Building and occupant characteristics as predictors of temperature-related illness in American households

Preliminary, unpublished results

1. Would a HVI with detailed information about the building be more accurate at predicting the risk of health hazards? If so, by how much?
2. Which building and occupant characteristics contribute most to predicting the risk of a health hazards?

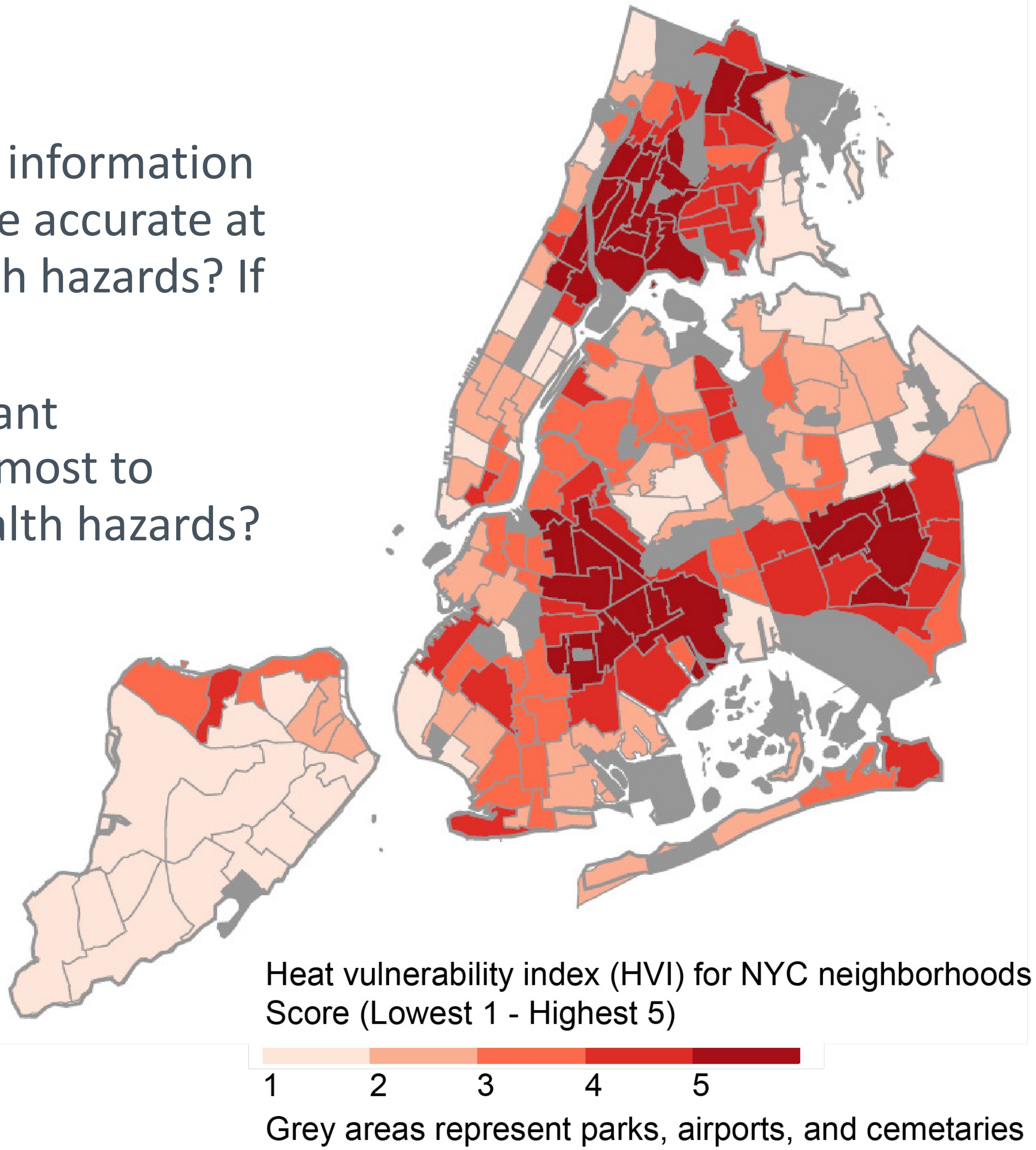


Image source: NYC Climate Resiliency Design Guidelines

Residential Energy Consumption Survey (RECS)

Source	U.S. Energy Information Administration (EIA)
Year	2015 and 2020
Sample size	5,700 and 18,500 households respectively
Sampling	Random, complex multi-stage area probability
Scope	U.S. households occupied as primary residence
Objective	Energy demand forecasting, energy efficiency planning

Temperature-related illness

In the last year, did anyone in your household need medical attention because your home was **too cold**?

