Computer vision methods to improve population data

Stephane Helleringer
Johns Hopkins University
Bloomberg School of Public Health

2019-12-04

Computer vision

Rapidly expanding field of research in computer science and engineering.

It seeks to train computers to extract information from images (e.g., photographs, video recordings) in order to recognize objects, persons, and/or activities.

Computer vision is often seen as a precursor to artificial intelligence.

Computer vision for population data

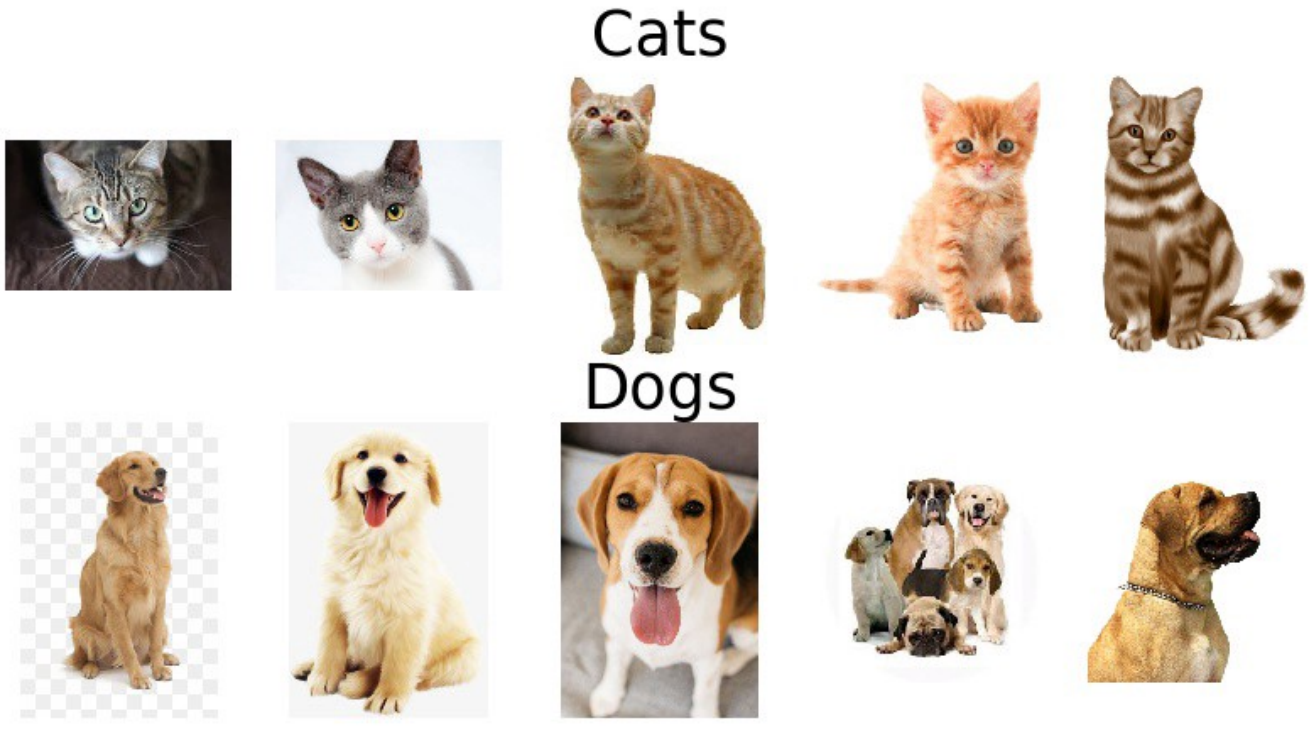
Fairly limited use so far. Examples include:

- Interpret satellite images for population mapping
- Google Street View for measuring neighborhood distress
- Gather demographic data on social media

Computer vision: General approach

- Constitute training dataset of labelled images
- Train computer system in recognizing object/trait of interest
  - Relate features extracted from image to the label
- Use trained model on unlabelled images

Computer vision: General approach



An application: age measurement in LMICs

Computer vision widely used to estimate age of persons in variety of contexts ("automatic age estimation").

