Artificial intelligence for the practical assessment of nutritional status in emergencies

Ben Watkins 1 | Lameck Odallo 2 | Jenny Yu 3

1Department of International Health and Sustainable Development, School of Public Health and Tropical Medicine, Tulane University, New Orleans, USA
2International Department, American Institutes for Research, Nairobi, Kenya
3Tulane University, School of Public Health and Tropical Medicine (International Health and Sustainable Development), New Orleans, Louisiana, USA

Correspondence
Ben Watkins, Department of International Health and Sustainable Development, School of Public Health and Tropical Medicine, Tulane University, 1440 Canal Street, Suite 2210, New Orleans, LA 70112 504-988-539, USA. Email: ben.watkins@open.ac.uk

Funding information
UNICEF

Abstract
This paper describes a novel method for detecting child malnutrition based on artificial intelligence and facial photography. Estimates of severe and moderate acute malnutrition in children are critical for rapid emergency responses. However, the two traditional measurement methods, mid-upper arm circumference (MUAC) and weight-for-height (WFH), are impractical in conflict and catastrophic disaster situations. They require well-trained enumerators, cumbersome equipment, and close supervision. The Method for Extremely Rapid Observation of Nutritional Status (MERON) addresses the problem, using simple facial photographs. Facial features are extracted to predict Body Mass Index (BMI) in adults and Weight for Height Z Score (WFHZ) in children under five. MERON correctly predicts adult BMI classification with 78% accuracy. A variant of the model, trained on a sample of 3167 children in Kenya, successfully classified 60% of cases. On most measures, MERON was easier and more culturally acceptable to use than the traditional measurement methods. If MERON were to be trained and validated on a larger sample, with more extreme cases, it would provide a practical solution to a recurrent humanitarian problem.

KEYWORDS
acute malnutrition, artificial intelligence, data collection, emergency nutrition, machine learning, mid-upper-arm circumference, survey methods

1 Introduction

This paper describes the MERON model, which is an AI tool for the detection of acute malnutrition in children. MERON addresses the need for a practical, lightweight, and culturally acceptable alternative to current measurement methods, which are error prone and inconvenient in poorly supervised or insecure settings. Our primary contribution is to show how AI, and specifically facial detection/recognition technologies, can be used as a lightweight predictor of malnutrition, based on simple facial snapshots of children. We believe it is the first attempt at using facial images to detect undernutrition in children. The MERON architecture includes tight integration with Open Data Kit (ODK) tools for field data capture, and an API that returns a diagnostic classification based only on the facial photo. The model was tested using data from a large-scale nutritional survey in Kenya and compared to traditional methods of anthropometric measurement that entail taking the child’s physical measurements.

The work was conducted while the researchers were employed at Kimetrica LLC.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2024 The Authors. Expert Systems published by John Wiley & Sons Ltd.

https://doi.org/10.1111/exsy.13550
This paper describes the technology and reports on the MERON field testing and performance from the perspectives of predictive accuracy, practicality, cost, and cultural acceptability.

A UNICEF/WHO/World Bank study in 2017 found that nearly 51 million children under the age of five were wasted and 16 million were severely wasted. Children suffering from wasting have weakened immunity are susceptible to long-term developmental delays and face an increased risk of death, particularly when wasting is severe. They require urgent feeding, treatment, and care to survive (UNICEF/WHO/World Bank Group, 2018). Interventions to tackle malnutrition require timely and accurate information on child malnutrition status. The common anthropometric measurements used in emergencies to help determine undernutrition are Middle Upper Arm Circumference (MUAC), Weight for Height/Length (WFH), Weight for Age (WFA), and Height/Length for Age (HFA) Z scores. However, the current measurement techniques are cumbersome and require skilled personnel.