In the last year, did anyone in your household need medical attention because your home was **too hot**?

Used sample weights to estimate population prevalence as part of results

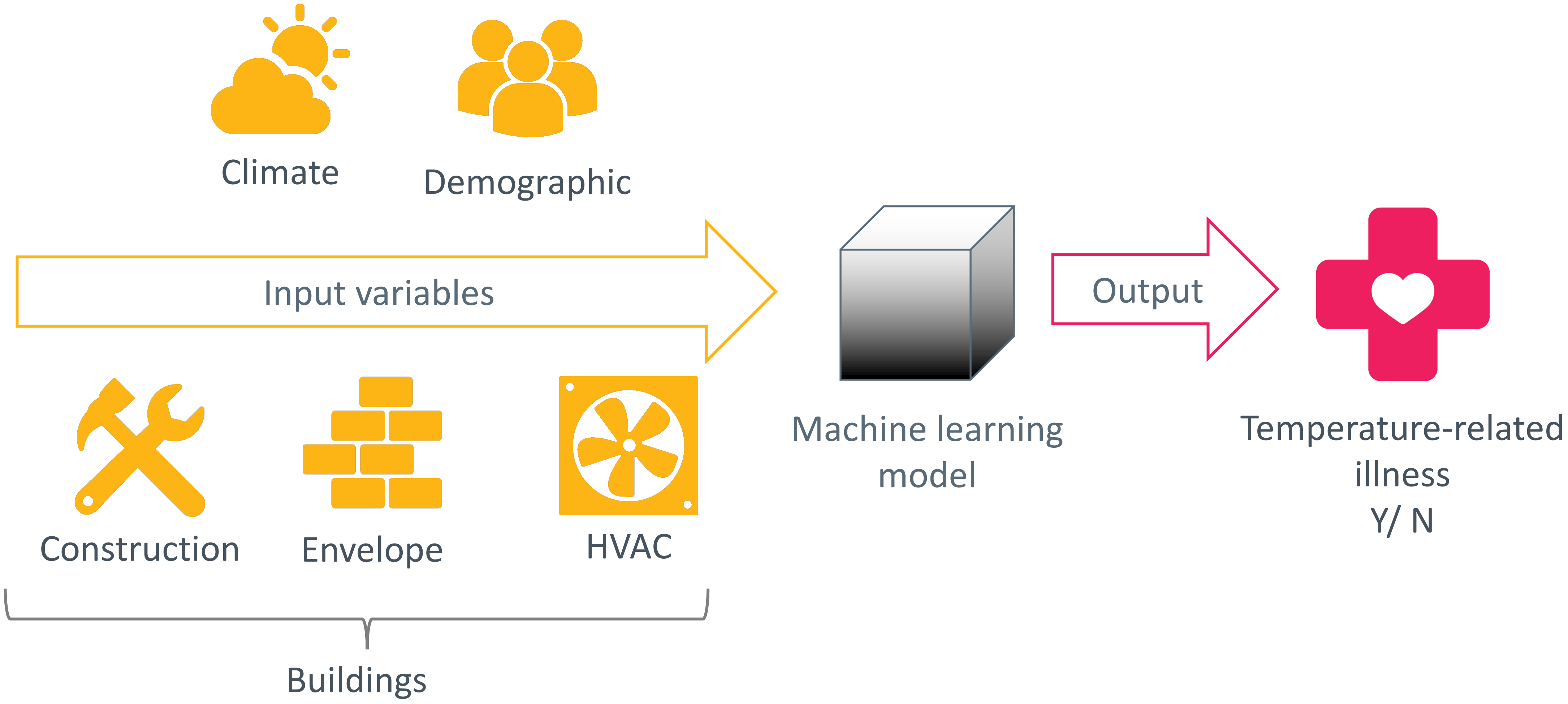
Combined 2015 and 2020 data sets in machine learning modeling

