Machine Learning Supports Automated Digital Image Scoring of Stool Consistency in Diapers

*Thomas Ludwig, †Ines Oukid, *Jill Wong, *Steven Ting, ‡Koen Huysentruyt,
*Puspita Roy, *Agathe C. Foussat, and ‡Yvan Vandenplas

ABSTRACT

Background/Aims: Accurate stool consistency classification of non–toilet-trained children remains challenging. This study evaluated the feasibility of automated classification of stool consistencies from diaper photos using machine learning (ML).

Methods: In total, 2687 usable smartphone photos of diapers with stool from 96 children younger than 24 months were obtained after independent ethical study approval. Stool consistency was assessed from each photo according to the original 7 types of the Brussels Infant and Toddler Stool Scale independently by study participants and 2 researchers. A health care professional assigned a final score in case of scoring disagreement between the researchers. A proof-of-concept ML model was built upon this collected photo database, using transfer learning to re-train the classification layer of a pretrained deep convolutional neural network model. The model was built on random training (n = 2478) and test (n = 209) subsets.

Results: Agreements between study participants and both researchers were 58.0% and 48.5%, respectively, and between researchers 77.5% (assessable n = 2366). The model classified 60.3% of the test photos in exact agreement with the final score. With respect to the 4-class grouping of the 7 Brussels Infant and Toddler Stool Scale types, the agreement between model-based and researcher classification was 77.0%.

Conclusion: The automated and objective scoring of stool consistency from diaper photos by the ML model shows robust agreement with human raters and overcomes limitations of other methods relying on caregiver reporting. Integrated with a smartphone application, this new framework for photo database construction and ML classification has numerous potential applications in clinical studies and home assessment.

Key Words: machine learning, smartphone application, stool consistency

(*JPGN* 2021;72: 255–261)

Stool consistency scoring is of critical importance in daily pediatric practice to assess gastrointestinal health and disorders. Although fundamental and seemingly straightforward, there is to date no reliable, simple to use, and universally accepted method to assess stool consistency. Stool consistency is also a frequent source of parental insecurity that causes them to seek professional advice. This has been found to result in increased use of health care resources (1–3). Unsurprisingly, it has been reported that it is challenging for parents to correctly assess stool consistencies of children (4), and apparently this accounts also for reporting during consultations with health care providers.

Notwithstanding, stool consistency is an easily accessible parameter and accordingly a frequent and relevant clinical study outcome. It has, for example, been proposed to be included in the core outcome set in clinical studies for functional constipation in children up to 18 years of age (5).

The consistency of stools reflects differences in water content and has been shown to depend on the colonic microbial ecosystem and transit time, where firm stools are correlated with slower transit (6–9). It has also been concluded that stool consistency is a decisive confounder in gut microbiota analysis (9). This implies that the assessment needs to be robust and reproducible.

The criterion standard for the assessment of stool consistencies in toilet-trained adults and children is the Bristol Stool Form Scale (BSFS). A study has, however, proven that only fair agreement exists between the BSFS and parental assessment of stool consistencies of infants and non–toilet-trained toddlers (10). For young children up to the age of 120 days after birth, the Amsterdam Infant Stool Scale (AMS) has been proposed (11). A modified version of the BSFS is only applicable for children 8 years and older (12). Besides the age gap for this population, it has been shown that AMS, BSFS, and a modified version of the BSFS are not user-friendly for the assessment of children (13). In addition, another study on the AMS concluded a lower reproducibility of the stool consistency assessment as compared to the first publication (14). Most recently, in an attempt to overcome limitations of the most commonly used existing stool form scales, the Brussels Infant and Toddler Stool Scale (BITSS) has been developed to provide a visual tool to describe and classify stool consistency for non–toilet-trained infants and toddlers (4,15). In addition, a clinical study reported that the BITSS enables a reliable assessment of stool consistencies from photos in line with the direct observation of fresh stools (16).

Received July 27, 2020; accepted October 9, 2020.

From the *Danone Nutricia Research, Precision Nutrition D-lab, Biopolis, Singapore, the †Danone Research, Palaiseau, France, and the ‡KidZ Health Castle, UZ Brussel, Vrije Universiteit Brussel, Brussels, Belgium.

Address correspondence and reprint requests to Professor Yvan Vandenplas, MD, PhD, KidZ Health Castle, UZ Brussel, Vrije Universiteit Brussel, Brussels, Belgium (e-mail: yvan.vandenplas@uzbrussel.be).

Supplemental digital content is available for this article. Direct URL citations appear in the printed text, and links to the digital files are provided in the HTML text of this article on the journal's Web site (*www.jpgn.org*).

www.clinicaltrials.gov ClinicalTrials.gov registration number: NCT03402555.

Danone Nutricia Research provided the funding to conduct the study.

T.L. and I.O. contributed equally.

T.L., J.W., S.T., P.R., and A.C.F. are employees of Danone Nutricia Research, and I.O. is an employee of Danone Research. K.H. and Y.V. declare no competing interests for this study.