Obtaining accurate weight and length measurements is challenging, especially when the child is wriggling and upset. Appendix A provides a summary of the current measurement protocols, based on UNICEF guidance, for measuring weight, height, length and MUAC. The logistical and practical challenges of anthropometric measurement are associated with high measurement error. According to a study on severe malnutrition in hospitalized children in rural Kenya (Berkley et al., 2005), weight for height/length measurement is often problematic and not undertaken in practice because of a shortage of properly calibrated and functioning scales. WFH measurement protocols require correct recording of two separate values and then looking up a third value on a chart, potentially introducing transcription errors. It is also difficult to measure height in children who are ill or stressed in crowded wards. MUAC measurements are also prone to multiple errors. Even the choice of tape measure influences the results, as Rana et al. (2021) show. In some contexts, climate and/or cultural preferences make it inappropriate or unacceptable to weigh children without clothing. According to UNICEF protocols, a standard maximum of clothing should be agreed and standardized across children. UNICEF advises that an average estimated weight for clothing should be subtracted from all the weights measured, while recognizing that this will introduce some degree of error into the results. If the child is wriggling and the needle on the hanging scale does not stabilize, UNICEF’s protocols indicate that weight is estimated using the value situated at the midpoint of the range of oscillations (UNICEF, n.d.).

Experiments using trained medical staff or students and undertaken in clinical conditions find variation in MUAC and WFL Z score observations. Mwangome et al. (2012) assess the inter-observer variability and accuracy of MUAC and weight for length z-score (WFLZ) among infants less than 6 months old performed by community health workers (CHWs) in Kilifi District, Kenya. The study finds strong correlations between the MUAC observations of health workers and of trainers, but considerable variation in WFLZ measurements. Saeed et al. (2005) assess inter and intra-observer variability in MUAC measurements of under-five children by community health nurses (CHN) in the Northern Region in Ghana. They show a high degree of consistency in observations if the same student records observations on the same child over a three-day period, but high variation in observations between students on the same child.

In field survey conditions, higher measurement errors can be anticipated. Perumal et al. (2020), demonstrate “substantial heterogeneity” in anthropometric data quality across a review of international Demographic and Health Surveys (DHS), using a range of quality indicators. The DHS are typically conducted with high levels of training and supervision. It is particularly challenging to recruit, train, and supervise skilled survey personnel in conflict situations. There is a clear need for a lightweight and easy-to-use alternative.

Our research assesses a two-part hypothesis: (1) that sufficient features can be detected from facial photography to allow a characterization of facial morphology; and (2) that facial morphology is influenced by underlying changes in the distribution of body fat. Both hypotheses are broadly supported by the current literature.

Major advances have been attained in the methods for detecting facial features from photos and characterizing morphology, primarily for the purpose of facial recognition. In their seminal study on facial recognition, Lawrence et al. (1997) propose a hybrid neural-network solution in developing a computational model for facial recognition. The combination of local image sampling and a convolutional neural network allowed for rapid and consistently better classification performance than other existing facial recognition models at the time. Facial recognition technology has improved drastically since then, with growing datasets and deep learning techniques greatly improving performance. Zhou et al. (2015) show how larger datasets can improve facial recognition, but they also highlight the major challenges of data bias, and low false positive criteria in real-world applications.

A second strand of literature investigates the association between certain facial characteristics and an individual’s BMI score, although Windhager et al. (2013) note that the literature is not extensive. Coetzee et al. (2010) show a strong correlation between adult BMI and four facial features. Lee, Do, and Kim (2012) show that it is possible to predict BMI status using facial characteristics. Their model predicts normal and overweight females according to BMI using facial features and they identify more than 40 facial characteristics that significantly differ between normal and overweight individuals. In an expansion of this study, Lee, Jang, and Kim (2012) use facial characteristics extracted from subject image data to predict BMI for males and females. Wen and Guo (2013) predict BMI for subjects of different ages and ethnicity using facial features extracted the MORPH II dataset. A study comprising 11,347 adult Korean men and women ranging from 18 to 80 years old, concludes that features in the buccal area of the face were the best indicators of normality and visceral obesity in males and females (Lee & Kim, 2014).

Windhager et al. (2013) assess the effect of body fat on numerous facial features on a sample of 22 adolescent girls and find that body fat explains 8% of the variation in facial characteristics ($p = 0.047; 10,000 permutations$). They also find, as would be expected, a strong correlation between body fat and BMI, independent of age ($r = 0.864, p 0.001$). There is a strong theoretical basis for expecting a correlation between facial
morphology and body fat and hence BMI. Levine et al. (1998) suggest that an upper distribution of fat is more predictive of the metabolic complications associated with obesity than the actual degree of overweight. They find a significant positive correlation between the areas of buccal fat, or fat stored in the cheekbones, and visceral fat which is stored in the abdomen. Facial morphology, and specifically the curvature of the cheek, is therefore a potential indicator of body fat and adult BMI.

