

04/13 update

- Preliminary results on a small, **balanced** subsample (5000 cases, 5000 controls).
- Training set of 9000 IDs, test set of 1000 IDs.
- Imputed missing variables with median.
- Lab data (blood tests, FOBT) and colonoscopies converted to longitudinal summaries (mean,max,min,...).
- Lab data restricted to time interval from ‘indexdate - 3 years’ to ‘indexdate - 1 year’.
- No ICD code information, no medication information, no smoking status records (smoking status at index is in demographic variables).
- Logistic regression.

| Predictors ($p = \#$ of predictors) | Training AUC | Test AUC |
|---|--------------|----------|
| Demographic ($p = 10$) | 0.814 | 0.818 |
| Charlson ($p = 18$) | 0.705 | 0.693 |
| Demographic + Charlson ($p = 28$) | 0.863 | 0.862 |
| Labs ($p = 210$) | 0.824 | 0.824 |
| Demographic + Charlson + Labs ($p = 238$) | 0.913 | 0.910 |

04/20 update

- Preliminary results on a small, **balanced** subsample (5000 cases, 5000 controls).
- Training set of 9000 IDs, test set of 1000 IDs.
- Imputed missing variables with median.
- **Sets of predictors:**
 - Demographic data includes race, age, weight, etc. (**Demographic**).
 - Charlson score inputs plus GERD diagnosis (**Charlson**).
 - Lab data (blood tests, FOBT) and colonoscopy data converted to longitudinal summaries (mean,max,min,...) (**Labs**).
 - Medication (H2R and PPI) converted to longitudinal summaries (mean,max,...) (**Meds**).
- Lab and medication data restricted to time interval from ‘indexdate - 3 years’ to ‘indexdate - 1 year’.
- Logistic regression.

| Predictors ($p = \#$ of predictors) | Training AUC | Test AUC |
|---|---------------------|-----------------|
| Demographic ($p = 10$) | 0.814 | 0.818 |
| Charlson ($p = 18$) | 0.705 | 0.693 |
| Meds ($p = 10$) | 0.604 | 0.598 |
| Demographic + Charlson + Labs + Meds ($p = 240$) | 0.915 | 0.912 |

04/27 update

- Baseline logistic regression on entire sample (no blood labs, no medication data).
- Full sample contains $n = 6,649,108$ observations, with $n_{\text{control}} = 6,637,713$ and $n_{\text{case}} = 11,395$.
- Imputed numeric variables with medians, imputed smoking status at random, proportional to non-missing in sample (45% current, 41% former, 14% never).

| Predictors ($p = \#$ of predictors) | Training AUC | Test AUC |
|--------------------------------------|--------------|----------|
| Demographic ($p = 10$) | 0.670 | 0.675 |
| Charlson ($p = 18$) | 0.684 | 0.705 |
| Demographic + Charlson ($p = 28$) | 0.760 | 0.770 |

Smoking status missingness.

In the entire sample:

| Smoking Status | Current | Former | Never | Missing |
|----------------|---------|---------|--------|---------|
| Count | 1696052 | 1521806 | 537727 | 2893523 |
| Probability | 0.26 | 0.23 | 0.08 | 0.44 |

Among cases:

| Smoking Status | Current | Former | Never | Missing |
|----------------|---------|--------|-------|---------|
| Count | 3901 | 3546 | 594 | 3354 |
| Probability | 0.34 | 0.31 | 0.05 | 0.29 |

Among controls:

| Smoking Status | Current | Former | Never | Missing |
|----------------|---------|---------|--------|---------|
| Count | 1692151 | 1518260 | 537133 | 2890169 |
| Probability | 0.25 | 0.23 | 0.08 | 0.44 |

05/04 update

- Preliminary results on a balanced **random sample** of 5000 cases and 5000 controls.
- Most missing variables imputed with their median, SmokeStatus imputed at random proportional to non-missing.
- Training on 9000 observations, testing on 1000 observations.
- Baseline logistic regression fit with subsets of predictors, and all predictors.
- Random forest fit with all predictors.

