Experience: Barriers and Opportunities of Wearables for Eating Research

Mahdi Pedram
Northwestern University
Chicago, Illinois, USA

Glenn Fernandes
Northwestern University
Chicago, Illinois, USA

Christopher Romano
Northwestern University
Chicago, Illinois, USA

Boyang Wei
Northwestern University
Chicago, Illinois, USA

Sougata Sen
BITS Pilani, Goa Campus
Pilani, Goa, India

Josiah Hester
Georgia Institute of Technology
Atalanta, Georgia, USA

Nabil Alshurafa
Northwestern University
Chicago, Illinois, USA

ABSTRACT

Wearable devices have long held the potential to provide real-time objective measures of behavior. However, due to challenges in real-world deployment, these systems are rarely tested rigorously in free-living settings. To reduce this challenge for future researchers, in this paper, we describe our experience developing several generations of a multi-sensor, neck-worn eating-detection system that has been tested with 130 participants across multiple studies in both laboratory and free-living settings. We describe the challenges faced in the development and deployment of the system by (1) presenting example deployment details captured either by the sensing system or the ground truth collector and (2) using structured interviews and surveys with developers and stakeholders of the system, collecting qualitative data on their experience. We performed thematic analysis and provided detailed lessons learned explaining factors that impact the experience of building and deploying such a wearable in a free-living setting, reducing challenges for future researchers. We believe that our experience will help future researchers develop successful mobile health (mHealth) systems that translate into reliable free-living deployments.

CCS CONCEPTS

• **Computer systems organization → Sensor networks**; • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; *User studies.*

KEYWORDS

Case study, wearable sensors, eating detection

ACM Reference Format:
Mahdi Pedram, Glenn Fernandes, Christopher Romano, Boyang Wei, Sougata Sen, Josiah Hester, and Nabil Alshurafa. 2023. Experience: Barriers and Opportunities of Wearables for Eating Research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3544549.3573841

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3573841

1 INTRODUCTION

A major challenge within the field of mobile health (mHealth) is developing wearable sensing systems that are energy efficient, accurate, and unobtrusive. The promise of mHealth is twofold. Wearable sensor systems have utility as commercial lifestyle technologies that empower users by providing accurate and timely information about their health. The appeal of estimating simple health metrics, such as steps, heart rate, and blood pressure, to consumers is evident in the success of commercial smartwatches that allow users to monitor these data themselves. mHealth technologies offer the potential for higher-level features, such as cataloging behavioral events or predicting health trajectories from long-term behavioral patterns. Wearable systems also have value as clinical instruments that can collect measurements in a greater variety of locations and timescales than current technologies allow. The work of behavioral health clinicians (e.g., clinical dietitians, smoking cessation specialists) is complicated by the unreliability of extant behavioral quantification tools such as self-report and guided recall, both of which suffer inaccuracy due to forgetting, non-adherence, and dishonesty. mHealth wearables offer an alternative to these methods that can provide clinicians with accurate, timely, quantitative reports of their patients' health risk behaviors at minimal cost to the patient and clinician alike.

While researchers have demonstrated the applicability of mHealth systems to myriad health topics—such as eating behavior [2, 6–9, 12, 16, 19, 24, 25, 27, 31, 32], fluid intake [15], sleep quality [10, 13, 17, 18, 21, 23], and mental health [22, 26, 29, 30]—adoption into clinical practice remains to be seen. In our extensive work on the passive capture of eating behaviors, we have found several barriers preventing the translation of these systems from bench to bedside. One such barrier is *sample selection.* The more specific a study's inclusion criteria are to the system's use case, the more challenging it becomes to recruit subjects. mHealth researchers must face this challenge head-on, as testing a system on a sample that does not represent the target population compromises any resulting claim of translational potential. For instance, we have previously explored

how eating research on non-obese samples often fails to generalize to the obese population [5].

Another barrier is *transdisciplinarity*. Creating a wearable mHealth system requires medical investigators to devise the research objective; software, electrical, and mechanical engineers to design the firmware, hardware, and enclosure, respectively; computer scientists to build and train the sensor analysis models; research staff to collect the necessary data to train and test the system; and information technologists to manage data storage and sharing. A successful mHealth group must unite this wide array of specializations under a common goal. Some development tasks can be avoided by outsourcing or building on top of commercial technologies, but these shortcuts must be used wisely, as they come at the cost of experimental flexibility, as data collection will be constrained to common measures, analysis, and sensor modalities.

Yet another barrier is *realism*. Although optimal for collecting data and ground truth (i.e., incontrovertible evidence of the presence or absence of a behavioral feature that our device purports to have detected), controlled lab studies often fail to generalize to the real world. Conversely, although higher in external validity, free-living or "in-wild" studies substantially complicate the data and ground truth collection procedures. Both study types have their place in the mHealth development cycle, and researchers must know when and why to apply them. Furthermore, the mere presence of a sensing device in-vivo may alter participant behaviors, reducing realism in even the most naturalistic of settings, as we have explored in previous work [3]. The material and personnel costs of development can be off-putting to researchers and engineers who see little research utility relative to the immense rigor required to validate the system. But such rigor in evaluation is necessary to increase clinicians' and behaviorists' trust in the potential of these devices to be used in the real world. We present our experiences in development of a neck-worn eating detection system to enable future researchers to navigate the process efficiently. We describe the steps taken in developing multiple generations of the system and the experiences gathered during the development and deployment of the system across multiple in-lab and free-living studies for more than half a decade. Our approach in evaluating the multiple generations of the neck-worn device is shown in Figure 1.

