Comorbidity and medical costs among Medicare beneficiaries with HIV in California: A case study in predictive data analysis

Kodi B. Arfer
14 Mar 2017
CHIPTS Methods Seminar, UCLA

In collaboration with Arleen Leibowitz, Jennifer Gildner, Katherine Desmond, David Zingmond, and Scott Comulada

Thinking inside the Box

> All models are wrong but some are useful.
>
> —George E. P. Box

Useful how?

What if instead of trying to infer properties of the true model, we focus on predictions?

What is prediction?

- Predicting specific individual values
  - E.g., "The annual outpatient costs of a 52-year-old white female Medicare beneficiary with HIV and lymphoma will be $4,303."
  - As opposed to, e.g., "Medicare beneficiaries with HIV will have higher outpatient costs if they have lymphoma."
- We estimate the predictive accuracy of a model by comparing estimated costs to true costs, per case.
  - E.g, The estimated cost is $4,303 but the true cost is $6,405. So the absolute error is $2,101.
    - The mean absolute error, across all cases, assesses the model's overall accuracy.
  - We use cross-validation to avoid the optimistic bias of overfitting.

Advantages of prediction

- Model truth no longer even matters.
  - A wronger model might predict better.
- A very large variety of models can be considered, and in a directly parallel way.
- Parsimony is automatically rewarded.
  - Simpler models are hurt less by overfitting.
- The analytic results are of immediate practical interest.

The current study

- We consider 2010 insurance claims data from 9,767 people with HIV and Medicare in California.
  - 74% also have Medicaid.
- Care for these people is expensive (median cost $35k)…
  - …but varies a lot (1st quartile $25k, 3rd quartile $52k).
- 64% of subjects have at least one of 27 coded comorbid conditions.
  - Using these should help predict individual costs.

Independent variables (IVs)

Has Medicaid, age, gender, race, is disabled, lives in an urban area, visits a high-volume HIV provider

Comorbidities:

- Congestive heart failure
- Cardiac arrhythmias
- Valvular disease
- Peripheral vascular disorders
- Hypertension, uncomplicated
- Hypertension, complicated
- Paralysis
- Other neurological disorders
- Pulmonary circulation disorders
- Chronic pulmonary disease
- Diabetes, uncomplicated
- Diabetes, complicated
- Hypothyroidism
- Renal failure
- Liver disease
- Peptic ulcer disease
- Lymphoma
- Metastatic cancer
- Solid tumor without metastasis
- Rheumatoid arthritis
- Coagulopathy
- Coagulopathy hemophilia
- Blood loss anemia
- Deficiency anemia
- Obesity
- Weight loss
- Fluid and electrolyte disorders

Dependent variables (DVs)

We conduct separate analyses for each of:

- Outpatient costs
- Inpatient costs
- Drug costs

Obstacle #1: skew

- A minority have very high costs.
  - 8% of subjects cost > $100k
  - 2% of subjects cost > $200k
- If we assess predictions with squared error, the extreme values will dominate.
- We use absolute error instead.
  - Instead of OLS, which minimizes squared error (by finding the conditional mean), we use quantile regression, which minimizes absolute error (by finding the conditional median).
- We also log-transform the DVs before fitting models.
  - The predictions are antilogged.

Obstacle #2: Medicaid

- 74% also have Medicaid, which could lead to radically different cost patterns.
- It would be nice if we could just treat having Medicaid as another IV, but can we?
- Let's find out by including this question in the model comparison.
- We pit **single** models, which use Medicaid as an ordinary IV, against **twinned** models, which internally use completely separate regression models for people who do vs. don't have Medicaid.

Model-comparison strategy

For each DV, we vary two aspects of the models:

- Single vs. twinned models
- What IVs are included:
  - Nothing (a trivial model)
  - Demographic IVs only (i.e., everything but comorbidities)
  - All IVs

The performance of the trivial models provides a baseline measure of accuracy.

The performance of the models with all IVs, compared to the models with demographic IVs only, shows how much comorbidities have predictive value **over and above** that of age, race, etc.

Results for outpatient costs

IVs	Twin?	MAE
trivial	single	$6,565
trivial	twinned	$6,555
demographic only	single	$6,495
demographic only	twinned	$6,487
all	single	$6,146
all	twinned	$6,165

- Demographic variables help a little, but comorbidities help much more.
- When we have all IVs, the extra complication of twinned models doesn't help; it only increases MAE a bit.
  - It probably overfit.

