APPENDIX

ILLUSTRATION FIGURE FOR COMPARISON OF MTL AND STL.

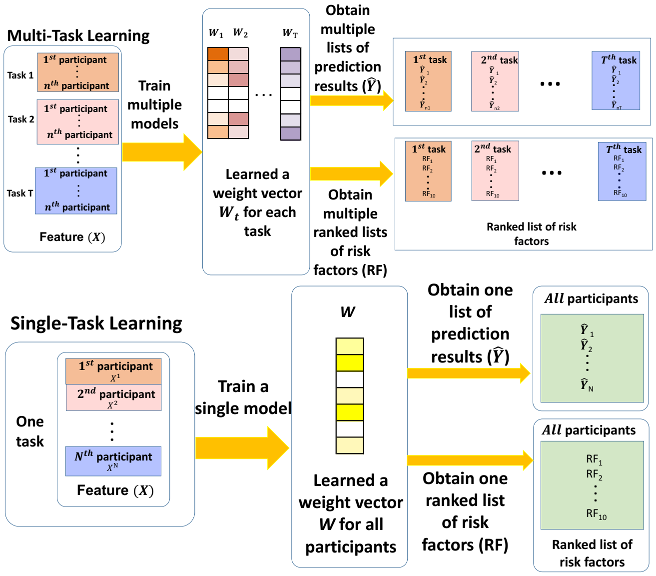

Fig. 2: **MTL trains multiple models simultaneously to obtain multiple lists of prediction results and multiple ranked lists of risk factors, i.e., one ranked list of risk factors for each subpopulation, whereas STL trains a single one-size-fits-all model to obtain a ranked list of risk factors for all subpopulations that ignores the data heterogeneity. Note that, in the learned weight vector, color box indicates a higher value of feature weight and white box means the weight is zero.**

PROXIMAL GRADIENT DESCENT ALGORITHM FOR $l_{2,1}$-NORM REGULARIZATION PROBLEM.

Algorithm 1: Proximal gradient descent algorithm for $l_{2,1}$-norm regularization problem.

Input: A set of feature matrices $\{X_1, X_2, \cdots, X_T\}$ and target value matrix Y for all T tasks, Initial coefficient matrix $W^{(0)}$, λ

Output: $\bar{W}$

Initialize: $W^{(1)} = W^{(0)}$, $d_{-1} = 0$, $d_0 = 1$,$\gamma_0 = 1$,$i = 1$;
repeat
  Set $\alpha_i = \frac{d_{i-2}-1}{d_{i-1}}$, $S^{(i)} = W^{(i)} + \alpha_i(W^{(i)} - W^{(i-1)})$;
  for $j = 2, 1, \cdots$ **do**
    Set $\gamma = 2^j \gamma_{i-1}$;
    Calculate $W^{(i+1)} = \pi_P(S^{(i)} - \frac{1}{\gamma} g^{'}(S^{(i)}))$;
    Calculate $Q_\gamma(S^{(i)}, W^{(i+1)})$;
    if $g(W^{(i+1)}) \leq Q_\gamma(S^{(i)}, W^{(i+1)})$ **then**
      $\gamma_i = \gamma$, **break** ;
    end
  end
  $d_i = \frac{1+\sqrt{1+4d_{i-1}^2}}{2}$;
  $i = i + 1$;
until *Convergence of* $W^{(i)}$;
$\bar{W} = W^{(i)}$;

DATA DISTRIBUTION OF 10 PROTOCOLS.

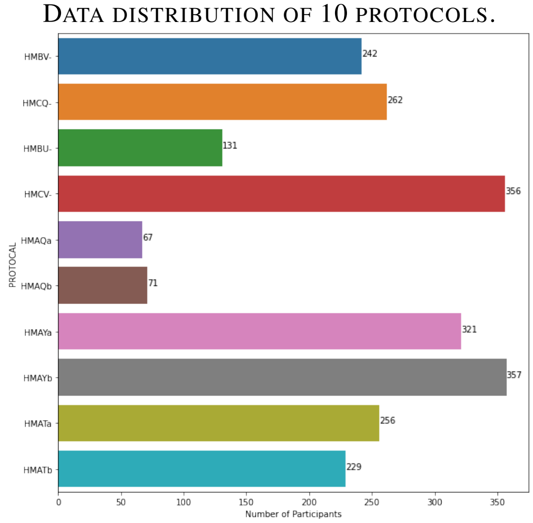

Fig. 3: **Data distribution of 10 protocols.**