DOI: 10.1097/MPG.0000000000003007

Machine learning (ML) offers powerful tools for automated, objective medical image recognition and classification, supporting clinical tasks such as screening and diagnosis. Here we report the outcomes of a study conducted to build a database of photos of stools from non–toilet-trained children and first insights into the performance of an image recognition model built for the objective, reproducible, and reliable scoring of stool consistency.

METHODS

This observational study was conducted between May and December 2018 in the United States of America. An independent institutional review board (IntegReview IRB, Austin, TX) reviewed and approved the study protocol and documents used for informed consent before study initiation. The study was conducted according to the principles and rules laid down in the Declaration of Helsinki, the International Conference for Harmonization of Technical Requirements for Pharmaceuticals for Human Use guidelines, and the Council for International Organizations of Medical Sciences. All participants provided informed electronic consent before any study-related activities were initiated. This study was registered at ClinicalTrials.gov (identifier NCT03402555). All co-authors had access to the study data and reviewed and approved the final manuscript.

The study was conducted using a smartphone technology–enabled virtual clinical study site. Participants interacted with study staff via a smartphone application (ClaimIt app, version 1, Obvio Health, Orlando, FL, USA). All study data were collected via the app running on participants' smartphones.

Study Population

The study population comprised mothers 18 years or older and their healthy infants or toddlers (aged 0–24 months at the time of enrolment) residing in the United States, who used or committed to use disposable diapers for the duration of the study. Participants were recruited via Internet advertisements on Facebook Ads and Google AdWords. The advertisements contained live links redirecting potential participants to a Webpage providing study information and a registration form to collect contact information. After completing the registration form, potential participants received an email link to an electronic prescreening questionnaire. Based on the prescreening, eligible participants were given access to download the ClaimIt app and complete the electronic informed consent process. After providing informed consent within the app, participants completed a Screening Questionnaire and took 3 screening photos of their child's stool using their smartphone via the app and scored them according to instructions provided. Participants were enrolled in the study if they successfully took good quality screening photos, could correctly score the photos as judged by the principal investigator of the study (a health care professional), and met all other inclusion criteria (in-home access to reliable Internet connection and smartphone; the infant/toddler consumes age-appropriate standard food) and none of the exclusion criteria (significant condition that might interfere with study compliance; likely to be noncompliant with the protocol).

Study Design

Enrolled participants were required to submit an empty diaper photo at "baseline" and then to upload at least 1 stool photo and corresponding BITSS score per day for 30 days. The participants were given instructions via the app on how to take photos and score the stool consistency according to the BITSS. Compliance was monitored by the study team and principal investigator and contact was made with the participants electronically or, if needed, by telephone. At the end of the study, participants were required to submit a second empty diaper photo and complete a questionnaire about the usability of the app.

All stool photos were also independently scored by 2 independent researchers. Both were mothers not enrolled in the study and who had no participant contact.

Study Assessments

Brussels Infant and Toddler Stool Scale

The BITSS was developed as an instrument to assess stool consistency in non–toilet-trained children (4,15). The original BITSS consists of 7 color photographs of diapers containing stools of infants or toddlers who are not toilet-trained, further categorized into a 4-category scale as hard (types 1–3), formed (type 4), loose (types 5 and 6), and watery (type 7) (Fig. 1).

Stool Photo Collection

Participants were given training via the app on how to take stool photos, including correct positioning of the diaper within the photo. The uploaded photos were subjected to a quality control process that excluded duplicate photos, and those of poor quality (blank, blurry, heterogeneous lighting). Photos submitted by ineligible participants and participants who withdrew consent were also excluded from the final set of evaluable photos (Fig. 1).

