scientific reports

Check for updates

OPEN

Validation studies of the FLASH-TV system to passively measure children's TV viewing

Anil Kumar Vadathya[1], Tatyana Garza[2], Uzair Alam[2], Alex Ho[3], Salma M.A. Musaad[2], Alicia Beltran[2], Jennette P. Moreno[2], Tom Baranowski[2], Nimah Haidar[3], Sheryl O. Hughes[2], Jason A. Mendoza[4,5], Ashok Veeraraghavan[1], Joseph Young[1], Akane Sano[1] & Teresia M. O'Connor[2]✉

TV viewing is associated with health risks, but existing measures of TV viewing are imprecise due to relying on self-report. We developed the *F*amily *L*evel *A*ssessment of *S*creen use in the *H*ome (FLASH)-TV, a machine learning pipeline with state-of-the-art computer vision methods to measure children's TV viewing. In three studies, lab pilot ($n=10$), lab validation ($n=30$), and home validation ($n=20$), we tested the validity of FLASH-TV 3.0 in task-based protocols which included video observations of children for 60 min. To establish a gold-standard to compare FLASH-TV output, the videos were labeled by trained staff at 5-second epochs for whenever the child watched TV. For the combined sample with valid data ($n=59$), FLASH-TV 3.0 provided a mean 85% (SD 8%) accuracy, 80% (SD 17%) sensitivity, 86% (SD 8%) specificity, and 0.71 (SD 0.15) kappa, compared to gold-standard. The mean intra-class correlation (ICC) of child's TV viewing durations of FLASH-TV 3.0 to gold-standard was 0.86. Overall, FLASH-TV 3.0 correlated well with the gold standard across a diverse sample of children, but with higher variability among Black children than others. FLASH-TV provides a tool to estimate children's TV viewing and increase the precision of research on TV viewing's impact on children's health.

Keywords Television, Gaze estimation, Screen media, Machine learning, Face detection

Children use multiple forms of screen media, such as TVs, videogames, and mobile devices, for a variety of reasons including entertainment, education, and communication. Studies have linked greater screen media use among children with higher rates of obesity[1], internalizing and externalizing behavior problems[2], worse academic performance[3], and shorter sleep duration[4], among other issues. These studies have raised awareness about children's screen use as a public health concern[5]. A majority of studies investigating the role of screen media use on children's health and developmental outcomes has measured screen use by self- or parent proxy-reports[6]. Such studies often rely on one or a few questions regarding children's typical screen use. Self- or proxy-reported use of screen media tends to be less accurate compared to assessment tools that utilize technology to passively track screen use on a specific screen or device for a defined study period[7]. This has resulted in a call for better methods to measure children's screen use across screen devices[8]. Several groups have been working on more accurate approaches for measuring children's mobile device use[7,9–11], however, tools are also needed to measure children's TV viewing which may have differing effects than use of other screens[3].

The Family Level Assessment of Screen use in the Home (FLASH)-TV 1.0[12] was developed to use machine learning to passively assess children's TV viewing in their natural home setting. FLASH-TV 1.0 faces the audience and captures images in front of a TV by a webcam mounted on or nearby the TV, and sequentially analyzes them through face detection, recognition, and gaze estimation machine learning algorithms. The face detector identifies all faces in the image; face recognition identifies the target child being studied from other faces detected; gaze estimation uses the target child's facial image to categorize whether the child is watching TV; and FLASH-TV aggregates this gaze data in 5-second epochs. FLASH-TV analyzes the video images almost instantaneously in near real time so the images can be deleted right away, after storing the processed output (whether the child of interest is present when the TV is on, and whether the child is gazing or not gazing on

[1]Department of Electrical & Computer Engineering, Rice University, Houston, TX, USA. [2]USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA. [3]Rice University, Houston, TX, USA. [4]Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA. [5]General Pediatrics, Department of Pediatrics, University of Washington, Seattle, WA, USA. ✉email: teresiao@bcm.edu

TV). This enhances participant privacy by not storing identifiable information (images of people or the room). To further enhance privacy and security the stand-alone version of FLASH-TV is not connected to Wi-Fi.

The objective of this study was to refine FLASH-TV 1.0 to FLASH-TV 2.0 to enhance its accuracy and test FLASH-TV 2.0 for accuracy, sensitivity, specificity, and agreement in comparison to gold standard human-labeled video data. For this, we piloted the FLASH-TV 2.0 with 10 participants and started the in-lab validation study ($n=15$). FLASH-TV 2.0 performance was poor on the initial validation sample. We pivoted FLASH-TV 2.0 to 3.0 to improve the performance on all the data collected from our pilot ($n=10$), in-lab ($n=30$), and in-home ($n=20$) studies. Our revised objective was to further refine FLASH-TV 2.0 to FLASH-TV 3.0 and assess its accuracy, sensitivity, specificity, and agreement as compared to gold-standard human-labeled video data collected in the same task-based protocols in our pilot, in-lab, and in-home validation studies. Here, we describe the revisions to the machine learning algorithms for FLASH-TV 2.0 and 3.0. The new versions of FLASH-TV were assessed in three independent studies: a pilot observational lab study, an observational lab validation study, and a validation study in the child's home. A secondary objective was to assess the performance of FLASH-TV among children of different racial and ethnic groups and both sexes to identify if the accuracy of the tools developed were race or sex dependent, which has been shown to be a challenge with many face recognition machine learning approaches[13].

Results

Table 1 provides the demographics of the primary or target children who took part in each of the three studies.

FLASH 2.0

Table 2 demonstrates the performance of FLASH-TV 2.0 on the pilot test sample ($n=10$) and initial half of the lab validation test sample ($n=15$). FLASH-TV 2.0 had reasonable performance in the pilot test with accuracies > 80% and relatively high kappa (0.68). However, FLASH-TV 2.0 did not perform as well among the initial half of the sample from the lab validation with 77.7% accuracy and lower kappa, Intra-Class Correlation (ICC), sensitivity, and specificity. As described in the methods, improvements were made to the algorithm resulting in the new FLASH 3.0.

