in three orientations: roll, pitch, and yaw) for each of the four sensors (left hip, right hip, left ankle, right ankle) with the corresponding timestamp and body position code.

Classification training and prediction was conducted on a windowed dataset that summarized the raw, 62.5 Hz motion signals within 4-s windows. Overlapping moving windows captured 4-s (250 samples) of data starting each second, which is a common unit of analysis in prior studies of human activity classification (Airaksinen et al., 2020; Franchak et al., 2021; Nam & Park, 2013). The 250 samples of data in each window were aggregated into a variety of motion features—summary statistics that could be fed into the machine learning model. The minimum, maximum, 25th percentile, 75th percentile, mean, median, skew, kurtosis, standard deviation, and sum were computed for each signal (e.g., right hip X linear acceleration, left ankle pitch angular acceleration). The 10 summary statistics and 24 sensor signals generated 240 columns of motion features that described movement within each window. Furthermore, a series of cross-sensor and cross-orientation summaries (such as the correlation, magnitude, and difference between pairs of sensors) added an additional 196 columns of motion features. The 436 total motion features corresponded to a single body annotation code for each 4-s window. Windows were only used for training/testing if they contained a single body position for $> 75\%$ (3 s) of time within the window to ensure that motion signals could be tied to a clear example of each behavior.

Windowed datasets were used for machine learning classification and validation. For each analysis reported in the results, a subset of data were defined as a "training" set and another, independent portion of the data were defined as a "testing" set. Different training/testing set combinations were used to address different goals; here we describe the general procedure of how a classifier was derived from a training set. Random forest models (Breiman, 2001) were trained to predict the human coded body position label for each window from the set of 436 motion features using the *randomForest* package in R (Liaw et al., 2002). Random forest models use random subsets of features and random subsets of data across many iterations; this random subsetting creates models that are less likely to be overfit. The *tidymodels* set of packages in R (Kuhn & Wickham, 2020) was

used to tune the "mtry" and "ntrees" hyperparameters, which control the number of columns sampled and the total number of trees created when modeling. We use an mtry parameter of 44 and and ntree parameter of 550, however, changing hyperparameters had little impact on model performance. Just like a linear regression model, the resulting random forest model could later be applied to a set of testing data with the *predict* function. Validation analyses below will compare the human-coded annotations and random forest model predictions for testing data that are independent from training data, providing a test of whether models can successfully generalize.

Results

We report four sets of results based on 34 full-day testing sessions resulting in a total of 301.74 hours of movement recording. First, we focus on the accuracy of group and individual models trained and validated on the proximal comparison period. This period is unique in having the highest degree of temporal synchrony between video and motion data, allowing us to assess the accuracy of individual body position events. Second, we applied models trained from data during the proximal comparison to predict infant behavior during the distal comparison, when coarser synchrony between video and motion data was available. This distal comparison is crucial, because it provides the first ever test of how well body classification models predict behavior beyond the initial, supervised period. Data from the first two sets of analyses confirm that body position classification models are accurate across a long period of time, suggesting that data from the entire day should be valid. In the third set of results we examine the data quality of full-day recordings, and in the fourth set of results we test whether full-day recordings capture well-established age differences in body position experiences.

Assess the Proximal Accuracy of Body Position Classification Models

The first set of analyses use data from the proximal comparison, in which video annotations and motion data were tightly synchronized (within 30-60 ms). This high

degree of synchronization makes it possible to link human-coded body position annotations to each 4-s window of motion data, providing ground truth data for model training and testing. As in past work (Franchak et al., 2021), we compared two types of models: *group models* and *individual models*. To assess the accuracy of body position classification for each recording session, we held back the last 25% of a session's proximal comparison data as the testing set. The testing set was never used as training data. A group model was trained for each session using an aggregated dataset containing the first 75% of all **other** sessions' proximal comparison data. This leave-one-out cross-validation tested the generalization of the model to a recording session that was not used at all in the training set. In contrast, individual models were trained using the first 75% of a recording session, then tested against the last 25%. The individual model tests whether earlier training data generalize to later testing data within an individual participant's recording. In both cases, the testing set was the same, allowing for direct comparison between the two modeling approaches. Figure 2, Table 1, and Table 2 summarize the performance of group and individual models using standard metrics for classification.

Overall Accuracy. Overall accuracy (Figure 2A) represents the proportion of 4-s windows in the testing set in which the model prediction matched the human annotation of body position. Overall accuracy for group models ($M = 0.85$) was slightly lower than accuracy for individual models ($M = 0.92$). Although overall accuracy from our unsupervised, in-home data collection did not match the near-perfect accuracy (.95-.98) found in prior in-lab studies (Airaksinen et al., 2020; Franchak et al., 2021), both models approached the level of agreement found between two human coders ($M = .905$). Most likely, lower accuracy in the current study results from the more variable and complex behavior observed in an unsupervised setting rather than from a difference in the quality of the classification model. Visual inspection of Figure 2A shows that accuracy values were heavily skewed, with most approaching perfect accuracy but a few participants with very poor accuracy. Looking at the median performance suggests that the difference between

models was not as wide for the typical participant (group median accuracy = 0.89; individual median accuracy = 0.93. Indeed, it is notable that the worst-case accuracy from group models (minimum = 0.50) was considerably lower than in individual models (minimum = 0.73). Possibly, individual models accounted for idiosyncrasies in behavior or inconsistencies in sensor placement for those sessions that group models could not generalize to.

Cohen's Kappa. Strong overall accuracy can be misleading when class prevalence is unbalanced. For example, most infants spent longer sitting than prone, so good performance for sitting classification could mask poorer performance in prone classification. Accordingly, we report Cohen's kappa, a commonly-used metric that penalizes missing rare events (Figure 2B), and provide classification metrics for each individual body position (Table 2) to account for imbalance in body position rates within and between individuals. Similar to overall accuracy, kappa values were strong for both model types with group kappas ($M = 0.75$) somewhat worse compared with individual kappas ($M = 0.82$). Guidelines for interpreting kappa statistics (Landis & Koch, 1977) consider 0.81–1.00 "Almost Perfect," 0.61–0.80 "Substantial," 0.41–0.60 "Moderate," 0.21– 0.40 "Fair," and 0–0.20 "Slight to Poor", indicating that agreement for most group and individual model predictions fell in the Substantial to Almost Perfect range.

