An exploratory analysis of patient episode data

Henry Wilde

June 6, 2018

Cardiff University School of Mathematics

Motivation

Motivation

- Observe and understand cost variation
- Identify important slices in the data
- Develop methods for examining slices of the data
- Analyse their impact on costs and resource consumption

Structure and origin

This slide should be replaced with a graphic.

Structure and origin

This slide should be replaced with a graphic. All data is provided by the Cwm Taf University Health Board.

Structure and origin

This slide should be replaced with a graphic. All data is provided by the Cwm Taf University Health Board.

We have:

- 2,447,299 patient episodes
- 1,946,545 patient spells
- 865,421 individual patients

Structure and origin

This slide should be replaced with a graphic. All data is provided by the Cwm Taf University Health Board.

We have:

- 2,447,299 patient episodes
- 1,946,545 patient spells
- 865,421 individual patients

Each row is made up of roughly 250 attributes, including:

Structure and origin

This slide should be replaced with a graphic. All data is provided by the Cwm Taf University Health Board.

We have:

- 2,447,299 patient episodes
- 1,946,545 patient spells
- 865,421 individual patients

Each row is made up of roughly 250 attributes, including:

- personal identifiers and demographic information

Structure and origin

This slide should be replaced with a graphic. All data is provided by the Cwm Taf University Health Board.

We have:

- 2,447,299 patient episodes
- 1,946,545 patient spells
- 865,421 individual patients

Each row is made up of roughly 250 attributes, including:

- personal identifiers and demographic information
- condition and treatment indicators (HRG, OPCS4, ICD10)

Structure and origin

This slide should be replaced with a graphic. All data is provided by the Cwm Taf University Health Board.

We have:

- 2,447,299 patient episodes
- 1,946,545 patient spells
- 865,421 individual patients

Each row is made up of roughly 250 attributes, including:

- personal identifiers and demographic information
- condition and treatment indicators (HRG, OPCS4, ICD10)
- other clinical references

Structure and origin

This slide should be replaced with a graphic. All data is provided by the Cwm Taf University Health Board.

We have:

- 2,447,299 patient episodes
- 1,946,545 patient spells
- 865,421 individual patients

Each row is made up of roughly 250 attributes, including:

- personal identifiers and demographic information
- condition and treatment indicators (HRG, OPCS4, ICD10)
- other clinical references
- cost components

Summative analysis

Not strictly necessary to have this up as a slide. Better to just say it during previous slide.

Summative analysis

Not strictly necessary to have this up as a slide. Better to just say it during previous slide.

- Our data is skewed towards low-cost, short-stay episodes

Summative analysis

Not strictly necessary to have this up as a slide. Better to just say it during previous slide.

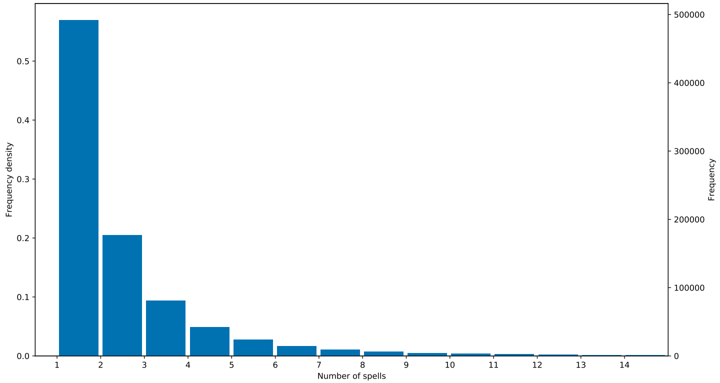
- Our data is skewed towards low-cost, short-stay episodes
- This extends to the spell level with largely one or two-time visits from patients

Not strictly necessary to have this up as a slide. Better to just say it during previous slide.

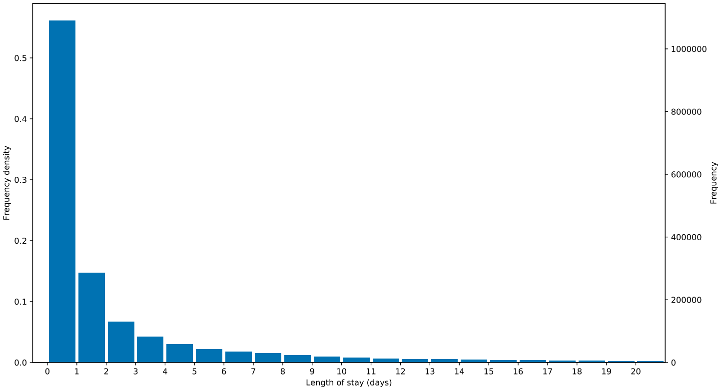
- Our data is skewed towards low-cost, short-stay episodes
- This extends to the spell level with largely one or two-time visits from patients
- Long and heavy tails are present in our costs and lengths of stay