Can it help address age misreporting in population data from LMICs?

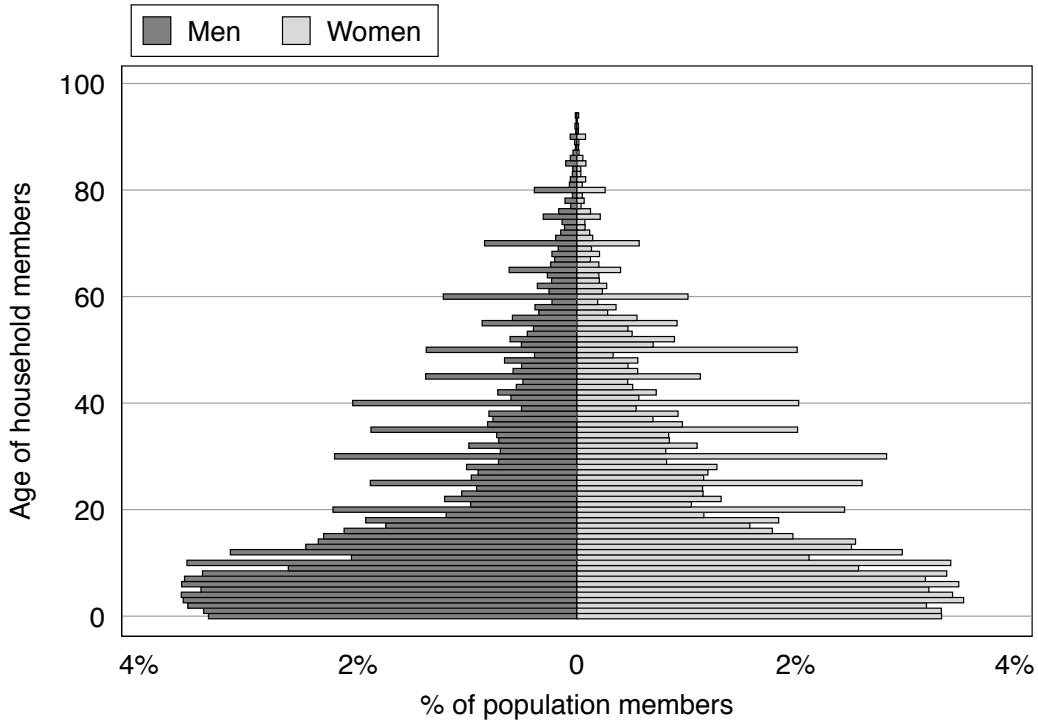
Age data in LMICs

Long-standing problems in age distributions.