As in past work (Airaksinen et al., 2020; Franchak et al., 2021), all body positions were accurately classified even though performance varied somewhat between positions. As Table 2 shows, mean kappa statistics were strongest for prone (group $M = 0.860$, individual $M = 0.841$) and supine (group $M = 0.764$, individual $M = 0.912$). Sitting performance fell in the middle, and was considerably worse for group models than individual models (group $M = 0.702$, individual $M = 0.887$). Held (group $M = 0.726$, individual $M = 0.727$) and upright (group $M = 0.673$, individual $M = 0.741$) performance was the least accurate, however, average performance was still within the "Substantial" range.

Sensitivity and Positive Predictive Value. Beyond accuracy, it is important to establish that classification is neither too sensitive nor too conservative in choosing one position versus another. Sensitivity refers to the proportion of events of a given position that were correctly identified (e.g., out of 100 human-coded sitting windows, how many of those windows did the model correctly classify as sitting?). High sensitivity means that events are unlikely to be missed. In contrast, positive predictive value (PPV) refers to the proportion of events classified for a given position actually belonged to that position (e.g., if the model said a baby was upright during 100 windows, how many of those windows were indeed human-coded sitting events?). High PPV means that we can be confident in the event label. Table 2 shows the sensitivity and PPV by body position class for group and individual models. For group models, sensitivity and PPV were similar: They were highest for supine, prone, and sitting (the most accurately identified class) and lowest for upright and held. All values were sufficient ($> .8$). Results for individual models were similar to group models, with the exception of a somewhat lower sensitivity score for held. Overall, the results suggest a reasonable balance between sensitivity and PPV among different body position classes for both model types. Table 1 shows the overall sensitivity and PPV across classes.

Measure the Distal Accuracy of Body Position Classification Models

In the first set of results, we showed that group and individual models trained from data during the proximal comparison period were accurate at classifying body position in the same recording. In the next analysis, we examine long-term performance by testing how accurately models trained from the proximal period could predict body position during the distal period—the 90-minute long recordings that followed the proximal recordings (Figure 1). A single group model was created using all sessions' proximal period training data (rather than group models leaving out a single session); the same individual models were used. As previously mentioned, distal videos had only coarse temporal

synchrony with motion recordings (misalignment of up to ~ 5 s in either direction). This misalignment precluded calculating accuracy based on the proportion of matching events, so we used another technique following past work (Franchak et al., 2021; Yao et al., 2019). For each distal comparison, we summed the amount of time infants were predicted to be in each of the 5 body position categories from the model and compared that to the summed time for the body positions based on human coding. Below, we report correlations between the model-predicted and human-coded aggregated body position time across the entire distal period as well as within finer, 10-minute bins.

Whereas the first set of analyses used all 34 sessions, this was not possible in the distal comparison. Because the start of the visit was scheduled during a time the infant was awake, it was common for a nap to be needed following the proximal period. Nine sessions were excluded because the infant was either napping or otherwise not on camera during the entire 90-minute distal recording. Three additional sessions were excluded because a caregiver accidentally turned off the video camera ($n = 1$) or left the house ($n = 2$). This left 22 sessions with usable distal comparison data.

Overall Agreement During the Distal Comparison. Figure 3 and Table 3 summarize the overall agreement during the distal comparison period. For each session, we calculated the actual prevalence of each body position as a percentage of the time that the infant was on video and awake using human annotated body position (x-axis on Figure 3). Predicted prevalence was calculated the same way for group and individual model predictions, omitting the off-camera and nap periods to make a direct comparison. Overall—across participants and across body position classes—agreement was strong: The correlation between group model predictions and human-coded prevalence was $r = 0.80$, and the correlation between individual model predictions and human-coded prevalence was $r = 0.91$. As in the proximal comparison, agreement varied somewhat between body positions; in particular, agreement for held was poor. Unlike in the proximal period, some body positions were better predicted by group models and others by individual models.

Visual inspection of Figure 3 indicated two extreme outliers, which we marked by a gray square and a gray diamond. We closely investigated each of the two outliers to understand why their agreement was below the standard of the other sessions. The "gray square" outlier had significant confusion between sitting and supine classification. Reviewing the video indicated that this participant spent a long period of time in a seating device that was reclined almost exactly at 45 degrees, making it difficult to determine if the infant was sitting or supine. The infant also spent a long time in mother's arms in an ambiguous supine/sitting position that was hard to classify. This participant's proximal accuracy was also poor because similar ambiguous appeared during the initial recording. In contrast, the "gray diamond outlier" had strong proximal accuracy, with confusion only arising in the distal period between upright and held categories. Reviewing the video showed that all disagreements occurred when the infant was in a baby walker; human coders scored this as "upright" but the models predicted it as "held". Most likely, the infant's movements in the baby walker were more similar to how a baby moved while carried, and unlike how most infant's moved while walking upright.

What is notable about both outliers is that disagreements were restricted to a particular border case (supine vs. sitting; upright vs. held); accuracy for other classes remained strong. This suggests that their poor performance came as a result of spending a long time in an ambiguous position, not the result of the entire model failing to generalize to the later time period (or an error in sensor placement, such as if the parent removed the leggings and put them on backwards after a diaper change). To better capture the typical level of agreement, we report all correlations in Table 3 excluding the two outliers. Overall agreement among the non-outlier sessions was excellent for both group models ($r = 0.95$) and individual models ($r = 0.96$). With the exception of held, within-class agreement was strong for the other four body positions.

Short-Timescale Agreement during the Distal Comparison. Collecting real-time data that can be synchronized to other sources is a key goal in our development of

long-form body position classification. Although overall aggregate agreement in the distal comparison was strong, it is important to show that similarly strong agreement is found within a shorter timescale. We repeated agreement analysis, correlating human-coded body position prevalence with model-predicted prevalence, after dividing the distal comparison period into nine 10-minute bins (marked by vertical dashed lines in Figure 1). Infants had varying numbers of 10-minute bins depending on much time they were awake and on camera. Bins were included only if there was > X minutes of usable data.

