**Supplementary Table 1**. Metrics used for assessing algorithmic fairness used in this study

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Base metrics  | Definition | Meaning | Comparison metric | Interpretation |
| Accuracy | (TP+TN)/(TP+FP+TN+FN) | The proportion of patients correctly classified by the model (range: 0-1; higher score means better performance).  | Accuracy equality | Is the model more accurate on one group than another?Ratio=1: fairRatio<1: unfavorable to unprivileged groupRatio>1: favorable to unprivileged group  |
| Sensitivity (recall, true positive rate) | TP/(TP+FN) = 1 - FNR | The proportion of patients classified as case by the model among true cases (range: 0-1; higher score means better performance) | Equal opportunity  | Are future incidences of asthma exacerbation detected equally between two groups? (Or, equivalently, are future incidences of asthma exacerbation missed equally between two groups?)Ratio=1: fairRatio<1: unfavorable to unprivileged groupRatio>1: favorable to unprivileged group |
| False positive rate | FP/(FP+TN) | The proportion of patients falsely classified as case among those who are not cases, which is same as 1-specificity (range: 0-1; higher score means worse performance)  | Predictive equality  | Do both groups share an equal burden of unnecessary worry from false positives?Ratio=1: fairRatio<1: favorable to unprivileged groupRatio>1: unfavorable to unprivileged group |
| Positive predictive value (Precision) | TP/(TP+FP) | The proportion of true cases among those classified as cases by the model (range: 0-1; higher score means better performance) | Predictive parity  | Are predictions on both groups equally useful for clinicians, or does one group have a higher proportion of false positives among predicted positives?Ratio=1: fairRatio<1: unfavorable to unprivileged groupRatio>1: favorable to unprivileged group |
| Unweighted average of FPR and FNR  | [FP/(FP+TN) + FN/(TP+FN)]/2 | Average between FPR (predictive equality) and FNR (1-sensitivity). Range: 0-.5 (higher score means worse performance) | Balanced error rate | (Interpretable as an average of equal opportunity and predictive equality)Ratio=1: fairRatio<1: favorable to unprivileged groupRatio>1: unfavorable to unprivileged group |
| TP: true positives; FP: false positives; TN: true negatives; FN: false negatives; FPR: false positive rate; FNR: false negative rate |

|  |
| --- |
| **Supplementary Table 2.** The reported health outcomes predicted by HOUSES  |
| **A. Chronic conditions and mortality** |
| **Adults:** 1) Risk of rheumatoid arthritis (RA) and post-RA mortality,76 2) risks of coronary heart disease, asthma, diabetes, hypertension, mood disorder,77 Post–Myocardial Infarction mortality,78 3) multiple chronic conditions,79 4) post-Glioma mortality80, 5) renal transplantation outcome68, 6) ICU mortality,81 7) persistent depressive symptoms and remission of depressive symptoms,82 **Children:** 8) Risks of asthma, epilepsy, mood disorders,83 9) overweight,33,84 8) asthma outcome67, 10) multiple complex chronic conditions,47 11) serum 25 (OH) D concentration85 |
| **B. Acute conditions** |
| **Adults:** 12) Risk of accidental falls86, 13) osteoporotic fracture incidence87**Children:** 14) Risk of invasive pneumococcal disease,88 15) risk ofbronchiolitis, pneumonia, urinary tract infection, adverse childhood experiences83 |
| **C. Other health outcomes** |
| **Adults:** 16) All-cause hospitalization,79 17) smoking status,89 18) detecting survey biases,53 19) end of life care access (advance directives, social work consultation)90**Children:** 20) Low birth weight, household smoking status,33,84 21) pertussis vaccine up-to-date status,91 22) self-rated health92, 23) HPV vaccine (initiation and completion) up-to-date status93  |

**Unique features of HOUSES index as an individual-level socioeconomic (SES) measure are described below:**

Validity: As summarized in Supplementary table 2, HOUSES has shown to predict a broad range (39 different) of health outcomes in both adults and children reported in 23 publications. In this, as described above, HOUSES predicts the risk of graft failure among kidney transplant recipients while individual-level SES measure (educational level) and aggregate-level SES measure failed to predict the risk of graft failure94 suggesting HOUSES might outperform the conventional SES measures such as individual educational level and aggregate-level SES measures. This observation might not be unexpected given the significant misclassification(20-35%) of individual-level SES measures by aggregate level SES measures64,65 (ecological fallacy66) and differential frequency of missing values of self-reported educational level or study outcomes between higher vs. lower SES group defined by HOUSES.53,79

Precision: HOUSES is formulated as z-score which is much more precise in measuring individual-level SES measures, compared to the commonly used conventional SES measures such as self-reported categorical educational levels or Medicaid eligibility as a binary variable. For example, educational level measures typically categorize college graduate or graduate school graduate as the highest SES measure but as described above, incomes or earning power within each educational category significantly varied (e.g., income range from $39,312 to $86,684 for college graduates). Binary Medicaid eligibility is subject to greater within-group heterogeneity resulting in imprecision and failure to detect the association. For example, our previous work showed HOUSES predicts the burden of 5 common chronic diseases (coronary heart disease, diabetes mellitus, hypertension asthma, and mood disorder) in the US in a dose-response manner (the lower SES by HOUSES, the greater burden of diseases).77 Importantly, such dose-response trends continued to be observed when we further categorized the highest SES group (Q4) into quartiles suggesting the association of SES with health outcomes in gradient not binary nor categorical as suggested by Michael Marmot95,96 and other studies.28 Thus, the current SES measures with imprecision might not be suitable for AI fairness research.

Objectivity: HOUSES does not rely on self-report but is based on publicly available assessment data (real property data) derived from individual housing features (e.g., value, size, etc) and thus, is an objective measure. Thus, formulating HOUSES index does NOT require patient contact which is a major advantage of HOUSES index enabling large-scale studies. However, the current SES measures rely on self-reported measures. The widely used aggregate level measures such as Area Deprivation Index (ADI) or similar indices are based on self-reported measures from a small sample (typically about 1%) for American Community Survey. Medicaid eligibility is an objective measure, but it is cumbersome as the eligibility for patients frequently changes and the duration of enrollment period for Medicaid can be varied (without knowing this info, Medicaid eligibility status at a given point in time might be inaccurate).

Scalability: Assessment data of each County is available electronically, and thus, we have been able to directly link/geocode address info to assessment data to formulate HOUSES index and it is standardized at a county level as a geographic unit for economic activity and housing market unlike other unstandardized SES measures. Currently, with support from the NIH, our team has developed a cloud-based software which enables us to expand HOUSES to the State of MN. We are planning to expand HOUSES to the US in the near future. As briefly discussed in the manuscript, as HOUSES index is a geocoded index, it allows investigators to perform a geospatial analysis which might be a useful feature of HOUSES which may potentially enhance performance of AI algorithms as it detects unrecognized geographic risk factors (e.g., mobile home community is associated with risk of adverse childhood experience72, low income apartment is associated with the risk of COVID-19),70 and geographic risk factors for pertussis outbreak71).