DRAFT

Omer Nivron

August 2018

1 Report outline

I first define what are specializations in the current context. According to the definition I layout the methodological analysis designed to uncover specializations which includes scoping and cleanup of data followed by clustering using the EM algorithm. Subsequently, I explain the automatic way chosen to characterize the clusters derived from the EM algorithm and alternative approaches. Finally I explain how the latter transforms into features and present the clusters and their associated features.

2 What is specialization?

Definition - specialization
A particular body region/organ (or a set) in which someone has performed 100+ diagnostics with specified technique (MRI, CT, ...) and specifications (1 view, 2 view, contrast, ...)

- The choice of 100 diagnostics is arbitrary and with more domain knowledge any other bar could be set (or different bars for different procedures)
- According to the definition above we could keep all procedures in place and mark unique specializations per radiologist. However, for the purposes of exploration I would like to find grouping of radiologists and the characterization of each group. This could be also interesting to check if for each group a different predictive model applies.

3 Methodological analysis

3.1 Scoping

1. Filter out from raw table "hcpcs code" not starting with "7" - none radiology codes.

2. Use only the "npi", "$hcpcs_code$", "$line_srvc_cnt$" and "$hcpcs_description$" columns from the raw table.

3. (After transformation - explained below) remove all columns with max argument smaller than 100. Since our definition cuts a specialization at 100, if no radiologist made the bar then we can safely remove it and reduce complexity.

3.2 Cleanup

The cleanup stage is a lot about exploration of missing values (NA), unrealistic values, wrong formats, consistency, duplicates and consumable values for specific usage.

Through analysis (not attached) we see that for the chosen scope and purpose duplicates, consumable values and formatting are of greatest importance. However for a full future solution all checkups should be included as updates/extensions of the raw data might introduce these issues.

3.2.1 Consumable data

The "$line_srvc_cnt$" has a pipe character in some cells which we want to drop and sum the values on both pipe sides. Further we format the column to numeric values (originally strings).

A similar pipe issue appears in the "$hcpcs_description$" column, where the description after the pipe is a copy of the description before the pipe. In this case, we drop the pipe character and all subsequent text. The format of a string stays the same.

3.2.2 Duplicates

After the "$hcpcs_description$" is cleaned we check if we find any codes that their description is the same. See table here-under for all duplicates.

	hcpcs_code	hcpcs_description
41	71020	X-ray of chest, 2 views, front and side
42	71021	X-ray of chest, 2 views, front and side
43	71022	X-ray of chest, 2 views, front and side
57	72020	X-ray of spine, 1 view
61	72070	X-ray of middle spine, 3 views
62	72072	X-ray of middle spine, 3 views
65	72081	X-ray of spine, 1 view
164	74010	Imaging of abdomen
165	74020	Imaging of abdomen
191	74270	X-ray of large bowel with contrast
192	74280	X-ray of large bowel with contrast
202	74450	Radiological supervision and interpretation X-...
203	74455	Radiological supervision and interpretation X-...
245	76120	Imaging of organ
246	76125	Imaging of organ
256	76641	Ultrasound of one breast
257	76642	Ultrasound of one breast
258	76700	Ultrasound of abdomen
259	76705	Ultrasound of abdomen
267	76856	Ultrasound of pelvis
268	76857	Ultrasound of pelvis
298	77080	Bone density measurement using dedicated X-ray...
299	77081	Bone density measurement using dedicated X-ray...
301	77085	Bone density measurement using dedicated X-ray...
321	78012	Nuclear medicine imaging for thyroid uptake me...
323	78014	Nuclear medicine imaging for thyroid uptake me...
360	78707	Nuclear medicine study of kidney with assessme...
361	78708	Nuclear medicine study of kidney with assessme...
362	78709	Nuclear medicine study of kidney with assessme...
368	78805	Nuclear medicine study of radioactive material...
370	78807	Nuclear medicine study of radioactive material...