Ensemble models for FLASH-TV 3.0

Table 3 compares the performance of FLASH-TV 3.0 with and without low-resolution regularization (described in methods) in the initial half of the sample of the lab validation study. The low-resolution regularization, proposed by Xu et al.[14], improved the robustness of our model when there was poor resolution of facial images. The FLASH-TV 3.0-unregularized model had higher specificity, whereas the FLASH-TV 3.0-regularized model had higher sensitivity. When the outputs of these models are linearly combined as described in methods, the

	Overall	Pilot test	Lab	Home
Parent-sibling triads (n)	60	10	30	20
Target Child, n (%)	60 (100)	10 (100)	30 (100)	20 (100)
Age (years), mean (SD)	8.5 (2.1)	8.4 (1.4)	8.3 (2.5)	8.7 (1.9)
Sex (female), n (%)	23 (38)	6 (60)	14 (47)	3 (15)
Race and ethnicity, n (%)				
Non-Hispanic White	14 (23)	2 (20)	8 (27)	4 (20)
Hispanic White	14 (23)	3 (30)	8 (27)	3 (15)
Non-Hispanic Black	13 (22)	0 (0)	6 (20)	7 (25)
Hispanic Black	4 (7)	1 (10)	3 (10)	0 (0)
Asian	7 (12)	2 (20)	3 (10)	2 (10)
Other (mixed or Hispanic other)	8 (13)	2 (20)	2 (6)	4 (20)
Parent, n (%)	60 (100)	10 (100)	30 (100)	20 (100)
Education, n (%)				
High school	4 (7)	1 (10)	2 (7)	1 (5)
Technical school	3 (5)	0 (0)	2 (7)	1 (5)
Some college	7 (11)	2 (20)	3 (10)	2 (10)
College	22 (37)	1 (10)	13 (43)	8 (40)
Graduate school	24 (40)	6 (60)	10 (33)	8 (40)
Income (US $), n (%)				
<30,000	9 (15)	2 (20)	6 (20)	1 (5)
>30,000 to <50,000	11 (18)	1 (10)	5 (17)	5 (25)
>50,000 to <70,000	11 (18)	0 (0)	6 (20)	5 (25)
>70,000	28 (47)	6 (60)	13 (43)	9 (45)
Not known	1 (2)	1 (10)	0 (0)	0 (0)

Table 1. Demographics of participants in the three independent samples.

Dataset	Gold standard TV time, minutes (SD)	FLASH predicted TV time, minutes (SD)	Kappa, PABAK (SD)	ICC	Accuracy (%)	Sensitivity (%)	Specificity (%)
Pilot test ($n=10$)	29.02 (10.78)	29.77 (10.46)	0.68 (0.14)	0.96	83.80 (7.24)	81.71 (19.29)	84.22 (6.85)
Lab validation test, initial sample ($n=15$)*	27.75 (10.36)	25.33 (10.81)	0.54 (0.21)	0.46	77.71 (10.41)	69.34 (22.78)	83.61 (8.50)

Table 2. FLASH-TV 2.0 performance on two independent samples. SD - standard deviation. PABAK - prevalence and bias-adjusted Kappa, compares the gaze and no-gaze of FLASH-TV and gold standard at 5-second epoch, nested within each participant. ICC-intraclass correlation, compares the overall duration of TV watching of FLASH-TV and gold standard per participant. *Here we used the initial $n=15$ from our full validation test set of 30 participants. Comparisons of FLASH-TV and gold-standard do not include video segments labeled as uncertain by the team.

FLASH-TV version	Gold standard TV time, minutes (SD)	FLASH predicted TV time, minutes (SD)	Mean absolute error, minutes (SD)	Kappa, PABAK (SD)	Accuracy (%)	Sensitivity (%)	Specificity (%)
3.0-unregularized	27.75 (10.40)	22.57 (11.69)	7.72 (6.12)	0.65 (0.21)	82.70 (10.71)	69.11 (23.58)	91.20 (6.55)
3.0-regularized	27.75 (10.40)	27.29 (10.76)	5.86 (6.72)	0.61 (0.21)	80.49 (10.69)	76.28 (20.17)	83.74 (10.44)
3.0 (ensemble model)	27.75 (10.40)	25.54 (11.32)	4.44 (6.72)	0.67 (0.21)	83.28 (10.41)	75.81 (21.43)	87.64 (7.53)

Table 3. Comparison of FLASH-TV 3.0 models with low-resolution regularization and linear ensemble models on initial lab validation test set ($n=15$). SD - standard deviation. PABAK - prevalence and bias-adjusted Kappa, compares the gaze and no-gaze of FLASH-TV and gold standard at 5-second epoch, nested within each participant. Comparisons of FLASH-TV and gold-standard do not include video segments labeled as uncertain by the team.

Study	Gold standard TV time, minutes (SD)	FLASH predicted TV time, minutes (SD)	Kappa, PABAK (SD)	ICC	Accuracy (%)	Sensitivity (%)	Specificity (%)
Pilot sample ($n=10$)	29.02 (10.78)	29.66 (10.08)	0.66 (0.15)	0.81	82.73 (7.51)	83.97 (13.04)	81.77 (10.26)
Lab validation ($n=30$)	26.19 (11.15)	25.09 (11.05)	0.70 (0.16)	0.74	85.08 (8.07)	77.84 (17.30)	88.38 (6.54)
Home validation ($n=19$)*	22.75 (15.04)	24.79 (14.50)	0.74 (0.13)	0.91	87.10 (6.71)	81.74 (20.29)	85.95 (8.77)
Combined sample ($n=59$)	25.56 (12.46)	25.77 (12.05)	0.71 (0.15)	0.86	85.33 (7.58)	80.14 (17.60)	86.48 (8.21)

Table 4. Performance of FLASH-TV 3.0 algorithm on three independent samples and the combined sample. SD- standard deviation. PABAK- prevalence and bias-adjusted Kappa, compares the gaze and no-gaze of FLASH-TV and gold standard at 5-second epoch, nested within each participant. ICC-intraclass correlation, compares the overall duration of TV watching of FLASH-TV and the gold standard per participant. Comparisons of FLASH-TV and gold-standard do not include video segments labeled as uncertain by the team (mean 2.5%, SD 10% of video segments). *1 participant in the home validation study was not included because of the time synchronization issues of the FLASH-TV device with the world clock. This precluded aligning the timestamps of gold standard with the FLASH estimates.

ensemble model provided a better balance of high sensitivity and specificity, while also slightly increasing accuracy, compared to the individual models. We refer to this version of the model as FLASH-TV 3.0 from here onwards.

FLASH 3.0 validation results

Table 4 shows the performance of FLASH-TV 3.0 in the three independent validation studies. One participant from the home validation study had problems with the FLASH-TV system's clock losing synchronization with the world clock, which was not detected until after data collection was complete since FLASH-TV is not connected to Wi-Fi to enhance privacy. This prevented that data from being aligned with the validation video. However, the rest of the 19 participants do have validation data for this study. The engineers on the team subsequently added an external battery for the system's alternative real-time clock which, once synchronized,

maintains the clock on its own. Our combined data from the three samples (last row of Table 4) demonstrated a strong performance with a mean accuracy of 85%, moderate mean kappa of 0.71, and good mean ICC of 0.86. For the three individual study samples, FLASH-TV 3.0 achieved a minimum of 82% accuracy, kappa of 0.66, and ICCs of 0.74. The home sample, which involved more naturalistic screen use by children, had higher metrics for Kappa, ICC, and accuracy in comparison to the pilot and lab validation samples indicating FLASH-TV 3.0 was able to perform well in real-life settings. However, as noted in the methods, four children from the home validation study had more than 10% of the video labeled as uncertain, and those segments were not included in comparisons. Figure 1 shows the scatter plot of FLASH-TV estimated time vs. the gold standard TV viewing time. Most of the points lie close to the line of equality ($y = x$), indicating agreement between both. Also, FLASH-TV does not have a bias towards over or underestimation as points are approximately equally distributed above and below the line of equality. There are two outliers (bottom right in Fig. 1) where FLASH-TV significantly underestimates the TV viewing time. These were not part of the 4 children where humans labeled > 10% of the video as uncertain. FLASH-TV struggled for these outliers during the low-lighting segments of the in-lab validation study. The relative position of the child and the lighting created situations where the images only captured the silhouette of the child obscuring details for face recognition and gaze estimation.

