This is the data and computer programs for "A Robust Approach for Electronic Health Record-Based Case-Control Studies with Contaminated Case Pools" by Guorong Dai, Yanyuan Ma, Jill Schnall, Jinbo Chen and Raymond J. Carroll..

The file *sepsis_data.csv* contains the data for the analysis in Section 5. The data have already been preprocessed so that they can be directly plugged into the estimating equation proposed in the article. All the continuous covariates have been standardized.

simulation_example.R and *sepsis.R* contain the codes for the simulations and data analysis in Sections 4 and 5, respectively.

ccs_functions.R and *ccs.f90* contain the functions used for numerical study. The file *ccs.f90* needs to be compiled first. The main function in *ccs_functin.R* is *solver*, whose arguments and outcomes are explained by the comments above the function. Here are some remarks:

1. The covariate matrix in association and phenotyping models, i.e., x and w, are allowed to be different.
2. In the case pool $r=1$ or 0 means a case is validated or not. In the control pool one could set r always equal to one. Therefore the length of r equals the sample size N.
3. The length of d is N. Although d is unknown when $r=0$, one could set d to be any value for cases which have not been validated. Those values do not affect the results but make the argument d meet the requirement of the function.
4. There is no need to set other arguments for the analysis in the current paper. This function includes the implementation of estimators for more complicated problems, which are not considered in the current paper.
5. The useful outputs are *$ theta* and *$ covariance*, which are the slope vector of the association model and the covariance matrix of the whole estimator *(theta0, theta ^ T, alpha0, alpha ^ T) ^ T*. One could take the submatrix of *$ covariance* that is the covariance of *$ theta*.