Table of duplicate combination of procedure code and description

(After transformation - explained below) we want to drop the duplicated procedures. I do this by picking the first hcpcs code from the duplicated set and sum the other procedures into this procedure. The assumption that these procedures should be summed up, should be further tested in future solution. Firstly, it should be verified further with experts whether these procedures are indeed the same. Secondly, it has to be questioned if the duplicated procedures are over-counting or not and hence should be summed or not.

3.2.3 Transformation

The raw table has to be transformed in order for every physician to have a vector representing its procedure's counts. This is a must for any clustering or matrix factorization technique. Hence, I transform the data to a table that is same as the one provided in the Readme background (see figure 1). i.e.

- Table: Number of times a radiologist provided a specific procedure.
- Rows: Physician ID.
- Columns: Procedure ID.

	71020	72110	73060	76830
1265413959	100	200	10	20
1255366340	80	250	10	10
1326089913	40	40	140	150
1871517128	30	30	150	150

Figure 1: ()

3.3 Clustering

A good starting point for grouping the radiologists could be using a clustering algorithm like: k-means, EM, spectral or matrix factorization techniques like PCA and NMF.

Unlike matrix factorization techniques which extract a hidden dimensional space representing specialties, the EM algorithm uses the original dimensions of the data so mapping specializations to what they actually describe might be easier. Further, the EM algorithm allows probabilistic (flexible) labeling of radiologists rather then hard assignments like in k-means and provides a richer representation of cluster variance. From the above reasons, I concentrate my solution with the EM. However, with longer time scope, it would be very interesting to examine the other techniques and see which provides the biggest end value.

3.3.1 Understanding EM - technical brief

Expectation maximization (EM) algorithm is an unsupervised technique for clustering - a probabilistic view to the classic k-means algorithm. If the data

labels were known, we could have used maximum likelihood estimation to reveal the parameters (mean, covariance, ...) that represent the distribution of the cluster.

Since the labels are unknown, we represent them by a multinomial (say Z) hidden variable and we can insert Z to the likelihood equation by summing over all of its values. We then can multiply and divide the log-likelihood equation by a distribution in Z. This results in equation of the form:

$$\sum_i log \sum_{z_i} Q(z_i) P(x, z; \theta) \div Q(z_i) \tag{1}$$

Through Jensen's inequality we can say the last equation is bigger than:

$$\sum_i \sum_{z_i} log Q(z_i) P(x, z; \theta) \div Q(z_i) \tag{2}$$

And the condition for equality between equation 1 and 2 results in the expectation step:
setting $Q(z_i) \propto P(x, z; \theta)$, this is followed by the maximization step, were we take the derivatives of all parameters and equate to zero. We repeat this process until convergence is achieved - The EM is monotonically increasing. For more detailed information visit http://cs229.stanford.edu/notes/cs229-notes8.pdf

3.3.2 Training EM

For training the EM algortihm we need to specify few parameters. First we need to decide which distributions will be fitted to clusters, initial states for these distribution parameters and number of clusters we want to optimize for.

For simplicity and under the time constraints we choose the common parameters to be the normal distributions with means intialized to the k-means solution and co-variances randomly initialized. But with longer time different initial parameters and distributions (t-distribution) should be tested.

The number of appropriate clusters is chosen by the model with the lowest AIC score - which is an equation trying to balance between the number of clusters to the log-likelihood score. I run the fitting for clusters ranging from 1-40 and choose lowest AIC model - that's also why this step takes around 20 minutes to run. The number 40 is arbitrary and can be adjusted in the train EM.py function. If the raw data is changed this part has to be fitted again and potentially will find different number of clusters as the appropriate separation.

It should be noted that in general, the EM should also include a testing step where new data is clustered and results are measured by a metric such as the ROC curve to choose the best initial parameters and model. I, however drop the testing phase under the current exploration.

Another trade-off of time that I do is choosing the labels with highest probability to represent radiologists and disregarding the probabilistic nature of the EM output. In a more comprehensive solution this attribute should be taken into account or otherwise a different algorithm can be chosen.

4 Feature extraction

The output of the EM algorithm in this case gives us 26 groups each is represented by a 361 vector (same as number of raw features minus duplicates) of mean values and a 361X361 covariance matrix.