End of Study Usability Questionnaire

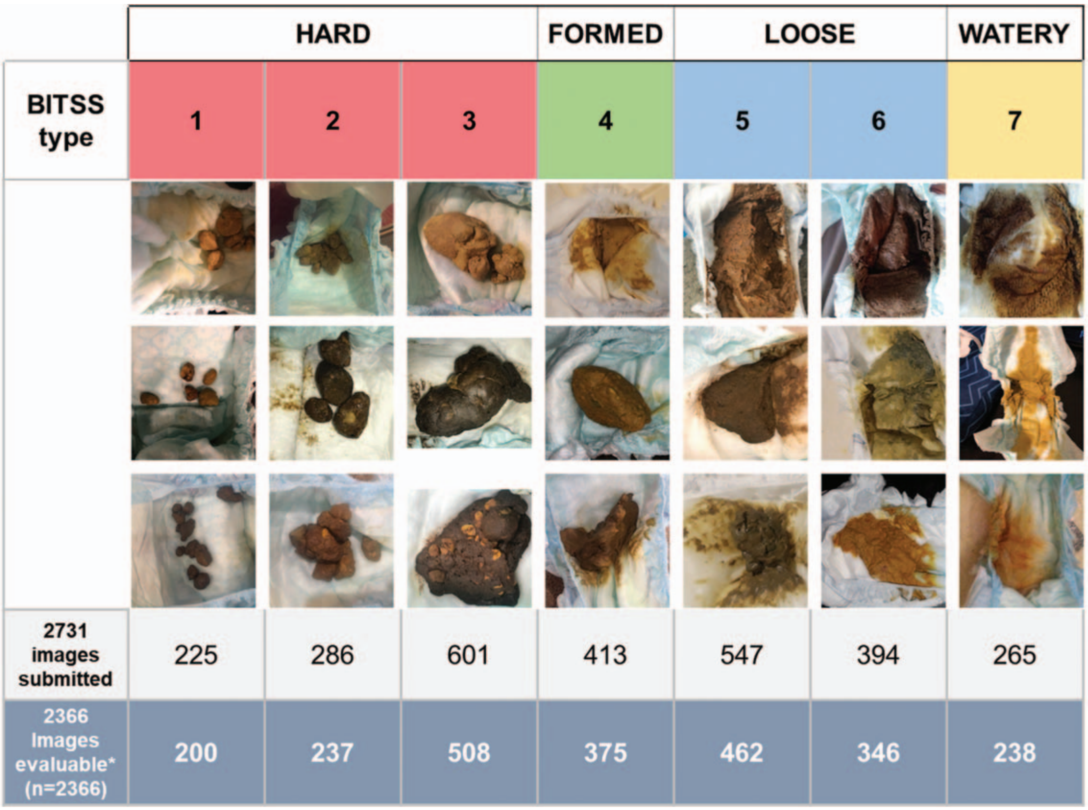
This questionnaire comprised 8 questions and assessed the usability of the app in this study. Six questions were on the ease of use for photo-taking, app malfunction, participation in other clinical studies, use of the app, and scoring of stool consistency. Two questions were whether an automated tool to score stool photos would be useful if a doctor wanted to know about their child's stool type, and whether their child's stool pattern is, or was, a concern to them.

Machine Learning

The above described collection of labeled photos was used as input for the ML model. In cases of scoring disagreements between the 2 independent researchers' scores, a final stool consistency score was assigned by a health care professional.

Significant progress in the field of image classification has been made possible by the availability of large annotated datasets and advances in deep convolutional neural networks. Due to the limited size of the stool photo dataset (>100-fold fewer images than datasets typically used to train deep neural network models for image recognition), it was not feasible to consider complete training of such a model for stool consistency prediction. Instead, our proof-of-concept project assessed the feasibility of adapting an existing model for classifying images (MobileNet (17), an efficient convolutional neural network for mobile vision applications) for the stool photo classification task. This transfer learning approach assumes that the selected source domain (images on which the existing model was trained) is sufficiently related to the target domain (stool photos).

MobileNet's streamlined architecture uses depthwise separable convolutions to build light-weight deep neural networks (17), and consists of a set of stacked layers, ending with the final classification layer. We took advantage of knowledge learned by the MobileNet model, which was pretrained using the ImageNet



	HARD			FORMED	LOOSE		WATERY
BITSS type	1	2	3	4	5	6	7
2731 images submitted	225	286	601	413	547	394	265
2366 Images evaluable* (n=2366)	200	237	508	375	462	346	238

FIGURE 1. Example photos submitted for each BITSS stool type. The 7 BITSS types are grouped into 4 categories: hard (1–3), formed (4), loose (5–6), and watery (7). *365 photos were excluded: 56 photos of poor quality (blurry, blank, duplicate photo, heterogeneous lighting, background issues); 9 photos submitted by ineligible/withdrawn participants; 300 pre-enrolment screening photos. BITSS = Bristol Infant and Toddler Stool Scale.

(18) database of more than 1 million images, and re-trained the classification layer to recognize the BITSS score from a stool photo. We modified an initial workflow (available at the TensorFlow for Poets git repository (19)) that uses an open source python library (TensorFlow, v.1.71) to perform *k*-fold crossvalidation (where $k = 10$), maximizing the use of the limited training dataset. Each step used *k*-1 folds from the training set, reading the associated layer values and feeding them into the classification layer. At each step, the model's prediction was compared with the final researcher's BITSS score, and the results of this comparison were used to update the final layer's weights through backpropagation.

Only the final classification layer of our model was re-trained; all other layers of the network were kept "frozen," retaining the optimal performing values and configurations defined during the initial training phase on the ImageNet database. Each stool photo had to be read several times in each fold of the crossvalidation, and the values of the layers before the classification layer had to be generated multiple times with the same output. To optimize calculation times and overcome this bottleneck of regenerating layer values, the values representing the last layer preceding the classification layer were generated upstream and cached.

The training was performed on a secured cloud services platform (Amazon Web Services, Instance Amazon EC2, p3.2× large). This training used a learning rate of 0.001 and ran 650 epochs, the threshold where the model reached a plateau. Cross-entropy was used as the loss function for model optimization and the F1 measure was the metric used to assess model performance (20). The customized model (poobao_v1.1) based on Mobilenet_v1 was trained on 2478 labeled photos chosen by random sampling, and a holdout set of 209 photos was used to assess final model performance. The code based on the TensorFlow for Poets script using Python (2.7.12) and TensorFlow (v.1.71) is available at *https://github.com/poobao/stool_consistency*. The repository contains the updated re-training script used to train the Poobao v1.1 model.