Table 4 shows the agreement correlation coefficients for group and individual models, including and excluding the outliers identified in the previous section. Performance at a short timescale was similar to performance overall: Overall agreement after excluding outliers was excellent for both group ($r = 0.92$) and individual models ($r = 0.94$). Within-position correlations were weakest for held and strongest for upright regardless of the model type. Agreement for prone was better for group models, whereas sitting and supine were better predicted by individual models.

To describe the observed amount of prediction error in 10-minute bins, we subtracted the predicted duration (in minutes) for each body position in each bin from the human coded duration in that bin to create a prediction difference score. A score of 0 would indicate no error; positive differences indicate that the model overestimated the amount of time in a position, whereas negative differences indicate underestimation. Figure 4 plots the mean prediction difference for each session for each body position. The gray shaded area marks ± 1 minute of prediction error. The vast majority of session-averaged predictions fall within 1 minute of error without a clear bias towards overestimation or underestimation (with the previously-identified outliers as exceptions). For group models, we calculated the percentage of 10-min bins across participants that had errors < 1 minute: 94.68% for held, 80.85% for supine, 94.68% for prone, 77.66% for sitting, and 94.68% for upright. For individual models, the percent of 10-min bins with < 1 minute of error was:

92.94% for held, 88.24% for supine, 88.24% for prone, 83.53% for sitting, and 96.47% for upright. Taken together, these results suggest that most estimates across models, sessions, and body positions had small amounts of prediction error even after a long delay from when models were trained.

Goal 3: Examine the data quality of full-day home recordings

The start time ranged from 08:55 to 13:20 with a median of 10:20. Although the equipment was always dropped off in the morning,

Goal 4: Assess the suitability of full-day predictions for capturing age differences in body position

Discussion

References

Adolph, K. E., & Tamis-LeMonda, C. S. (2014). The costs and benefits of development: The transition from crawling to walking. *Child Development Perspectives*, *8*, 187–192. https://doi.org/10.1111/cdep.12085

Adolph, K. E., Vereijken, B., & Denny, M. A. (1998). Learning to crawl. *Child Development*, *69*, 1299–1312. https://doi.org/10.1111/j.1467-8624.1998.tb06213.x

Airaksinen, M., Gallen, A., Kivi, A., Vijayakrishnan, P., Häyrinen, T., Ilén, E., Räsänen, O., Haataja, L. M., & Vanhatalo, S. (2022). Intelligent wearable allows out-of-the-lab tracking of developing motor abilities in infants. *Communications Medicine*, *2*(1). https://doi.org/10.1038/s43856-022-00131-6

Airaksinen, M., Räsänen, O., Ilén, E., Häyrinen, T., Kivi, A., Marchi, V., Gallen, A., Blom, S., Varhe, A., Kaartinen, N., et al. (2020). Automatic posture and movement tracking of infants with wearable movement sensors. *Scientific Reports*, *10*(1), 1–13.

Arif, M., & Kattan, A. (2015). Physical activities monitoring using wearable acceleration sensors attached to the body. *PLoS ONE*, *10*, e0130851.

Armstrong, B., Covington, L. B., Hager, E. R., & Black, M. M. (2019). Objective sleep and physical activity using 24-hour ankle-worn accelerometry among toddlers from low-income families. *Sleep Health*, *5*(5), 459–465.

Barbaro, K. de. (2019). Automated sensing of daily activity: A new lens into development. *Developmental Psychobiology*, *61*(3), 444–464.

Barbaro, K. de, & Fausey, C. M. (2022). Ten lessons about infants' everyday experiences. *Current Directions in Psychological Science*, *31*(1), 28–33. https://doi.org/10.1177/09637214211059536

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science*, *22*, e12715.

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do North American babies hear? A large-scale cross-corpus analysis.

Developmental Science, *22*, e12724.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Bruijns, B. A., Truelove, S., Johnson, A. M., Gilliland, J., & Tucker, P. (2020). Infants' and toddlers' physical activity and sedentary time as measured by accelerometry: A systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity*, *17*(1), 14.

Chen, Q., Schneider, J. L., West, K. L., & Iverson, J. M. (2022). Infant locomotion shapes proximity to adults during everyday play in the u.s. *Infancy*. https://doi.org/10.1111/infa.12503

Cliff, D. P., Reilly, J. J., & Okely, A. D. (2009). Methodological considerations in using accelerometers to assess habitual physical activity in children aged 0–5 years. *Journal of Science and Medicine in Sport*, *12*(5), 557–567.

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J., & Bergelson, E. (2020). A thorough evaluation of the language environment analysis (LENA) system. *Behavior Research Methods*, *53*(2), 467–486. https://doi.org/10.3758/s13428-020-01393-5

Deng, W., Trujillo-Priego, I. A., & Smith, B. A. (2019). How many days are necessary to represent an infant's typical daily leg movement behavior using wearable sensors? *Physical Therapy*, *99*(6), 730–738. https://doi.org/10.1093/ptj/pzz036

Franchak, J. M. (2019). Changing opportunities for learning in everyday life: Infant body position over the first year. *Infancy*, *24*, 187–209.

Franchak, J. M. (2020). The ecology of infants' perceptual-motor exploration. *Current Opinion in Psychology*, *32*, 110–114.

Franchak, J. M., Kretch, K. S., & Adolph, K. E. (2018). See and be seen: Infant-caregiver social looking during locomotor free play. *Developmental Science*, *21*, e12626.

Franchak, J. M., Scott, V., & Luo, C. (2021). A contactless method for measuring full-day, naturalistic motor behavior using wearable inertial sensors. *Frontiers in Psychology*, *12*.

https://doi.org/10.3389/fpsyg.2021.701343

Galland, B. C., Taylor, B. J., Elder, D. E., & Herbison, P. (2012). Normal sleep patterns in infants and children: A systematic review of observational studies. *Sleep Medicine Reviews*, *16*(3), 213–222. https://doi.org/10.1016/j.smrv.2011.06.001

Hager, E., Tilton, N., Wang, Y., Kapur, N., Arbaiza, R., Merry, B., & Black, M. (2017). The home environment and toddler physical activity: An ecological momentary assessment study. *Pediatric Obesity*, *12*(1), 1–9.