2 THEORETICAL CHALLENGES IN BEHAVIOR DETECTION

General considerations: Wearable sensors return a stream of values corresponding to some directly measured quantity from which researchers can infer events. For instance, photoplethysmographic devices measure heart rate and pulse rate monitoring by using a light source and a photodetector. The light source emits light to a tissue and the photodetector measures the reflected light, which is proportional to blood volume variations and is then mapped to pulse and heart rate. In moving away from biometrics and toward detecting behaviors, researchers look for signals uniquely perturbed by the behavior of interest and find a way to encode a detection logic into the device. For instance, pedometers can count steps by recognizing the abrupt spike a step causes in an on-body inertial sensor stream.

Machine understanding of complex behaviors: Not all behaviors are relatively straightforward to capture. To detect more complex behaviors, we take a compositional approach in which the system understands the behavior emerging from multiple, easier-to-sense behavioral or biometric features. In the case of eating, we determine the following component of eating behavior for sensor detection: *bites, chews, swallows, feeding gestures,* and *leans*. For instance, we predict eating if we detect bites, chews, swallows, feeding gestures, and a forward lean angle in close temporal proximity. However, if we detect only swallows, feeding gestures, and a *backward* lean angle, we would instead predict drinking.

Confounding: Detection systems return false positives when receiving input that closely resembles but is not caused by the behavior of interest. For example, intentionally or accidentally, pedometers can be fooled into returning a false positive by shaking the sensor unit. In the case of eating, we anticipated that other hand-to-mouth gestures would likely confound feeding gesture detection, such as smoking, non-food ingestion (e.g., pills), and answering a phone call. To combat confounding, compositionality is helpful, as both multimodal sensing and compositional definitions of behavior increase the robustness (i.e., resilience to confounding) of classification systems. Confounding behaviors can be identified by reasoning a priori, deploying the system with real users, and examining any false positives.

Body variability: Wearable sensors measure the body, which is highly variable between individuals. Body shape impacts sensor orientation. Other body features can also cause problems, such as large beards that limit the use of neck-worn optical sensors. Again, researchers can anticipate body variability (i.e., induced sensor failures) and design around them, and identify system-breaking body features by deploying the system with real users and examining any resulting unintelligible data streams. This challenge creates a fork in the road: researchers must decide whether to design a one-size-fits-all solution or to design modular form factors that can be configured to match a given body type.

3 METHODS

We interviewed 20 team members, including investigators, project managers, hardware engineers, clinicians, data analysts, and study coordinators, to identify the critical points in the conducted research related to the development of the neck-worn eating detection system. About 60% of the interview participants worked on the eating monitoring systems for over a year, and 45% of the participants were familiar with the system design. Qualitative data were analyzed using thematic analysis [11]. First, two authors read all transcripts during an open coding period and independently generated code lists. Then, they met to create an initial codebook informed by our research questions. The codebook comprised lower-level codes sorted into themes. We further analyze responses from each team according to their role (e.g., hardware engineers). Overall we conducted two controlled in-lab studies (Studies 1 and 2), and two free-living (in-wild) naturalistic studies (Studies 3 and 4). We provide details about the evolution while describing each study in this section. Table 1 presents the overview of the studies we conducted for eating detection.

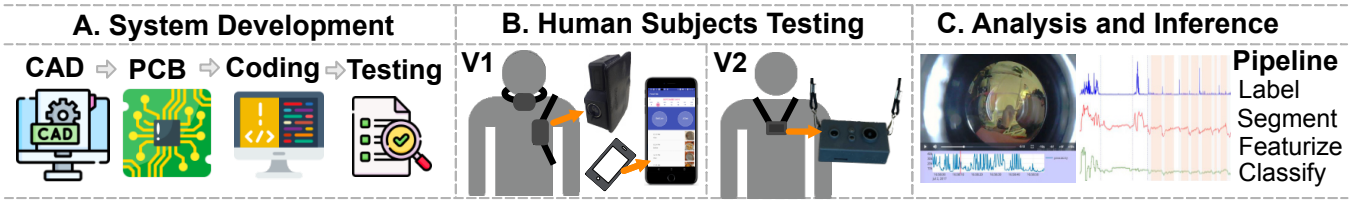

Figure 1: Deployment details of our necklace for automatic dietary monitoring in four studies. We present (A) multiple hardware generations; (B) ground truth collection devices (self-report or video); in our final iteration, the hardware and ground truth device merged into one; and (C) the pipeline for inferring the data and verifying against the ground truth video. Eventually, people will wear only the necklace, and the system will support real-time interventions and feedback to clinicians. CAD, computer-aided design; PCB, printed circuit board.