Statistical Analyses

No formal sample size calculation was performed. The estimated number of photos needed per class was based on tests with a previous ML photo recognition task. The primary objective of the study was to create a database with at least 200 photos per BITSS type (1–7) and 200 photos of empty diapers. The recruitment target of 100 participants was estimated based on the number of photos required, assuming a withdrawal rate of 15%.

For each evaluable photo, the participant's score was compared with each independent researcher's score for agreement, calculated using percentage agreement (exact match between scores) and mean level of disagreement. The mean level of disagreement was defined as the mean of the difference in scores, when disagreements occurred. For instance, given a participant score of 2 and independent researcher score of 4, the magnitude and direction of disagreement for that photo is -2.

Bubble plots were created to visualize the level of agreement between participants and each independent researcher and the 2 independent researchers. Bubble sizes were directly proportional to the number of score matches (agreement).

Frequency and percentages were calculated for eligible questions in the End of Study Usability Questionnaire. Free text responses for questions and subquestions were summarized by listing answers/answer themes and calculating the frequency of responses where applicable.

RESULTS

Participants

Out of 113 parent-child pairs screened, 96 enrolled and 91 participating pairs completed the study (Fig. 2A); 76 unique pairs completed their first 30-day study period, and 15 completed their second, third, or fourth 30-day study period. Due to scarceness of photos of hard stools, participants who predominantly submitted such photos during the 30-day period were offered to re-enroll for additional 30-day study periods.

There were 5 withdrawals during the study (4 during their first 30-day study period, 1 during their second 30-day study period). No children were withdrawn due to illness lasting longer than 7 days, surgery or hospitalization for more than 24 hours.

The mean age of the children at enrolment was 13.3 months (range 0–24 months) and 52% were girls. The mean age of the participating mothers was 30 years (range 21–48 years).

Photos

A total of 2731 photos of stools were submitted via the app (Fig. 2B). Figure 1 shows the distribution of submitted photos across the 7 BITSS types, as scored by the study participants. As expected, since this study was conducted on healthy infants, more photos were submitted for types 3, 4, 5, and 6 than for types 1, 2, and 7. Pre-enrolment screening photos (300 photos) and those submitted by ineligible/withdrawn participants (9 photos) were excluded. A further 56 photos were excluded due to poor quality, resulting in a final set of 2366 evaluable photos provided by 91 participant pairs.

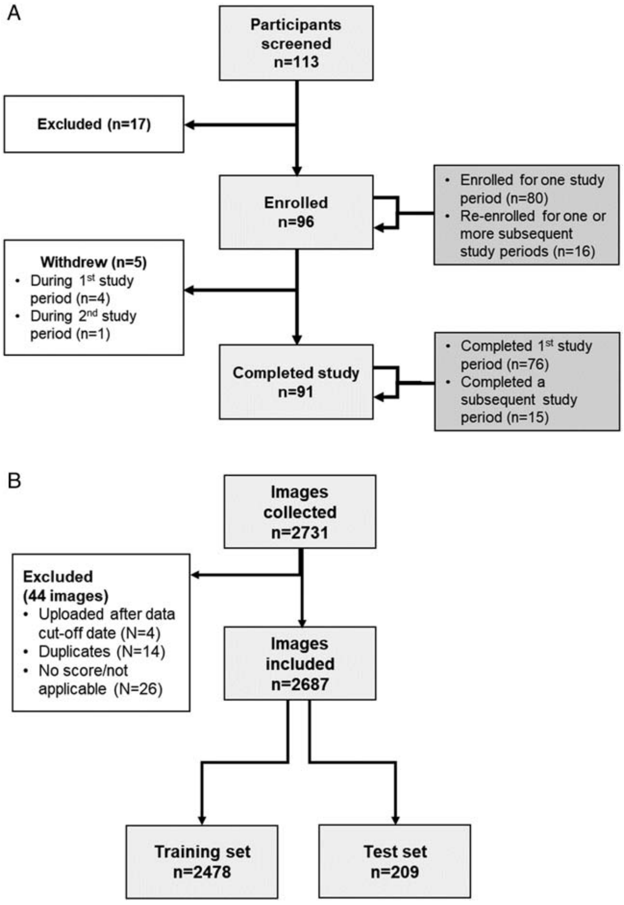

A
Participants screened n=113
Excluded (n=17)
Enrolled n=96
• Enrolled for one study period (n=80)
• Re-enrolled for one or more subsequent study periods (n=16)
Withdrew (n=5)
• During 1st study period (n=4)
• During 2nd study period (n=1)
Completed study n=91
• Completed 1st study period (n=76)
• Completed a subsequent study period (n=15)
B
Images collected n=2731
Excluded (44 images)
• Uploaded after data cut-off date (N=4)
• Duplicates (N=14)
• No score/not applicable (N=26)
Images included n=2687
Training set n=2478
Test set n=209

FIGURE 2. A, Study participant flow. B, Image flow for machine learning.