| Model/predictors ($p = \#$ of predictors) | Training AUC | Test AUC |
|--|--------------|--------------|
| Logistic regression | - | - |
| Demographic ($p = 10$) | 0.669 | 0.676 |
| Charlson ($p = 18$) | 0.689 | 0.659 |
| Meds ($p = 10$) | 0.515 | 0.513 |
| Colonoscopies ($p = 2$) | 0.513 | 0.502 |
| Labs ($p = 200$) | 0.671 | 0.621 |
| All ($p = 240$) | 0.814 | 0.778 |
| Random forest | - | - |
| All ($p = 240$) | 0.985 | 0.844 |

Random forest variable importance. Most important blood lab measurements (all means of measurements over the 2 year window):

1. Hematocrit value (CBC labs)
2. MCH value (CBC labs)
3. ALT value (LFT labs)
4. Alk. Phos. value (LFT labs)
5. AST value (LFT labs)
6. White blood cells (WBC) value (CBC labs)
7. Glucose value (BMP labs)

Next steps:

- Better approaches to impute missing variables (esp. SmokeStatus).
- Improved non-linear classifiers for subsample.
- Logistic regression baselines for the full sample of 6M observations.

05/11 update

- Baseline logistic regression on full sample (no medication data).
- $n = 6,649,108$ observations, with $n_{\text{control}} = 6,637,713$ and $n_{\text{case}} = 11,395$.
- Most missing variables imputed with their median, SmokeStatus imputed at random proportional to non-missing.
- For lab information, only means of each measurement, no longitudinal information so far.
- **05/27 update:** Charlson and Demographic+Charlson updated to exclude Cancer and Metastatic Carcinoma.

| Model/predictors ($p = \#$ of predictors) | Training AUC | Test AUC |
|--|--------------|----------|
| Demographic ($p = 10$) | 0.670 | 0.675 |
| Charlson ($p = 16$) | 0.684 | 0.705 |
| Demographic + Charlson ($p = 26$) | 0.760 | 0.770 |
| Colonoscopies ($p = 2$) | 0.511 | 0.504 |
| Labs ($p = 34$) | 0.604 | 0.601 |
| Meds ($p = 10$) | 0.523 | 0.512 |

Next steps:

- Process medication data.
- Approaches to impute missing variables (esp. SmokeStatus).
- Non-linear classifiers for subsample, plots/graphics to interpret predictor effects.

05/18 update

- Medication data processed for full sample, new line (red) added to last week's table of baseline logistic regression AUCs.
- Gradient boosting implemented for the random sample of 5000 cases and 5000 controls.

| Model/predictors ($p = \#$ of predictors) | Training AUC | Test AUC |
|--|--------------|--------------|
| Logistic regression | - | - |
| All ($p = 240$) | 0.814 | 0.778 |
| Random forest | - | - |
| All ($p = 240$) | 0.985 | 0.844 |
| Gradient boosting | - | - |
| All ($p = 240$), no interactions | 0.900 | 0.847 |
| All ($p = 240$), 2-way interactions | 0.924 | 0.857 |
| All ($p = 240$), 3-way interactions | 0.937 | 0.864 |

Same gradient boosting results without Charlson score inputs or GERD at index:

| Model/predictors ($p = \#$ of predictors) | Training AUC | Test AUC |
|--|--------------|----------|
| Gradient boosting | - | - |
| No Charlson inputs ($p = 222$), no interactions | 0.881 | 0.811 |
| No Charlson inputs ($p = 222$), 2-way interactions | 0.898 | 0.820 |
| No Charlson inputs ($p = 222$), 3-way interactions | 0.909 | 0.817 |