Herzberg, O., Fletcher, K. K., Schatz, J. L., Adolph, K. E., & Tamis-LeMonda, C. S. (2021). Infant exuberant object play at home: Immense amounts of time-distributed, variable practice. *Child Development*, *93*(1), 150–164.

Kadooka, K., Caufield, M., Fausey, C. M., & Franchak, J. M. (2021, April). Visuomotor learning opportunities are nested within everyday activities. *Paper Presented at the Biennial Meeting of the Society for Research in Child Development*.

Karasik, L. B., Kuchirko, Y., Dodojonova, R. M., & Elison, J. T. (2022). Comparison of U.S. And Tajik infants' time in containment devices. *Infant and Child Development*, *31*(4). https://doi.org/10.1002/icd.2340

Karasik, L. B., Tamis-LeMonda, C. S., & Adolph, K. E. (2011). Transition from crawling to walking and infants' actions with objects and people. *Child Development*, *82*, 1199–1209. https://doi.org/10.1111/j.1467-8624.2011.01595.x

Karasik, L. B., Tamis-LeMonda, C. S., & Adolph, K. E. (2014). Crawling and walking infants elicit different verbal responses from mothers. *Developmental Science*, *17*, 388–395. https://doi.org/10.1111/desc.12129

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child Development*, *85*, 1503–1518. https://doi.org/10.1111/cdev.12206

Kuhn, M., & Wickham, H. (2020). Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles. *Boston, MA, USA*.

Kwon, S., Zavos, P., Nickele, K., Sugianto, A., & Albert, M. V. (2019). Hip and wrist-worn accelerometer data analysis for toddler activities. *International Journal of Environmental Research and Public Health*, *16*(14), 2598.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Lobo, M. A., Hall, M. L., Greenspan, B., Rohloff, P., Prosser, L. A., & Smith, B. A. (2019). Wearables for pediatric rehabilitation: How to optimally design and use products to meet the needs of users. *Physical Therapy*, *99*(6), 647–657.

Luo, C., & Franchak, J. M. (2020). Head and body structure infants' visual experiences during mobile, naturalistic play. *PLoS ONE*, *15*, e0242009.

Majnemer, A., & Barr, R. G. (2005). Influence of supine sleep positioning on early motor milestone acquisition. *Developmental Medicine and Child Neurology*, *47*, 370–376.

Malachowski, L. G., Salo, V. C., Needham, A. W., & Humphreys, K. L. (2023). Infant placement and language exposure in daily life. *Infant and Child Development*. https://doi.org/10.1002/icd.2405

Mendoza, J. K., & Fausey, C. M. (2021). Everyday music in infancy. *Developmental Science*, *24*(6). https://doi.org/10.1111/desc.13122

Mendoza, J. K., & Fausey, C. M. (2022). Everyday parameters for episode-to-episode dynamics in the daily music of infancy. *Cognitive Science*, *46*(8). https://doi.org/10.1111/cogs.13178

Micheletti, M., Yao, X., Johnson, M., & Barbaro, K. de. (2022). Validating a model to detect infant crying from naturalistic audio. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01961-x

Nam, Y., & Park, J. W. (2013). Child activity recognition based on cooperative fusion model of a triaxial accelerometer and a barometric pressure sensor. *IEEE Journal of*

Biomedical and Health Informatics, *17*, 420–426.

Patel, P., Shi, Y., Hajiaghajani, F., Biswas, S., & Lee, M.-H. (2019). A novel two-body sensor system to study spontaneous movements in infants during caregiver physical contact. *Infant Behavior and Development*, *57*, 101383. https://doi.org/10.1016/j.infbeh.2019.101383

Perry, L. K., Prince, E. B., Valtierra, A. M., Rivero-Fernandez, C., Ullery, M. A., Katz, L. F., Laursen, B., & Messinger, D. S. (2018). A year in words: The dynamics and consequences of language experiences in an intervention classroom. *PLOS ONE*, *13*(7), e0199893. https://doi.org/10.1371/journal.pone.0199893

Preece, S. J., Goulermas, J. Y., Kenney, L. P. J., & Howard, D. (2009). A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Transactions on Biomedical Engineering*, *56*, 871–879.

Räsänen, O., Seshadri, S., Lavechin, M., Cristia, A., & Casillas, M. (2020). ALICE: An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behavior Research Methods*, *53*(2), 818–835. https://doi.org/10.3758/s13428-020-01460-x

Ren, X., Ding, W., Crouter, S. E., Mu, Y., & Xie, R. (2016). Activity recognition and intensity estimation in youth from accelerometer data aided by machine learning. *Applied Intelligence*, *45*(2), 512–529.

Schneller, M. B., Bentsen, P., Nielsen, G., Brønd, J. C., Ried-Larsen, M., Mygind, E., & Schipperijn, J. (2017). Measuring children's physical activity. *Medicine & Science in Sports & Exercise*, *49*(6), 1261–1269. https://doi.org/10.1249/mss.0000000000001222

Soska, K. C., & Adolph, K. E. (2014). Postural position constrains multimodal object exploration in infants. *Infancy*, *19*, 138–161. https://doi.org/10.1111/infa.12039

Stewart, T., Narayanan, A., Hedayatrad, L., Neville, J., Mackay, L., & Duncan, S. (2018). A dual-accelerometer system for classifying physical activity in children and adults. *Medicine and Science in Sports and Exercise*, *50*(12), 2595–2602.

Tamis-LeMonda, C. S., Custode, S., Kuchirko, Y., Escobar, K., & Lo, T. (2018). Routine language: Speech directed to infants during home activities. *Child Development*, *90*(6), 2135–2152. https://doi.org/10.1111/cdev.13089

Thurman, S. L., & Corbetta, D. (2017). Spatial exploration and changes in infant-mother dyads around transitions in infant locomotion. *Developmental Psychology*, *53*, 1207–1221.