Scoring by Study Participants Versus Independent Researchers

The level of agreement between study participants and independent researchers was 58.0% (standard deviation [SD] 14.9%) (independent researcher 1; Fig. 3A) and 48.5% (SD 11.8%) (independent researcher 2; Fig. 3B). The mean level of disagreement with the participant ratings was −0.33 (SDs 1.14 and 1.23) for both independent researchers, meaning that both independent researchers scored photos 0.33 points lower than participants, on average.

Agreement with independent researcher 1 was highest for photos that participants scored as BITSS type 7 (75.2%), whereas for independent researcher 2, agreement was highest for photos rated by participants as BITSS type 4 (63.2%).

Agreement between the independent researchers was moderately high (77.5%, SD 5.8%), with a mean level of disagreement of 0.21 (SD 1.08) (Fig. 3C). On average, independent researcher 2 scored 0.21 points higher than independent researcher 1 when the scores were not the same. The BITSS types with the highest level of agreement between the independent researchers were types 1 (89.9%), 4 (80.6%), and 7 (79.5%).

Human Scoring Versus Machine Learning Model Predictions

The metric used to assess the performance of the model has been detailed elsewhere (20). Figure 4A summarizes the procedure used to train and test the ML model. Ten 10-fold crossvalidations were performed and the average F1 measure obtained on the 2478 training photos was 57.5% (SD 3.2%) Training-validation loss curves (Supplementary Fig. S1, *http://links.lww.com/MPG/C80*) indicated an acceptable degree of fit using the model obtained with a final learning process on the 2478 training photos (learning rate 0.001). Of the holdout test set containing 209 photos, this final trained model classified 60.3% in exact agreement with the researchers' final scoring. The individual F1-measure per BITSS score type is shown in Supplementary Table S1, *http://links.lww.com/MPG/C81*, and indicates similar performance across all 7 BITSS types.

A confusion matrix for the final model's predictions on the holdout test photos is shown in Figure 4B. If the 7 BITSS types were grouped into the 4 classes identified in the validation study (hard stools: types 1–3; formed: type 4; loose: types 5 and 6; watery: type 7) (4,15), the agreement between model-predicted and researcher's final scores was 77.0%.

Smartphone App: Ease of Use and Photo Scoring

Of 87 participants who completed the ease-of-use questionnaire, 89.6% answered ''easy'' or ''very easy'' to the overall ease of use of the application, with 95.4% answering that it was easy to take the photos, and 89.7% responding that it was not difficult to score the stool photos. Of those responding that it was difficult to score