Results for drug costs

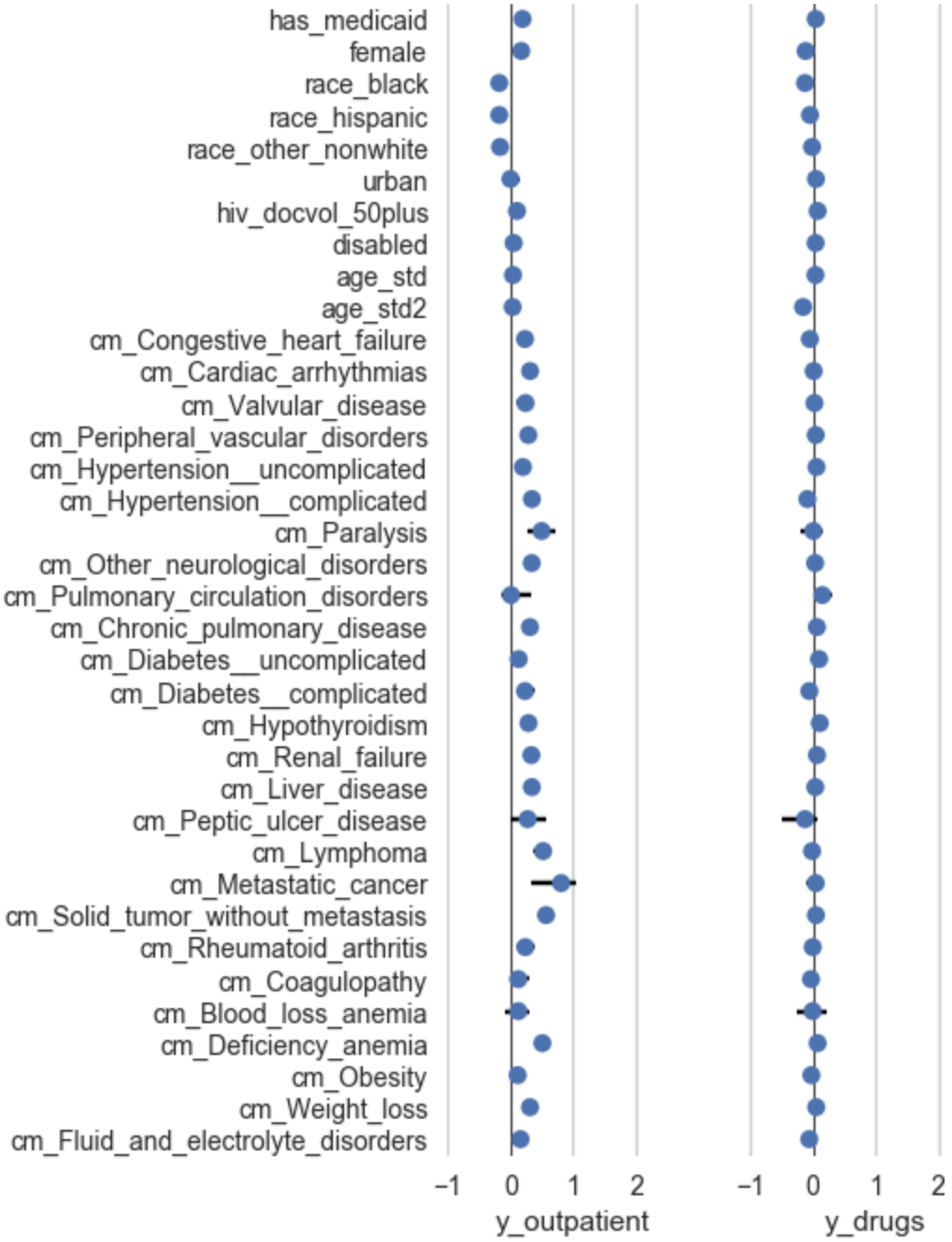
IVs	Twin?	MAE
trivial	single	$12,943
trivial	twinned	$12,932
demographic only	single	$12,701
demographic only	twinned	$12,683
all	single	$12,581
all	twinned	$12,595

- Similar results, but less dramatic (smaller improvements, proportional to baseline).
  - Drug costs are largely antiretrovirals (ARVs), which all subjects share. Comorbidities can't help much with predicting variability in ARV costs.