Warlaumont, A. S., Sobowale, K., & Fausey, C. M. (2021). Daylong mobile audio recordings reveal multitimescale dynamics in infants' vocal productions and auditory experiences. *Current Directions in Psychological Science*, *31*(1), 12–19. https://doi.org/10.1177/09637214211058166

Wass, S., Phillips, E., Smith, C., Fatimehin, E. O., & Goupil, L. (2022). Vocal communication is tied to interpersonal arousal coupling in caregiver-infant dyads. *eLife*, *11*. https://doi.org/10.7554/elife.77399

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*, 2143–2152.

West, K. L., & Iverson, J. M. (2021). Communication changes when infants begin to walk. *Developmental Science*, *24*(5). https://doi.org/10.1111/desc.13102

Yao, X., Plötz, T., Johnson, M., & Barbaro, K. de. (2019). Automated detection of infant holding using wearable sensing: Implications for developmental science and intervention. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(2), 1–17.

Table 1

Summary statistics for model performance metrics shown separately for group and individual models.

	Group			Individual		
Metric	Median	Mean	SD	Median	Mean	SD
Overall Accuracy	0.894	0.846	0.131	0.933	0.916	0.072
Kappa	0.768	0.746	0.161	0.849	0.821	0.143
Sensitivity	0.856	0.825	0.127	0.847	0.841	0.119
Pos Pred Value	0.826	0.810	0.125	0.928	0.899	0.107

Table 2

Model performance metrics for each body position category, shown separately for group and individual models.

		Group			Individual		
Metric	Position	Median	Mean	SD	Median	Mean	SD
Kappa	Supine	0.907	0.764	0.295	0.983	0.912	0.166
	Prone	0.968	0.860	0.259	0.942	0.841	0.246
	Sitting	0.816	0.702	0.297	0.915	0.887	0.127
	Upright	0.707	0.673	0.281	0.822	0.741	0.236
	Held	0.732	0.726	0.209	0.826	0.727	0.277
Sensitivity	Supine	1.000	0.905	0.180	1.000	0.954	0.129
	Prone	1.000	0.894	0.234	0.974	0.849	0.272
	Sitting	0.910	0.811	0.257	0.964	0.915	0.136
	Upright	0.837	0.730	0.297	0.891	0.786	0.253
	Held	0.852	0.772	0.228	0.773	0.702	0.312
Pos Pred Value	Supine	0.995	0.828	0.292	1.000	0.932	0.134
	Prone	0.987	0.877	0.236	1.000	0.892	0.210
	Sitting	0.896	0.802	0.261	0.972	0.945	0.079
	Upright	0.839	0.739	0.283	0.923	0.825	0.228
	Held	0.852	0.794	0.240	0.943	0.899	0.134

Table 3

Correlations between human-coded and model-predicted body position durations across the entire long delay period. Correlations are provided within each posture and overall, and computed separately using group and individual models with and without outlier participants.

	With Outliers		Without Outliers	
Position	Group	Individual	Group	Individual
Held	0.02	0.04	0.73	0.60
Prone	0.97	0.86	0.97	0.84
Sitting	0.79	0.97	0.91	0.95
Supine	0.88	0.98	0.94	0.97
Upright	0.63	0.83	0.99	0.95
Overall	0.80	0.91	0.95	0.96

Table 4

Correlations between human-coded and model-predicted body position durations using 10-minute bins during the distal comparison. Correlations are provided within each posture and overall, and computed separately using group and individual models with and without outlier participants.

	With Outliers		Without Outliers	
Position	Group	Individual	Group	Individual
Held	0.51	0.46	0.67	0.63
Prone	0.96	0.90	0.96	0.89
Sitting	0.72	0.93	0.89	0.92
Supine	0.76	0.96	0.88	0.93
Upright	0.91	0.93	0.98	0.96
Overall	0.80	0.94	0.92	0.94

Table 5

Summary of age differences in full-day body position for younger (4- to 7-month) and older (11- to 14-month) infants. Values shown are the mean percent of time for each body position averaged across infants in each group. Standard deviations are shown in parentheses. Descriptive statistics are shown separately for group and individual models.

	Group		Individual	
Position	Younger	Older	Younger	Older
Upright	7.7% (9.3)	18.6% (7.4)	10.4% (13.6)	18.7% (8.4)
Sitting	24.9% (11.5)	44.4% (10.1)	18.8% (16.3)	46.9% (13.3)
Prone	14.4% (13.8)	14.4% (6.0)	12.5% (10.2)	16.9% (10.6)
Supine	38.5% (24.0)	14.0% (8.4)	40.2% (31.6)	10.0% (9.5)
Held	12.7% (7.1)	8.5% (5.4)	17.1% (20.8)	7.4% (7.6)