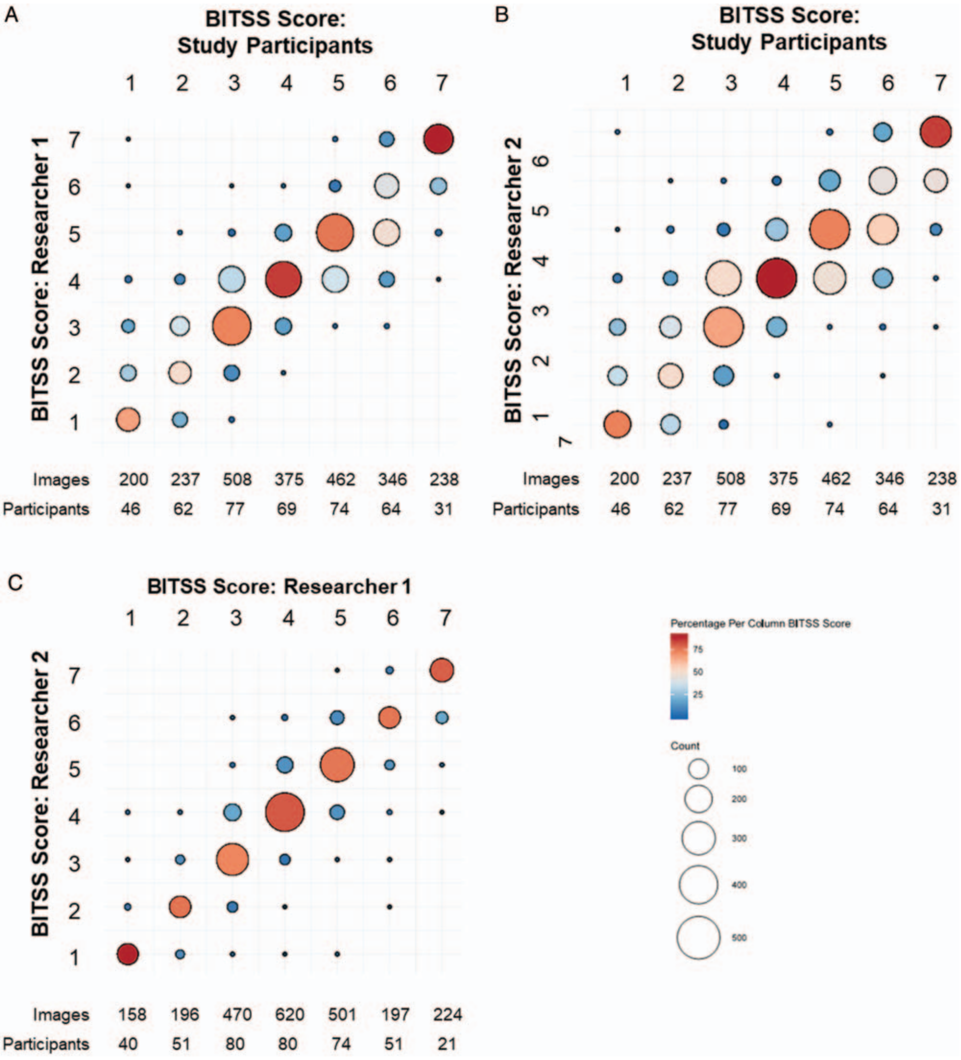


FIGURE 3. Bubble plots depicting agreement between participants' and the 2 independent researchers' scoring of photos across the 7 BITSS types. Bubble sizes were directly proportional to the number of score matches (agreement). The percentage of matches per BITSS type along the *x*-axis is depicted using a color gradient. A darker red color represents a higher percentage of scored photos among total images scored by participants at a given BITSS score (a score on the *x*-axis), whereas a darker blue color represents a lower percentage. BITSS = Bristol Infant and Toddler Stool Scale.

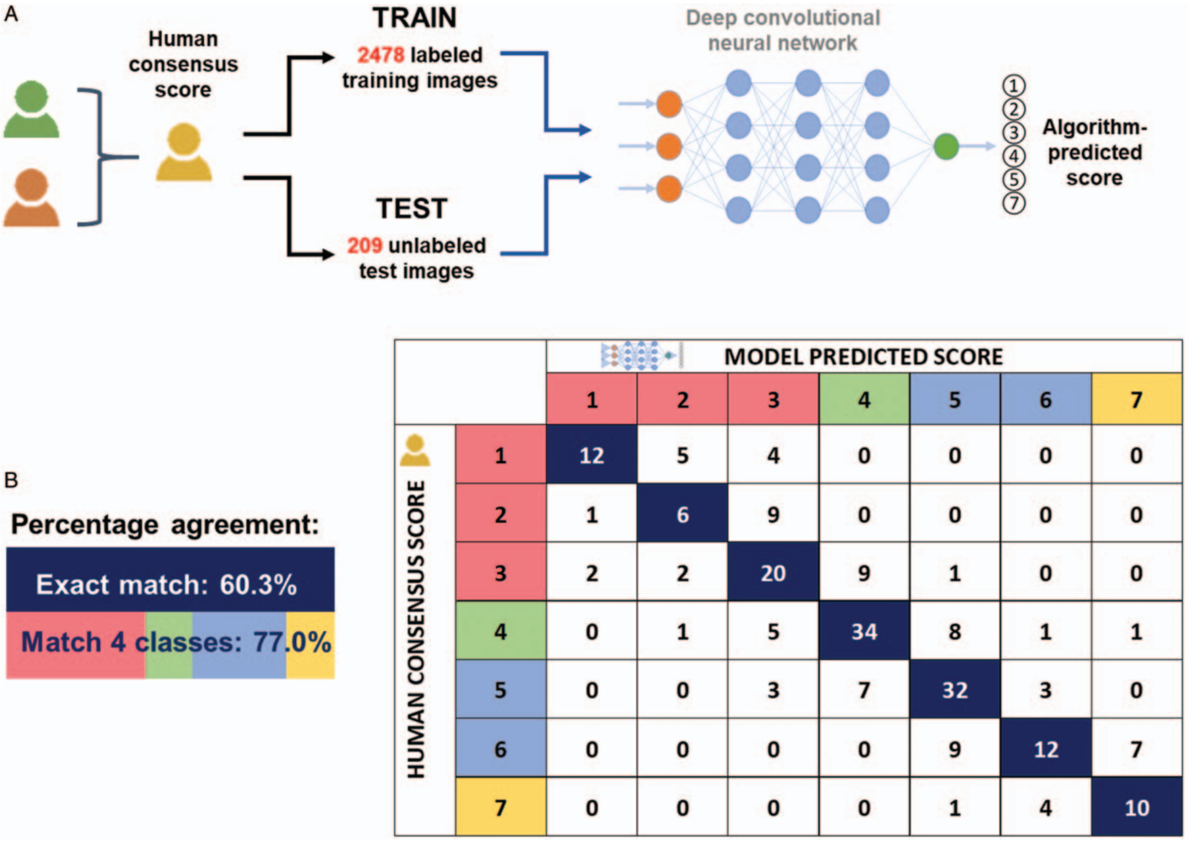
the photos, most reported it was ‘‘because sometimes the stool seemed in between 2 examples’’. On whether it would be helpful to have an automated tool to score photos if the doctor wanted to know about their child's stool type, 90.8% responded ‘‘yes.’’

DISCUSSION

In this work, we demonstrate the feasibility of using ML-based image recognition in a clinical application relevant to a very broad population: diaper-wearing children. The assessment of stool consistency is a source of worry for many parents and a topic frequently discussed during healthcare provider consultations. We developed an easy-to-use mobile device–based workflow that facilitates collection of photos and manual annotation, and used it to establish a photo collection of used diapers. The photo dataset was used to generate a novel ML model running on users' smartphones for the assessment of stool consistency in healthy infants/toddlers. Finally, the model's performance was evaluated alongside that of human raters in predicting infant/toddler stool consistency from photos, a nontrivial task even for clinicians (4).