Distributions of attributes

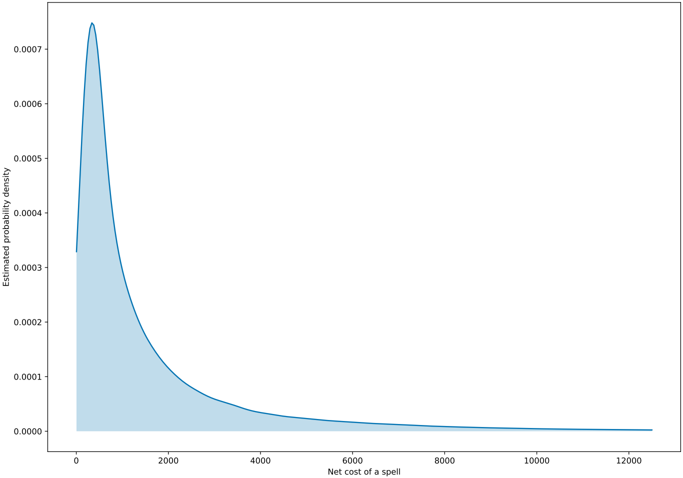
Number of spells



Length of stay (spell-wise)



Net cost



Estimated probability density
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0
2000
4000
6000
8000
10000
12000
Net cost of a spell

Summative analysis

Other areas of interest to us are:

- Other clinical measures
- Demographic variables
- Interactions between variables

Clinical variables

We will investigate:

- the number of diagnoses
- the number of procedures

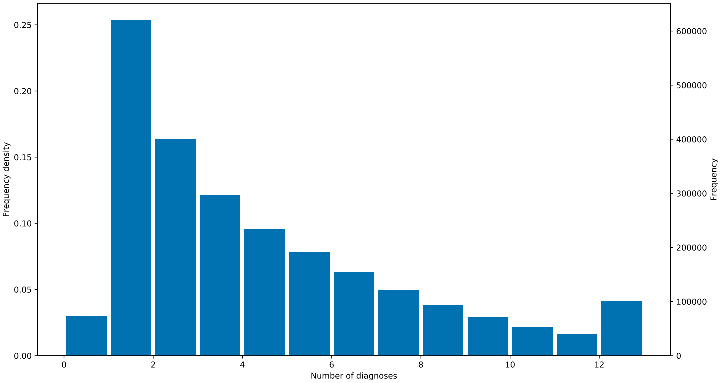
Clinical variables

We will investigate:

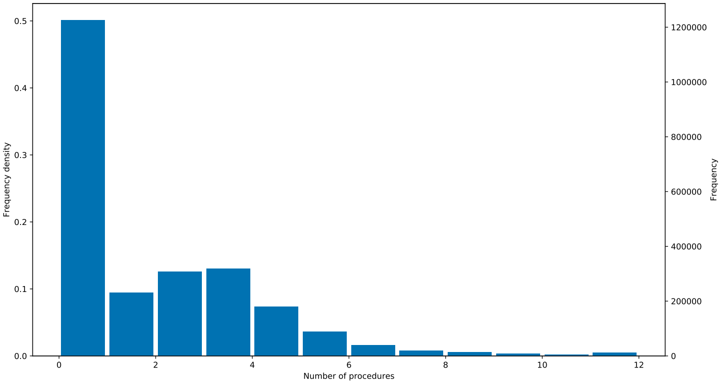
- the number of diagnoses
- the number of procedures

These contribute to comorbidity rates and presumably costs.

Number of diagnoses



Number of procedures



Demographic analysis

As it stands, demographic information is not well-recorded in the data.

Demographic analysis

As it stands, demographic information is not well-recorded in the data.

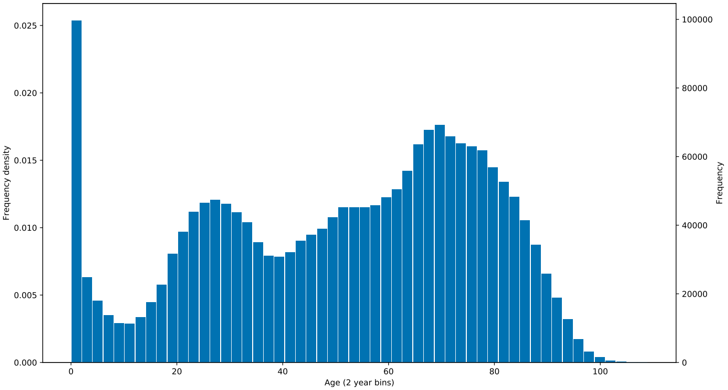
- Gender is strictly binary and not recorded for all patients or episodes

Demographic analysis

As it stands, demographic information is not well-recorded in the data.

- Gender is strictly binary and not recorded for all patients or episodes
- Limited geographic information is encoded in the GP practice code of the patient

Demographic analysis



Correlation

Measuring variation and importance

Measuring variation

This slide is unnecessary and confusing to non-mathematicians. Variance is not scale invariant and did lead to misconceptions.

Measuring variation

This slide is unnecessary and confusing to non-mathematicians. Variance is not scale invariant and did lead to misconceptions.