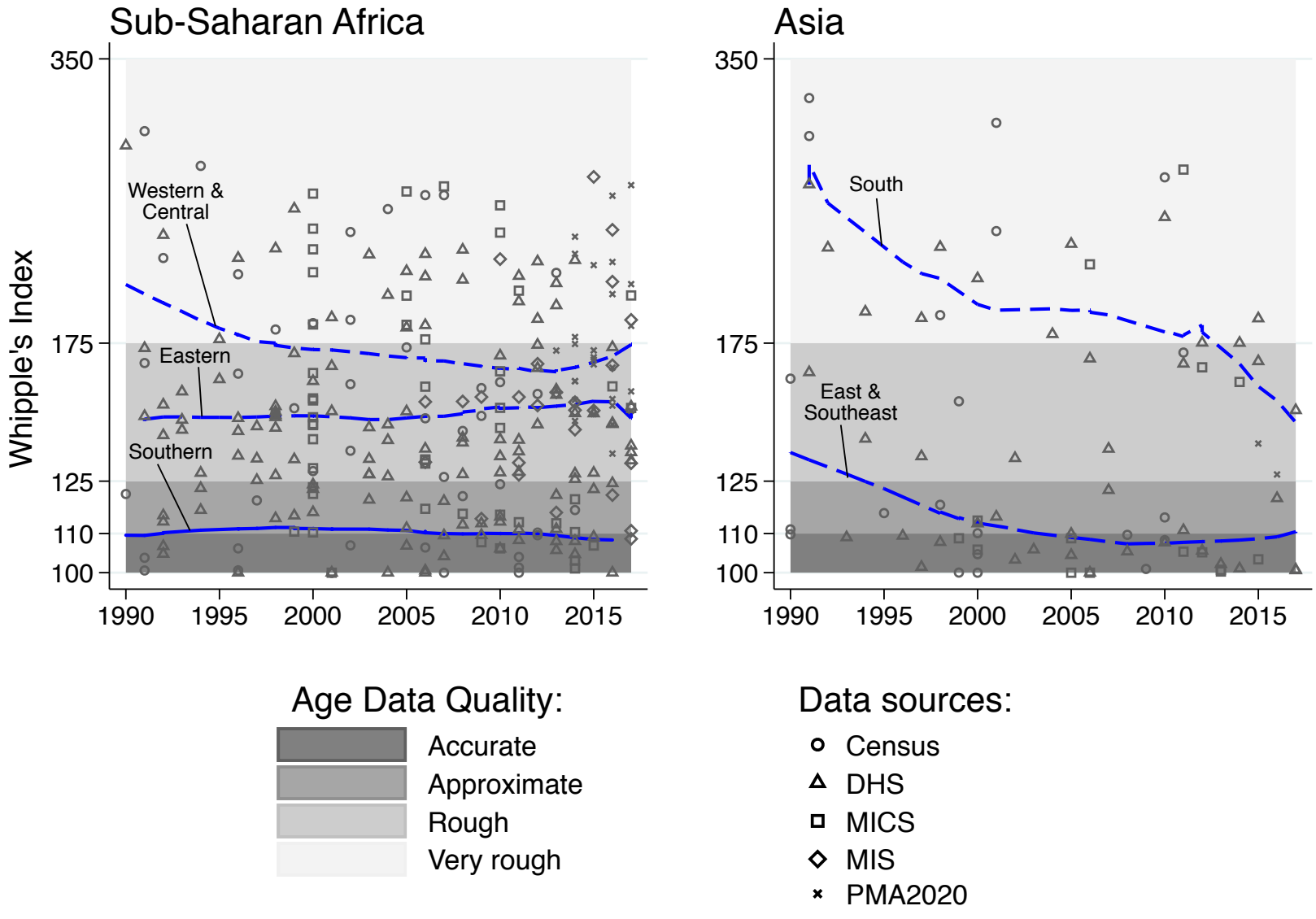
Due to variety of causes (low birth registration, numeracy, interviewer behaviors)