Results stratified by child race and ethnicity

The top half of Table 5 compares the performance of FLASH-TV 3.0 across children in the lab validation study, stratified by race and ethnicity. Because ICC is particularly sensitive to the negative impacts of small sample sizes, the bottom half compares the statistics for the combined samples from the pilot, lab, and home validation studies to provide estimates with larger samples across groups. In the lab validation study ($n = 30$) FLASH-TV 3.0's performance was relatively high across categories with high accuracies (> 80%) and moderate Kappa

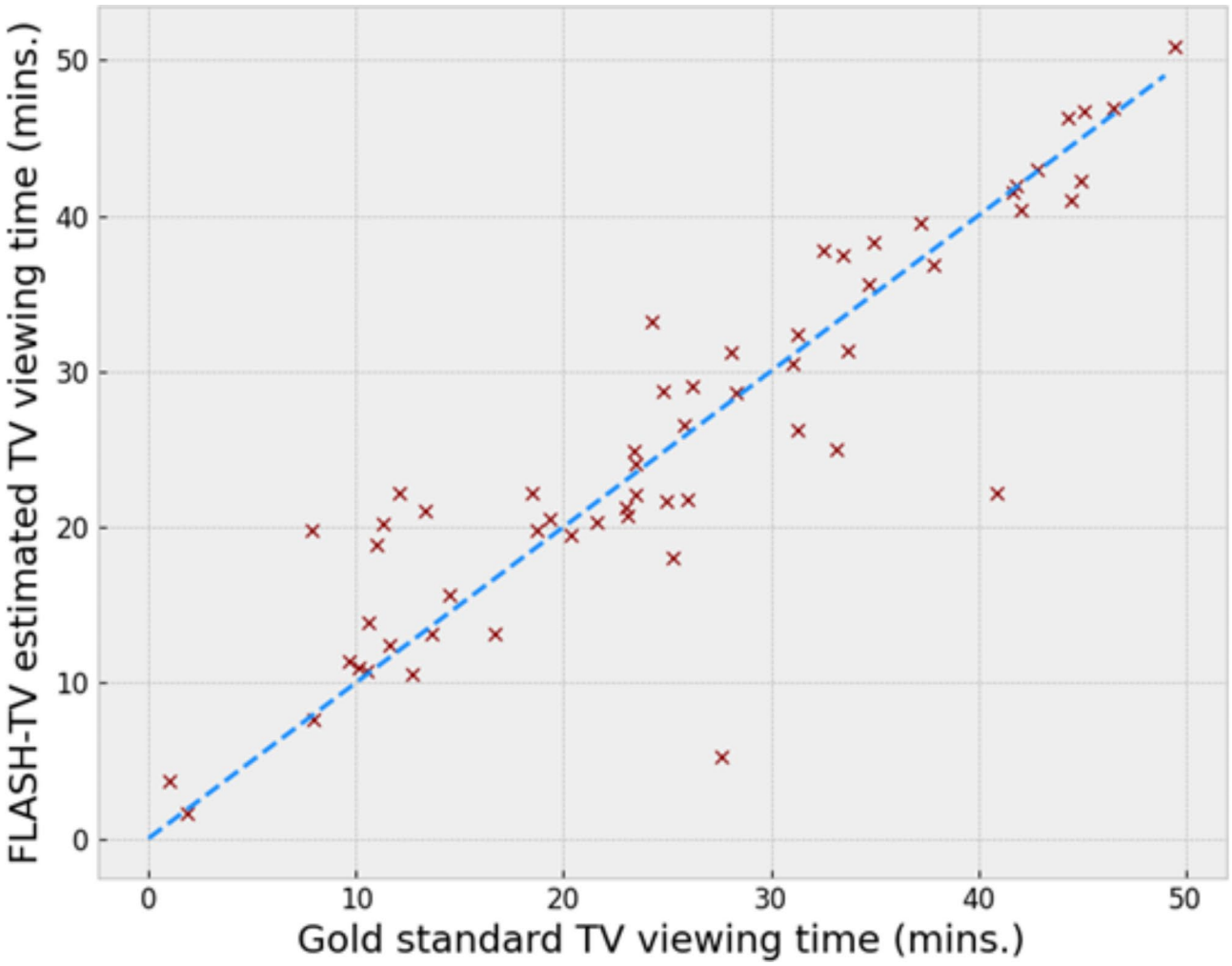

Fig. 1. Scatter plot of FLASH-TV estimation vs. Gold standard TV viewing time. The scatter plot compares the FLASH-TV estimate vs. the Gold standard for TV viewing time (mins) for the combined sample of children ($n = 59$). The points on the dotted line, the line of equality ($y = x$), indicate perfect agreement between both. FLASH-TV does not have a significant bias towards under ($y < x$) or overestimation ($y > x$) as points are approximately equally distributed around on both sides of the line.

Race or Ethnicity	Gold standard TV time, minutes (SD)	FLASH predicted TV time, minutes (SD)	Kappa, PABAK (SD)	ICC	Accuracy (%)	Sensitivity (%)	Specificity (%)
Lab Observation Validation Study, stratified by race and ethnicity ($n=30$)							
Black ($n=9$)	25.33 (10.11)	21.75 (10.19)	0.62 (0.22)	0.37	80.82 (11.22)	69.84 (25.03)	88.15 (6.09)
Hispanic White ($n=8$)	28.91 (11.56)	28.04 (10.66)	0.74 (0.09)	0.95	87.12 (4.57)	81.86 (11.13)	89.53 (4.67)
Non-Hispanic White ($n=8$)	24.89 (12.62)	25.26 (12.40)	0.76 (0.10)	0.97	87.81 (5.21)	83.89 (10.45)	88.75 (4.74)
Other ($n=5$)	25.47 (12.83)	26.08 (12.88)	0.70 (0.17)	0.98	85.12 (8.48)	76.16 (15.97)	86.36 (12.24)
Combined sample ($n=59$) from pilot study ($n=10$), lab validation study ($n=30$) and home validation protocol ($n=19$)							
Black ($n=17$)	23.55 (11.43)	22.63 (10.84)	0.67 (0.20)	0.80	83.64 (9.89)	72.40 (25.99)	88.01 (7.12)
Hispanic White ($n=14$)	28.20 (12.81)	27.76 (12.52)	0.77 (0.09)	0.98	88.39 (4.56)	85.68 (10.22)	88.58 (5.57)
Non-Hispanic White ($n=14$)	22.22 (12.00)	22.65 (11.98)	0.72 (0.14)	0.89	85.98 (7.09)	79.86 (11.69)	87.22 (7.95)
Other ($n=14$)	28.70 (13.74)	30.69 (12.22)	0.67 (0.14)	0.82	83.68 (6.91)	84.27 (13.45)	81.76 (10.56)