Definition

Let μ, σ^2 denote the population mean and population variance of some population respectively. Then we define the *coefficient of variation*, denoted by C_v, to be:

$$C_v = \frac{\sigma}{\mu}$$

Measuring variation

This slide is unnecessary and confusing to non-mathematicians. Variance is not scale invariant and did lead to misconceptions.

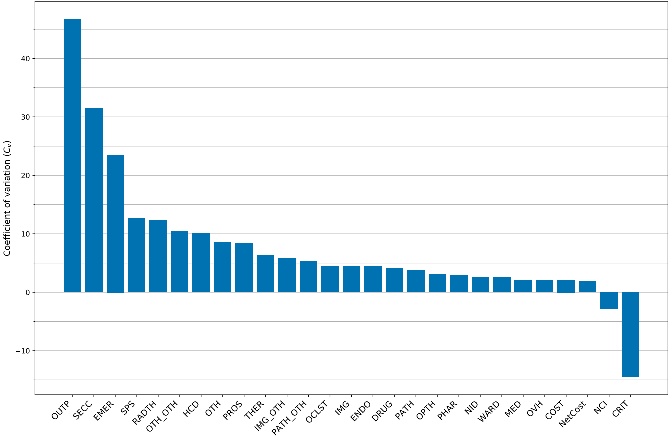
Definition

Let μ, σ^2 denote the population mean and population variance of some population respectively. Then we define the *coefficient of variation*, denoted by C_v, to be:

$$C_v = \frac{\sigma}{\mu}$$

The coefficient of variation is scale invariant, and allows us to see the relative variation in each of our cost components.

Measuring variation

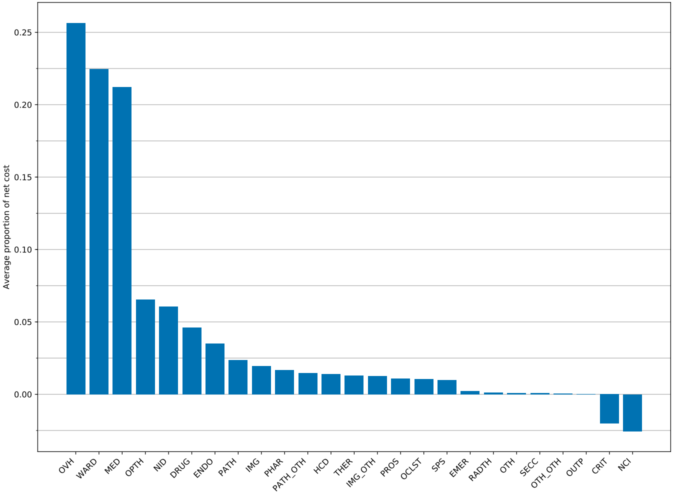


Are these actually important?

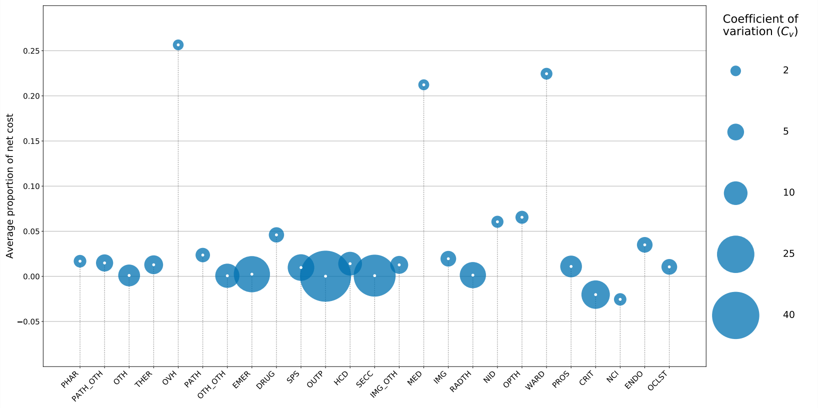
Despite the relative variation of our cost components being whatever value, does it matter to the actual cost?

Let us investigate their contribution to the final cost.

Cost component contribution



Visualising relative importance



Taking a slice: diabetic patient analysis

General methods for slice analysis

Too wordy? Given some slice of the data, we want to:

- Examine cost variations and general surface-level statistics
- Determine components and variable relationships of interest
- Consider the relative ‘cost’ of the patients in this slice
- Contrast this against its complement and the general dataset

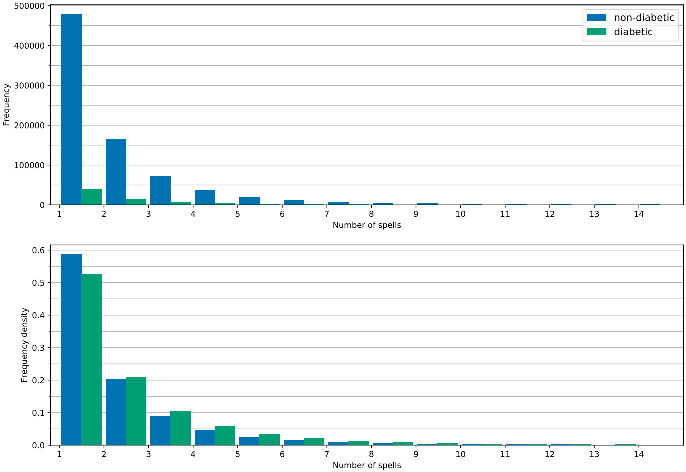
Diabetic patient analysis

This is a known area of interest to the health board.