Most obvious signal is age heaping: excess numbers og ages ending in -0 or -5

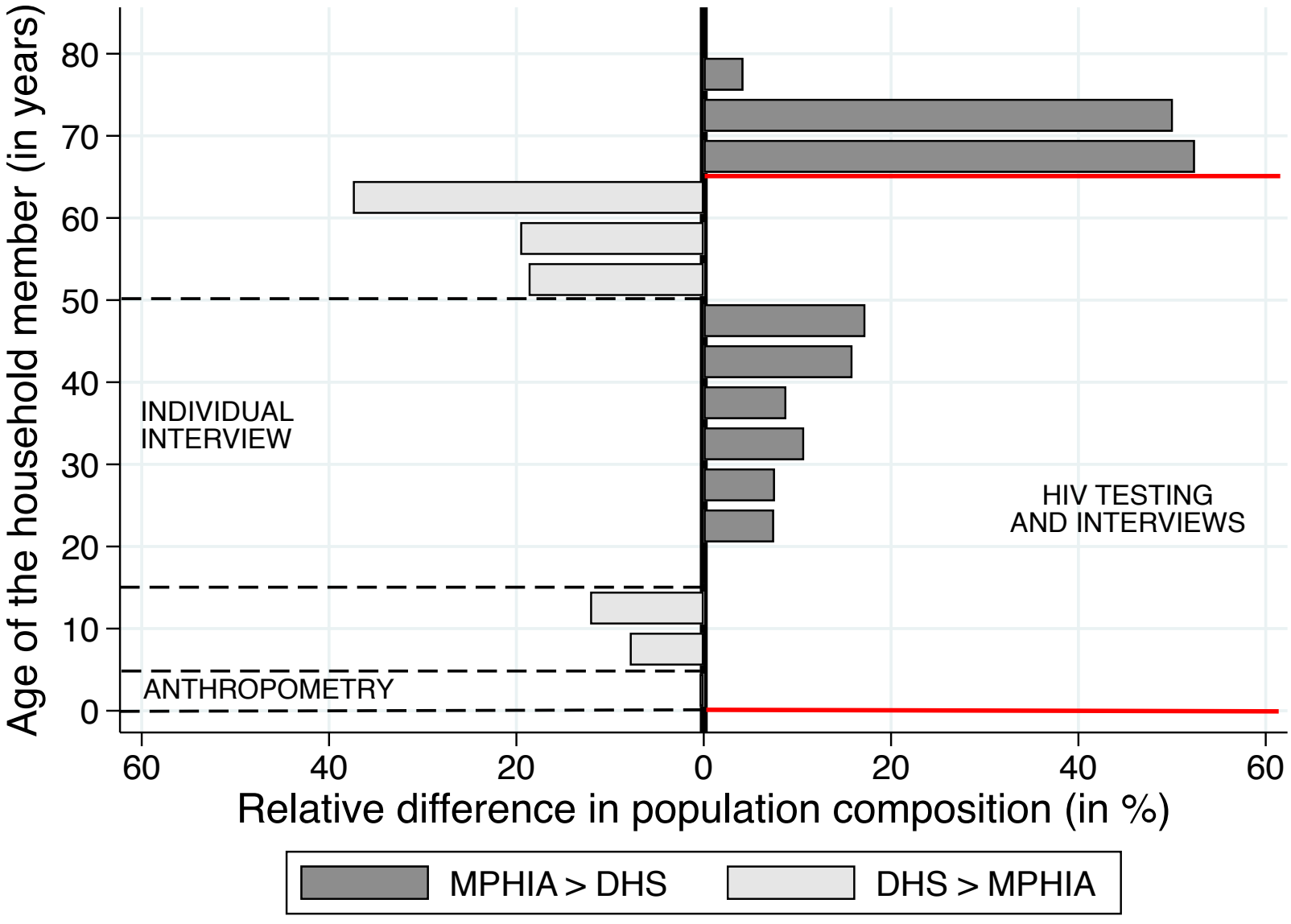
- 2017 Nigeria MICS



Age heaping in LMICs: limited improvements



Other distortions of age data



Test in Senegal

- Can we use computer vision to:
  - Determine inclusion in surveys? (Classification problem)
  - Estimate age in single years? (Estimation problem)

Test in Senegal: assembling the training dataset

- There are publicly available photographic databases, and pre-trained algorithms for age prediction.
- Most training datasets primarily contain photos from individuals of European or Asian descent.
  - Prediction across racial/ethnic groups (much) more inaccurate
- More diverse training datasets only contain individuals of African descent from US or UK.
  - Potential differences in facial aging patterns