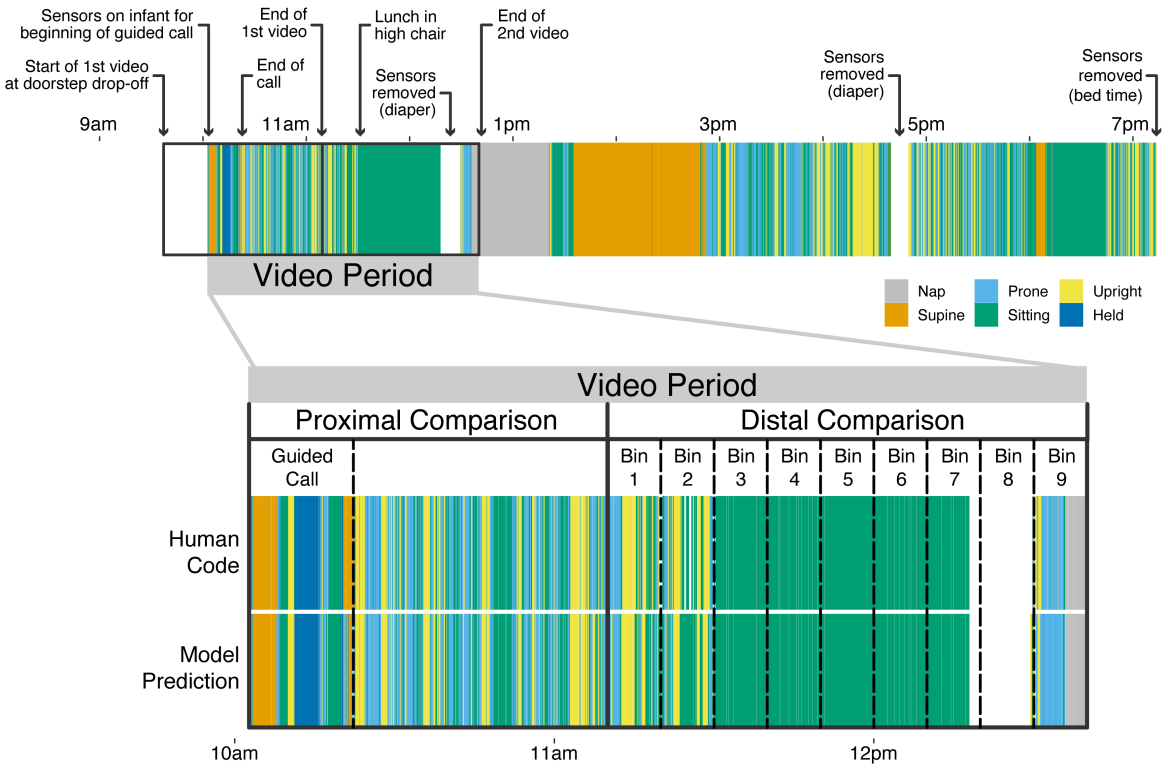

Figure 1. Example Timeline Caption.