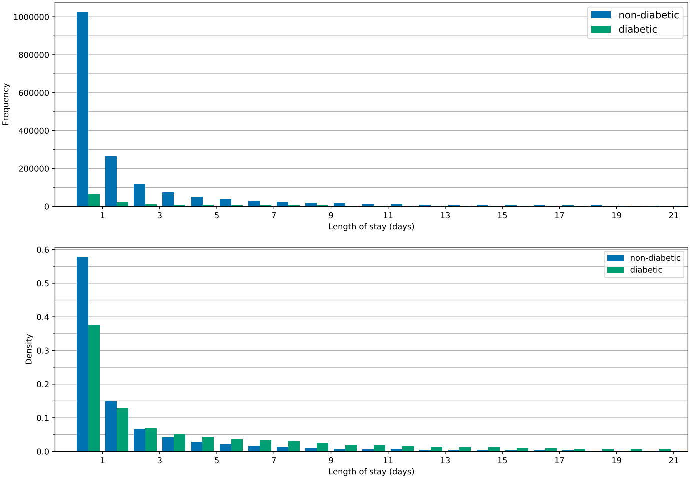
Diabetic patients make up 10.8% of all the episodes in the dataset, and roughly 8.7% of the unique patients in the dataset.

Here we consider patients to be ‘diabetic’ if they have diabetes flagged as either a primary or secondary condition in their episode.

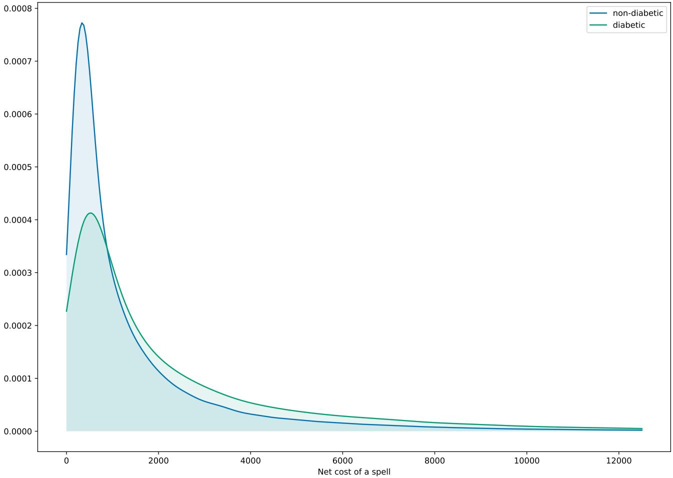
Number of spells



Length of stay

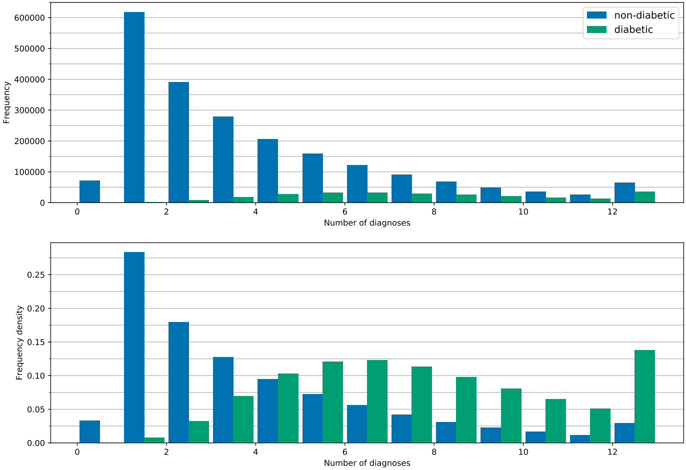


Net cost

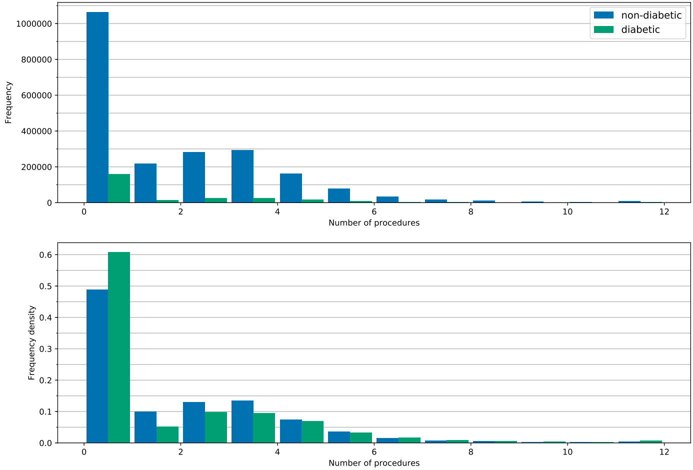


0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
2000
4000
6000
8000
10000
12000
Net cost of a spell
non-diabetic
diabetic