Study	Study Type	Participants (obese), n #	Sensing Modalities	Ground Truth	Data (hrs)	Target	Result
1 (Alshurafa et al. [4])	In-lab	20	Piezo	Mobile App	3	Swallow	87.0%
2 (Kalantarian et al. [20])	In-lab	30	Piezo, Accelerometer	Mobile App	5	Swallow	86.4%
3 (Zhang et al. [33])	In-wild	20 (10)	Proximity, Ambient, IMU	Wearable Camera	470	Eating	77.1%
4	In-wild	60 (60)	Proximity, Ambient, IMU	Wearable Camera, Mobile App	5600	Eating	TBD

Table 1: Studies included in the analysis.

3.1 Study 1: Exploring the Possibility of Eating Detection in Lab

We explored various combinations of sensor modalities and behavioral proxies with a neck-based approach. To detect swallows, piezoelectric sensors, which record vibrations from the electrical charge produced when the sensor is deformed, were embedded in a necklace (to ensure the material fit snugly to the neck) and used to record eating episodes of in-lab participants. By training classification algorithms on the resulting voltages and inertial data streams, we were able to detect whether a person had swallowed a solid (F = 0.864) or liquid (F = 0.837) with relatively high accuracy. By applying spectral analysis, we were able to determine whether a consumed beverage had been hot or cold (F = 0.9) and we could detect several types of solid foods (F = 0.8) via spectral analysis of their swallow patterns.

3.2 Study 2: Generalization of the Eating-Detection System

Through Study 1, we demonstrated the feasibility of detecting swallows using a single sensing modality. To further improve the reliability of the system when deployed in a free-living environment, we reduced false positives by detecting undesirable movements using an accelerometer. Therefore, to avoid false positives, we added an accelerometer mounted to the left of the piezoelectric sensor, detecting extraneous motions and preventing false nutritive swallows' recordings. Another concern was ensuring the proper sensor placement on the neck. We ensured the proper placement of the accelerometer to detect confounding factors such as head movements and throat muscle movements while speaking. Study 2 included the development and testing (in the laboratory) of this improved system to ensure reliable data collection in free-living environments in the future. Using data collected during the study, we extracted features and developed a sensor fusion classification model to detect swallows. Despite achieving promising results, as seen in Table 1, several issues still hindered the device's deployment in a free-living setting. The first issue was that the neckband form factor required the neckband to be sufficiently tight to enable good contact of the sensor with the skin. This proved to be uncomfortable for the user during the in-lab sessions. The second issue was that the neckband needed to be placed precisely within an error margin of 0.8 cm, to ensure accurate detection of swallows. The third issue was that every time participants took the device off and put it on again, the quality of the signal changed substantially as its position changed. These issues prevented us from deploying the necklace in free-living settings, and we needed to address these issues in the next system design.

3.3 Study 3: Longitudinal Free-Living Study

We also explored neck-based chewing detection and found two critical sensor-proxy pairings (see Table 1, Study 3). First, we aimed an IR proximity sensor at the lower jaw to encode vertical jaw movements. Measured continuously, we found chewing produced a highly periodic waveform, as 'chew-cycles' are evenly spaced across time, and the jaw-proximity signal (which roughly resembles a sine wave) was consistent in amplitude. The periodicity of chewing proved—and has continued to prove—highly useful in the automatic differentiation of chewing from other activities that involve repetitive jaw movement (e.g., talking). Our second proxy was the forward lean angle provided by the gyroscope sensor of the necklace's IMU. In previous studies, participants generally leaned forward while eating and backward while drinking. Applying a forward lean threshold to the detection logic of the necklace system allowed us to predict chewing with even greater confidence. Having established plausible wrist- and neck-based approaches to eating detection, we endeavored to apply our system in free-living settings. The principal challenge was establishing ground truth. Ground truth in the wild could theoretically come from the participants themselves, but this would return us to methodological

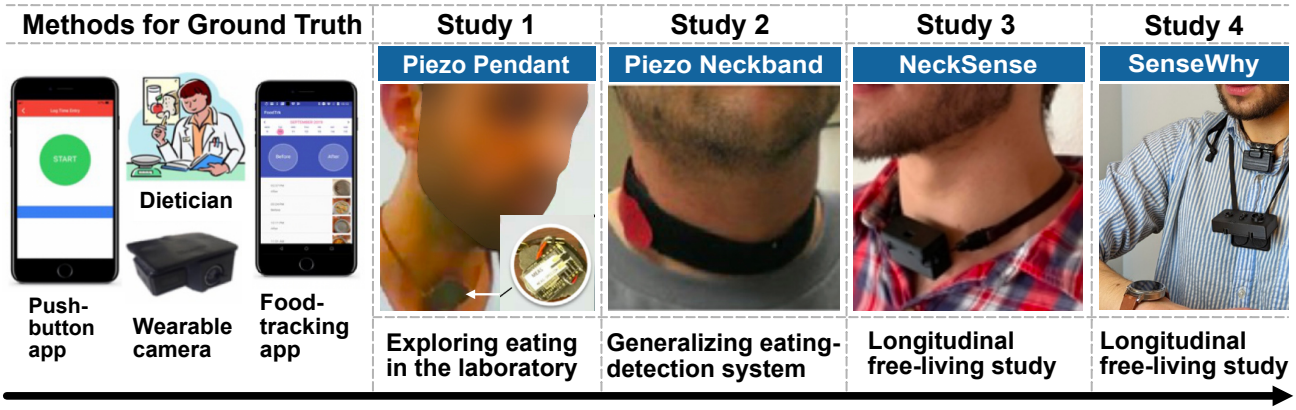

Figure 2: Evolution of the necklace hardware and ground truth systems (in Studies 1 to 4) and timeline of the studies. The studies and results caused us to rethink the necklace hardware and sensor, the mechanical design, and the ground truth device to enable longer-term free-living evaluation. This culminated in the final system, with hardware and ground truth system merged into one.