Much of the work has been focused on markers of attractiveness in adults, and research on child facial morphology is less common. However, Huang et al. (2015) create a predictive model of children's body weight using age and morphological facial features. The model uses a three-layer feed-forward artificial neural network model and an age-based, Centre for Disease Control inferred median body weight along with three facial feature distances derived from facial images. The model predicts a child's body weight with a correlation coefficient of 0.94 and a mean prediction error of 0.48.

MERON combines a classification model and a regression model which individually detect a set of optimized facial features and patterns. These features are constructed from a predetermined set of facial landmarks. The patterns are self-derived from the model, based on training data.

Section 2 summarizes the MERON method and includes a description of the data capture methods and fieldwork as well as a review of the MERON application architecture. Section 3 presents the results of the first round of testing including both statistical performance metrics and details of cultural acceptability and practicability. The paper concludes in Section 4, with recommendations for additional field research to improve reliability.

2 | METHODS

2.1 | Data

The research followed a two-stage process. First, we established whether it was possible to extract sufficient features from facial photography to classify BMI bands and estimate BMI. The second stage of the testing focused on whether the method could be applied to children under difficult field conditions. The methodology and data collection tools were designed to address the following research questions:

1. How does the accuracy of MERON compare with traditional measurements in a field setting?
2. How cost-effective is MERON compared with traditional measurements in a field setting?
3. How feasible is it to collect data required for MERON in different field settings?

A successful proof of concept was achieved based on the analysis of thousands of facial images of adults in North America using the MORPH database. The database was developed by the University of North Carolina Wilmington for researchers studying adult age-progression. Ricanek and Tesafaye (2006) provide a comprehensive overview of MORPH. While this database was helpful in the initial testing of MERON, a different sample was required for understanding if MERON could be applicable to children under five.

The second model was trained using images and anthropometric data from under-five children in food insecure areas of Kenya. Data were collected between January and February 2018 in the diverse counties of Isiolo, Turkana, Marsabit, and Tana River, each with a high risk of under-five undernutrition. In collaboration with the Ministry of Health (MoH) and UNICEF Kenya, Kimetrica took photographs of children's faces and concurrently assessed anthropometrics as part of the Standardized Monitoring and Assessment of Relief and Transitions (SMART) surveys. SMART is an interagency initiative that was launched in 2002 by a network of humanitarian organizations aimed at improving nutrition surveys in emergencies.

Testing MERON in the context of a SMART survey allowed us to evaluate the cost, feasibility and cultural acceptability of using the method in realistic field conditions. Evidently, the data were not of quality that could be expected from a more tightly supervised clinical setting. Data collection began once ethical clearances had been obtained and the MERON methodology had been validated by Kenya's National Nutrition Technical Working Group. Ethical clearance was obtained from AMREF's government-appointed Ethics and Scientific Review Board.

A Kimetrica team participated in each SMART survey to ensure that photographic data was captured correctly and could be associated with the anthropometric records. The team also observed the reactions of children and their parents to both methods. Ministry of Health supervisors carried out back-checks on 10% of the total sample with anthropometric measurements for quality control purposes. They also collected qualitative data on household perceptions and preference of photos versus the traditional methods.

Strict measures were enforced to protect child and parent privacy and to prevent malicious use of the photographs. The system only saves a vector that corresponds to semantic abstraction of the original image after it goes through computational operations (i.e., convolution, non-linearity function, etc.) and dimensionality reduction. This vector encoding cannot be used to reconstruct the pixel image of the raw photo. Vectorisation occurred shortly after data capture, allowing the image file to be deleted from all devices and servers. Parents accompanied their children for all photo taking.
SMART survey team leaders were trained on the MERON photo capture. The SMART survey data had to be synced daily, so the photographs were not embedded in the SMART survey form. Given the concern of syncing photos in low bandwidth environments, it was agreed to de-link the photos from the SMART survey forms, and to sync the photos to secure AWS servers on completion of each survey round, using encrypted file transfer. SMART survey enumerators and team leaders were trained on the simultaneous capture of SMART survey data and photographs to ensure the correct matching of the photograph and anthropometric data for each child. The SMART team leaders were trained by Kimetrica on the following photo capture requirements in Box 1.

Photos were captured concurrently by UNICEF/MoH enumerators through the KoBo photo capture form (https://kobo.kimetrica.com/enketo/x/#YYyu) and were then matched and linked using a Pandas script.