Notable coefficients

Since we logged the DVs, the postulated effects are multiplicative, not additive.

- Hemophilia predicts 20× outpatient costs and 3× drug costs
- Metastatic cancer predicts 2.25× outpatient costs
- Lymphoma, deficiency anemia, and paralysis predict 1.6× outpatient costs


has_medicaid
female
race_black
race_hispanic
race_other_nonwhite
urban
hiv_docvol_50plus
disabled
age_std
age_std2
cm_Congestive_heart_failure
cm_Cardiac_arrhythmias
cm_Valvular_disease
cm_Peripheral_vascular_disorders
cm_Hypertension__uncomplicated
cm_Hypertension__complicated
cm_Paralysis
cm_Other_neurological_disorders
cm_Pulmonary_circulation_disorders
cm_Chronic_pulmonary_disease
cm_Diabetes__uncomplicated
cm_Diabetes__complicated
cm_Hypothyroidism
cm_Renal_failure
cm_Liver_disease
cm_Peptic_ulcer_disease
cm_Lymphoma
cm_Metastatic_cancer
cm_Solid_tumor_without_metastasis
cm_Rheumatoid_arthritis
cm_Coagulopathy
cm_Blood_loss_anemia
cm_Deficiency_anemia
cm_Obesity
cm_Weight_loss
cm_Fluid_and_electrolyte_disorders
−1
0
1
2
y_outpatient
−1
0
1
2
y_drugs

Obstacle #3: zero-inflated inpatient costs

Only 26% of subjects have nonzero inpatient costs (i.e., were hospitalized).

It would be nice to predict probability of nonzero costs, instead of just amounts.

We try a two-stage model:

- Probability: logistic regression to predict whether the subject has any inpatient costs.
- Amount: quantile regression, only among those with nonzero costs, to predict the amount.

Model-comparison strategy

For simplicity, we omit the old twinned models.

We vary the IVs of each of the two stages (probability and amount) just like before:

- Nothing (a trivial model)
- Demographic IVs only (i.e., everything but comorbidities)
- All IVs

When the probability model is trivial, we ignore it and use the amount model only, as for the earlier DVs.