4.1 From cluster to specialties

There is not one way solution for extracting the most important features representing each cluster. Few methods include hand engineering - looking at cluster means and variances across dimensions (possibly applying some tests like ANOVA to distinguish among clusters, plotting boxplots histograms etc.) and deciding based on the above and additional insider insights what a specific cluster represents. Another interesting approach is to train a classifier like random forest on radiologists (since after the EM fitting we have labels) and extract the most important features as the ones that were most important for the classification algorithm to achieve high accuracy. I have chosen a simple method that takes a cluster representative - the radiologist that is closest to the cluster mean as fitted by the EM algorithm. This has the downside that if the variance in any dimension is large then the representative might be a bad approximation for the cluster specialties. Remedies could be to use any of the aforementioned methods.

The representative is chosen by taking the euclidean distance between cluster center and any radiologist in the cluster, choosing the one with smallest distance. Then we look at the values of each procedure for the representative: if the value is bigger than 100 (according to specialty definition) we tag the procedure as a specialty and add it to a list representing the cluster.

4.2 From specialties to features

According to the definition I have provided above it makes sense to keep all unique procedures derived in the sub-section above (From cluster to specialties). If we mine more the description column we might be able to describe in an easier language-wise manner the specialty (e.g. MRI neck), but this both asks for bigger scope and also might lose some predictive power (for example, dropping the specialty of doing a scan with contrast). Moreover, the procedure code can be readily matched to its description. So, if cluster number 7 has specialty in procedure "70100", this procedure will become a feature for all radiologist. Whereas, if cluster number 5 does not have "70100" as a specialty, all radiologists in cluster 5 will get 0 values for this specialty.

To enrich the features, once a value of a procedure is bigger than 100 (for the cluster representative) we put the value in the complete distribution of all radiologists from all clusters and give a score equivalent to the quantile position of the representative value. For example if cluster 7 representative has a count of 350 for procedure "70100" and out of all radiologists 350 is the 81.6 percentile, all cluster 7 members will get 81.6 under this feature. This process helps to identify

levels of expertise among specialists and also gives more predictive flexibility due to the continuous nature of the scoring.

My solution consists of 59 features with scoring with two extra columns - one representing the "npi" - physician identifier and the other the label of the radiologist's respective cluster.

5 Cluster specialty results

The clusters and their respective specialties are presented by label order ascending here-under. This gives us some insight into how our EM algorithm chose to cluster. Qualitatively, many of the clusters seem to be working as we hope them to - clustering specialists in brain or abdomen together. Or clustering together radiologists with no specialty according to the definition above.

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
51	CT scan chest with contrast
53	CT scan of blood vessels in chest with contrast
70	X-ray of lower and sacral spine, minimum of 4 ...
73	CT scan of upper spine
111	X-ray of shoulder, minimum of 2 views
133	X-ray of hip with pelvis, 2-3 views
143	X-ray of knee, 4 or more views
150	X-ray of foot, minimum of 3 views
163	X-ray of abdomen, single view
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
258	Ultrasound of abdomen
261	Ultrasound behind abdominal cavity, limited
286	Computer analysis of screening mammogram to as...

	hcpcs_description
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
258	Ultrasound of abdomen
286	Computer analysis of screening mammogram to as...

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
69	X-ray of lower and sacral spine, 2 or 3 views
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
256	Ultrasound of one breast
258	Ultrasound of abdomen
285	Computer analysis of diagnostic mammogram
286	Computer analysis of screening mammogram to as...
292	Screening digital tomography of both breasts
298	Bone density measurement using dedicated X-ray...

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
50	CT scan chest
111	X-ray of shoulder, minimum of 2 views
133	X-ray of hip with pelvis, 2-3 views
142	X-ray of knee, 3 views
148	X-ray of ankle, minimum of 3 views
150	X-ray of foot, minimum of 3 views
159	MRI scan of leg joint
163	X-ray of abdomen, single view
164	Imaging of abdomen
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
256	Ultrasound of one breast
258	Ultrasound of abdomen
286	Computer analysis of screening mammogram to as...
298	Bone density measurement using dedicated X-ray...