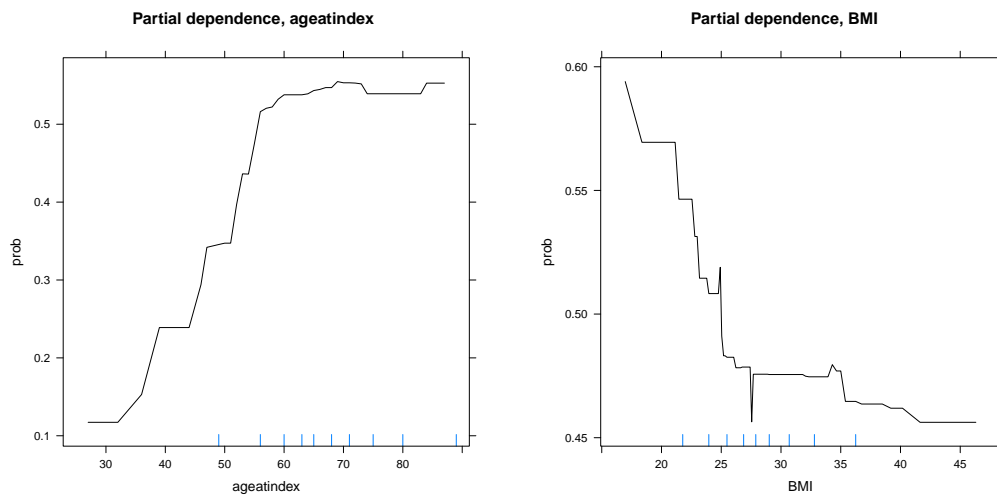
Next steps

- Scaling up gradient boosting, dealing with unbalanced classes in full sample.
- Plots/graphics to interpret black box models.

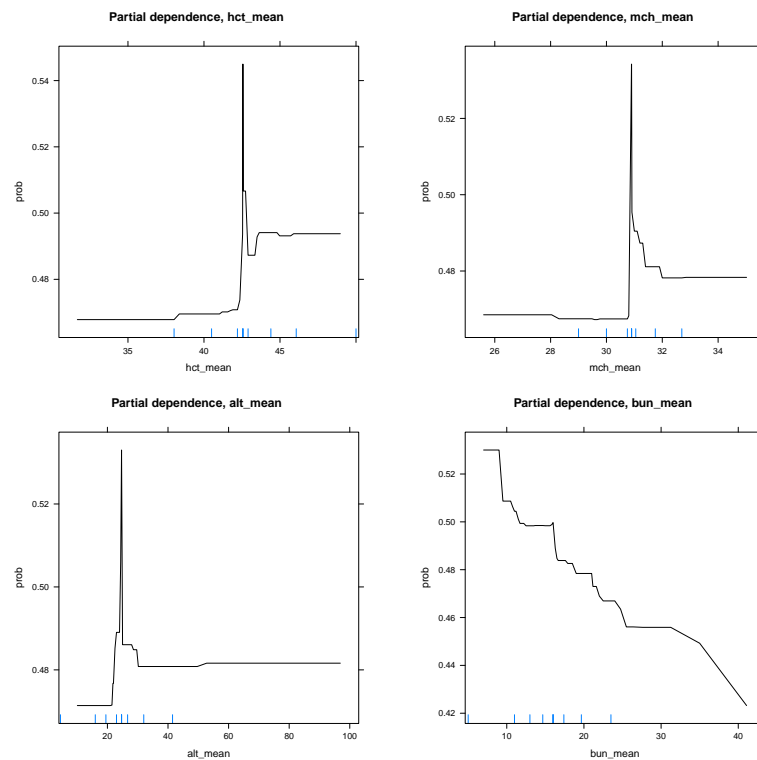
05/25 update

- Results now without Cancer and Metastatic Carcinoma inputs to Charlson score.
- Implemented scalable **xgboost** model, test AUC **0.860**.