Table 5. FLASH-TV 3.0 results stratified by race and ethnicity in the lab validation study ($n=30$) and combined sample ($n=59$) from the pilot study, lab validation and home validation. SD - standard deviation. PABAK - prevalence and bias-adjusted Kappa, compares the gaze and no-gaze of FLASH-TV and gold standard at 5-second epoch, nested within each participant. ICC-intraclass correlation, compares the overall duration of TV watching of FLASH-TV and gold standard per participant. Comparisons of FLASH-TV and gold-standard do not include video segments labeled as uncertain by the team (mean 2.5%, SD 10% of video segments).

(>0.6). The ICCs were in the good range (>0.8) except for the Black children ($n=9$), for whom it was low at 0.37. There were two outliers in the Black sample, for whom FLASH-TV 3.0 did not perform well (see outliers in Fig. 1), missing most of their TV viewing. This is illustrated with a low ICC, which is sensitive to outliers and small sample size. The ICC for the sample of Black children without the two outliers was 0.90. In the combined sample ($n=59$), we achieved $\geq 83\%$ accuracy, Kappas in the moderate range, ranging from 0.67 to 0.77, and ICC ranging from good to excellent, 0.80–0.98, including for Black children (accuracy 83.6%, ICC 0.8, and Kappa of 0.67).

Results stratified by gender

We analyzed the performance of FLASH-TV 3.0 on our combined sample ($n=59$) stratified by gender. Our combined sample size ($n=59$) has 36 male and 23 female children. FLASH-TV 3.0 has an accuracy of 86.2% and 84.0% for male and female participants, respectively. The sensitivity was 83% and 76% for males and females, respectively, with a specificity of 86% and 87% respectively. The Kappa was 0.72 for males and 0.68 for females. The ICC was 0.92 and 0.79, respectively. The lower sensitivity and ICC among female children may be due to the relatively low sample size amplifying the effect of a single outlier among the female sample. Excluding this outlier, the sensitivity is 0.80, and the ICC is 0.89, achieving similar performance to the male participants.

Discussion

FLASH-TV uses three sequential neural network machine learning algorithms to identify children's TV viewing and provides time-stamped TV-viewing output over a specific study period. Three independent validation studies among diverse samples of children demonstrated that FLASH-TV had moderate Kappa (0.66–0.74), moderate-to-excellent ICCs (0.74–0.91) and high accuracy (82.7–87.1%) in identifying the duration of children's TV viewing. Importantly, the home validation study demonstrated good performance on the key-performance metrics compared to a gold standard, suggesting that FLASH-TV provides a valid estimate of children's TV viewing in the home. Previous literature reports that most self- or parent-reported measures of children's screen use have correlations less than 0.5 compared to gold standard assessments[7]. FLASH-TV has a much higher correlation of >0.7 with the gold standard, suggesting it is more valid for TV viewing measurements than self-report tools[7]. The time-stamped data output from FLASH-TV will allow researchers to combine the TV viewing data with other time-stamped data, such as mobile device use or accelerometer data, to assess how TV-viewing may influence other behaviors, health, or developmental outcomes.

In recent years, machine learning approaches have expanded into healthcare and social and behavioral research. For example, vision-based machine learning programs have been developed to detect COVID-19-induced pneumonia from patients' computed tomography (CT) scans[15], diagnosing breast cancer through mammogram images[16], and diagnosing different types of melanoma[17], with similar or better metrics to specialized physicians' diagnosis and often at improved efficiency. Many applications of machine learning for behavioral research involve the analysis of text using natural language processing (NLP), which is homogenized due to the necessitated standardization of behavioral research report structures, data formats, and word choices[18]. However, we have demonstrated with FLASH-TV that machine learning can also be used to analyze non-homogenized non-text data by processing video data from varying environments. This shows that machine

learning can help improve the versatility of current behavioral datasets and assist with projecting estimated behavioral consequences.

Others have also employed machine learning approaches for measuring youth's use of different screens. The Screenomics approach captures screenshots of a mobile device and includes machine learning in computational machinery for processing the images and text captured in the screenshots to identify duration and content of mobile device use for adolescents[9]. In another study, head-mounted video cameras captured videos, while computational models of head movements identified when a person was watching TV with a positive predictive value of 0.87 and sensitivity of 0.82 in naturalistic settings[10]. Both of these approaches create challenges due to privacy issues (Screenomics) or requiring wearables that may interfere with daily life activities (head-mounted camera).

Among the numerous applications of machine learning across the mentioned fields, the improved efficiency, accuracy, and accessibility were common advantages. However, alongside these strengths, vision based machine learning applications for face recognition include a frequent flaw, specifically racial or gender bias[13,19]. Studies of machine learning face recognition software from IBM, Microsoft, and Megvii (Face++) revealed significant deficits in recognition accuracy for females compared to males and darker skin tones relative to lighter skin tones[20]. For example, Microsoft had a very low error rate (<0.1%) for lighter males with a 20.8% error rate for darker-skinned females. In another report, Amazon Rekognition had a very low error (<0.1%) for lighter skinned male populations, with a contrasting 31.4% error rate for darker-skinned females[21]. The race biases seen in face recognition machine learning algorithms are believed to occur due to either data-driven problems or scenario-modeling issues[13]. Data-driven problems often relate to the data sets used to train convolutional neural networks for face recognition and how diverse or representative these are for the target sample to whom the algorithm is later applied. Scenario modeling issues are a result of (1) the threshold functions applied for identification set by the "user" and may need to differ by racial and/or ethnic group; and (2) the lack of diverse samples comparing dissimilar faces in terms of race, age, and gender demographic homogeneity of the different identity distribution[13]. To tackle the dataset issues, our methods were trained on diverse large-scale datasets WebFace260M[22] and ETHXGaze[23]. For scenario modeling, the AdaFace[24] approach adaptively adjusts the decision threshold based on image quality and difficulty.