In total, 4075 images were captured. While all enumerators were comfortable using tablets and had been trained on the photo capture requirement, some data quality issues arose. In 908 images, faces were not detected for several reasons including:

- Facial detection model was unable to detect a face in an image because of low contrast;
- No face could be detected in an image;
- Face was out of focus in an image;
- Face was not aligned in an image causing partial obstruction of face.

For the cases of low contrast, we were able to recover most facial images using the method of Contrast Limited Adaptive Histogram Equalisation (CLAHE), which applies a transformation function to each pixel based on neighbouring pixels rather than the entire image. The transformation function is proportional to the cumulative distribution function of neighbouring pixels and adjusts the pixel intensity to enhance contrast. CLAHE limits the contrast amplification to reduce the amount of noise that is introduced. We were unable to address the other issues through data processing. However, the Kimetrica team subsequently designed an app to analyse image quality at the point of data capture, and upload only the images that are suitable for feature detection.

2.2 | MERON model architecture

MERON employs three distinct models to predict malnutrition. The first model uses VGGFace base architecture, which was pre-trained on 2 million face images, to extract facial features encoded in the vector form \( n = 2048 \). The technology selection criteria included performance, computational cost, and ease/cost of deployment. The VGGFace model was selected as the best performing lightweight option available at the time of MERON development. The base architecture offers several models and the team selected ResNet50. The output of VGGFace is a vector that corresponds to semantic abstraction of the original image after it goes through computational operations (i.e., convolution, nonlinearity function, etc.). This vector does not map to one specific facial feature but acts as an abstract representation of multiple features. A subsequent neural

---

**BOX 1  Photo capture requirements**

- **Distance from the subject:** The enumerator taking the photo must stand at a distance of at least 25 cm from the subject.
- **Frame of the photo:** The frame of the photo should capture the entire face, neck and shoulders.
- **Position of face:** The subject should be facing forward and looking straight at the camera. The eyes should be open.
- **Facial expression:** The subject should not smile and the mouth should be closed. The expression should be neutral, where possible.
- **Remove all obstruction:** The head should not be covered (no hats) unless it is for religious or medical reasons, no headgear, no jewellery, nothing should be covering the face. Remove any hair covering the face or the eyes. The photo must not contain any other objects or people.
- **Background:** Ideally, the background should be solid, preferably a wall, of any shade or colour. If in a room, the room should be well lit. Use the photo cloth provided.
- **Photo colour:** The photo must not be in black and white, it must be a colour photo.
- **Photo orientation:** Landscape not portrait.
- **Photo resolution:** The resolution must be at a minimum 1.5 megapixels.
- **Format:** The format must be in JPEG format.
- **Camera settings:** The camera should be on the following settings: macro, no flash, auto-focus and grid lines. With the grid line setting on, the screen is divided into nine squares. Enumerators should ensure that the photo is taken in the centre/middle.
network layer is trained on the feature vector combined with metadata to optimize overall prediction accuracy. It is open source and can be found at: https://github.com/kimetrica/MERON_model/blob/master/meron/model/single_model.py.

A second model reduces the dimensionality of the feature vectors \( n = 512 \) using an auto encoder. The third model is a fully connected neural network, which applies a trainable neural network layer to predict malnutrition status through a classification model and WHZ through a regression model. Figure 1 provides an overview of the MERON workflow.

Figure 2 represents the detailed schematic of VGGFace feature extraction. The input is the pixel array of the photo and passes through several layers of convolution and max pooling operations. The last pooling layer is merged with metadata and serves as the input vector for the Neural Network layer as shown in Figure 1.

As the input image progresses through the convolution and max pooling layers, the feature maps generated become increasingly higher-level abstractions. At the last max pooling layer, the feature vector is combined with subject metadata before being processed by the Neural Network layer where nonlinearity is applied to the input. The Neural Networks are optimized separately for the classification and the regression outputs. An API linked to the fully connected model has been built that allows the user to post an image, along with the required metadata. The API returns a WHZ-score and malnutrition category, and it can be found at: https://meron.kimetrica.com/.

3  |  RESULTS

3.1  |  Model results

We quantify the classification model performance by precision, recall and \( F_1 \) scores. Estimates are based on a testing dataset of 838 observations, from a total usable sample of 3167 observations (26%).