Partial dependence plots (from xgboost model). Demographic variables:



Most influential blood lab variables:



Missingness rates for blood lab variables (where ‘missing’ means zero measurements of that variable in the 2 year window):

| Proportion missing | controls | cases |
|---------------------------|----------|-------|
| A1c_mean (A1C Labs) | 0.51 | 0.52 |
| bun_mean (BMP Labs) | 0.17 | 0.30 |
| hct_mean (CBC Labs) | 0.20 | 0.50 |
| CRP_mean (CRP Labs) | 0.95 | 0.94 |
| alkphos_mean (LFT Labs) | 0.23 | 0.47 |
| chol_mean (Lipid Labs) | 0.19 | 0.32 |

08/20: now with the new data, 4 year window from index-1 to index-5

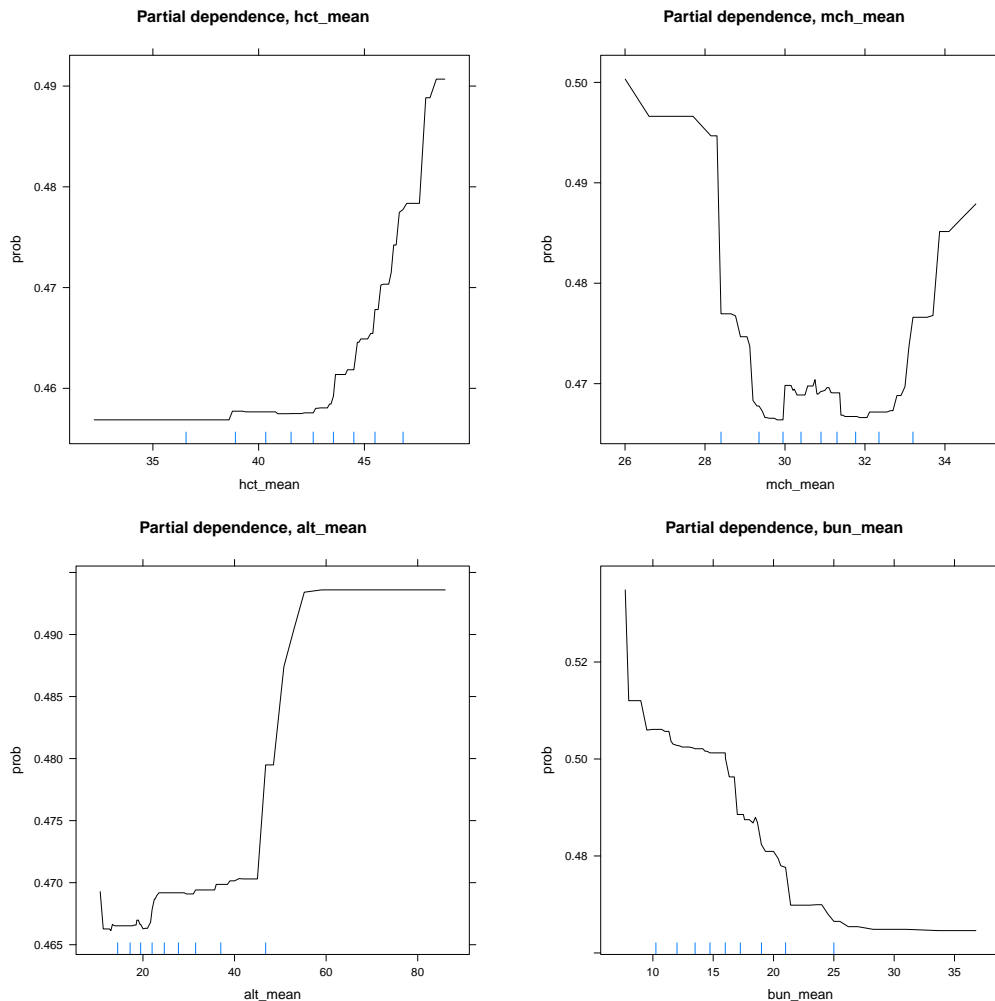
| Proportion missing | controls | cases |
|---------------------------|----------|-------|
| A1c_mean (A1C Labs) | 0.43 | 0.47 |
| bun_mean (BMP Labs) | 0.11 | 0.26 |
| hct_mean (CBC Labs) | 0.12 | 0.46 |
| CRP_mean (CRP Labs) | 0.93 | 0.91 |
| alkphos_mean (LFT Labs) | 0.15 | 0.42 |
| chol_mean (Lipid Labs) | 0.12 | 0.29 |

uniformly fewer missing, but rates are not cut in half, moreover heterogeneity of missingness between cases and controls is not fixed.