The sample stratified by race and ethnicity in the lab observation study and the diversity of our combined sample allowed us to preliminarily investigate potential race bias of FLASH-TV. We demonstrated that FLASH-TV performs relatively well across the races and ethnicities in our samples, but had higher variability of performance among Black children (i.e. larger standard divisions for Kappa and accuracies among the Black sample). In reviewing the individual data of participants across the studies, FLASH-TV worked well on most Black children. However, the outliers for whom FLASH-TV did not perform as well were more likely to be Black, suggesting some race bias in FLASH-TV. Some of the low-lighting situations tested in the study protocols, along with the relative position of the child to the light, created instances where only silhouettes of the child were visible, obscuring the details of their faces. It is also noteworthy that researchers labeling the video data for the gold standard were more likely to use the "uncertain" label for Black children, suggesting that the gold standard also failed in some TV viewing scenarios among Black youth. In reviewing these cases of low performance of FLASH-TV compared to gold standard and cases in which our research team labeled data as "uncertain," both mostly happened in difficult scenarios in which the light source within the room or sunlight streaming in from a window caused silhouetting or shadowing of the child's face. Because these data were captured by a camera, the gold standard is subject to limitations associated with the camera's ability to perform under challenging light conditions, as demonstrated by the use of "uncertain" labels in some of the videos. While others have commented that the newer deep convolutional neural network algorithms for face recognition used in FLASH 3.0 have reduced race bias, this bias is often exacerbated in challenging situations such as certain lighting conditions[13].

Our gender bias analysis indicated FLASH-TV performs somewhat worse for detecting TV gaze for girls than boys. However, further investigation found that the gender difference was primarily due to one outlier. FLASH-TV had a lower performance (< 70% accuracy) for this participant due to challenging low-light conditions. Analysis that excluded the outlier found similar performance metrics for boys and girls. The slightly lower sensitivity and ICC with the female participants may therefore be due to the outsized effect of the outlier and the small sample size. It is of note that the outlier was a Black youth, so it is not clear if race difference in performance for some participants may be confounding the sex differences reported.

While some caution is warranted when using FLASH-TV in diverse samples of children, the FLASH-TV system has a higher correlation for TV-viewing to gold standard across Black children than the typical correlations reported for self- or parent-reported assessment tools for children's TV viewing[7]. This suggests it is better than these tools for measuring TV viewing even among Black children. Further research is needed to reduce any race bias in the system, or to identify any children (regardless of race or ethnicity) for whom FLASH-TV may be more likely to miss TV-viewing. Increasing the accuracy of the FLASH-TV system in difficult lighting conditions would help alleviate some of these challenges.

These studies have several limitations, including small sample sizes for each of the validation studies. However, combining the data across the three studies provided more robust estimates of FLASH-TV's performance metrics. Additionally, we only have gold standard labeling of the video data for the target child's gaze on the TV, the final step of FLASH-TV output. The gold standard labeling did not involve information about the faces themselves, or which face the target child/sibling/parent were. This limits our ability to identify which of the three iterative steps of FLASH-TV (face detection, face verification, or gaze detection) fails for some outliers when FLASH-TV does not perform as well. Future studies will investigate the feasibility and acceptance of using FLASH-TV devices to measure children's TV viewing in their homes.

Conclusion

Overall FLASH-TV had moderate to excellent correspondence to gold standard assessment across multiple metrics among a diverse sample of children, with higher variability found among Black children than others. FLASH-TV is an important new tool for measuring the duration and timing of TV-viewing among children in naturalistic settings, while preserving privacy when employed in family homes. FLASH-TV can help improve researchers' assessment of TV viewing patterns and eventually shed light on how these behaviors may impact other health-related outcomes among children.

Methods

FLASH-TV 1.0 was a standalone algorithm pipeline that analyzed video data after data collection on lab computing resources[12,25]. All three steps of FLASH-TV 1.0, face detection, face recognition, and gaze estimation, were designed iteratively until design criteria were met. For FLASH-TV 2.0 and 3.0, our goal was to develop a complete stand-alone system using portable hardware that would collect the video frames and process it in real-time. This would allow us to collect data with ease in the participants' homes while maintaining the privacy components of FLASH-TV. The new versions of FLASH-TV were evaluated under deliberately challenging conditions such as living room variations (relative location of furniture, framed faces in the room, and lighting conditions) and two different TV sizes, with appropriate sample sizes to inform the validity of the system to capture children's TV viewing. The protocols comply with the ethical principles for human subject research and in accordance with the Declaration of Helsinki. The protocol was reviewed and approved by the Baylor College of Medicine Institutional Review Board (H-40556), with a reciprocal authorized agreement with Rice University. Families who met the study eligibility criteria engaged in an informed consent review over the phone prior to attending their data collection visit. If the parent agreed to take part in the study, the child's assent was obtained in person at the data collection visit by explaining the study in a developmentally appropriate manner. Parents and children were offered the opportunity for questions over the phone and in person before obtaining written informed consent at the time of the study visit to ensure transparency and informed consent, with the ability to revoke consent at any time.

FLASH-TV 2.0 enhancements

FLASH-TV 2.0 maintained the same initial two machine learning algorithms for face detection[26] and face verification[27] used in FLASH-TV 1.0[12]. For gaze estimation, FLASH-TV 1.0[12] used the Gaze360[28] approach, which was not as accurate in low-light conditions or when the target child was laying down to watch TV (causing rotated head-positions or extreme angles of gaze). To improve gaze estimation for FLASH-TV 2.0 during such situations, we combined the Gaze360 data of 200k samples with the ETHXGaze data set[23], a recently published large-scale gaze estimation dataset with more than a million samples from 110 subjects with diverse racial and ethnic backgrounds[23]. For each participant in the dataset, ETHXGaze captured more than 500 gaze directions under 16 different lighting conditions and included procedures for aligning all faces up-right using facial landmarks. These variations captured several scenarios similar to the FLASH-TV data regarding lighting conditions and extreme gaze directions.

We trained the FLASH-TV 2.0 gaze estimator using combined data from the ETHXGaze and Gaze360 datasets. In each iteration, we sampled with 0.7 probability from the Gaze360 data and the remaining 0.3 from the ETHXGaze data to create the batch for training. More samples are drawn from Gaze360 as the dataset was more similar to the FLASH-TV requirements regarding facial image resolution and relative distance between the camera and the child. We trained the combined dataset method for 100k iterations. Although FLASH-TV 2.0 performed well in our initial pilot testing sample ($n=10$; see details in results), it did not perform as well based on interim analysis of the first half ($n=15$) of the lab validation study (see results). In reviewing the FLASH-TV 2.0 output of the first half of the validation study data set, we noted patterns in FLASH-TV errors. For example, the face detector was not as accurate in detecting faces when a child laid on the sofa because the face was not upright. Accuracy also decreased when the light was low, due to low-resolution images. This led to FLASH-TV 2.0 not identifying the child, resulting in false negatives for TV viewing.