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
50	CT scan chest
51	CT scan chest with contrast
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
256	Ultrasound of one breast
285	Computer analysis of diagnostic mammogram
286	Computer analysis of screening mammogram to as...
292	Screening digital tomography of both breasts
298	Bone density measurement using dedicated X-ray...

	hcpcs_description
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
69	X-ray of lower and sacral spine, 2 or 3 views
111	X-ray of shoulder, minimum of 2 views
133	X-ray of hip with pelvis, 2-3 views
141	X-ray of knee, 1 or 2 views
142	X-ray of knee, 3 views
150	X-ray of foot, minimum of 3 views
163	X-ray of abdomen, single view
258	Ultrasound of abdomen
298	Bone density measurement using dedicated X-ray...

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
50	CT scan chest
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
258	Ultrasound of abdomen

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side

	hcpcs_description
41	X-ray of chest, 2 views, front and side
85	MRI scan of lower spinal canal
258	Ultrasound of abdomen
298	Bone density measurement using dedicated X-ray...

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
133	X-ray of hip with pelvis, 2-3 views
150	X-ray of foot, minimum of 3 views
163	X-ray of abdomen, single view
173	CT scan of abdomen and pelvis with contrast

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
163	X-ray of abdomen, single view
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
258	Ultrasound of abdomen
286	Computer analysis of screening mammogram to as...
292	Screening digital tomography of both breasts

	hcpcs_description
41	X-ray of chest, 2 views, front and side
58	X-ray of spine of neck, 2 or 3 views
59	X-ray of upper spine, 4 or 5 views
66	X-ray of spine, 2 or 3 views
69	X-ray of lower and sacral spine, 2 or 3 views
70	X-ray of lower and sacral spine, minimum of 4 ...
90	X-ray of pelvis, 1 or 2 views
111	X-ray of shoulder, minimum of 2 views
115	X-ray of elbow, 2 views
119	X-ray of wrist, minimum of 3 views
121	X-ray of hand, 2 views
122	X-ray of hand, minimum of 3 views
123	X-ray of fingers, minimum of 2 views
133	X-ray of hip with pelvis, 2-3 views
136	X-ray of both hips with pelvis, 3-4 views
140	X-ray of femur, minimum 2 views
141	X-ray of knee, 1 or 2 views
142	X-ray of knee, 3 views
143	X-ray of knee, 4 or more views
147	X-ray of ankle, 2 views
148	X-ray of ankle, minimum of 3 views
149	X-ray of foot, 2 views
150	X-ray of foot, minimum of 3 views
151	X-ray of heel, minimum of 2 views
153	CT scan leg
159	MRI scan of leg joint
248	3D radiographic procedure with computerized im...
296	Imaging of 2 or more joints, single view
298	Bone density measurement using dedicated X-ray...

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
50	CT scan chest
69	X-ray of lower and sacral spine, 2 or 3 views
90	X-ray of pelvis, 1 or 2 views
111	X-ray of shoulder, minimum of 2 views
133	X-ray of hip with pelvis, 2-3 views
141	X-ray of knee, 1 or 2 views
149	X-ray of foot, 2 views
150	X-ray of foot, minimum of 3 views
163	X-ray of abdomen, single view
166	Imaging of abdomen and chest
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
258	Ultrasound of abdomen
260	Ultrasound behind abdominal cavity
261	Ultrasound behind abdominal cavity, limited
267	Ultrasound of pelvis

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
50	CT scan chest
73	CT scan of upper spine
111	X-ray of shoulder, minimum of 2 views
133	X-ray of hip with pelvis, 2-3 views
141	X-ray of knee, 1 or 2 views
150	X-ray of foot, minimum of 3 views
163	X-ray of abdomen, single view
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
258	Ultrasound of abdomen
260	Ultrasound behind abdominal cavity

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
50	CT scan chest
61	X-ray of middle spine, 3 views
69	X-ray of lower and sacral spine, 2 or 3 views
73	CT scan of upper spine
78	CT scan of lower spine
111	X-ray of shoulder, minimum of 2 views
122	X-ray of hand, minimum of 3 views
133	X-ray of hip with pelvis, 2-3 views
141	X-ray of knee, 1 or 2 views
150	X-ray of foot, minimum of 3 views
163	X-ray of abdomen, single view
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
174	CT scan of abdomen and pelvis before and after...
258	Ultrasound of abdomen
260	Ultrasound behind abdominal cavity
298	Bone density measurement using dedicated X-ray...