06/08 update

- Re-coded BMI/weight measurements?
- Re-run xgboost on balanced subsample with NAs, test AUC **0.880**, but it is using missing values to identify cases.
- Source of missingness? More detailed breakdown for LFT Labs.

Updated partial dependence plots for blood lab variables. Using new xgboost model on subsample.



Detailed breakdown of missingness for LFT labs. (Prediction window is second and third years before index)

| Mean # of labs | controls | cases |
|--------------------------|----------|-------|
| First year before index | 1.87 | 2.42 |
| Second year before index | 1.47 | 1.52 |
| Third year before index | 1.42 | 1.34 |

Detailed breakdown for the prediction window (index minus 3 years to index minus 1 year):

| Proportion | controls | cases |
|-------------------|----------|-------|
| At least 1 lab | 0.83 | 0.69 |
| At least 2 labs | 0.67 | 0.59 |
| At least 3 labs | 0.45 | 0.44 |
| At least 4 labs | 0.30 | 0.31 |
| At least 5 labs | 0.18 | 0.21 |

Next Steps:

- Imputation with xgboost
- Implement xgboost for entire sample

06/22 update

- Set up code on GitLab (waljee-zhu-ml-projects/hosea-project).
- Received updated sample with recoded BMIs. BMIs are more likely to be missing for cases (14%) than controls (8%).
- Continuing to code imputation following Deng and Lumley (2021).

BMI comparison on the original sample data:

| Original BMI | < 20 | $\in (20, 25]$ | $\in (25, 30]$ | $\in (30, 35]$ | $\in (35, 40]$ | > 40 |
|---------------------------|--------|----------------|----------------|----------------|----------------|--------|
| $\mathbb{P}(Control BMI)$ | 0.9960 | 0.9978 | 0.9985 | 0.9985 | 0.9986 | 0.9986 |
| $\mathbb{P}(Case BMI)$ | 0.0040 | 0.0022 | 0.0015 | 0.0015 | 0.0014 | 0.0014 |

With the re-coded sample data:

| New BMI | < 20 | $\in (20, 25]$ | $\in (25, 30]$ | $\in (30, 35]$ | $\in (35, 40]$ | > 40 |
|---------------------------|--------|----------------|----------------|----------------|----------------|--------|
| $\mathbb{P}(Control BMI)$ | 0.9983 | 0.9985 | 0.9985 | 0.9984 | 0.9981 | 0.9979 |
| $\mathbb{P}(Case BMI)$ | 0.0017 | 0.0015 | 0.0015 | 0.0016 | 0.0019 | 0.0021 |

06/29 update

- Now using updated data (new BMIs, Charlson scores).
- For subsample, comparison of baseline logistic regression and xgboost models for different imputation approaches:
 1. **No imputation:** leave NAs in the data, learn a classification for subjects with missing data (only compatible with xgboost).
 2. **Median imputation:** impute all variables with median/most common class.
 3. **Regression imputation:** impute by (linear) regression of each predictor variable on the others (similar to MICE).
 4. **Random sample imputation:** impute each predictor by sampling at random from the non-missing entries.

AUC results for xgboost/logistic regression for different imputation approaches. Results still on balanced subsample.

| Imputation method/model | Training AUC | Test AUC | Test AUC (complete records) |
|---------------------------------|--------------|----------|-----------------------------|
| No imputation | - | - | - |
| xgboost | 0.977 | 0.942 | 0.643 |
| Median imputation | - | - | - |
| logistic regression | 0.834 | 0.810 | 0.585 |
| xgboost | 0.970 | 0.857 | 0.623 |
| Regression imputation | - | - | - |
| logistic regression | 0.805 | 0.760 | 0.557 |
| xgboost | 0.948 | 0.815 | 0.705 |
| Random sample imputation | - | - | - |
| logistic regression | 0.724 | 0.689 | 0.724 |
| xgboost | 0.919 | 0.729 | 0.745 |

Regression imputation still has ‘regression to mean’ effect for imputed values, imputed values have smaller variance than the true values (with observation error). Maybe try a version that imputes a rank then resamples from observed values.