FLASH-TV 3.0 enhancements

To address these limitations of FLASH-TV 2.0, FLASH-TV 3.0 included enhancements for all three sequential machine learning steps in the image processing: face detection, face recognition and gaze detection. For face detection, we replaced the YoLo[26] based face detector used in FLASH 1.0 and 2.0 with the RetinaFace[29] detector which utilized a multi-pronged approach to detect faces using various supervision methods – face recognition, facial landmark predictions and 3D reconstruction. Unlike a typical face detector trained to identify only the regions containing the faces, RetinaFace uses additional supervision to identify landmark regions and estimates the 3D geometry of the face. This extra supervision enabled the detector to generalize better, improving face detection in new data sets. RetinaFace for FLASH-TV reduced face detection errors from 5 to 3%. The facial landmarks predicted by the face detector were used in the later steps of face recognition and gaze estimation to align the faces up-right to enhance the accuracies of those steps. To detect faces more accurately when they were not upright, for example when a child was lying on the sofa, we explicitly rotated the images (clockwise or anti-clockwise by 90 degrees) for each input frame. We ran the face detector on the original input frame and the rotated images to avoid missing these faces. Note that any faces detected in either of the images are retained and fed to the next step of face verification.

For face verification, we replaced DeepFace[27] with AdaFace[24,30], which used adaptive margin loss to enhance the discriminative power of the face recognition model. AdaFace applies adaptive margin loss to a derived measurement of the image quality of face features, which leads to better results on low quality images and

more separation between face identities in the latent space. We also added a face-alignment component which corrected for any rotation in the detected face and set it to a vertically aligned frontal face. Although we did not find much improvement in sensitivity in detecting the target child with AdaFace, our positive predictive value when the target child had no gaze on TV improved from 88 to 96%. This helped reduce the false positives when there was no gaze from the child. Such false positives would result in errors in the TV viewing time for the target child as another person's TV time is counted towards the target child's viewing time. Our refinements reduced the average amount of time that we were not able to detect and identify the target child by 60%.

To improve FLASH-TV 3.0 gaze detection, we increased the model complexity by changing the backbone of the Gaze360 neural network architecture from ResNet18[31] to ResNet34[31], which had about twice the parameters. We refer to this model as FLASH-TV 3.0-unregularized, which was trained in the same way as FLASH-TV 2.0, combining Gaze360 and ETHXGaze datasets for the training input. One of the challenges that FLASH-TV faced was the poor resolution of captured facial images. The poor resolution limited the details around the eye region, making gaze estimation difficult. To address this, we followed the low-resolution regularization approach of Xu et al.[14] During the training, for each input batch of images to the gaze estimation network, we simulated a low-resolution version. Then, a regularization loss (mean square error) was used to ensure the network features and the gaze angle output for both the high and low-resolution versions were similar. This regularization ensured that the network performance was robust to resolution variations. We refer to this regularized model as FLASH-TV 3.0-regularized. This was trained similarly to the earlier model but with an additional regularization loss weighted with 0.09 strength[14]. The v3.0-unregularized model achieved higher specificity, and the latter regularized model had higher sensitivity (see Results). To obtain the best of both models, we built an ensemble model that linearly combined the gaze vector outputs of v3-unregularized and v3-regularized with optimal weights of 0.6 and 0.4, respectively, chosen to get the best accuracy, sensitivity, and specificity (see Results).

Pilot, lab validation and home validation studies

Three independent 60-minute task-based studies were completed: an observational lab pilot test ($n=10$), an observational lab validation study ($n=30$), and a home validation study at the end of a 3-day home-based feasibility study ($n=20$). The three studies allowed comparison of the FLASH-TV system in detecting whether the target child was watching TV compared to video data labeled for child's gaze on the TV by study staff (gold standard). We developed the task-based protocols (Supplementary Table S1 for example of lab observation study) to be similar to protocols used to assess energy expenditure during different sedentary behaviors and physical activities[32,33]. This allowed assessment of a variety of different screen use behaviors, from different locations in the room, and postures while watching TV.

Recruitment and sampling procedures. Parent, child, and sibling triads were recruited for the pilot and lab validation studies, while parent-child dyads were enrolled in the home observation study. Recruitment occurred across Houston using flyers posted or handed out in local pediatric and other medical clinics, on general community announcement boards (physical and electronic), around the Texas Medical Center, local community centers/groups, shopping areas, Young Men's Christian Associations (YMCAs), health fairs, churches, and apartment buildings. Families with children in the age group from the United States Department of Agriculture (USDA) / Agriculture Research Service (ARS) Children's Nutrition Research Center at Baylor College of Medicine volunteer data base and the NIH Research Match were also notified about the study.

The pilot test consisted of 10 parent child triads and used the same procedures as described below for the lab study to ensure procedures worked as intended. Given that no major issues were encountered, data from that sample is included to assess FLASH-TV 3.0 validity. To ensure a diverse sample for the lab validation study, the sample of 30 triads was stratified by child race and ethnicity with a goal of enrolling at least eight non-Hispanic White, eight Hispanic-White, and eight non-Hispanic or Hispanic Black youth. The remaining six triad-slots were open and could include children from any of the above or other racial ethnic groups, such as Asian, American Indian or Alaska Native, Native Hawaiian or Pacific Islander, or Multi-racial. For each family, a target child (aged 6–11 years old), their sibling (6–14 years old), and parent were enrolled with the parent providing consent for themselves and their children to participate, and the children providing assent. The home validation study among parent-child dyads was conducted at the end of a 3-day home-based feasibility study with children 5–12 years old and their parent. The feasibility results will be reported separately; here we focus on reporting the comparison of FLASH-TV to gold standard staff labeled video data for a 60-minute validation protocol conducted in the family's home where we collected other video data in addition to FLASH-TV.

Lab-based procedures used in the pilot test and lab validation studies. For each family in the pilot and lab validation study, the observation room was set up to resemble a living room, but the appearance of the room varied for each family by changing the location of the furniture, photos of faces, rugs and/or other decorations. The size of the TV (30 inches vs. 55 inches) and position of the TV in room varied as well to simulate different locations families may place their TV at home. Prior to starting the protocol, families were presented with a selection of age appropriate DVDs to choose from that would be played on the television or monitor. The participants were allowed to use any apps on their personal mobile devices that were brought to the laboratory with no content restraints.