Table 1. Observations of temperature-related illness in RECS

	2015	2020	Total
Heat-related	39	76	115
Cold-related	54	120	174
Any temperature	81	171	252
None	5,605	18,496	24,101

Preliminary, unpublished results

Predictive model



Input variables

Climate	Demographics	Buildings		
		Construction	Envelope	HVAC
Degree-days	Non-white	Construction age	Thermally massive wall	System type
	Over 65	Apartment	Thermally massive roof	Energy insecurity
	Lives alone	Mobile home	Insulation	Fans
	Large households		Infiltration	Off-grid
	Poverty		Windows per room	
	Unemployed		Glazing type	
	Education level			
	Renting			
	Pays utility or fuel			

Machine learning model building

- Binary classification model
  - Input: building and household characteristics
  - Output: temperature-related morbidity (Y/ N)
- Model fitting
  - 80/20 training/test data split
  - Training data subsampled to address class imbalance
  - Hyperparameter tuning with repeated 5-fold cross validation
  - Compare 8 machine learning algorithms
    - Generalized linear model, multivariate adaptive regression spline, penalized discriminant analysis, bagged classification and regression trees, random forest neural net.
- Uncertainty
  - 30 bootstrap iterations, with different training/test data splits

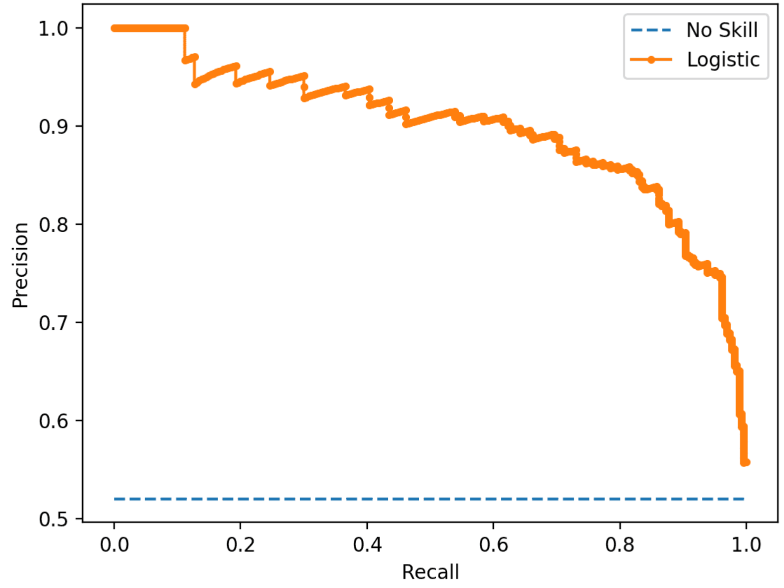
Machine learning model building contd.

- Pre-processing
  - Checked for zero or near-zero variance, which removed two demographic variables
    - Large households and pays utility/fuel
  - Checked for correlated variables and linear combinations, but none removed
  - Normalize numerical inputs to have zero mean and unit variance
- Imbalanced class handling:
  - Class-weights – higher penalty for misclassifying homes with temperature-related illness
  - Up-sampling – over sample minority class
  - SMOTE and ROSE – generate synthetic data in minority class and under sample majority class
  - Performance metrics suited for imbalanced data: balanced accuracy, recall, precision.

Machine learning model performance contd.

- Based on the confusion matrix:
  - Precision: TP/TP+FP
  - Recall: TP/TP+FN (same as sensitivity)
- The PR curve plots these values at different thresholds
- The area under the PR curve summarizes the curve into one metric
- Because neither precision or recall use the number of TN, this metric is well suited for imbalanced data

		Predicted	
		Yes	**No**
Actual	**Yes**	True positive	False negative
	No	False positive	True negative



Comparison of input features groups

- Train one set of models with Climate + Demographics variables and another with Climate + Demographics + Buildings variables
- Compared results with paired t-test (based on bootstrapped iteration)
- Consider $p < 0.05$ statistically significant
- For statistically significant results, calculate Cohen's d for effect size
  - $0.4 \leq |d| < 1.15$ for recommended minimum practical effect
  - $1.15 \leq |d| < 2.7$ for moderate effect
  - $|d| \geq 2.7$ for strong effect