The precision is given by:

\[
\text{Precision} = \frac{tp}{tp + fp}
\]

FIGURE 1  Workflow of MERON to generate predicted values.

FIGURE 2  Neural network layer.
The recall is given by:

$$\text{Recall} = \frac{tp}{tp + fn}$$

The F1 score is given by:

$$F1 \text{ Score} = \frac{2 \times (\text{recall} \times \text{precision})}{(\text{recall} + \text{precision})}$$

True positive (tp) indicates that the model correctly predicts that the child belongs to an under-weight category from an image. tn (true negative) means that the model correctly predicts that the child is not under-weight, and false positives and negatives are the incorrect model predictions. The recall score measures the proportion of true positives that is correctly predicted. The precision score captures the proportion of all predicted positives that are true positives. The F1 score is a weighted average of the precision and recall. We take the true condition of malnutrition classification as indicated by the WFH or WFL Z-score. A WFH or WFL score greater than –2 is considered normal. If the WFH or WFL score is between –2 and –3 the individual is moderately acutely malnourished (MAM) while a score of less than –3 indicates that the individual is severely acutely malnourished (SAM). Our model results are compared to these classifications. The precision and recall scores are broken down for each malnutrition category in Table 1.

The results can also be shown through a confusion matrix (see Figure 3). The vertical axis represents the actual classification of the image while the horizontal axis represents model classification. If, for example, the model accurately classifies every single image, then all the off-diagonal terms are zero. The confusion matrix shows the number of observations in each category, from the total testing sample of 838 observations.

The results show that the model predicts the normal category with the highest accuracy followed by MAM. SAM accuracy is low, reflecting the distribution of the data used to train the model (see below). Image augmentation was applied to the under-represented MAM and SAM pictures to up-sample those categories for model training. Even after optimizing the model hyperparameters, the lack of sample data on MAM and SAM children was a constraint to accuracy. During model optimization, there was a trade-off between accuracy in the normal class and accuracy of the under-represented malnutrition classes. As the standard MUAC measurements were taken during the SMART surveys, we can compare our model with MUAC as a predictor of WFL/WFH classifications. Table 2 shows the precision and recall score for malnutrition classification using MUAC.

**TABLE 1**  Precision and recall scores using MERON.

<table>
<thead>
<tr>
<th>Score</th>
<th>Normal</th>
<th>MAM</th>
<th>SAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.78</td>
<td>0.30</td>
<td>0.1</td>
</tr>
<tr>
<td>Recall</td>
<td>0.60</td>
<td>0.37</td>
<td>0.27</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.68</td>
<td>0.33</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**FIGURE 3**  Confusion matrix.
While MUAC out-performs MERON in the prediction of normal values, MERON more accurately predicts MAM and SAM in this sample.

In addition to the classification model, we developed a regression model to predict the subject’s WFH or WFL, using the same input data. Unfortunately, it was not possible to optimize the model with the available time and resources. The scatter plot of predicted WFH versus actual WFH scores is shown in Figure 4.

The model $R^2$ value between MERON predicted and actual values is only 0.10 and a similar result was obtained predicting WHZ score from the MUAC.

Additional training data will improve the predictive skill of both the classification and the regression models. The primary limitation of the training data is that the proportion of MAM and SAM observations was low compared to previous surveys in the same areas. Large-scale nutrition interventions in the months prior to the sample survey appear to have been largely successful in reducing the prevalence of child under-nutrition, to well below the normal seasonal levels. Hence, the distributions were unusually clustered within the “normal” range. To improve tuning and predictive accuracy, more MAM and SAM observations are required. A minimum of 1000 images is necessary for training each class. In our sample, out of the total 3167 usable images collected only 117 images are associated with SAM children, while 471 correspond to the MAM children. The next round of testing will therefore require a larger sample with considerably more MAM and SAM cases.

Model performance may also be improved by including additional variables. For example, mobile applications can estimate the height of objects (e.g., Smart Measure and EasyMeasure). These types of applications can be used to estimate the height of a subject in a picture. Moreover, adding age and gender detection to the model is likely to improve performance. As mentioned, the MERON architecture allows for flexible inclusion of additional subject attributes/metadata.

### 3.2 Cost and feasibility

Kimetrica calculated the marginal cost of MERON as a diagnostic tool compared to anthropometry, based on cost data from the SMART survey in Tana River County sub-sample. Survey costs including vehicle hire, conference facilities, enumerator per diems, and equipment are largely the same across all four counties with only minor variations, so the findings are generally applicable to the sample as a whole.