Before starting the lab-based protocol, the FLASH-TV system was trained to identify the target child compared to the sibling and parent during a short set-up procedure. This involved the child, sibling and parent standing or sitting in different locations of the room with signs identifying who they were (target child, sibling, or parent). The family then engaged in a 60-minute task-based protocol with multiple sections (Supplementary Table S1). The protocol began with four 1.5-minute sections in which the target child or sibling were asked to step out of the camera view to ensure FLASH no longer captured the child of interest when not present, varying between dim and bright lighting. These sections were followed by two 15-minute sessions of free play, varying between dim and bright lighting. During the two free play sessions, the families were asked to behave as they

would want while a movie was playing, with the option of also using a mobile device simultaneously. After that, there were three 3-minute sections in which the lighting and seating locations of the participants were changed and the mobile device was out of reach. Following this, there were four 3-minute sections in which the mobile device was required to be used and the user was varied. Finally, there was a 5-minute section where the entire family watched TV with disposable facial masks over the nose and mouth. The last 5-min period was not used in analysis in this study.

Task-based procedures used in the home validation study. The FLASH-TV system was trained using similar procedures as in the lab validation study to differentiate the target child compared to the parent and one sibling, if available. If a third household member was not available, an online stock photo was used. There were no stipulations to the screen media content consumed in the home feasibility study due to the naturalistic nature of the home environment. The families were instructed to maintain their typical screen use habits. The 60-minute observation protocol consisted of 15 min in which the target child actively viewed the TV from different areas of the room, turned off the TV but remained in the room, and left the room while the TV was still on. This was followed by a 45-minute free play session when the child was instructed to use their mobile device and/ or watch TV as they normally would. The free play session provided an option for the child to leave the room for periods of time. Other family members were allowed to be present during the observation if they wished.

Gold standard data capture and video labeling. A separate high-definition video camera identical to one used by FLASH was placed next to the FLASH system and collected video data for labeling. The video images from all three studies (pilot, lab, and home validation) were labeled by study staff who were trained and certified for labeling. Duration-labeling was used[12], with one of four labels linked to each family's video clips, until the behavior changed. The four labels included: (1) target child watching screen, (2) target child not watching screen, (3) target child out of frame, or (4) uncertain. The 'target child out of frame' label was used 6.7% of the time, and the protocol included a 3 min section when the child was asked to leave the room to ensure the system would accurately identify the child as not present. The uncertain label was used when the camera did not provide clear images of the target child's face and eye gaze and staff could not ascertain whether the child was watching TV or not. This occurrence could be caused by extreme lighting, such as the child sitting next to bright light from a window creating a shadow that obscured the face, etc. The "uncertain" label was never used for 37 participants of 59 for whom we had available video data, 10 participants had the uncertain label used < 1% of the time, and 8 participants had the uncertain label used between 1% to < 10% of the time. However, 4 out of 59 participants had more than 10% of the video labeled as uncertain (one Hispanic White and three Black youth). All four participated in the home validation study. Out of these four, only one had a significant portion (70% of the video) labeled uncertain. The other three had less than 25% of video labeled as uncertain. 10% of the video for each family was double labeled by two trained staff to assess inter-rater reliability ($\kappa = 0.88$, SD 0.12). These labeled video data were the gold-standard for comparison with the FLASH-TV estimates.

Statistical analysis

Demographic characteristics were summarized using mean (SD) or frequencies and percent overall and by study (pilot, lab validation, home validation). Agreement between gaze and no-gaze of FLASH-TV and gold standard was assessed, but video segments labeled as uncertain were not included in the performance metrics and TV watching time analysis. Video labeled as "out of frame" was classified as no gaze. Comparison of FLASH-TV output and gold standard at 5-second epochs was assessed using the prevalence and bias adjusted Kappa (PABAK) statistic[34]. A Kappa ranging from 0.21 to 0.39 was considered minimal, 0.40–0.59 weak, 0.60–0.79 moderate, and ≥ 0.80 excellent[35]. Epochs of TV gaze data were summed per participant to generate a continuous TV duration variable (minutes of TV viewing). The validity of TV viewing data between FLASH-TV and gold standard was assessed using the intraclass correlation coefficient (ICC). ICC was calculated using a generalized linear mixed effects model[36] due to the clustering of epochs within participant. Model specification included a random intercept, nesting of epochs within participants, a lognormal distribution for the outcome of minutes of TV viewing and no fixed effects. ICC values ranging from 0.50 to 0.79 were considered moderate, 0.80–0.89 good and ≥ 0.90 as excellent[37]. ICC was calculated in Statistical Analysis System (SAS) version 9.4 (SAS Institute, Incorporation, Cary, North Carolina).

Data availability

The datasets generated during and/or analyzed during the current study are not publicly available due to HIPAA and Human Subjects privacy constraints on the video data. De-identified processed data (not image data itself) can be made available under reasonable requests to the corresponding author within the Baylor College of Medicine Institutional Review Board approved intended uses and with appropriate formal institutional data sharing agreement established. Our algorithm pipeline code with the latest FLASH-TV 3.0 additions is publicly available on GitHub at https://github.com/anilrgukt/flash-tv-scripts, along with the instructions for the setup.

Received: 17 May 2024; Accepted: 25 November 2024
Published online: 30 November 2024

References

1. Stiglic, N. & Viner, R. M. Effects of screentime on the health and well-being of children and adolescents: a systematic review of reviews. *BMJ Open.* **9**, e023191. https://doi.org/10.1136/bmjopen-2018-023191 (2019).
2. Eirich, R. et al. Association of screen time with internalizing and externalizing behavior problems in children 12 years or younger: a systematic review and meta-analysis. *JAMA Psychiatry.* **79**, 393–405. https://doi.org/10.1001/jamapsychiatry.2022.0155 (2022).
3. Adelantado-Renau, M. et al. Association between screen media use and academic performance among children and adolescents: a systematic review and meta-analysis. *JAMA Pediatr.* **173**, 1058–1067. https://doi.org/10.1001/jamapediatrics.2019.3176 (2019).