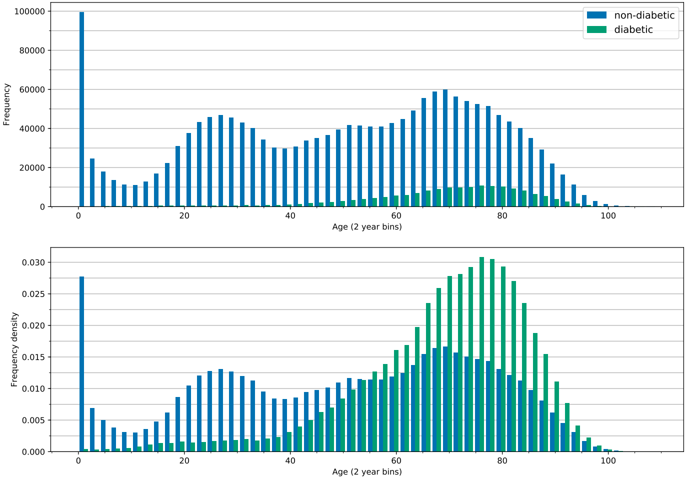
Number of diagnoses



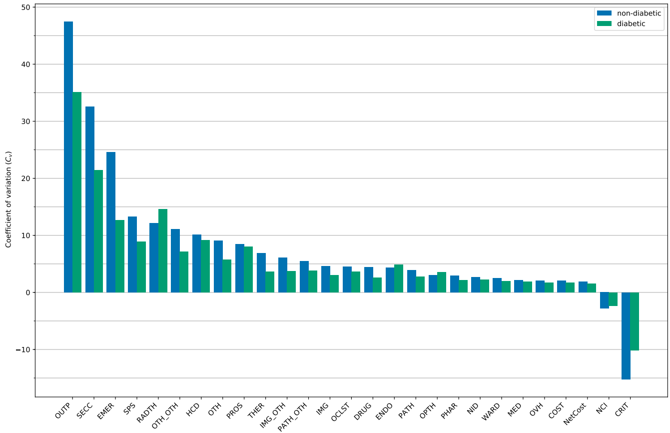
Number of procedures



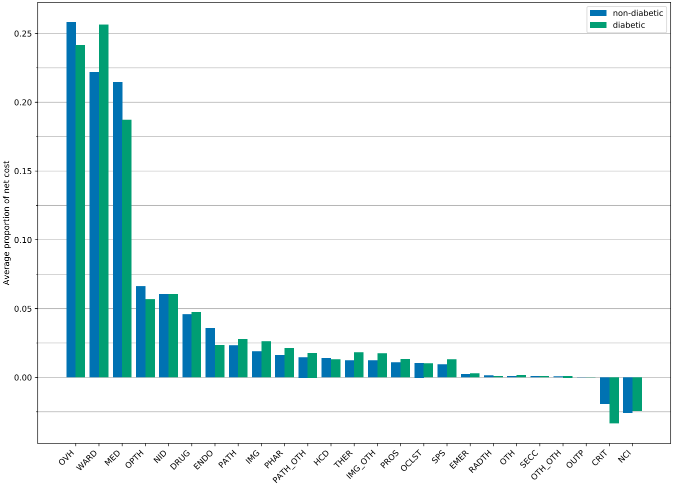
Demographic analysis



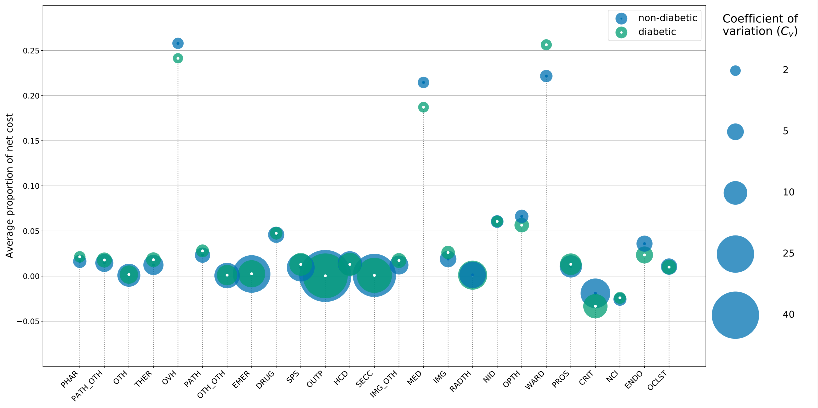
Cost variation



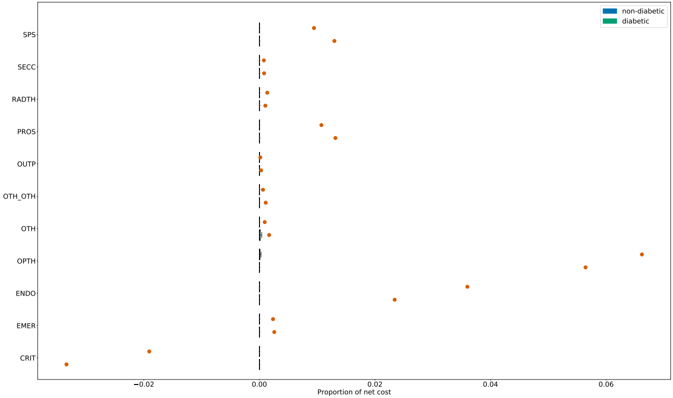
Cost component contribution