The total cost of the anthropometry component of the SMART survey for 621 children sampled in Tana River was $16,808.12, implying a per observation cost of $27.70. The cost of administering MERON was calculated under two conditions: using a car and using a motorcycle. The cost of administering MERON in Tana River is $14.06 per child, implying a saving of $13.64 per sampled child.

Car hire and fuel generally often constitute the largest proportion of survey costs. Bulky anthropometric equipment cannot be transported without cars, whereas MERON requires only a smartphone, and enumerators are able to travel on motorcycle rather than by car.

<table>
<thead>
<tr>
<th>Score</th>
<th>Normal</th>
<th>MAM</th>
<th>SAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.97</td>
<td>0.12</td>
<td>0.06</td>
</tr>
<tr>
<td>Recall</td>
<td>0.75</td>
<td>0.34</td>
<td>0.30</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.85</td>
<td>0.18</td>
<td>0.1</td>
</tr>
</tbody>
</table>

**TABLE 2** Precision and recall scores using MUAC.

**FIGURE 4** Predicted WFH versus actual.
second major cost saving is in labour time as the traditional methods require two enumerators per observation, but MERON requires only one.

MERON saves costs primarily by reducing enumeration time. Based on Kimetrica’s experience conducting anthropometric assessments, it takes 12.5 min per observation on average to set up equipment, prepare a child, and take weight, height and MUAC measurements. It took approximately 129 h to assess the 621 children. In contrast, the MERON process took approximately 5 min per child and 52 h in total, a total time saving of 78 h. For this sample size, the hourly cost is $133, so the time savings are substantial. While the costs are dependent on the particularities of the sample area, anthropometric protocols are relatively standard and take a similar enumeration time everywhere. It is likely, therefore, that the time savings will be similar in other geographical settings.

3.3 | Cultural acceptability

Most caregivers in all four counties stated that it is culturally acceptable to take a photograph of their child to assess malnutrition. Figure 5 provides the response data.

In three out of four counties (Isiolo, Turkana and Marsabit) caregivers preferred photographs to traditional methods of assessing malnutrition. In Tana River only 17% of caregivers preferred photos. Data on the preferred method is summarized in Figure 6.

Caregivers stated they preferred photography over traditional methods because it is faster, involves no physical contact with enumerators, is less stressful for children, and it is less invasive and cumbersome. Respondents also mentioned that it is easier to take photo of disabled persons, the method is more hygienic and it is less likely to injure the baby than the suspended weighing scales. The main reason cited for preferring the traditional method was familiarity. It was also noted that with anthropometric methods, caregivers can immediately see the weight and length of the child, and this helps to track the progress of their child in terms of height and weight gain. Some respondents, especially in Tana River, could not understand or believe how a photo could determine the malnutrition status of their child.

4 | CONCLUSIONS

The preliminary results presented here suggest that the toolkit for rapid assessment of nutritional epidemiology can be enhanced by including an orthogonal metric, based on facial morphology from photography. The method is inexpensive—approximately half the cost per observation of...
WHZ and MUAC—and feasible in a wide range of field conditions. In the context in which it was tested, it also proved to be culturally acceptable. It does not require the skills of clinically trained enumerators and is lightweight and therefore potentially applicable in conflict areas and other environments where conventional anthropometry may be too conspicuous, too cumbersome, or dependent on scarce skills. The field research also yielded important insights into the architecture of the MERON software. Notably, we learnt that rigorous field level validation of photos was required, for which an app was developed.

Using WHZ and MUAC measures from 1832 surveys in 47 countries, Grellety and Golden (2016) demonstrate that the two methods lead to different children being classified as under-weight and in need of intervention. They conclude that neither method should be used exclusively. With adequate investment in training data, MERON could be included as a third and orthogonal indicator in conventional SMART and DHS surveys, to further reduce exclusion errors.

MERON is successful in BMI prediction for adults, but malnutrition classification for children was only moderately successful. Continuous WHZ predictions were weak, and equivalent to MUAC. This result can be attributed to the small training sample and to the relatively small number of MAM and SAM observations in the sample. In our view, however, the initial results are encouraging enough to test and refine the approach on a larger sample. Under clinical conditions, several studies have found a strong relationship between facial features and BMI for various age-groups and there is no inherent reason why, with a larger sample size, the precision cannot be augmented under field conditions.