Results

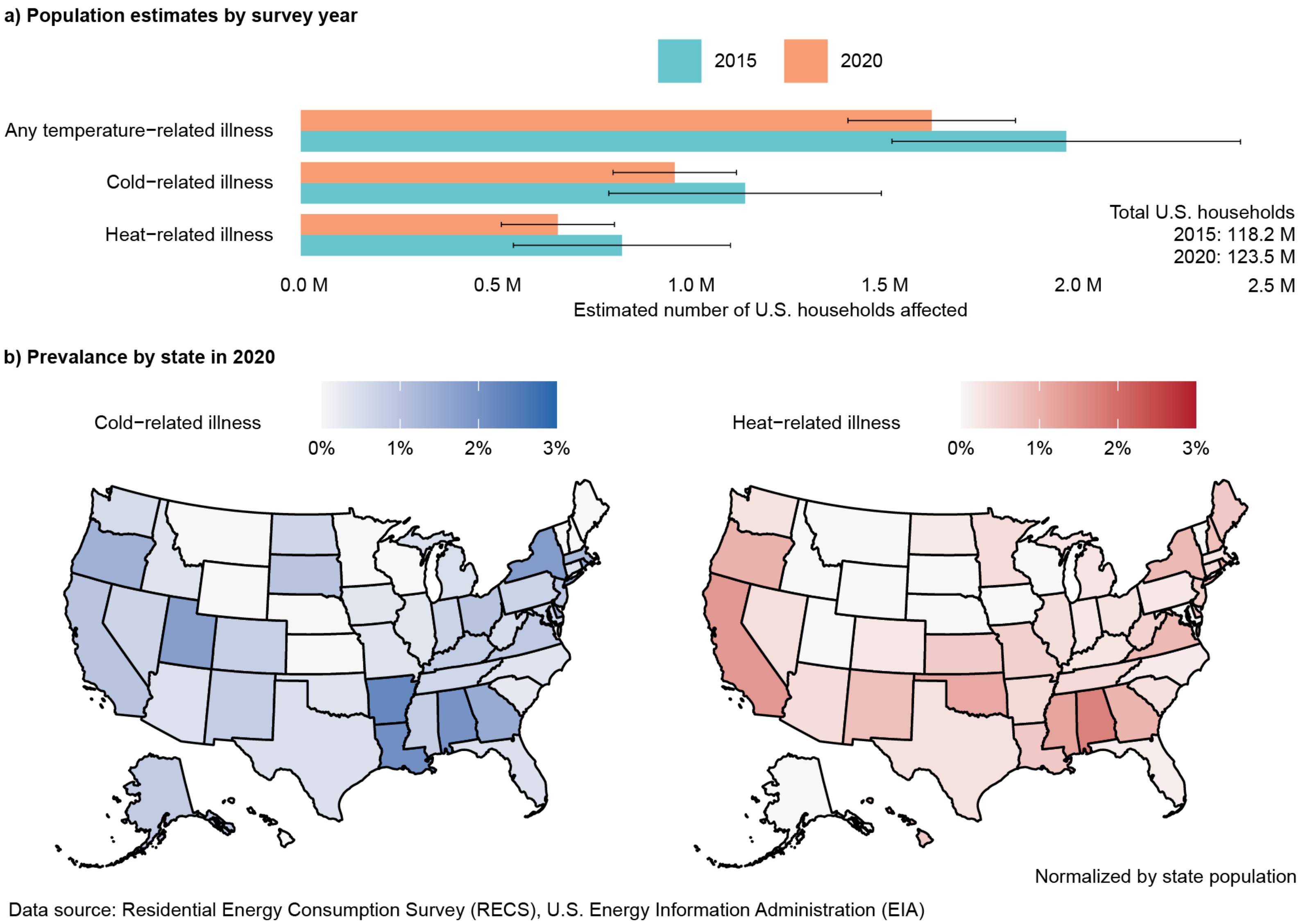


Figure 1: Prevalence of temperature-related illness in U.S. households

Preliminary, unpublished results

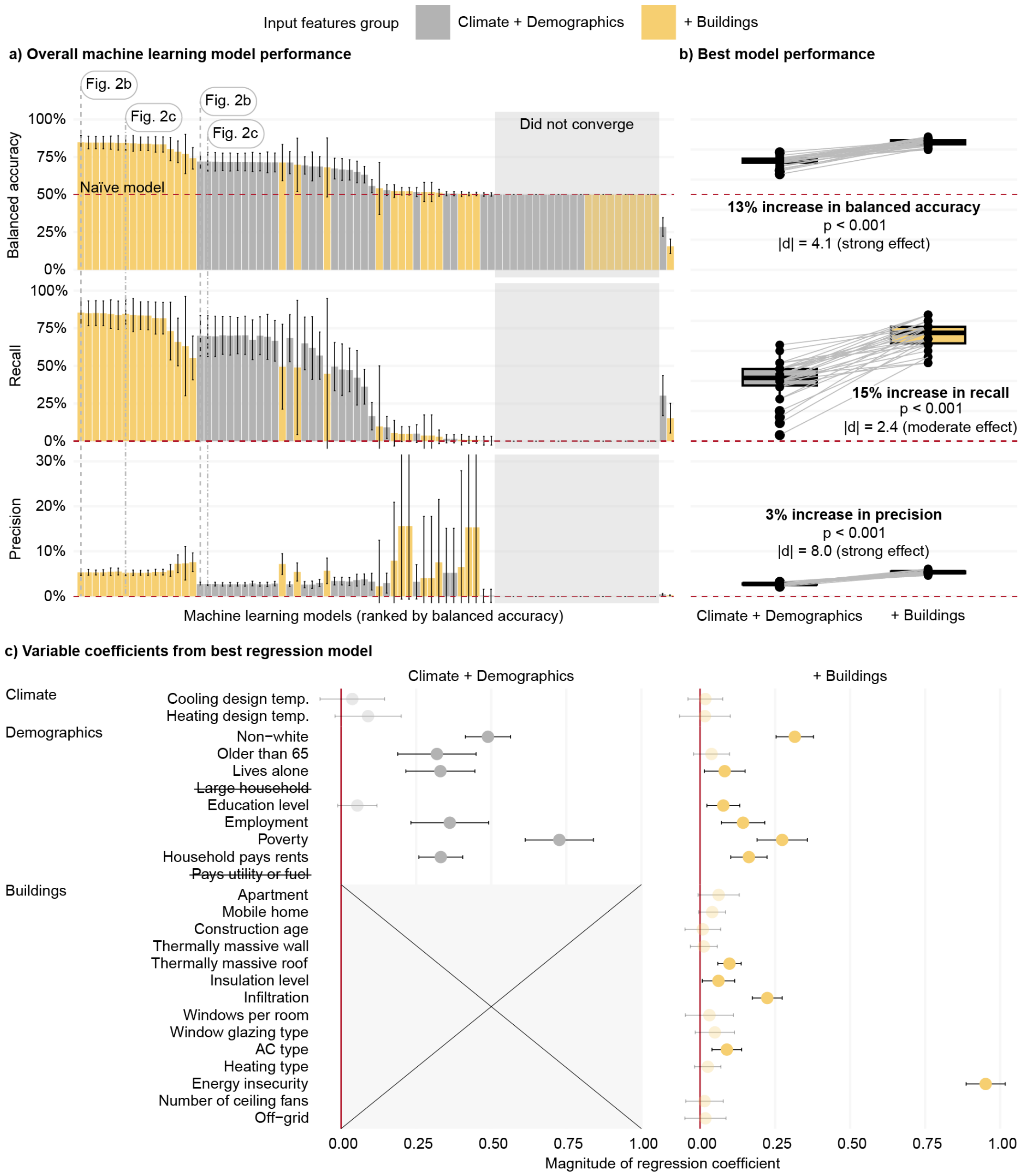


Figure 2: Predicting temperature-related illness

Preliminary, unpublished results

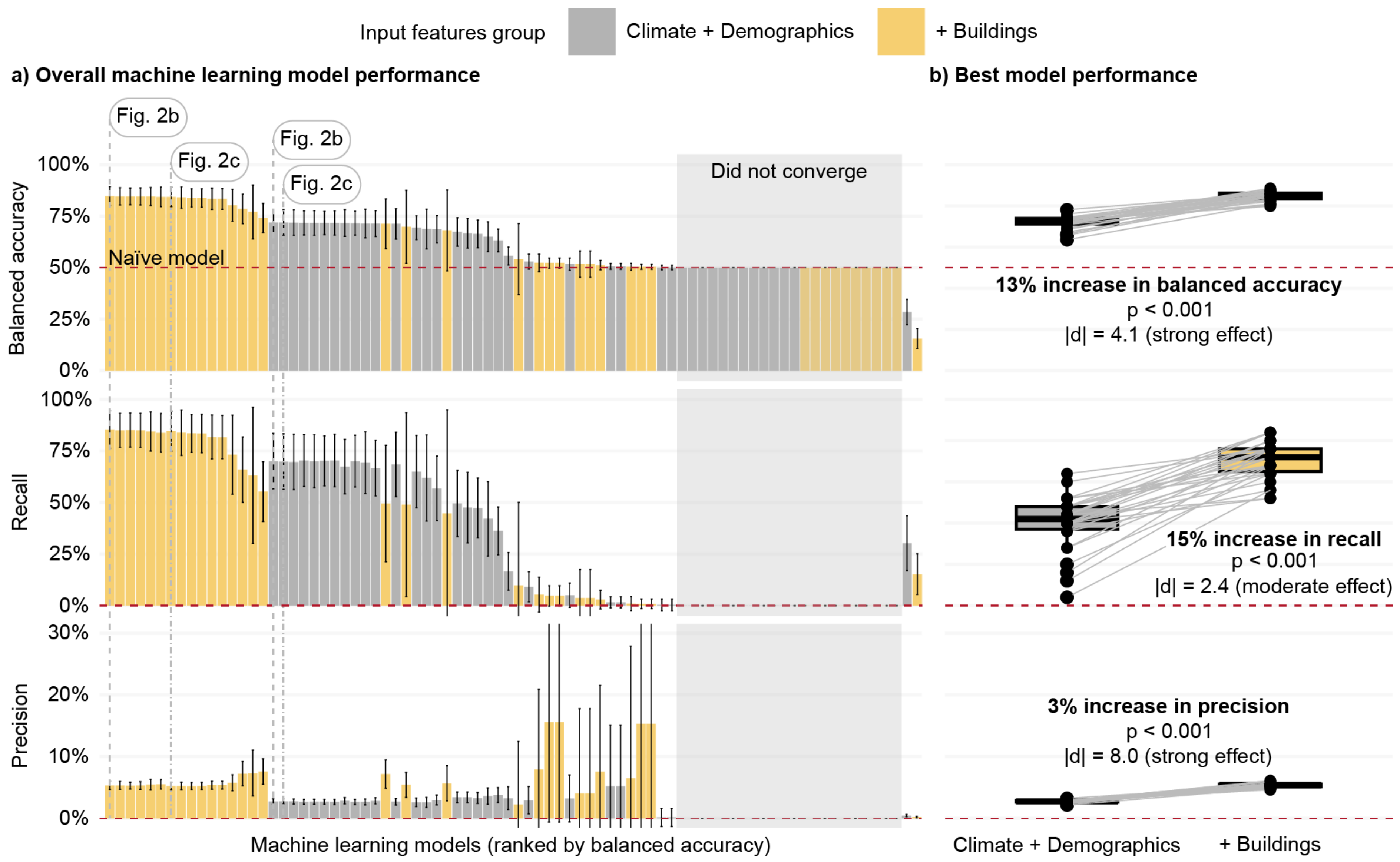