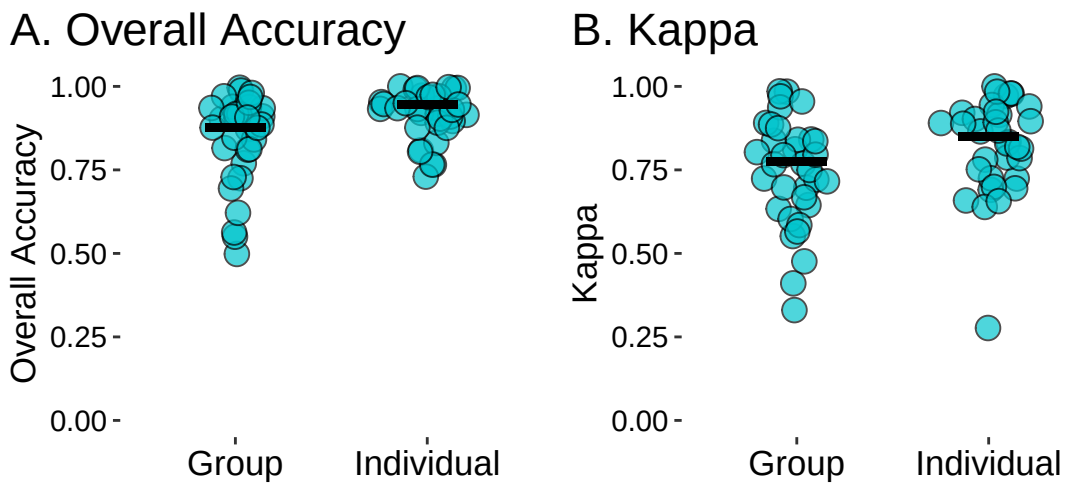

Figure 2. Metrics

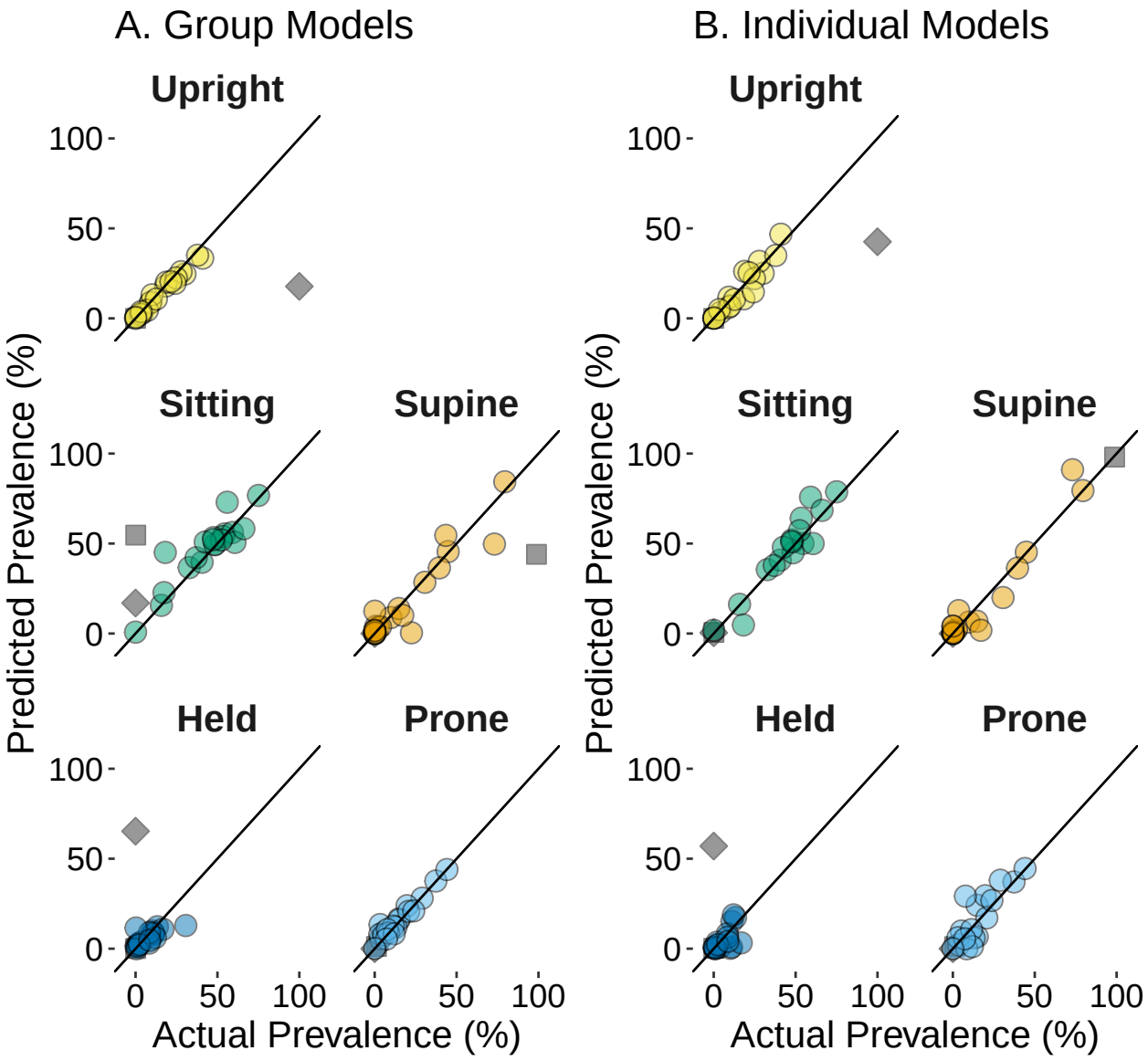

Figure 3. Overall agreement between human-coded body position and model-predicted body position in the distal comparison. Agreement for group models is shown in (A) and agreement for individual models is shown in (B). Plots are shown separately for each body position with a reference line that indicates perfect agreement; each point in a plot represent data for a single participant. The three outlier participants are plotted in dark gray, with a different shape marking each individual.

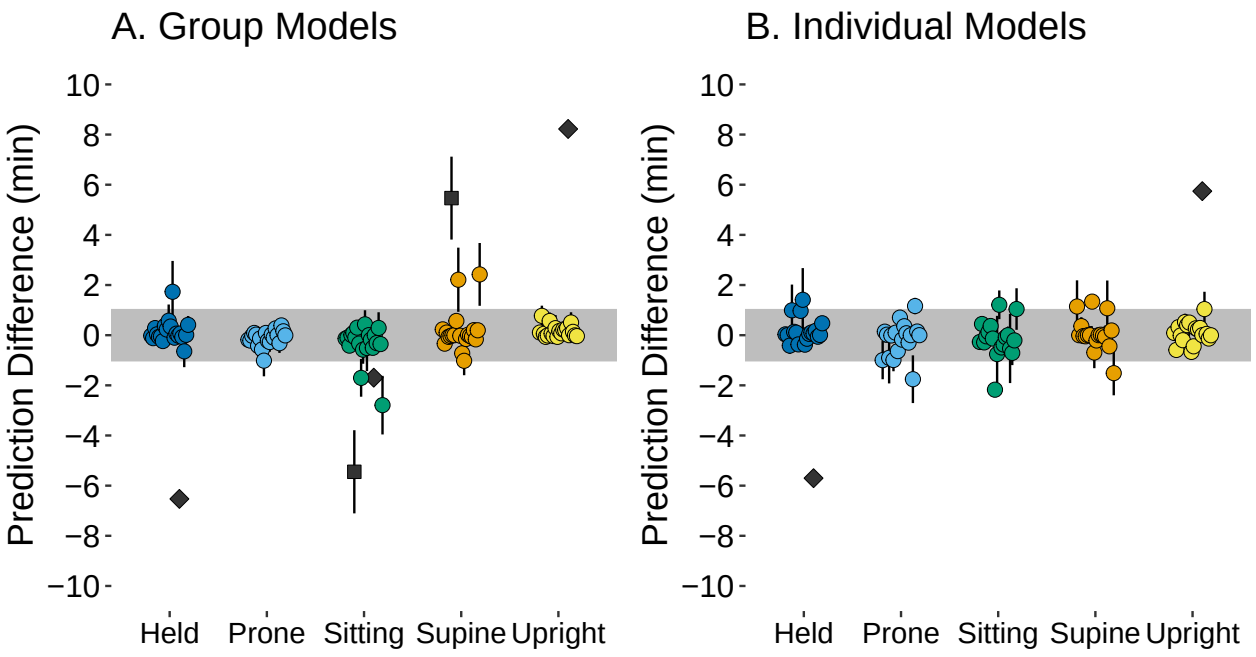

Figure 4. Prediction difference (difference in minutes between human-coded and model-predicted body position) for 10-minute bins in the distal comparison. Each point shows the mean and SE for a single participant for each body position, summarizing the prediction difference for each of their 10-minute bins. Points falling within the gray shaded region indicate that average prediction errors were less than 1 minute. Performance is plotted separately for (A) group models and (B) individual models. The three outlier participants are plotted in dark gray, with a different shape marking each individual.