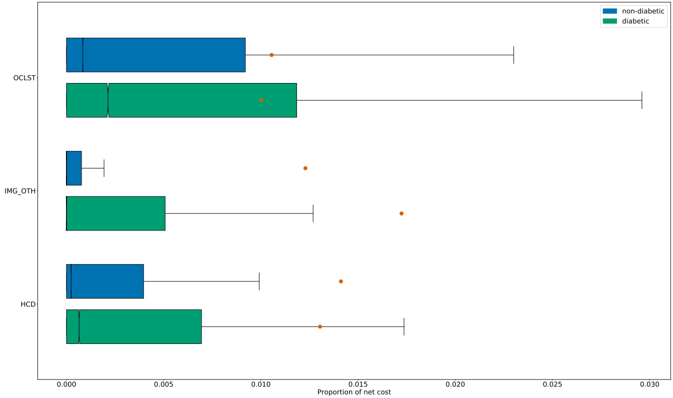
Relative importance



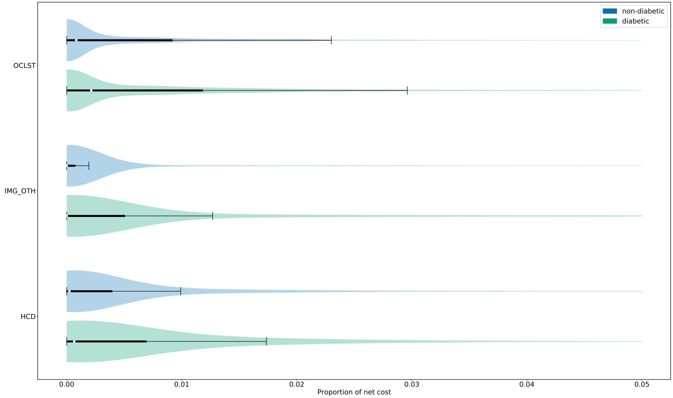
Cost component distributions (negligible)



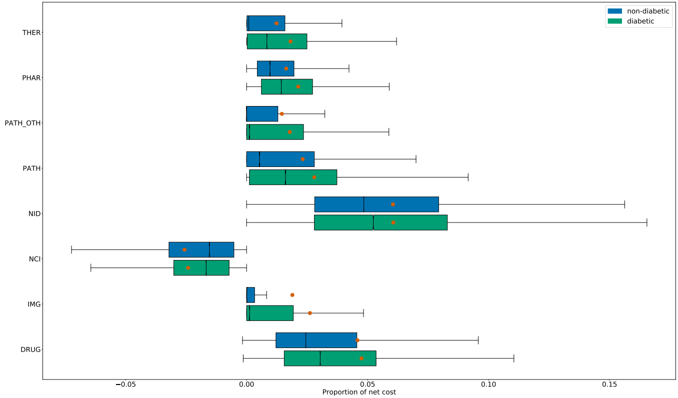
Cost component distributions (small)



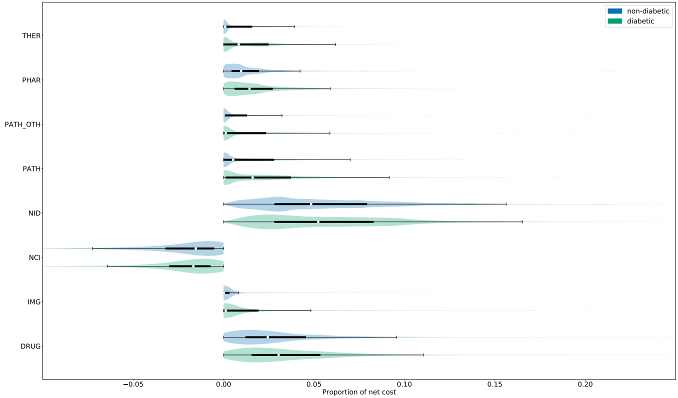
Cost component distributions (small)