Figure 2a and 2b: Machine learning model performance

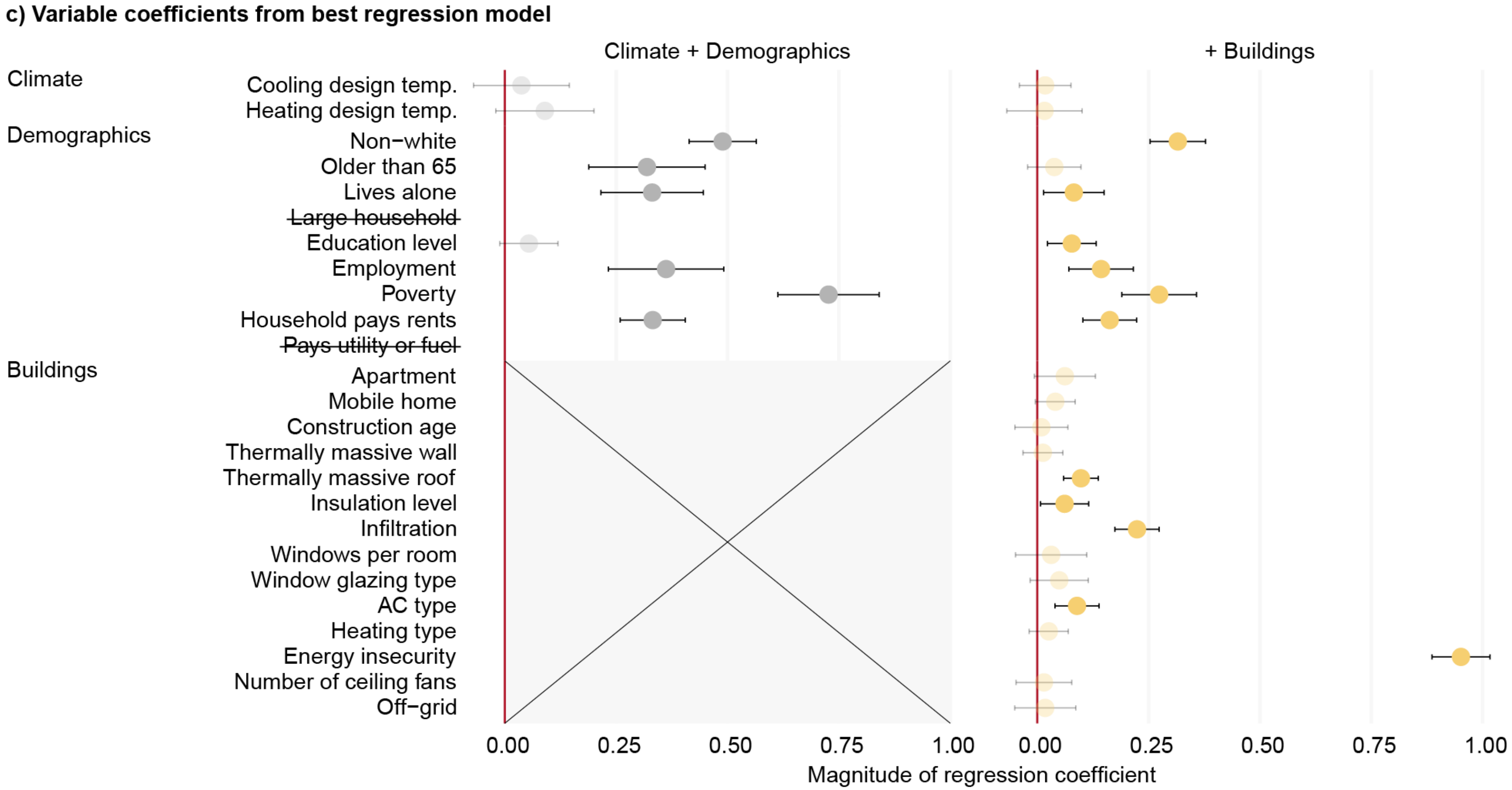


Figure 2c: Variable coefficients from top regression model

Discussion

- Study demonstrates that we can predict temperature-related illness to relatively high accuracy (up to 85%) but precision is low (around 5%)
- Therefore, we predict many false positives, but that could be due to temperature-related illness being underreported
- Energy insecurity is by-far the most important predictor of temperature-related illness, in line with heat death investigations in Maricopa County, Arizona
- This gives public health authorities a pathway to
  1. Prioritize data collection to identify at-risk households more accurately
  2. Design more effective interventions to prevent temperature-related illness (e.g. British Columbia launched a $10 million program last summer to distribute window AC units to vulnerable households)