square one, in which we are relying on the very unreliable measurements that we are trying to circumvent. Instead, we developed a wearable camera to be worn concurrently with one or more of our eating-detection devices and can provide video evidence of our eating-detection system's successes, failures, and confusions. We used a commercial camera (Qcam QSD-722) to capture ground truth. The camera was placed on the shoulder with the lens facing the participant's dominant hand to minimize the privacy concerns of bystanders. In addition, participants had the option of deleting video segments they did not want to share.

Following the pilot study, we validated the improved necklace through a two-day free-living study. We improved the battery life of the necklace to address challenges associated with deploying the system in a free-living setting that we faced in past studies.

We then evaluated the performance of the system on 117 meals collected across both studies. As a result, we achieved an average F1 score of 76.2% for eating detection, which showed an 8% improvement over using only proximity sensor data. However, we faced problems such as synchronization difficulties, privacy concerns regarding the ground truth camera, and user acceptability of the necklace. For example, bystanders to the participant's side were noticeable in the video footage. In addition, the audio recording was another major concern for the participants. These three issues collectively rendered 67.3% of the collected data unusable.

3.4 Study 4: SenseWhy Study

Lessons from Study 3 informed another system redesign. The necklace's neckband was loosened to improve comfort by allowing the necklace to rest at the base of the neck, and we added a frictional hinge to the enclosure that allowed us to change the orientation of the necklace as needed. The camera was moved to the chest with the sensor array facing upward toward the face to minimize privacy concerns, and a thermal module (MLX90640) was added to the array. We conducted the SenseWhy study to increase the generalizability of the method and explore other aspects, such as measuring calorie intake. First, we needed to address the challenges of the previous study, including privacy concerns, synchronization, and the wearability of the necklace. We designed the infrared activity oriented device (IR-AOD) for visual confirmation of activity. The IR-AOD is a wearable thermal and RGB camera developed to maximize information collection while minimizing user discomfort (both physical and psychologic) and privacy risks. A chief motivation for using thermal imaging as a second modality is that it provides information that augments and complements RGB camera data. We also use the thermal sensor to separate wearer pixels from background pixels and mask or obfuscate the background pixels to remove bystanders—a major concern of wearers. Therefore, we conducted another study to explore the user's privacy concerns while using wearable cameras [2]. As a result, sensors are directed toward the user's face and upper torso (Figure 2, right-most panel) for increased user privacy as determined in our "I can't be myself" paper. The device also has night vision capability based on an IR emitter. The custom-designed camera allowed us to confirm 98% of the eating episodes recorded by the wearable camera through visual confirmation.

Another consequence of participant unfamiliarity with wearable sensor devices is that, in some cases, device misuse or malfunction goes unnoticed by the participant. If the study team cannot recognize such instances, massive portions of study data can be lost. To address this, we implemented a 'heartbeat' system by which each device, upon being turned on, transmitted a timestamp through the study phone to the research server. The study team checked these messages and contacted participants to investigate if one or more heartbeats were missing.

4 LESSONS BY ROLE

In this section, we draw on our experience over the four studies and hardware iterations to extract the lessons learned and present the challenges faced by each team, including engineers and research coordinators.

4.1 Hardware Engineers

Device development begins with designing the printed circuit board (PCB), ensuring the physical size and shape of the device, as well as positioning of ports on the device, are ergonomic. The design process involves a trade-off between the microcontroller's computing power, the battery capacity, and the device's physical size. For example, facilitating complex data processing on the device requires a microcontroller with a higher clock frequency. However, a higher clock frequency will increase power consumption, implying a larger battery and device sizes. To collect reliable data throughout the day, striking a balance between the microcontroller, battery, and case design is critical. This requires rigorous testing on the hardware and firmware front, and endurance and interaction of different components in the wild.

Hardware test: Hardware tests consisted of conductivity tests and power consumption tests. We used oscilloscopes, multimeters, laser-based thermal camera systems, and power supplies to check all the power pins and data pins were correctly connected. Before any system-level test, avoid unintended surges that could potentially damage the device or your external programmer. For instance, our custom wearable camera battery charging circuit was unsuitable for use in a wearable device because of an undesirable rise in temperature while charging. Therefore, we reduced the charging rate to address this problem. Through this scenario, we learned that the system's temperature needs to be checked in different states with a thermal camera for temperature anomalies to ensure reliable and safe operation.

Firmware test: Firmware testing was conducted by flashing a test version of the firmware to ensure all onboard systems were functioning as expected. The testbench included microcontroller unit initialization, peripheral test, sample data collection, and a time-synchronization test. An error code should handle each issue for accelerated debugging in a non-ideal behavior in any of the testbench components. The testbench ensured all components were working as expected. For example, the thermal IR sensor did not function correctly during our test. We identified the cause as the thermal sensor requiring a steady power supply. As a result, we used a dedicated 3.3V regulator for the sensor to resolve the problem.