The use of ML for image recognition has been explored in a number of specialized areas in medicine including gastroenterology (21), for example, for distinguishing pathological versus healthy features in duodenal tissue (22). Other applications range from classification of skin lesions (23) to glaucoma diagnosis based on fundus images (24). Many applications have relied on specialized or invasive procedures (biopsy, endoscopy), or standardized imaging techniques and acquisition protocols (radiology, microscopy). In contrast, the use of smartphone photos, which are typically captured under highly variable conditions (resolution, lighting exposure, orientation), is less well explored in medical image recognition.

Deep learning methods have allowed important and rapid progress in the field of image recognition but have historically required massive datasets and computing power to perform well. This study addressed the challenge of automating a specific novel image recognition task, namely predicting stool consistency from a



	1	2	3	4	5	6	7
1	12	5	4	0	0	0	0
2	1	6	9	0	0	0	0
3	2	2	20	9	1	0	0
4	0	1	5	34	8	1	1
5	0	0	3	7	32	3	0
6	0	0	0	0	9	12	7
7	0	0	0	0	1	4	10

FIGURE 4. Machine learning for recognition of stool consistency from photos. A, Each photo was scored independently by 2 researchers. In cases of scoring disagreement between the independent researchers, this was resolved by a health care professional, to derive the final score or label for that photo. Labeled photos were used to train a deep convolutional neural network to recognize the BITSS scores of stool photos. B, Agreement between model-predicted and researcher's final BITSS scores for the test set of 209 stool photos. The 7 BITSS types are grouped into 4 categories: hard (1–3), formed (4), loose (5–6), and watery (7).

smartphone photo. This presented 2 major constraints. The first was a relatively small dataset of only 2687 photos, compared with hundreds of thousands of images typically used to train deep learning models from the ground up, such as inception models trained on the ImageNet dataset (18). The second constraint was the requirement to perform the inference step (predict stool consistency) on the user's mobile device in an offline manner within a reasonable run-time.

As a starting point, we chose a MobileNet model with a computationally efficient architecture optimized to run on mobile devices (17). Transfer learning was used to re-train the classification layer of the model, applying knowledge learnt from previous tasks to the new task of predicting stool consistency, where limited labeled data were available. This allowed the final model to be built using a small-sized training set of around 2500 smartphone photos taken by participants, without complex preprocessing, image segmentation, or domain-specific feature extraction.

On the holdout 209-photo test set, agreement between the independent researchers' final scores and the model's predictions was 60.3% across the 7 BITSS types, in approximately the same range as the observed agreement between independent researchers and study participants (48.5%, 58.0%). The model presented here was intentionally developed on the initial order of 7 BITSS types to retain flexibility for future adaptations of the scale, and potentially other stool scales. To this end, we also explored performance across higher-level grouping into 4 stool classes. Based on a multicenter, cross-sectional study, the original 7 color photos of the BITSS were grouped into a 4-category stool scale (4,15). If this was considered, the agreement between the algorithm and human raters was 77.0%. As a comparison, only 65% of photos were correctly classified by parents in the BITSS validation study (4). Even expert clinicians classified only 85% of the photos correctly, indicating the difficulties inherent in scoring. The BITSS scores consistency of stools, a criterion which study participants may have found it hard to limit ratings to, given the variety of other stimuli (color, smell) present. Differences in setting and timing of ratings, whereby participants rated photos 1 at a time versus independent researchers who scored batches of photos sequentially, may have affected scoring patterns. Compared with participants, the independent researchers each scored considerably more photos (>2000), although this does not eliminate all potential biases, this may have contributed to their more consistent and similar scoring patterns. The consistent scoring pattern of both researchers (Fig. 3) with a mean level of disagreement of 0.21 across all images also implies that the settling of the final stool consistency score by the healthcare provider was potentially less critical than initially expected in the design of the study.

Standardized scoring tools not only facilitate direct comparison of results from different clinical studies but, in real-world settings, can also provide parents and health care professionals with common categorizations and nomenclature for communication. From the results presented here, 89.6% of participants felt it was easy to use the smartphone app to photographically document their child's stool and score stool consistency using a visual scale, and 90.8% felt an automated scoring tool would be helpful to support discussions with their doctor. This implies that technological support for the automated and objective assessment of stool consistency could improve the exchange of information between health care professionals and parents. Most participants could submit stool photos and scores regularly using the app, supporting the usability of the platform for studies or home assessment.

The results of our pilot study show that automated assessment of stool consistency from smartphone photos is feasible and yields moderately high agreement with human raters. This method enables the quick and efficient re-analysis of complete datasets (eg, from clinical studies), which would not be possible with manual methods. For future use, the existing database of our study could readily be expanded with photos submitted for automated classification. Model performance could also be improved with an expanded photo database. It would also be of interest to validate the model by comparing its performance with that of expert clinicians.