Test in Senegal: assembling the training dataset

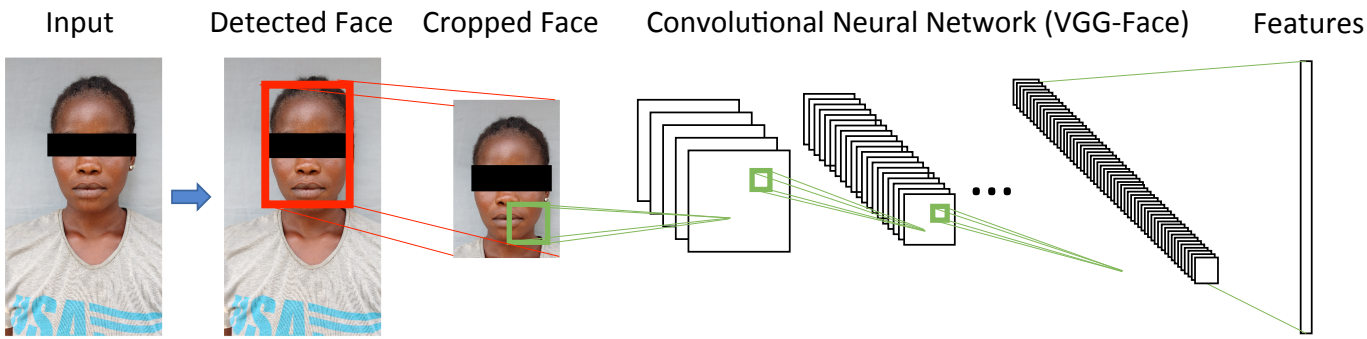
- Worked in Niakhar HDSS:
  - small population under demographic surveillance since 1962
  - Household visits every 2-4 months to record births and deaths
  - Gold standard age data for subset of the population
- Data collection focused on women aged 18 and older
  - Excluded those who moved into surveillance area after birth since 1962 (since precision of age data is more suspect)
  - Included those who were already residing in the area in 1962 at first census (and thus were aged $>$53 years old at time of study)

Test in Senegal: assembling the training dataset

- Trained two fieldworkers in using ODK to collect photographs
- Training emphasized issues related to lighting, to avoid over-exposed photos
- Used standardized background to help alleviate some of these issues
- In total, 353/515 women photographed

Test in Senegal: extracting age-related information

- Worked with team of engineers specializing in computer vision
- Pre-process images for comparison (face detection, cropping)
- Extract age-related information contained in image using pre-trained convolutional neural networks (VGG-Face)
- Yields >4,000 features per image



Test in Senegal: solving classification problem

We used support vector machine (SVM) to predict binary outcome (Below 50y or 50+)

- Find boundary that maximizes distance between the two groups
- Yields a score akin to probability of membership in a specific group
- Used leave-one-out approach for cross-validation

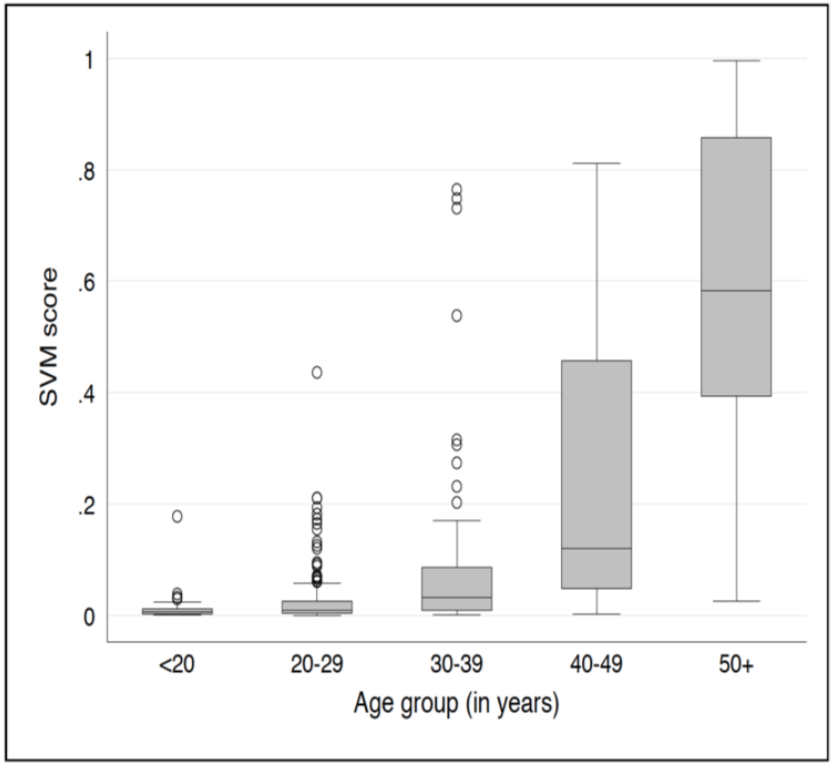


Figure 4: Probability of being aged 50 years and older according to the SVM analysis