07/06 update

- Source of missingness in the data? About 85% of cases missing Charlson score inputs (compared to about 47% of controls).
- Missingness in test set/use case?
- Updated (in red) last weeks results with a test set of 1000 complete records (226 cases, 774 controls).
 - Poor generalization to complete cases for no imputation, median imputation implies that those models are fitting to the missingness, patterns won't continue to hold with fully observed data.
 - Regression imputation still has 'regression to mean' effect for imputed values, imputed values have smaller variance than the observed values (observation error).
 - Good generalization of random sample imputation implies that there is no bias introduced from training on data with missingness.

07/13 update

- Should we evaluate the model on complete records or missing/imputed records?
- Recall: previous results with an additional test set of 1000 complete records (226 cases, 774 controls).
 - New comparison for **multiple** random sample imputation: imputes by sampling from non-missing entries several times (reps) for each missing value, training on the multiple imputed records.
- Memory issues fitting xgboost on the full data, implementing “batched” training.

AUC’s for xgboost/logistic regression for different imputation approaches. Results still on balanced subsample.

| Imputation method/ prediction model | Training AUC | Test AUC (imputed records) | Test AUC (complete records) |
|--|--------------|-------------------------------|--------------------------------|
| No imputation | - | - | - |
| xgboost | 0.977 | 0.942 | 0.643 |
| Median imputation | - | - | - |
| logistic regression | 0.834 | 0.810 | 0.585 |
| xgboost | 0.970 | 0.857 | 0.623 |
| Regression imputation | - | - | - |
| logistic regression | 0.805 | 0.760 | 0.557 |
| xgboost | 0.948 | 0.815 | 0.705 |
| Random sample imputation | - | - | - |
| logistic regression | 0.724 | 0.689 | 0.724 |
| xgboost | 0.919 | 0.729 | 0.745 |
| Multiple rand. samp. | - | - | - |
| xgboost, 10 reps | 0.909 | 0.742 | 0.781 |
| xgboost, 20 reps | 0.939 | 0.765 | 0.820 |
| xgboost, 30 reps | 0.948 | 0.768 | 0.781 |

Final two columns suggest two different evaluation metrics, ensuring the model performs well on **both** (A) new patients with imputed data and (B) new patients with fully observed blood labs, etc.

07/20 update

- Continuing to test/tune imputation approaches.
- Notes on decision curve analysis.

| Imputation method/ prediction model | Training AUC | Test AUC (imputed records) | Test AUC (complete records) |
|--|--------------|-------------------------------|--------------------------------|
| Separate class | - | - | - |
| xgboost | 0.977 | 0.942 | 0.643 |
| Median | - | - | - |
| logistic regression | 0.834 | 0.810 | 0.585 |
| xgboost | 0.970 | 0.857 | 0.623 |
| Regression | - | - | - |
| logistic regression | 0.805 | 0.760 | 0.557 |
| xgboost | 0.948 | 0.815 | 0.705 |
| Regression v2 | - | - | - |
| logistic regression | 0.769 | 0.706 | 0.564 |
| xgboost | 0.928 | 0.722 | 0.764 |
| Random sample | - | - | - |
| logistic regression | 0.724 | 0.689 | 0.724 |
| xgboost | 0.919 | 0.729 | 0.756 |
| Multiple rand. samp. | - | - | - |
| xgboost, 10 imputes \times 50 trees | 0.909 | 0.742 | 0.781 |
| xgboost, 20 imputes \times 25 trees | 0.906 | 0.750 | 0.778 |
| xgboost, 30 imputes \times 20 trees | 0.909 | 0.765 | 0.788 |
| xgboost, 100 imputes \times 5 trees | 0.898 | 0.762 | 0.806 |

07/27 update

- Continuing to tweak regression imputation approaches.
- Comparing parameters/number of imputations needed for multiple sample imputation.
- Notes on decision curve analysis.