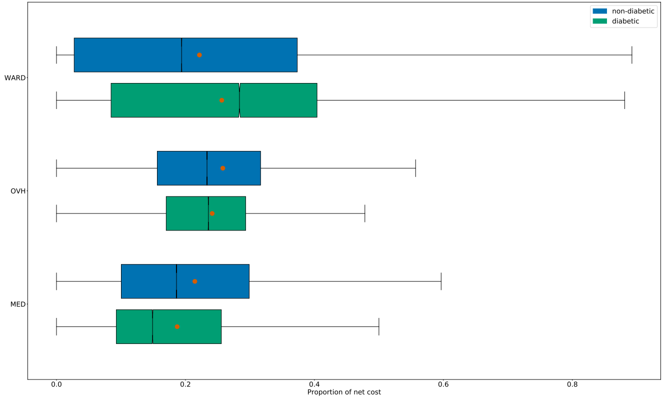
Cost component distributions (medium)



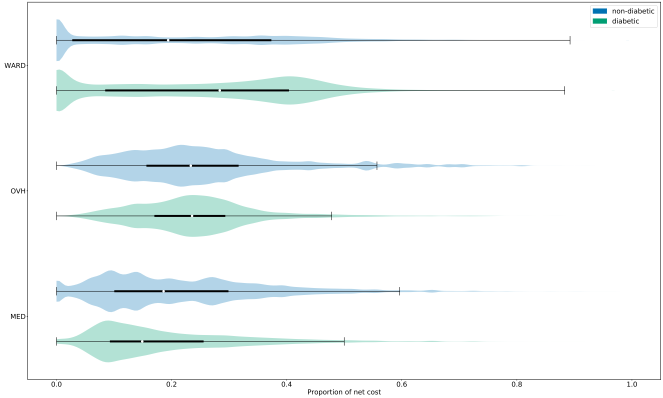
Cost component distributions (medium)



Cost component distributions (large)

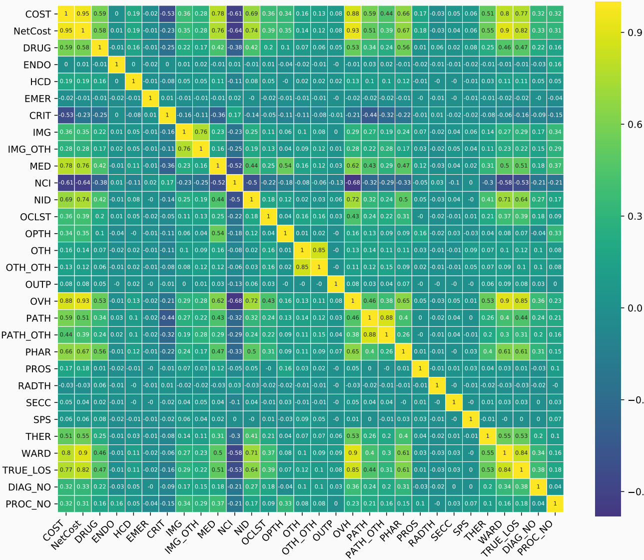


Cost component distributions (large)



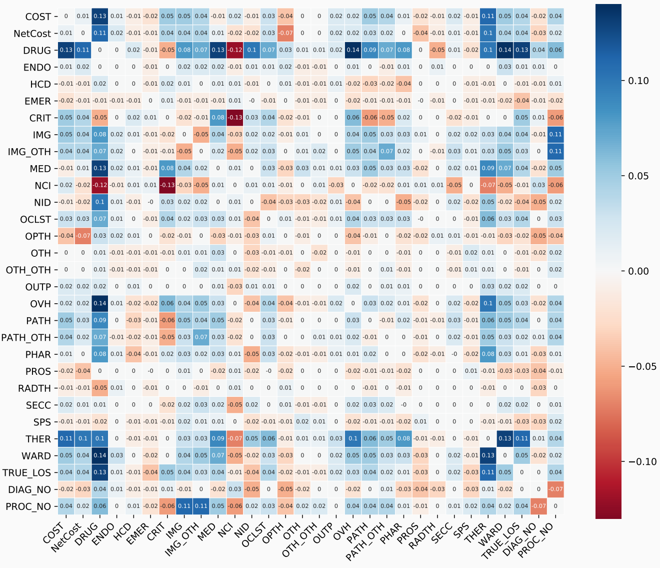
Correlation

Correlation coefficients for (diabetic) spell-level cost components and other clinical variables

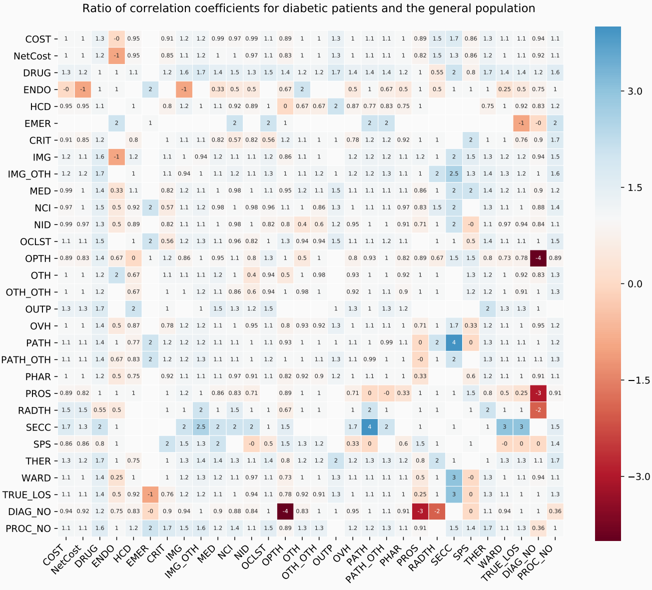


Correlation (differences)