Test in Senegal: solving classification problem

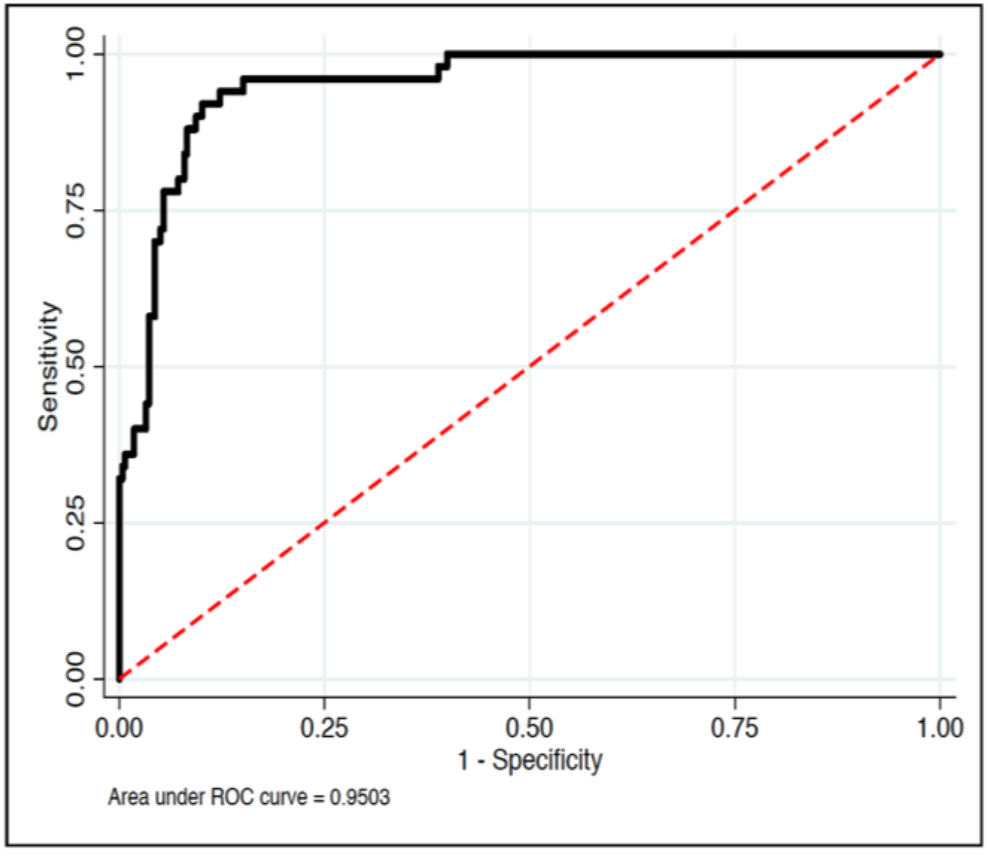


Figure 5: ROC analysis of the use of AAE to identify women of reproductive ages

ROC analysis: high predictive power of computer vision method.

Accuracy remains high even at the boundary of eligibility (i.e., limited sample to 40y and above)

Classification errors associated with poor health and smoking in multivariate analyses