AUTHOR CONTRIBUTIONS
BW led the overall methodological design, wrote the research proposals, and contributed to the drafting and review of the manuscript. LO provided research assistance, contributed to the literature review and to later drafts. JY undertook substantial data analysis and worked on training the MERON model.

ACKNOWLEDGEMENTS
While Eric Nussbaumer was working at Kimetrica, he led the development of the machine learning models and trained the various iterations of the MERON model. Eric was also responsible for leading the integration of the model with the Application Programming Interface (API) and compiling draft technical documentation. Anita Shah of Kimetrica managed the field work and early drafting of the results. Jesaja Everling and Hussein Lightwalla of Kimetrica were highly instrumental in developing the tools for field data management and for building the MERON API. Grainne Moloney of UNICEF provided an invaluable input to coordination to the field work and has been an outstanding advocate. Nancy Mock of Tulane University provided pertinent comments on draft versions, as well as insights during the early design stages. We would like to extend our thanks to colleagues at the Kenya Ministry of Health and AMREF for making the research possible. Melissa Earl provided additional research assistance and useful comments.

FUNDING INFORMATION
Funding for the development and training of the second MERON model and API, data collection in Kenya and the publication of this article was provided by the UNICEF Innovation Fund. Funding for the underlying architecture, first MERON model and support to the field testing of MERON in Kenya was provided from Kimetrica R&D funds.

CONFLICT OF INTEREST STATEMENT
The authors declare that they have no conflicting interests.

DATA AVAILABILITY STATEMENT
The datasets generated and/or analysed during the current study are not publicly available. Two key datasets were used: photographs of children and anthropometric data. A requirement under the ethical clearance obtained from the Kenyan authorities for the research precluded the use of photographs of children beyond training of the MERON model. All photographs had to be destroyed once the model was trained. The anthropometric data collected during the SMART surveys was provided to Kimetrica by the Ministry of Health, again for the sole purpose of training the MERON model. This data would have to be requested by any third party directly from the Ministry of Health.

INFORMED CONSENT
All caregivers signed informed consent forms before photographs of their children were taken. The ethical clearance applied to all the survey areas and the application clearly specified that cultural acceptability would be assessed in the course of the fieldwork.

ORCID
Ben Watkins https://orcid.org/0000-0002-7307-3433
REFERENCES


AUTHOR BIOGRAPHIES

**Ben Watkins** holds an MSc from Oxford University in Agricultural Economics. After many years working in international development and humanitarian operations, he is now an Adjunct Professor at Tulane University, School of Public Health and Tropical Medicine, and Chief Research Officer at Insight Commons. Much of his current research is on the applications of artificial intelligence to public choice.

**Lameck Odallo** holds a Masters in Public Health from BRAC James P Grant School of Public Health, Bangladesh. He currently works as a Senior Technical Assistance Consultant with expertise in the design and implementation of Third Party Monitoring Systems in non-permissive environments. His research interests are in predictive modeling of malnutrition and impact evaluation of food security interventions in post-conflict regions.

**Jenny Yu** received her MSc in Bioengineering from the University of Utah. She is an experienced data scientist currently working at Snapdocs, applying AI and machine learning for process automation. Her professional interest centers on machine learning, MLOps, and statistical analysis.
APPENDIX A: WEIGHT FOR HEIGHT AND MUAC MEASUREMENT PROTOCOLS

A height board is used to measure length or height to the nearest 0.1 cm. Children under 24 months (or up to 87 cm in height) are measured lying down. Children aged 24–59 months (or 87 cm and above) are measured standing up. To measure length, the board is placed on a flat, stable surface. The child is placed on the board, with the head against the fixed vertical part, and the soles of the feet near the cursor or moving part. The child lies straight in the middle of the board, looking directly up. The assistant holds the child's head firmly against the base of the board, while the measurer places one hand on the knees to keep the legs straight and places the child's feet flat against the cursor with the other hand. The measurer reads and announces the length to the nearest 0.1 cm. The assistant repeats, verifies the measurement, and then records it.