Endurance test: Endurance testing included charging a device from zero to full and then monitoring its functionality throughout a battery life cycle. During the test, we monitored the battery level, time drift of the real-time clock (RTC), and data acquisition rate. If an abnormality was observed in a component, it would undergo checks and, if necessary, replaced. This guaranteed reliable data collection for the researcher, and optimal interaction performance for the user, when the device was deployed in a free-living setting. For example, while running the endurance test, we noticed a higher battery drain rate in some of the devices. On further investigation, we identified the cause as the night vision circuit, which unexpectedly remained on, despite the IR LED being set to turn on, when the light intensity in the environment dropped below a certain threshold. The continuous 40mA draw of the night vision circuit resulted in reduced battery life for some devices. To recover the expected functionality, we replaced the faulty transistors. This enabled the auto-activation of the IR sensor only when the environment was dark.

Simulating and testing extreme circumstances: Endurance testing ensures devices are ready to be deployed; however, we still needed to be careful about deployment as users might not follow the user guide. We needed to ensure the device could operate reliably in unforeseen scenarios while deployed. We designed tests to simulate situations where the device could be misused, exposed to vibrations or physical shock, or could fail due to power issues. For example, to simulate an instance where most of the above could happen, we dropped a device from varying heights to test how well various systems were functioning.

4.2 Software and Machine Learning Engineers

The data collection process comprised: developing a protocol, instructing participants on device usage, deploying the devices, and collecting feedback from participants. Despite training participants during in-lab sessions, there still existed challenges at each stage of deployment.

Design: The quality at which data were acquired affected the battery life of the device. For example, acquiring data at higher rates resulted in faster battery drain. Tuning the device parameters can help strike a balance between this trade-off. During the SenseWhy study, we acquired data with a resolution of 360 pixels at a rate of 5 Hz, to ensure all-day battery life. However, the low resolution was not sufficient for ground truth validation. We learned that conducting a parametric analysis to quantify the trade-off could ensure reliable outcomes. The participants in the SenseWhy study wore multiple devices, which posed another challenge. Each device had its own clock; we needed to ensure that the time offset between the devices were accounted for, to ensure our ability to conduct multimodal analysis after data collection. We created sync events, which ensured our ability to align the data from the sensors after the study. To sync the RGB and thermal camera, we asked the device wearer to block the camera lens with their hand for 10 seconds and repeat the gesture three times. We used similar events to sync the necklace and wristband.

In-lab session: The ideal passive-sensing system requires no input nor specialized knowledge on the wearer's part. However, in early design iterations, it is often impractical to achieve this ideal, so some degree of active participation is required of the user. In such cases, participant understanding is critical to the success of high-quality data collection. During the SenseWhy study, although each participant was provided with detailed instructions regarding device use, such as wearing and powering the device on/off, we observed unexpected behaviors in free-living settings. For instance, on the first day of the in-lab session, participants were instructed to perform a sequence of events (covering the necklace with their hand, clapping their hands, and drinking from a bottle) to allow us to synchronize the devices. However, we frequently observed that participants covered the camera instead of the necklace. Therefore, it was essential to have detailed instructions, preferably with effective visualizations, that participants were able to review after training. In addition, during the training session, we learned that we must carefully observe these behaviors and document potentially incorrect gestures.

In-wild deployment: In a less-constrained in-wild environment without supervision, participants sometimes fail to follow the

standard instructions. We believe it is crucial to consider that people tend to repeat what they have learned from the in-lab session. We observed that some participants followed the incorrect instructions immediately after the in-lab session and made the same errors for the entire study. For instance, some participants did not perform the synchronization events and forgot to turn off the device after each day. These patterns were consistent for participants over the entire 2-week study. We learned that following up with the participants when shifting from in-lab to in-wild environments, especially the first few days, was crucial to help develop standardized behaviors. A reasonable way to do this would be to examine the video footage and data from all modalities on the first day of in-wild and send feedback to the participants to ensure they follow the standard instructions.

Labeling and annotation: Detecting eating activity from signal streams acquired by different sensors is accomplished through machine learning algorithms, mainly supervised learning techniques. Assembling a training dataset with labels requires a trained annotator. Several studies do this by video recording and manually labeling each frame as either chewing, eating, or other [1, 14]. In laboratory settings, annotating and labeling the signals is manageable. However, in-wild studies are more complicated, especially when a time offset exists between the sensor data and video footage. As a result, researchers rely on a wide range of solutions, from manual self-annotation on a smartphone [14, 28] to using a ground truth camera pointed at the participant's dominant arm or up at their chin [2]. At the very start of Study 3, our labeler only used video to identify chewing. This resulted in inaccurate label work even though we had two labelers to cross-validate the annotations. Therefore, we used multiple methods for labeling 'chewing' that were based on video and audio signals. Afterward, we verified our label with our FoodTrek App, which participants used to report and record their meals. This delivered higher-quality labels and made it less likely to miss any meal.