Test in Senegal: Predicting age in single years

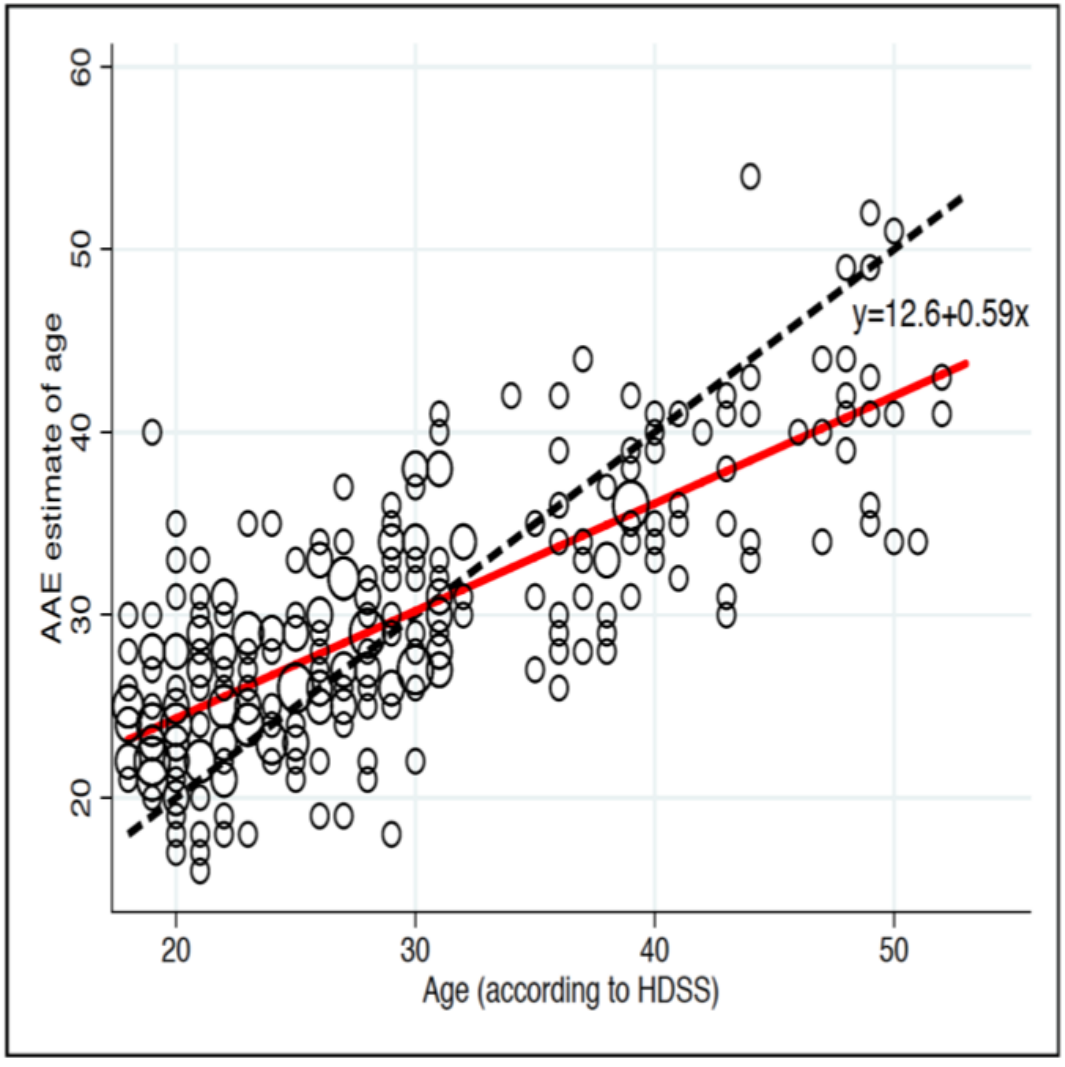


Figure 6: results from AAE models predicting age in completed years.

More challenging task given small dataset

MAE = 4.62 years

Comparable to error from automatic age estimation in HICs with training datasets of similar size.

- But some applications now get close to 2 years MAE.