Next steps:

- Re-process new data when available

| Imputation method/ prediction model | Training AUC | Test AUC (imputed records) | Test AUC (complete records) |
|--|--------------|-------------------------------|--------------------------------|
| Separate class | - | - | - |
| xgboost | 0.977 | 0.942 | 0.643 |
| Median | - | - | - |
| logistic regression | 0.834 | 0.810 | 0.585 |
| xgboost | 0.970 | 0.857 | 0.623 |
| Regression | - | - | - |
| logistic regression | 0.759 | 0.704 | 0.657 |
| xgboost | 0.902 | 0.742 | 0.757 |
| Single rand. samp. | - | - | - |
| logistic regression | 0.724 | 0.689 | 0.724 |
| xgboost | 0.919 | 0.729 | 0.756 |
| Multiple rand. samp. | - | - | - |
| xgboost, 10 imputes | 0.909 | 0.742 | 0.781 |
| xgboost, 20 imputes | 0.906 | 0.750 | 0.778 |
| xgboost, 30 imputes | 0.909 | 0.765 | 0.788 |
| xgboost, 100 imputes | 0.898 | 0.762 | 0.806 |

08/03 update

- Received new data, loading/processing files
- Still issues with missing Charlson scores (0's vs NAs), relationship between CaseControl indicator and number of visits.

08/17 update

- Processed longitudinal summaries for BMP labs. Missingness rates down but still different for cases and controls. Continuing to process new blood lab data.
- Incidence rates for some example Charlson inputs: Peptic Ulcer Disease, Renal Disease, GERD. Full tables for all inputs are on the next page.

| Variable (ignoring NA's) | Among cases | Among controls |
|-----------------------------------|------------------|------------------------|
| pu | 6.8% (120/1762) | 3.8% (142821/3803451) |
| RD | 11.8% (208/1762) | 12.1% (460302/3803451) |
| GERD | - | - |

| Variable (impute NA's as 0's) | Among cases | Among controls |
|--|--------------------|-------------------------|
| pu | 1.1% (120/11395) | 2.2% (142821/6637713) |
| RD | 1.8% (208/11395) | 6.9% (460302/6637713) |
| GERD | 17.9% (2044/11395) | 15.7% (1039206/6637713) |

Cross-classified by number of BMP lab results (0, 1 or 2+):

Peptic Ulcer Disease, **controls**:

| Number of BMP labs | 0 | 1 | NA |
|--------------------|-------|------|-------|
| 0 | 29.3% | 0.9% | 69.8% |
| 1 | 32.2% | 0.9% | 66.8% |
| 2+ | 61.5% | 2.5% | 36.0% |

Peptic Ulcer Disease, **cases**:

| Number of BMP labs | 0 | 1 | NA |
|--------------------|-------|------|-------|
| 0 | 8.6% | 0.7% | 90.8% |
| 1 | 10.7% | 0.4% | 88.8% |
| 2+ | 17.0% | 1.3% | 81.7% |

Full tables for all Charlson indicators (**red** indicates these variables have been dropped from past models):

| Variable (ignoring NA's) | Among cases | Among controls |
|-----------------------------------|------------------|-------------------------|
| CANCER | 30.5% (538/1762) | 18.3% (695558/3803451) |
| CHF | 12.3% (217/1762) | 11.8% (450090/3803451) |
| CTD | 2.9% (51/1762) | 3.0% (115492/3803451) |
| DEM | 1.4% (24/1762) | 2.3% (85704/3803451) |
| DIAB_C | 15.5% (273/1762) | 13.8% (525226/3803451) |
| HIV | 0.3% (6/1762) | 0.8% (30507/3803451) |
| MET_CAR | 6.6% (117/1762) | 1.3% (47696/3803451) |
| MLD | 9.1% (160/1762) | 8.0% (305462/3803451) |
| MSLD | 0.7% (13/1762) | 0.8% (29495/3803451) |
| PARA | 1.9% (33/1762) | 1.9% (72051/3803451) |
| RD | 11.8% (208/1762) | 12.1% (460302/3803451) |
| cd | 16.3% (287/1762) | 15.3% (582328/3803451) |
| copd | 43.1% (759/1762) | 35.9% (1365577/3803451) |
| diab_nc | 44.6% (786/1762) | 44.8% (1702629/3803451) |
| mi | 9.1% (161/1762) | 7.0% (264731/3803451) |
| pud | 6.8% (120/1762) | 3.8% (142821/3803451) |
| pvd | 20.1% (354/1762) | 16.0% (609425/3803451) |
| GERD | - | - |