4. Lund, L., Sølvhøj, I. N., Danielsen, D. & Andersen, S. Electronic media use and sleep in children and adolescents in western countries: a systematic review. *BMC Public. Health*. **21**, 1598. https://doi.org/10.1186/s12889-021-11640-9 (2021).
5. Sigman, A. Time for a view on screen time. *Arch. Dis. Child.* **97**, 935–942. https://doi.org/10.1136/archdischild-2012-302196 (2012).
6. Byrne, R., Terranova, C. O. & Trost, S. G. Measurement of screen time among young children aged 0–6 years: a systematic review. *Obes. Rev.* **22**, e13260. https://doi.org/10.1111/obr.13260 (2021).
7. Perez, O. et al. Validated assessment tools for screen media use: a systematic review. *PLoS One*. **18**, e0283714. https://doi.org/10.1371/journal.pone.0283714 (2023).
8. Saunders, T. J. & Vallance, J. K. Screen time and health indicators among children and youth: current evidence, limitations and future directions. *Appl. Health Econ. Health Policy*. **15**, 323–331. https://doi.org/10.1007/s40258-016-0289-3 (2017).
9. Ram, N. et al. A new approach for observing and studying individuals' digital lives. *J. Adolesc. Res.* **35**, 16–50. https://doi.org/10.1177/0743558419883362 (2020).
10. Zhang, Y. C. & Rehg, J. M. Watching the TV watchers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2**, 88. https://doi.org/10.1145/3214291 (2018).
11. Radesky, J. S. et al. Young children's use of smartphones and tablets. *Pediatrics* **146**, e20193518. https://doi.org/10.1542/peds.2019-3518 (2020).
12. Kumar Vadathya, A. et al. An objective system for quantitative assessment of TV viewing among children: FLASH-TV. *JMIR Pediatr. Parent.* **5**, e33569. https://doi.org/10.2196/33569 (2022).
13. Cavazos, J. G., Phillips, P. J., Castillo, C. D. & O'Toole, A. J. Accuracy comparison across face recognition algorithms: where are we on measuring race bias? *IEEE Trans. Biom Behav. Identity Sci.* **3**, 101–111. https://doi.org/10.1109/tbiom.2020.3027269 (2021).
14. Xu, X., Chen, H., Moreno-Noguer, F. & Jeni, L. A. De La Torre, F. 3D human pose, shape and texture from low-resolution images and videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 4490–4504. https://doi.org/10.1109/tpami.2021.3070002 (2022).
15. Wang, B. et al. AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI system. *Appl. Soft Comput.* **98**, 106897. https://doi.org/10.1016/j.asoc.2020.106897 (2021).
16. Pacilè, S. et al. Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool. *Radiol. Artif. Intell.* **2**, e190208. https://doi.org/10.1148/ryai.2020190208 (2020).
17. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118. https://doi.org/10.1038/nature21056 (2017).
18. Mac Aonghusa, P. & Michie, S. Artificial intelligence and behavioral science through the looking glass: challenges for real-world application. *Ann. Behav. Med.* **54**, 942–947. https://doi.org/10.1093/abm/kaaa095 (2020).
19. Coe, J. & Atay, M. Evaluating impact of race in facial recognition across machine learning and deep learning algorithms. *Computers* **10**, 113. https://doi.org/10.3390/computers10090113 (2021).
20. Buolamwini, J. & Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification in *Proceedings of Machine Learning Research - Conference on Fairness, Accountability and Transparency* Vol. 81 (eds. Friedler, S. A. & Wilson, C.) 77–91ML Research Press,. (2018).
21. Raji, I. D. & Buolamwini, J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products in *AIES '19: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 429–435. (2019). https://doi.org/10.1145/3306618.3314244 (Association for Computing Machinery, 2019).
22. Zhu, Z. et al. IEEE Computer Society,. Webface260m: A benchmark unveiling the power of million-scale deep face recognition in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10492–10502. (2021).
23. Zhang, X. et al. ETH-XGaze: a large scale dataset for gaze estimation under extreme head pose and gaze variation in Computer Vision – ECCV 2020 Vol. 12350 (eds Vedaldi, A., Bischof, H., Brox, T. & Frahm, J. M.) 365–381. https://doi.org/10.1007/978-3-030-58558-7_22 (Springer, (2020).
24. Kim, M. J., Jain, A. K., Liu, X. & AdaFace IEEE, : Quality adaptive margin for face recognition in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 18729–18738. (2022). https://doi.org/10.1109/CVPR52688.2022.01819
25. Kumar Vadathya, A. et al. Development of Family Level Assessment of screen use in the home for television (FLASH-TV). *Multimed Tools Appl.* **83**, 63679–63697. https://doi.org/10.1007/s11042-023-17852-y.c (2024).
26. Redmon, J. & Farhadi, A. IEEE, YOLO9000: Better, faster, stronger in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 6517–6525. (2017). https://doi.org/10.1109/CVPR.2017.690
27. Taigman, Y., Yang, M., Ranzato, M. A., Wolf, L. & Deepface IEEE, : Closing the gap to human-level performance in face verification in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Vol. 2014 1701–1708. (2014). https://doi.org/10.1109/CVPR.2014.220
28. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W. & Torralba, A. IEEE, Gaze360: Physically unconstrained gaze estimation in the wild in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 6911–6920. (2019). https://doi.org/10.1109/ICCV.2019.00701
29. Deng, J. et al. IEEE, : Single-stage dense face localisation in the wild in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 5202–5211. (2020). https://doi.org/10.1109/CVPR42600.2020.00525
30. Deng, J. et al. Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 5962–5979. https://doi.org/10.1109/tpami.2021.3087709 (2022).
31. He, K., Zhang, X., Ren, S., Sun, J. & Deep IEEE, residual learning for image recognition in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778. (2016). https://doi.org/10.1109/CVPR.2016.90
32. Puyau, M. R. et al. Energy cost of activities in preschool-aged children. *J. Phys. Act. Health*. **13**, 11–16. https://doi.org/10.1123/jpah.2015-0711 (2016).
33. Bitar, A., Fellmann, N., Vernet, J., Coudert, J. & Vermorel, M. Variations and determinants of energy expenditure as measured by whole-body indirect calorimetry during puberty and adolescence. *Am. J. Clin. Nutr.* **69**, 1209–1216. https://doi.org/10.1093/ajcn/69.6.1209 (1999).
34. Byrt, T., Bishop, J. & Carlin, J. B. Bias, prevalence and kappa. *J. Clin. Epidemiol.* **46**, 423–429. https://doi.org/10.1016/0895-4356(93)90018-v (1993).
35. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Med. (Zagreb)*. **22**, 276–282 (2012).
36. Nakagawa, S., Johnson, P. C. D. & Schielzeth, H. The coefficient of determination R(2) and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R Soc. Interface*. **14**, 2017021. https://doi.org/10.1098/rsif.2017.0213 (2017).
37. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012 (2016).

Acknowledgements

This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) of the National Institutes of Health (NIH) [grant number R01DK113269] and has resulted in additional work supported by Eunice Kennedy Shriver National Institute of Child Health and human Development [grant number P01HD109876]. This work was also supported by the United States Department of Agriculture/Agricultural

Research Service (USDA/ARS) [cooperative agreement 58-3092-0-001]. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or USDA.

Author contributions

TMO and AV developed the original plans for the studies and obtained funding. AKV, AH, and UA led the revisions to the FLASH-TV system under the supervision of AV. TMO and SH developed the protocols for video labeling. TMO, AV, TG, AB, JM, SMM, and TB developed the validation study protocols and TG, AB, AV, UA, NH implemented the protocols and processed the data. AKV and SMM conducted the analysis. AKV, TG, AB, UA, SMM, JPM, TB, NH, SH, JAM, AV, JY, AS, and TMO reviewed and interpreted the results. AKV and TMO drafted the manuscript with NH assisting with drafting part of the discussion. All authors reviewed, edited and approved the final version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-81136-0.

Correspondence and requests for materials should be addressed to T.M.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.