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
51	CT scan chest with contrast
111	X-ray of shoulder, minimum of 2 views
133	X-ray of hip with pelvis, 2-3 views
143	X-ray of knee, 4 or more views
163	X-ray of abdomen, single view
164	Imaging of abdomen
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
256	Ultrasound of one breast
258	Ultrasound of abdomen
286	Computer analysis of screening mammogram to as...
298	Bone density measurement using dedicated X-ray...

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
258	Ultrasound of abdomen
286	Computer analysis of screening mammogram to as...

	hcpcs_description
256	Ultrasound of one breast
286	Computer analysis of screening mammogram to as...
292	Screening digital tomography of both breasts
298	Bone density measurement using dedicated X-ray...

	hcpcs_description
37	MRI scan brain
39	MRI scan of brain before and after contrast
81	MRI scan of upper spinal canal
83	MRI scan of middle spinal canal
85	MRI scan of lower spinal canal
129	MRI scan of arm joint
159	MRI scan of leg joint

	hcpcs_description
15	CT scan head or brain
37	MRI scan brain
39	MRI scan of brain before and after contrast
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
50	CT scan chest
51	CT scan chest with contrast
53	CT scan of blood vessels in chest with contrast
69	X-ray of lower and sacral spine, 2 or 3 views
73	CT scan of upper spine
85	MRI scan of lower spinal canal
111	X-ray of shoulder, minimum of 2 views
119	X-ray of wrist, minimum of 3 views
122	X-ray of hand, minimum of 3 views
133	X-ray of hip with pelvis, 2-3 views
143	X-ray of knee, 4 or more views
148	X-ray of ankle, minimum of 3 views
150	X-ray of foot, minimum of 3 views
159	MRI scan of leg joint
163	X-ray of abdomen, single view
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
181	Imaging for evaluation of swallowing function
258	Ultrasound of abdomen
261	Ultrasound behind abdominal cavity, limited

Empty DataFrame Columns: Index(['hcpcs_description'], dtype='object') Index: Int64Index([], dtype='in

	hcpcs_description
41	X-ray of chest, 2 views, front and side
256	Ultrasound of one breast
286	Computer analysis of screening mammogram to as...
292	Screening digital tomography of both breasts

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
258	Ultrasound of abdomen
286	Computer analysis of screening mammogram to as...

	hcpcs_description
41	X-ray of chest, 2 views, front and side
85	MRI scan of lower spinal canal
129	MRI scan of arm joint
159	MRI scan of leg joint

	hcpcs_description
15	CT scan head or brain
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
50	CT scan chest
164	Imaging of abdomen
172	CT scan of abdomen and pelvis
173	CT scan of abdomen and pelvis with contrast
256	Ultrasound of one breast
258	Ultrasound of abdomen
286	Computer analysis of screening mammogram to as...
289	Screening mammography of both breasts
298	Bone density measurement using dedicated X-ray...

	hcpcs_description
15	CT scan head or brain
37	MRI scan brain
39	MRI scan of brain before and after contrast
40	X-ray of chest, 1 view, front
41	X-ray of chest, 2 views, front and side
53	CT scan of blood vessels in chest with contrast
73	CT scan of upper spine
81	MRI scan of upper spinal canal
85	MRI scan of lower spinal canal
173	CT scan of abdomen and pelvis with contrast

6 Additional code notes

I have tried to put the code in a high level framework, but it has to be noted that due to time constraints, essential pieces to make it into "near production code" are missing - such as logging and testing cases for all operations.