4.3 Mechanical Engineers and Designers

When designing the case and strip for the neck, we found that we needed to make it both look good and feel good. When asking participants their opinion about the device, they may focus more on the look, but what makes participants decide whether to wear it for long is comfortability. Therefore, we carefully chose silicon stripe and designed a natural curve to fit the user's neck. The study showed that most users ignored the device's presence after a long time, and the longer they wore the device, the more natural their behaviors became.

4.4 Research Coordinators

Research staff who administer study procedures are responsible for three key project elements.

Subject recruitment: Though it is best practice to avoid convenience sampling whenever possible, some population constraints are inherent to free-living research or are imposed by elements of study design. Our study designs require in-person laboratory visits to provide devices and teaching to participants, restricting our sample by proximity to the laboratory. Our location in Chicago, IL, minimizes the impact of this restriction. The population of the Chicago Metropolitan Area is 9.6 million, from which we crudely estimate a population with obesity of 3.1 million by applying the Illinois obesity rate of 32.5%. Groups with smaller target populations in their vicinity may need to eliminate location restrictions to achieve a representative sample in a reasonable amount of time. This can be done by mailing study equipment to participants, providing teaching via phone or video conference, and designing studies that require no in-lab data collection. The cost of these measures should be compared against the cost of running the study at its current rate; that is, spending on remote study administration should not exceed savings from the reduction of facilities and personnel costs.

Technology teaching: Participants should understand the function and purpose of the wearable, which is often difficult to convey when dealing with early-stage and prototype devices. A helpful metaphor is to describe machine learning in the language of human learning (e.g., "We want this device to know when the user is eating. So, we show the device what eating looks like.") Teaching participants for in-wild studies should err on the side of overkill, as a participant who understands the device well is less likely to experience technical issues. These participants should practice every step of using the device (equipping, turning on/off, charging, etc.) multiple times before they begin an in-wild collection.

In-field monitoring: The free-living setting is where device failures are most common and damaging to the integrity of the data. Participants can be monitored with device 'heartbeat' systems that remotely and intermittently notify the study team of the device's operational status. Research staff should receive education on troubleshooting procedures and be prepared to support participants who experience device failures. The start time and duration of each device failure should be documented so that resulting gaps in the dataset can be explained and considered in the analysis.

Inference: A wearable camera captures behavior that can be used as the ground truth. Trained labelers watch the video and tag (or label) sensor signals. However, we learned that obtaining objective measures of authentic behavior is complex. How a camera is mounted, what the field of view captures, and how the wearer reacts to their private moments being captured on video add to privacy concerns. In addition, the wearers raised concern for the privacy of bystanders. Mitigating these issues in ways that guarantee an acceptable level of privacy and utility to label the signals from the wild requires novel approaches. We explored and developed new strategies and solutions to this challenge.

Hawthorne effect: Some participants reported becoming more conscious of their in-wild eating due to the presence of the devices. However, other participants reported forgetting they were wearing the devices and noticed no increase in eating-based conscientiousness. This suggests that an in-wild study is not automatically immune to observation bias but that ergonomic device design can foster the collection of natural behavior through comfort of device.

5 INSIGHTS FROM STUDY TEAM

Through survey responses, we generated a codebook and then identified themes (see Table 2). From the coding process, we identified five major themes that highlight the challenges faced throughout

Theme	Code (n)
One sensing modality is insufficient	Sensor placement (5), Multiple modalities (5), Audio (3), Video (2), Other modalities(4)
Do not cut corners with hardware development	Form Factor(9), Industrial Grade(5), Existing Hardware(4)
You cannot out-engineer the wild	In-lab testing (6), Real-world (6), Customer service (5), Codesigning (2)
Just because you think you will see it, does not mean you will	Data quality (7), Time-stamping (6), Personalized model (3), Ground truth validity (3)
We should respect other cultures	Organization (9), Labor intensive (5), Cultural norms (4), Publication (4)

Table 2: Themes and codes from responses to open-ended questions. The number in parentheses indicates the number of participants

our multiple food detection and monitoring studies, providing potential guidelines for future researchers who wish to build food monitoring systems.

A1: ***A single sensing modality might not be sufficient for such a complex problem.*** Eating detection researchers continue to focus on searching for a single sensor that captures fine-grained eating-related information. However, eating involves a combination of independent sequences of action units, such as a hand-to-mouth gesture followed by chewing and swallowing, with potential pauses in between and some actions missing at times. For instance, when consuming yogurt, an individual may forgo chewing but not if it has pieces of granola. Another individual may consume a sandwich and forgo gesturing after the first bite (i.e., keeping the hands near the mouth). Moreover, even if such a sensor existed and could detect all eating episodes, the confounders discussed point to the added challenge of preventing false positives.

A2: ***It is not worth compromising hardware development.*** Due to the nature of our research world and our "publish or perish" culture, we sometimes cut corners that may be great for the short term but detrimental in the long run. For example, it took multiple years for our team to publish our paper on Study 2, and reviewers even felt after Study 2 that we needed to redo the study to address a lot of the device problems, which resulted in our deployment of an entirely new Study 3, with a new device. Many survey participants also reported needing to use expert(s) for industrial design.