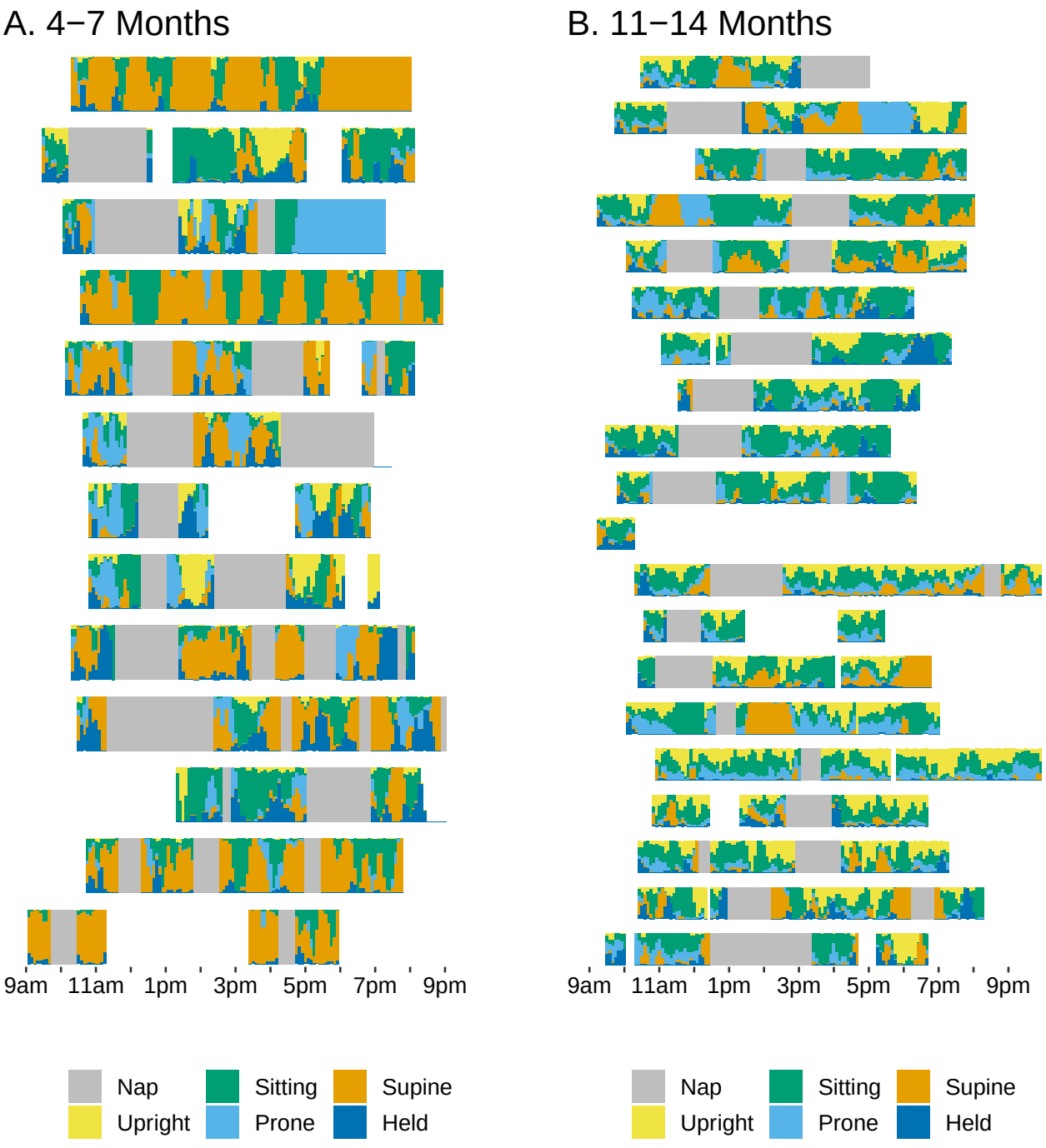

Figure 5. Timelines

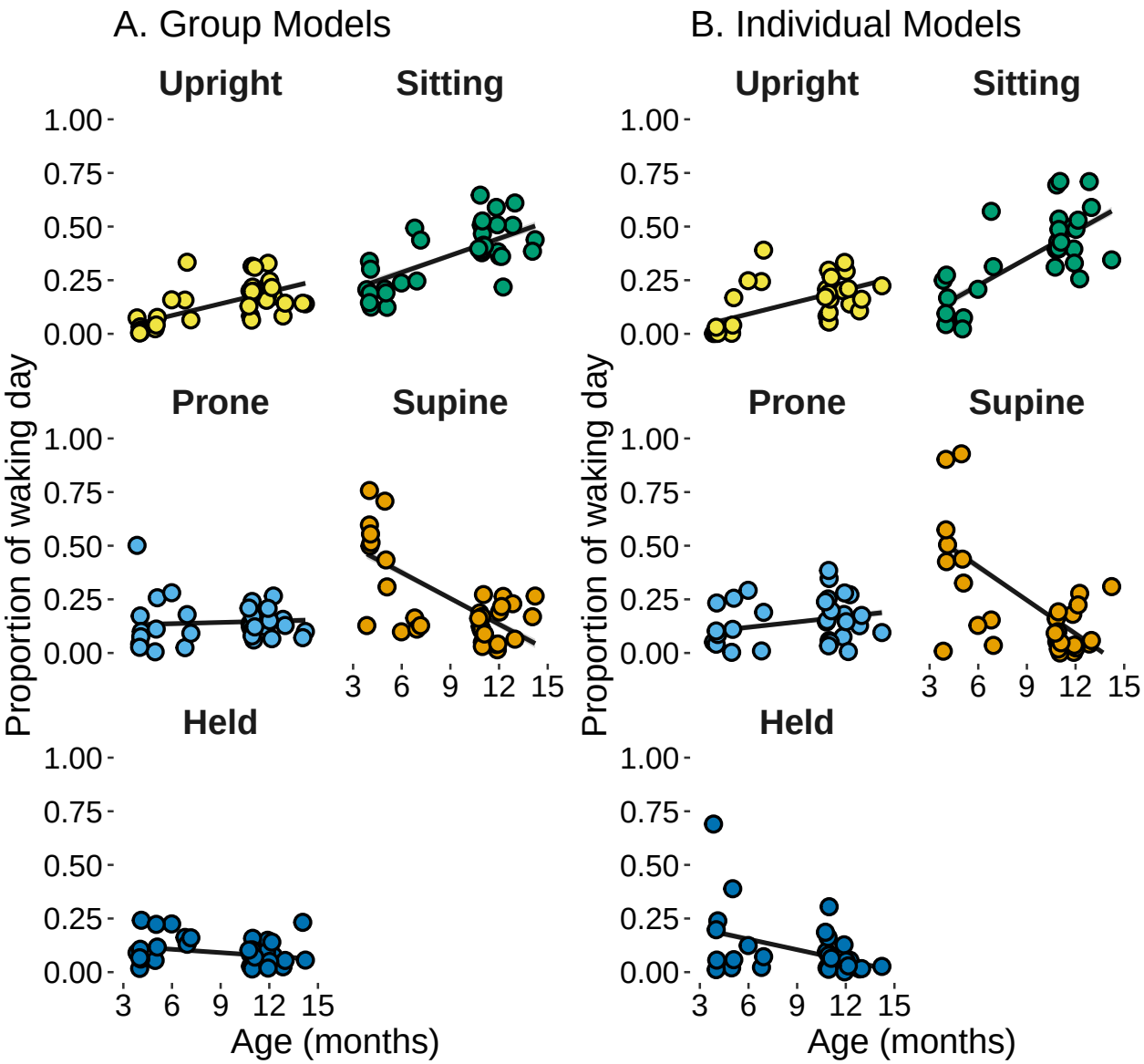

Figure 6. Age trends