Importantly, this study also illustrates possibilities for using smartphones to rapidly generate and annotate photo datasets relevant for clinical use. Standardized assessments could be recorded and tracked easily over time, to examine the interplay of nutrition and stool consistency in large population samples. As an example, there are currently very few studies on the physiological stool consistency of infants (25). In the future, real-time access to automated stool scoring could provide reassurance to concerned parents, or objective scores a physician can review before or during a consultation.

Acknowledgments: Manuscript drafting and editorial support were provided by G. Toh (Tech Observer Asia Pacific).

REFERENCES

1. Morley R, Abbott RA, Lucas A. Infant feeding and maternal concerns about stool hardness. *Child Care Health Dev* 1997;23:475–8.
2. Mahon J, Lifschitz C, Ludwig T, et al. The costs of functional gastrointestinal disorders and related signs and symptoms in infants: a systematic literature review and cost calculation for England. *BMJ Open* 2017;7:e015594.
3. Iacono G, Merolla R, D'Amico D, et al. Gastrointestinal symptoms in infancy: a population-based prospective study. *Dig Liver Dis* 2005;37:432–8.
4. Huysentruyt K, Koppen I, Benninga M, et al. The Brussels Infant and Toddler Stool Scale: a study on interobserver reliability. *J Pediatr Gastroenterol Nutr* 2019;68:207–13.
5. Kuizenga-Wessel S, Benninga MA, Tabbers MM. Reporting outcome measures of functional constipation in children from 0 to 4 years of age. *J Pediatr Gastroenterol Nutr* 2015;60:446–56.
6. Oozeer R, Van Limpt K, Ludwig T, et al. Intestinal microbiology in early life: specific prebiotics can have similar functionalities as human-milk oligosaccharides. *Am J Clin Nutr* 2013;98:561S–71S.
7. Scholtens PA, Goossens DA, Staiano A. Stool characteristics of infants receiving short-chain galacto-oligosaccharides and long-chain fructo-oligosaccharides: a review. *World J Gastroenterol* 2014;20:13446–52.
8. Takagi T, Naito Y, Inoue R, et al. Differences in gut microbiota associated with age, sex, and stool consistency in healthy Japanese subjects. *J Gastroenterol* 2019;54:53–63.
9. Vandeputte D, Falony G, Vieira-Silva S, et al. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* 2016;65:57–62.
10. Koppen IJN, Velasco-Benitez CA, Benninga MA, et al. Using the Bristol Stool Scale and parental report of stool consistency as part of the Rome III Criteria for functional constipation in infants and toddlers. *J Pediatr* 2016;177:44.e1–8.e1.
11. Bekkali N, Hamers SL, Reitsma JB, et al. Infant Stool Form Scale: development and results. *J Pediatr* 2009;154:521.e1–6.e1.
12. Lane MM, Czyzewski DI, Chumpitazi BP, et al. Reliability and validity of a modified Bristol Stool Form Scale for children. *J Pediatr* 2011;159:437.e1–41.e1.
13. Saps M, Nichols-Vinueza D, Dhroove G, et al. Assessment of commonly used pediatric stool scales: a pilot study. *Rev Gastroenterol Mex* 2013;78:151–8.
14. Ghanma A, Puttemans K, Deneyer M, et al. Amsterdam Infant Stool Scale is more useful for assessing children who have not been toilet trained than Bristol stool scale. *Acta Paediatr* 2014;103:e91–2.
15. Vandenplas Y, Szajewska H, Benninga M, et al. Development of the Brussels Infant and Toddler Stool Scale ('BITSS'): protocol of the study. *BMJ Open* 2017;7:e014620.
16. Berthold AA, Levy EI, Hofman B, et al. Real time versus photographic assessment of stool consistency using the Brussels Infant and Toddler Stool Scale: are they telling us the same? Pediatr Gastroenterol Hepatol Nutr 2020 (in press).
17. Howard AG, Zhu M, Bo C, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. *https://arxiv.org/abs/1704.04861*. Accessed June 3, 2020.
18. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vision* 2015;115:211–52.
19. TensorFlow. TensorFlow for poets 2. *https://github.com/googlecodelabs/tensorflow-for-poets-2*. Published June 2020. Accessed June 3, 2020.
20. Chawla NV. Data mining for imbalanced datasets: an overview. Data Mining and Knowledge Discovery Handbook. New York, USA: Springer US; 2009:875–886.
21. Le Berre C, Sandborn WJ, Aridhi S, et al. Application of artificial intelligence to gastroenterology and hepatology. *Gastroenterology* 2020;158:76.e2–94.e2.
22. Syed S, Al-Boni M, Khan MN, et al. Assessment of machine learning detection of environmental enteropathy and celiac disease in children. *JAMA Netw Open* 2019;2:e195822.
23. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
24. Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmol* 2019;137:1353–60.
25. Vandenplas Y, Ludwig T, Bouritius H, et al. Randomised controlled trial demonstrates that fermented infant formula with short-chain galacto-oligosaccharides and long-chain fructo-oligosaccharides reduces the incidence of infantile colic. *Acta Paediatr* 2017;106:1150–8.