A3: ***You cannot out-engineer the wild*** There was a disagreement between the survey participants on the need to spend more time on in-lab studies before going out into the wild. One survey respondent felt the problems could have been mitigated through in-lab studies. However, most thought spending time on in-lab studies was not fruitful since participants in the field could be more decisive. On the other hand, there was strong agreement on the need to provide proper technical support. In all studies, including Study 4, we needed to generate standard operating procedures to ensure the participants received immediate attention when a problem was detected. One survey respondent reported, "Planning pre-deployment carefully is the most important thing; it is better to delay the study by a week because of a problem or needed fix than let those problems go into the study. It is better to lose a week without participants involved than lose a week of participant data because of a problem."

A4: ***Just because you think you will see it, does not mean you will.*** Many reported problems with data quality, time synchronization, and reliability of collected data through live visualization of the data. But there was a concern with ground truth validity and its ability to capture the intended behavior. For example, the wearable camera could often visually confirm eating behavior reliability; however, detecting chewing from the camera or swallowing at times proved challenging. It seemed trivial when tested in the laboratory. However, this proved challenging given the varying scenarios of participants eating in the wild. This is why ensuring multiple methods of validating or confirming a behavior through sensors or logs is essential.

A5: ***We should respect other cultures.*** Culture is often different between the health sciences and engineers or computer scientists. But even within a discipline, cultural differences might vary between a hardware engineer, a graduate student, an undergraduate student, and a principle investigator, which is why a team must develop laboratory norms. It is also critical for investigators to face their fears. For example, investigators often want to hear the good news when a study starts, and unfortunately, they regret this at the end of the study because so much could have been done early on to mitigate problems. Faculty in the health sciences are used to hiring the right and experienced people for the job. However, in engineering and computer science, we often expect students to be able to wear multiple hats, running the study and building ML models and sensors. This often compromises the quality of the results, infuses bias, and creates an unhealthy culture where people cannot express their opinions. Survey participants felt organization and labor should be given to the person with the proper skill set, even if the grant or project needed to be extended. This requires a culture shift and greater resources for engineering faculty to hire study coordinators for their in-lab and in-wild experiments.

6 CONCLUSIONS

Mobile computing devices have been a boon to healthcare delivery, especially enabling preventive medicine applications like eating monitoring. This paper presents a longitudinal view of experiences and understanding of the development and deployment of eating-detection devices in the form of a necklace on 130 participants in two in-lab and two free-living studies. Many challenges were raised concerning device robustness for long-term use, ground truth collection difficulty, device failure, recruitment of varied populations, and wearability. From these experiences, our recommendations for future studies are to (1) recruit broadly rather than just with students, (2) bring the study from a laboratory to a free-living setting to understand how our technology interacts with naturalistic human behaviors, (3) understand that iterative co-design of functionality and usability is time consuming but worthwhile as participants are more likely to wear; and (4) explore new ground truth methods that are less labor intense for annotators and more friendly to users with regard to physical burden and privacy concerns.

REFERENCES

[1] Rawan Alharbi, Angela Pfammatter, Bonnie Spring, and Nabil Alshurafa. 2017. Willsense: adherence barriers for passive sensing systems that track eating behavior. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2329–2336.

[2] Rawan Alharbi, Tammy Stump, Nilofar Vafaie, Angela Pfammatter, Bonnie Spring, and Nabil Alshurafa. 2018. I Can't Be Myself: Effects of Wearable Cameras on the Capture of Authentic Behavior in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 90.

[3] Rawan Alharbi, Tammy Stump, Nilofar Vafaie, Angela Pfammatter, Bonnie Spring, and Nabil Alshurafa. 2018. I Can't Be Myself: Effects of Wearable Cameras on the Capture of Authentic Behavior in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (9 2018), 1–40. https://doi.org/10.1145/3264900

[4] Nabil Alshurafa, Haik Kalantarian, Mohammad Pourhomayoun, Shruti Sarin, Jason J Liu, and Majid Sarrafzadeh. 2014. Non-invasive monitoring of eating behavior using spectrogram analysis in a wearable necklace. In *2014 IEEE healthcare innovation conference (HIC)*. IEEE, 71–74.

[5] Nabil Alshurafa, Shibo Zhang, Christopher Romano, Hui Zhang, Angela Fidler Pfammatter, and Annie W. Lin. 2021. Association of number of bites and eating speed with energy intake: Wearable technology results under free-living conditions. *Appetite* 167 (2021), 105653. https://doi.org/10.1016/j.appet.2021.105653

[6] Yicheng Bai, Wenyan Jia, Zhi-Hong Mao, and Mingui Sun. 2014. Automatic eating detection using a proximity sensor. In *Bioengineering Conference (NEBEC), 2014 40th Annual Northeast*. IEEE, 1–2.

[7] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. 2017. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 37.

[8] Abdelkareem Bedri, Yuchen Liang, Sudershan Boovaraghavan, Geoff Kaufman, and Mayank Goel. 2022. FitNibble: A Field Study to Evaluate the Utility and Usability of Automatic Diet Monitoring in Food Journaling Using an Eyeglasses-based Wearable. In *27th International Conference on Intelligent User Interfaces*. 79–92.

[9] Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald Peterson, Kofi Odame, Kelly Caine, et al. 2018. Auracle: Detecting Eating Episodes with an Ear-mounted Sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 92.