For height, the measurer removes the child's shoes and any hair ornament or hair knots. The measuring board is mounted at a right angle between a level floor and against a straight, vertical surface. The child stands in the middle of the measuring board. An assistant firmly presses the child's ankles and knees against the board. The child looks straight ahead with his/her head parallel to the baseboard. Keeping the head in place, the headboard is pulled down to rest firmly on top of the head, compressing the hair. The measurer reads the measurements to the nearest 0.1 cm and the assistant then verifies the measurement before recording it.

Weight is measured to the nearest 100 g. The most commonly used scale is the hanging spring balance, which can hold weight up to 25 kg. Weight is also measured using an electronic scale such as UNICEF’s UNISCALE, which is more precise and allows a child to be measured in the mother/caregiver’s arms.

Before weight measurements are taken the procedure is explained to the caregiver and the scale is prepared. A mobile weighing scale is hooked on a tree or a stick and held by two people. The child is then placed in the weighing pants. Weighing pants are suspended from the lower hook of the scale, the scale is re-calibrated to zero and the child’s shoes, clothes, and jewellery are removed. The scale is read at eye level.

The UNISCALE is designed to allow a mother or caregiver to hold the child while the child is weighed. The scale is placed on a flat surface in a well-lit area. A hand is waved over the solar switch to turn on the scale. The mother stands on the scale first, without the child. After the mother’s weight appears on the display, the mother remains standing on the scale. The reading is reset to zero by waiving a hand over the solar switch. The child is passed to the adult and the weight of the child that is displayed on the scale is recorded. The WHO recommends that the UNISCALE’s accuracy be checked on a weekly basis by using known weights of 3, 5, 10 and 20 kg (World Health Organization, 2008). If these checks are not carried out and corrective re-calibration done diligently, then there is the potential risk of introducing systematic errors to weight measurements.

To measure MUAC, a flexible measuring tape is wrapped around the mid-upper arm, between the shoulder and elbow. With the left arm bent, the measurer uses a string to find the midpoint of the arm between the shoulder and the tip of the elbow. Once the middle of the left arm is located, either a graduated or a simple tri-coloured (red, yellow and green) band is wrapped around the midpoint and a measurement in millimetres is determined. The tricolored band however does not give an actual measure when tied around the middle of the arm, but a malnutrition classification based on the colour. Red indicates that the child is severely malnourished, while yellow and green signifies that the child is moderately malnourished or at risk and that the child is of normal weight respectively.
**APPENDIX B: ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMREF</td>
<td>African Medical and Research Foundation</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>BMI</td>
<td>Body Mass Index</td>
</tr>
<tr>
<td>CHN</td>
<td>Community health nurse</td>
</tr>
<tr>
<td>CHW</td>
<td>Community health worker</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional neural network</td>
</tr>
<tr>
<td>CLAHE</td>
<td>Contrast Limited Adaptive Histogram Equalization</td>
</tr>
<tr>
<td>ENA</td>
<td>Emergency nutrition assessment</td>
</tr>
<tr>
<td>HAZ</td>
<td>Height for age Z-score</td>
</tr>
<tr>
<td>ICC</td>
<td>Intraclass correlation</td>
</tr>
<tr>
<td>KII</td>
<td>Key Informant Interview</td>
</tr>
<tr>
<td>MAM</td>
<td>Moderate acute malnutrition</td>
</tr>
<tr>
<td>MERON</td>
<td>Method for Extremely Rapid Observation of Nutritional Status</td>
</tr>
<tr>
<td>MoH</td>
<td>Ministry of Health</td>
</tr>
<tr>
<td>MORPH</td>
<td>Craniofacial Longitudinal Morphological Face Database</td>
</tr>
<tr>
<td>MUAC</td>
<td>Mid-upper arm circumference</td>
</tr>
<tr>
<td>SAM</td>
<td>Severe Acute Malnutrition</td>
</tr>
<tr>
<td>SDG</td>
<td>Sustainable Development Goals</td>
</tr>
<tr>
<td>SMART</td>
<td>Standardized Monitoring and Assessment of Relief and Transitions</td>
</tr>
<tr>
<td>UNICEF</td>
<td>United Nations International Children's Emergency Fund</td>
</tr>
<tr>
<td>WAZ</td>
<td>Weight for age Z-score</td>
</tr>
<tr>
<td>WFH</td>
<td>Weight for height</td>
</tr>
<tr>
<td>WFLZ</td>
<td>Weight for length Z score</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
<tr>
<td>WFHZ</td>
<td>Weight for height Z score</td>
</tr>
</tbody>
</table>