Difference in correlation coefficients for diabetic patients and the general population



Correlation (ratio)



Measuring resource consumption

Too wordy. We will focus our definition of system 'cost' on three measures:

Diabetic patient resource consumption analysis

Too wordy. We will focus our definition of system ‘cost’ on three measures:

- Proportion of total daily admissions
- Average length of stay given admission date
- Proportion of net costs spent given admission date

Diabetic patient resource consumption analysis

Too wordy. We will focus our definition of system 'cost' on three measures:

- Proportion of total daily admissions
- Average length of stay given admission date
- Proportion of net costs spent given admission date

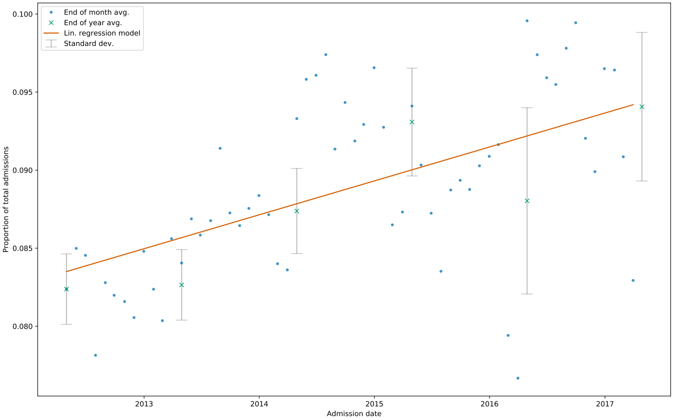
These are indicators of resources used and resources necessary.

This grouping by admission date will lead to a degree of misrepresentation in our plots.

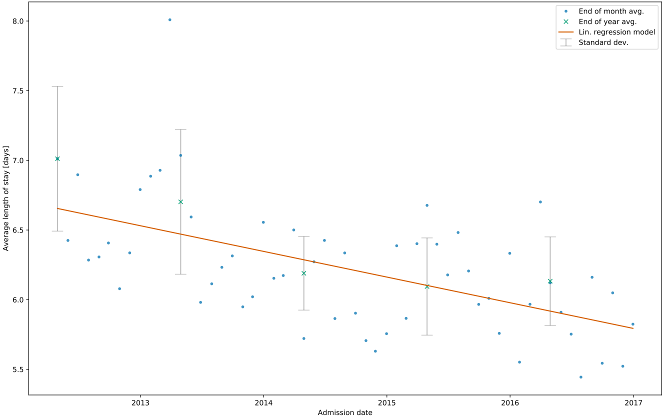
Allows us to investigate patterns developing over time.

Resource consumption

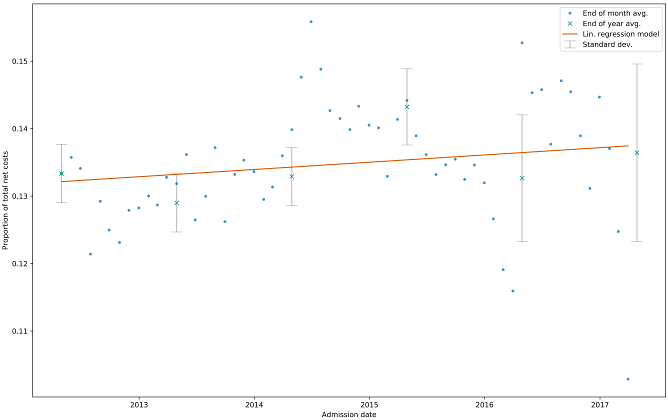
These plots need CIs and accompanying boxplots.



Resource consumption



Resource consumption



Conclusions

Conclusions

- Relative resource consumption by diabetic patients is consistent
- Cost components are less variant than - and are comparable in their distribution to - non-diabetic patients

What next?

Moving forward

Moving forward

- Resource consumption metric

Moving forward

- Resource consumption metric
- Perform clustering analysis to find inherent slices in the data

Moving forward

- Resource consumption metric
- Perform clustering analysis to find inherent slices in the data
- Incorporate external data
  - Decode GP practice codes for GeoPandas
  - Socio-economic analysis based on deprivation and geography
  - Temperature-based analysis

Moving forward

- Resource consumption metric
- Perform clustering analysis to find inherent slices in the data
- Incorporate external data
  - Decode GP practice codes for GeoPandas
  - Socio-economic analysis based on deprivation and geography
  - Temperature-based analysis
- Severity and comorbidity analysis
  - Average severity of secondary conditions given some primary condition
  - Using the comorbidity index as a class label in some predictive analysis