[10] Marko Borazio and Kristof Van Laerhoven. 2012. Combining wearable and environmental sensing into an unobtrusive tool for long-term sleep studies. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. ACM, 71–80.

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[12] Yang Chen and Ching Chiuan Yen. 2022. SLNOM: Exploring the sound of mastication as a behavioral change strategy for rapid eating regulation. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–6.

[13] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and Andrew T Campbell. 2013. Unobtrusive sleep monitoring using smartphones. In *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and ..., 145–152.

[14] Keum San Chun, Sarnab Bhattacharya, and Edison Thomaz. 2018. Detecting Eating Episodes by Tracking Jawbone Movements with a Non-Contact Wearable Sensor. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 4 (March 2018), 21 pages. https://doi.org/10.1145/3191736

[15] Keum San Chun, Ashley B Sanders, Rebecca Adaimi, Necole Streeper, David E Conroy, and Edison Thomaz. 2019. Towards a generalizable method for detecting fluid intake with wrist-mounted sensors and adaptive segmentation. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 80–85.

[16] Yujie Dong, Jenna Scisco, Mike Wilson, Eric Muth, and Adam Hoover. 2014. Detecting periods of eating during free-living by tracking wrist motion. *IEEE journal of biomedical and health informatics* 18, 4 (2014), 1253–1260.

[17] Peter Hillyard, Anh Luong, Alemayehu Solomon Abrar, Neal Patwari, Krishna Sundar, Robert Farney, Jason Burch, Christina Porucznik, and Sarah Hatch Pollard. 2018. Experience: Cross-Technology Radio Respiratory Monitoring Performance Study. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 487–496.

[18] Ryoichi Ishida, Yoshiharu Yonezawa, Hiromichi Maki, Hidekuni Ogawa, Ishio Ninomiya, Kouji Sada, Shingo Hamada, Allen W Hahn, and W Morton Caldwell. 2005. A wearable, mobile phone-based respiration monitoring system for sleep apnea syndrome detection. *Biomedical sciences instrumentation* 41 (2005), 289–293.

[19] Azusa Kadomura, Cheng-Yuan Li, Yen-Chang Chen, Koji Tsukada, Itiro Siio, and Hao-hua Chu. 2013. Sensing fork: Eating behavior detection utensil and mobile persuasive game. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 1551–1556.

[20] Haik Kalantarian, Nabil Alshurafa, Tuan Le, and Majid Sarrafzadeh. 2015. Monitoring eating habits using a piezoelectric sensor-based necklace. *Computers in biology and medicine* 58 (2015), 46–55.

[21] Anh Nguyen, Raghda Alqurashi, Zohreh Raghebi, Farnoush Banaei-kashani, Ann C. Halbower, and Tam Vu. 2016. A Lightweight and Inexpensive In-ear Sensing System For Automatic Whole-night Sleep Stage Monitoring. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM* (Stanford, CA, USA) *(SenSys '16)*. ACM, New York, NY, USA, 230–244. https://doi.org/10.1145/2994551.2994562

[22] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. 2011. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 385–394.

[23] Mahsan Rofouei, Mike Sinclair, Ray Bittner, Tom Blank, Nick Saw, Gerald DeJean, and Jeff Heffron. 2011. A non-invasive wearable neck-cuff system for real-time sleep monitoring. In *2011 international conference on body sensor networks*. IEEE, 156–161.

[24] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. 2010. Collecting complex activity datasets in highly rich networked sensor environments. In *Networked Sensing Systems (INSS), 2010 Seventh International Conference on*. IEEE, 233–240.

[25] Jaemin Shin, Seungjoo Lee, Taesik Gong, Hyungjun Yoon, Hyunchul Roh, Andrea Bianchi, and Sung-Ju Lee. 2022. MyDJ: Sensing Food Intakes with an Attachable on Your Eyeglass Frame. In *CHI Conference on Human Factors in Computing Systems*. 1–17.

[26] Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. 2017. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 715–724.

[27] Edison Thomaz, Irfan Essa, and Gregory D Abowd. 2015. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1029–1040.

[28] Edison Thomaz, Aman Parnami, Irfan Essa, and Gregory D. Abowd. 2013. Feasibility of Identifying Eating Moments from First-person Images Leveraging Human Computation. In *Proceedings of the 4th International SenseCam & Pervasive Imaging Conference* (San Diego, California, USA) *(SenseCam '13)*. ACM, New York, NY, USA, 26–33. https://doi.org/10.1145/2526667.2526672

[29] Rui Wang, Min SH Aung, Saeed Abdullah, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A Scherer, et al. 2016. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 886–897.

[30] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 3–14.

[31] Xu Ye, Guanling Chen, Yang Gao, Honghao Wang, and Yu Cao. 2016. Assisting food journaling with automatic eating detection. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 3255–3262.

[32] Shibo Zhang, Rawan Alharbi, Matthew Nicholson, and Nabil Alshurafa. 2017. When generalized eating detection machine learning models fail in the field. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM, 613–622.

[33] Shibo Zhang, Yuqi Zhao, Dzung Tri Nguyen, Runsheng Xu, Sougata Sen, Josiah Hester, and Nabil Alshurafa. 2020. Necksense: A multi-sensor necklace for detecting eating activities in free-living conditions. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 4, 2 (2020), 1–26.