Test in Senegal: Predicting age in single years

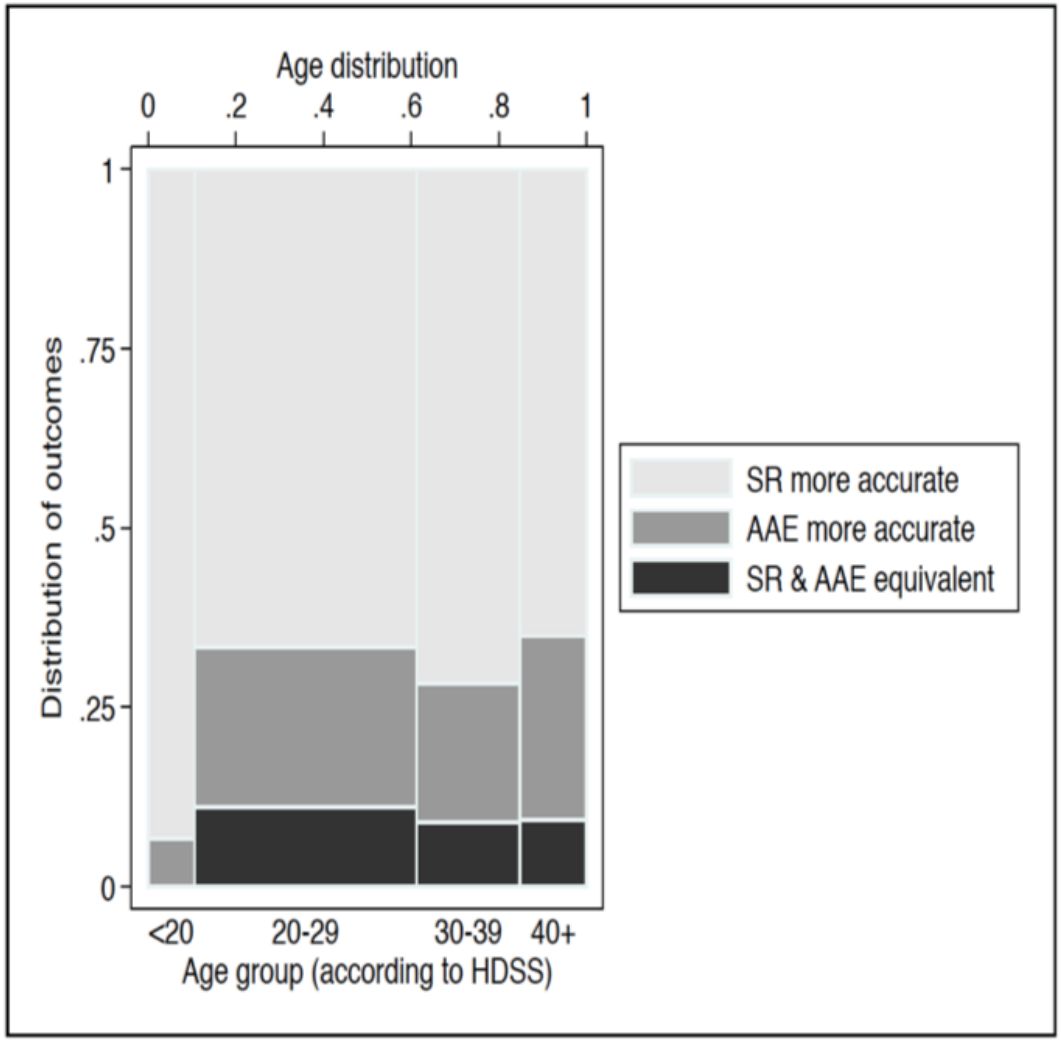


Figure 7: relative accuracy of AAE estimates and self-reported age data

High knowledge of age in Niakhar, but automatic age estimation still performs as well or better than self-reported data for a portion of population.

- Older age groups
- potential for combining SR and AAE estimates?

Lessons learned from test in Senegal

- Computer vision has potential to help address some measurement problems in LMICs.
- Can easily be integrated into:
  - Surveys that now overwhelmingly use camera-enabled devices (tablets, mobile phones)
  - CRVS/ID systems that collect photos/biometric data.
- Unlike other new data sources (e.g., social media), does not require much rethinking re: selectivity.
- Can be expanded to explore measurement of other items
  - Malnutrition or obesity
  - Health
  - Sensitive topics (e.g., smoking)

Lessons learned from test in Senegal

But there are major challenges:

- Training datasets:
  - Our test in Senegal was based on very small dataset.
  - Industry standards now larger by orders of magnitude.
- Multi-task learning:
  - Trying to estimate only age runs into issues linked with heterogeneity (e.g., larger errors with poor health)
  - New models for face analysis attempt to perform multiple tasks at once, might help improve inferences about age.
- Data privacy:
  - Photographs add PII to data collection
  - Develop data collection apps that do not store images