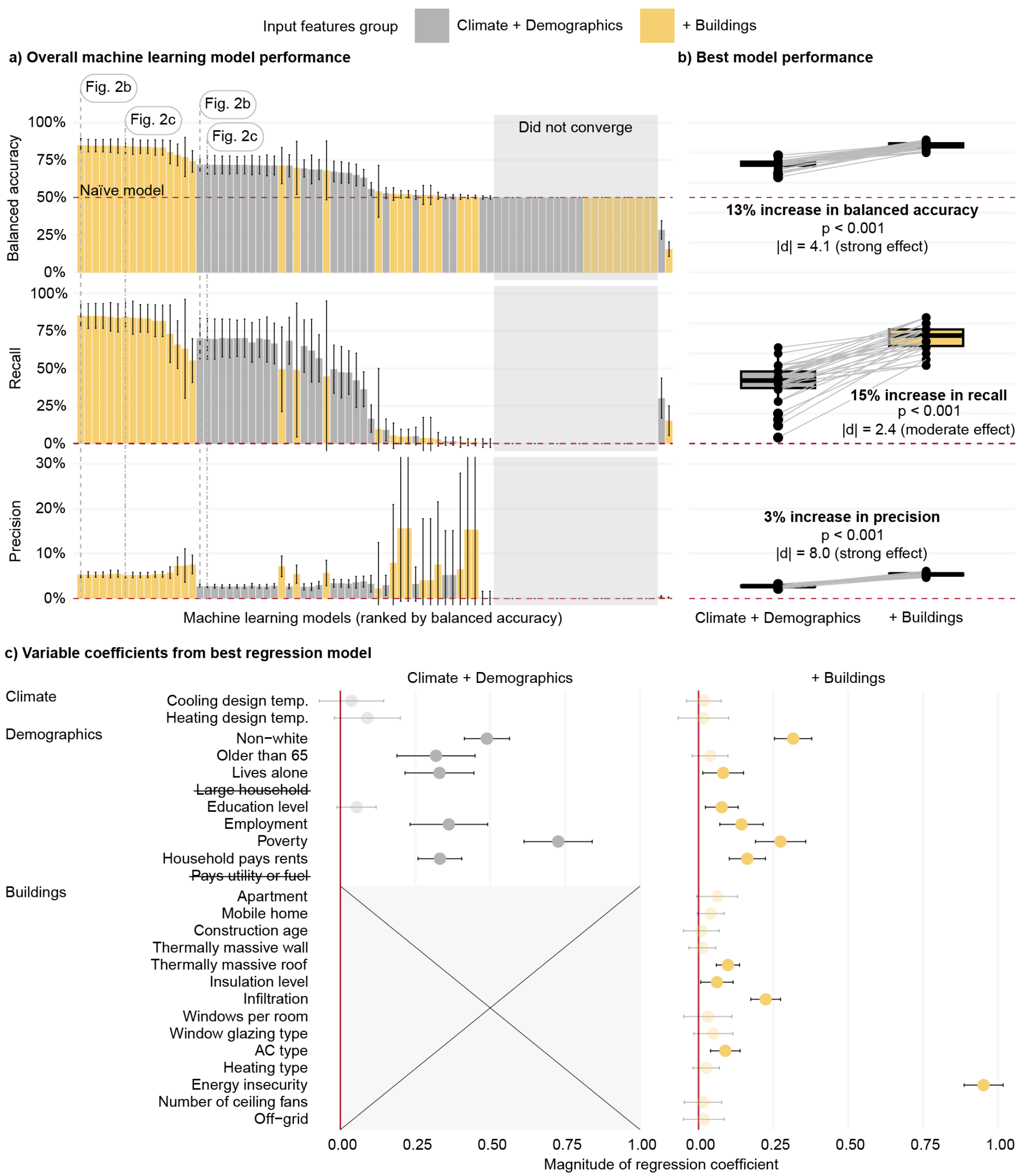
Limitations

- RECS only represents homes occupied as primary residence and most notably does not include nursing homes
- RECS is self-reported by household resident
- Our study focuses on predictive power, not causal relationships

Specific questions

1. How are climate variables such as design temperature calculated?
2. Have there been any studies validating the self-reported approach?
3. What is the background behind the questions regarding heat or cold-related illness?
4. Are there any noteworthy studies utilizing RECS data you recommend?

Thank you for your time!



Preliminary, unpublished results