Results for inpatient costs

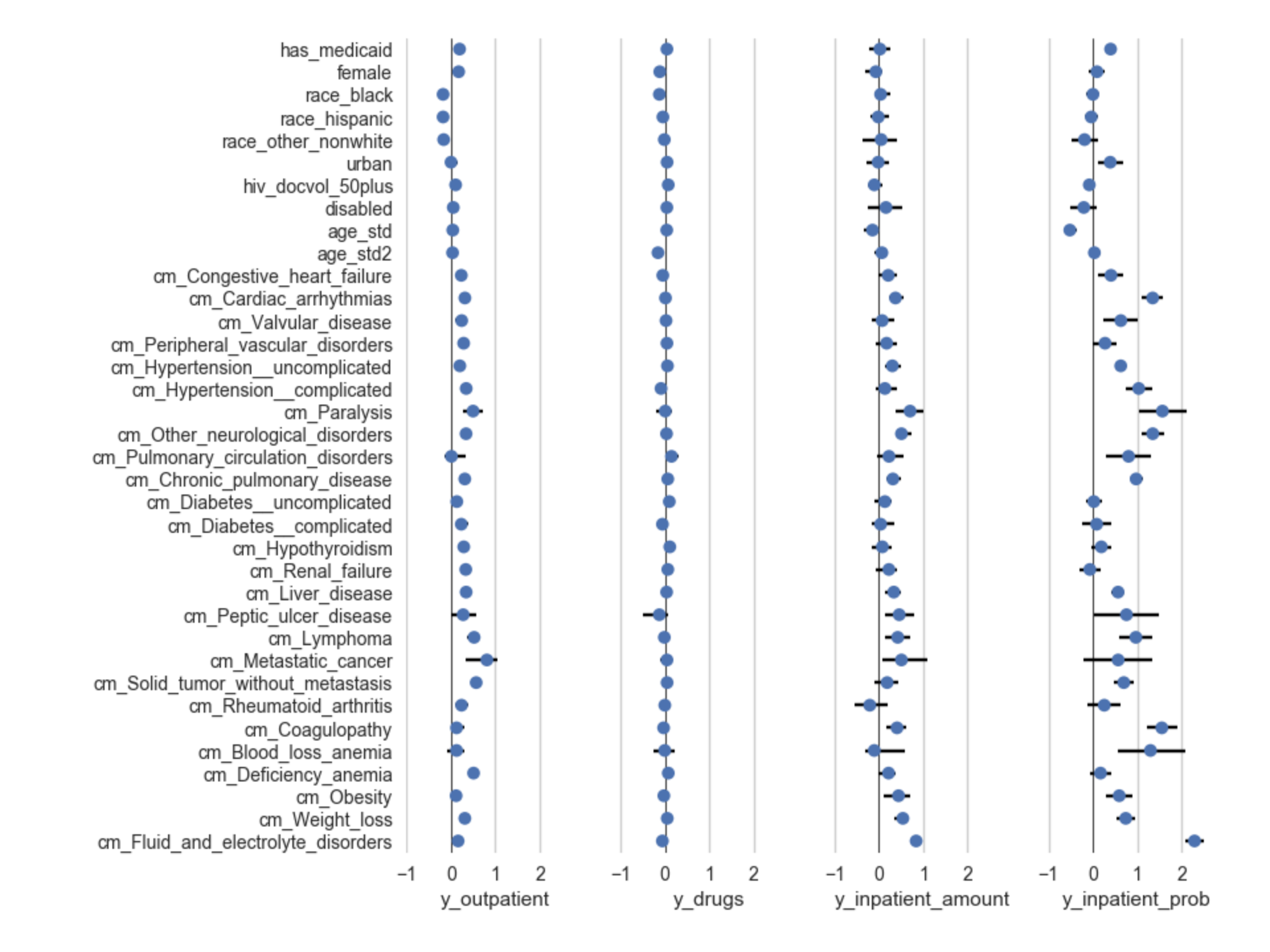
IVs, probability	IVs, amount	MAE
trivial	trivial	$8,489
trivial	demographic only	$8,489
trivial	all	$9,832
demographic only	trivial	$8,497
demographic only	demographic only	$8,505
demographic only	all	$8,501
all	trivial	$7,809
all	demographic only	$7,777
all	all	$6,792

The most complex model wins, by a substantial margin.

Without the probability model, we wouldn't have improved over baseline at all.

Notable coefficients

- Probability stage
  - Fluid and electrolyte disorders predict 10× odds of hospitalization.
  - Paralysis and coagulopathy (excl. hemophilia) predict 5× odds.
  - Cardiac arrhythmias and misc. neurological disorders predict 4× odds.
- Amount stage
  - Fluid and electrolyte disorders, hemophilia, and paralysis predict 2× costs.
  - Many other conditions predict at least 1.5× costs.


has_medicaid
female
race_black
race_hispanic
race_other_nonwhite
urban
hiv_docvol_50plus
disabled
age_std
age_std2
cm_Congestive_heart_failure
cm_Cardiac_arrhythmias
cm_Valvular_disease
cm_Peripheral_vascular_disorders
cm_Hypertension__uncomplicated
cm_Hypertension__complicated
cm_Paralysis
cm_Other_neurological_disorders
cm_Pulmonary_circulation_disorders
cm_Chronic_pulmonary_disease
cm_Diabetes__uncomplicated
cm_Diabetes__complicated
cm_Hypothyroidism
cm_Renal_failure
cm_Liver_disease
cm_Peptic_ulcer_disease
cm_Lymphoma
cm_Metastatic_cancer
cm_Solid_tumor_without_metastasis
cm_Rheumatoid_arthritis
cm_Coagulopathy
cm_Blood_loss_anemia
cm_Deficiency_anemia
cm_Obesity
cm_Weight_loss
cm_Fluid_and_electrolyte_disorders
−1 0 1 2
y_outpatient
−1 0 1 2
y_drugs
−1 0 1 2
y_inpatient_amount
−1 0 1 2
y_inpatient_prob

In review

- Predictive data analysis is a data-driven (rather than theory-driven) approach.
  - It has a machine-learning flavor, although it can use familiar statistical models.
- It can cope with any kind of "model" that produces predictions, no matter the internal structure, and compare all models side-by-side.
  - The model can be a black box.
- It automatically seeks a balance between excess complexity (overfitting) and oversimplification (underfitting).
- It uses metrics of model quality that are of substantive interest, not just means to an end.