| Variable (impute NA's as 0's) | Among cases | Among controls |
|--|--------------------|-------------------------|
| CANCER | 4.7% (538/11395) | 10.5% (69558/6637713) |
| CHF | 1.9% (217/11395) | 6.8% (450090/6637713) |
| CTD | 0.4% (51/11395) | 1.7% (115492/6637713) |
| DEM | 0.2% (24/11395) | 1.3% (85704/6637713) |
| DIAB_C | 2.4% (273/11395) | 7.9% (525226/6637713) |
| HIV | 0.1% (6/11395) | 0.5% (30507/6637713) |
| MET_CAR | 1.0% (117/11395) | 0.7% (47696/6637713) |
| MLD | 1.4% (160/11395) | 4.6% (305462/6637713) |
| MSLD | 0.1% (13/11395) | 0.4% (29459/6637713) |
| PARA | 0.3% (33/11395) | 1.1% (72051/6637713) |
| RD | 1.8% (208/11395) | 6.9% (460302/6637713) |
| cd | 2.5% (287/11395) | 8.8% (582328/6637713) |
| copd | 6.7% (759/11395) | 20.6% (1365577/6637713) |
| diab_nc | 6.9% (786/11395) | 25.6% (1702629/6637713) |
| mi | 1.4% (161/11365) | 4.0% (264731/6637713) |
| pud | 1.1% (120/11395) | 2.2% (142821/6637713) |
| pvd | 3.1% (354/11395) | 9.2% (609425/6637713) |
| GERD | 17.9% (2044/11395) | 15.7% (1039206/6637713) |

08/31 update

- Re-fit xgboost models on new subsampled data (compare to 07/27 results).
 $n_{\text{train}} = 9000$, $n_{\text{test}} = 1000$, $n_{\text{complete.test}} = 1000$.
- Compare (xgboost) model performance with different groups of variables included.
- Ongoing: a proxy for number of visits to better impute Charlson ICD codes?

Comparing different imputation approaches and prediction models:

| Imputation method/ prediction model | Training AUC | Test AUC (imputed records) | Test AUC (complete records) |
|--|--------------|-------------------------------|--------------------------------|
| Separate class | - | - | - |
| xgboost | 0.985 | 0.962 | 0.582 |
| Median | - | - | - |
| logistic regression | 0.853 | 0.846 | 0.583 |
| xgboost | 0.963 | 0.911 | 0.618 |
| Regression | - | - | - |
| logistic regression | 0.772 | 0.726 | 0.651 |
| xgboost | 0.911 | 0.750 | 0.798 |
| Single rand. samp. | - | - | - |
| logistic regression | 0.739 | 0.695 | 0.703 |
| xgboost | 0.916 | 0.780 | 0.833 |
| Multiple rand. samp. | - | - | - |
| xgboost, 10 imputes | 0.942 | 0.824 | 0.885 |
| xgboost, 20 imputes | 0.935 | 0.820 | 0.872 |
| xgboost, 30 imputes | 0.943 | 0.810 | 0.879 |
| xgboost, 100 imputes | 0.931 | 0.824 | 0.891 |

Comparing different included variables (imputing with single random sampling, fit with xgboost):

| Variables included | Training AUC | Test AUC (imputed records) | Test AUC (complete records) |
|------------------------------|--------------|-------------------------------|--------------------------------|
| Demographic ($p = 11$) | 0.743 | 0.700 | 0.663 |
| Charlson + GERD ($p = 16$) | 0.600 | 0.542 | 0.530 |
| Medications ($p = 10$) | 0.594 | 0.501 | 0.539 |
| Labs ($p = 202$) | 0.903 | 0.687 | 0.829 |