Paper02: Problematic Risk Adjustment in National Healthcare Safety Network Measures.

Summary

The papers and the figures in this paper are about fairly different things. The figures of this paper are looking at ICU versus no ICU for different SOI by hospital size (number of HF)

Replication notes

Paper 2 Journal:

Table 1 is a list of 5 NQF measures; however, I don't see how these are actually used in the analysis of the paper.

Table 2 shows hospital level coefficients. These are a definition, not calculated.

Table 3 is a list of diagnosis and procedure codes that they're using to define a cohort of encounters.

Table 4 and 5 are tables that can be replicated

What information do I need for this study:

ICU Admissions by Rec-Center-Code from revenue file where the list of codes is what they suggest.

They're grouping by hospital, so ORG\_NPI\_NUM or PROVNUM.

They're using heart failure as their cohort, specifically APR DRG

The explanation / exploration of APRDRG codes is discussed in its own section. This was more challenging than I realized when writing this initially.

Inclusion / Exclusion Criteria:

Claims from October 1, 2015 to September 30, 2016 (This is odd criteria because it starts at the ICD-9/10 transition rather than at the start of a year) This is the Claim\_Through column.

Excluded: discharge disposition 2

Excluded hospitals with fewer than 100 ICU cases (This is slightly complex because it’s convoluted in the text if they mean 100 ICU cases or 100 HF cases, or 100 HF ICU cases)

Identifying the cohort of heart failures:

They are looking at 1,374 hospitals, 430,397 cases, and 2706,16 ICU visits. I assume, they're using APRDRG to define heart failure; however, it could be Elixhauser or CCI so I will be checking (They were not)

They're using CMS 2017 INP for their findings as well.

Continuing to Figure 4

Figure 4 is

Dim 1: SOI level (1-4 From DRG)

Dim 2: ICU Charge (Boolean from rec-center-code)

And then for each of these they want:

Cases unique Claim\_No

ALOS They don't call this out specifically and there is no citation, so I am going to assume they mean the average time between admission and discharge.

%Died Discharge disposition 20

%High Insensitive This is looking to see if they have any procedures / diagnosis from their list in the previous figure.

Figure 5 is Figure 4 but grouped by hospital quartile by percentage of reported ICU use for CHF admissions.

What that means is, of the claims of CHF, what % of them were sent to the ICU.

The inconsistency so far is that it is not clear what data is being used.

The paper references Claims from FY 2016 in one place and 2017 claims dataset. So I will need to preform analysis on all of the combination of dates to try and generate a similar cohort.

I am expecting some variation because they used APRDRG 36.0 and I'm using 38.0 (initially), but that is a problem for the me of the future to solve probably. I don't know if it's going to be a huge difference and it's probably something that I'll need to figure out eventually; however, this is probably close enough for now.

Step 1, we need to identify the cohort as hospitals with more than 100 ICU visits. The paper defines a claim with an ICU visit as having a revenue center code of...

200 General classification for intensive care unit (ICU)

201 Surgical ICU

202 Medical ICU

203 Pediatric ICU

204 Psychiatric ICU

206 Intermediate ICU

207 Burn care

208 Trauma care

209 Other intensive care

210 General classification cardiac care unit (CCU)

211 Myocardial infarction care

212 Pulmonary care

213 Heart transplant

214 Intermediate CCU

219 Other coronary care

So I am going to use this code to identify our hospital cohort (Dev12). This query resulted in 3054 hospitals versus the paper’s 1374 hospitals.

It is possible that I did not apply the exclusion criteria early enough. In Dev13 I will add the patient discharge status.

In Dev13 I got 2979 hospitals. So maybe I am using the wrong claim date range. From their paper they state that “The research team obtained financial year 2016 (FY2016) Medicare fee-for-service claims spanning October 1, 2015, through September 30, 2016.” This implies the 2016 dataset.

“The team retained claims for hospitals paid under the IPPS and grouped claims under Version 36 of the All-Patient Refined Diagnosis-Related Group (APR-DRG) classification system.“

How is this defined. They do not provide any way to do this or any citation. I wonder if it's just an eccentricity of their dataset. I am using the SAF and they are using IPPS; however, a common error is to call the SAF INP dataset the IPPS. This is a mistake I made myself in my dissertation proposal. It could be that IPPS contains some information about hospitals using APR-DRG whereas I just calculated APR-DRG myself (One of the authors of their paper has an affiliation with 3m, so it’s likely they also calculated it themselves using the 3m software). Even though my cohort is larger I should get the same result with the same year of data, or atleast something quite similar. This is the moment that I am giving up on an "exact" replication for paper 2, I am just aiming for a statistical replication now- can I do the same process on the same general dataset and get the same results (Since I no longer believe I am using the same dataset)

So given that, I have identified my hospital cohort, now I need to get my claim cohort. I will start with the query in Dev14 as my initial cohort. This query resulted in 360,188 cases. Again, the target 1,374 hospitals, 430,397 cases, and 270,616 ICU visits. Well, that's not too far off; however, I wonder when they said 100 ICU visits if they meant 100 ICU Heart Failure visits. I need to edit the query above because I didn't include hospital ID.

I’m looking at 2976 hospitals total. If I just look at hospitals with more than 100 claims there are 1368. I suspect that this is what they did as I am casual 0.6% off, so that means I need to go adjust my query in order to perform the ICU filter step later. I don't think it will change anything but computers, SQL, all this stuff is magic. My logical brain tells me nothing will change; however, I just want to validate.

Actually, when I went and looked what I tested I tested this query

SELECT count([CLAIM\_NO])

,[ORG\_NPI\_NUM]

FROM [CAT1].[cms].[zzpaper2\_temp]

group by org\_npi\_num

having count([CLAIM\_NO]) >100

So I wasn't even looking at ICU claims, I was looking at Heart Failure claims... Did I misread the paper?

" The analysis was further restricted to hospitals having at least 100 ICU cases within the claims data."

Nope, I did not. I need to explore what's going on here further.

I altered the query by no longer inner joining on HospList. This increased the number of hospitals to 1369- no difference. I wonder how different the cases are in each case.

Paper

430397 cases

270616 ICU visits

1374 Hospitals

100 HF 2016

1369

285604

100 ICU --> 100 HF 2016

360188 cases -->285482

2976 --> 1368

So, this so with the 2016 data... I wonder if they really did the 2017 data.

In 2017 data I'm getting 1778 is orgs, either way I calculate it. So that's totally different.

I keep coming back to the dates. It's weird that they specify a dataset, and specify a date range. Maybe I just don't really understand that data included in the CMS dataset and that's my failing. The claims included in a dataset are claims that are submitted in 2016 for the 2016 dataset. They say the obtained "claims spanning October 1, 2015, through September 30, 2016. " and I understand why they picked October 1st 2015 because it's when ICD-10 was implemented. The issue is that all the claims in the 2015 claims dataset are ICD9, and all the claims in the 2016 claims dataset are ICD10, which, in my dataset, literally never happens. So, I wonder if my dataset is flawed. So, to test that I will go back to the very start. I suspect my approach was flawed. I was looking for ICD\_DGNS\_VRSN equal to 10, where as in reality I needed to find them equal to 0. This whole time, more or less, I've been using the wrong dataset. Well let’s start again, this time using the correct dataset.

Well all of that was a waste I suspect. As it turns out I was referencing the 2017 MBSF with the 2016 data, which makes it extremely odd that I got so close to their numbers.

I Updated the 2016 code to now reference the correct table, and I executed the 2015 code to get that dataset using claimsj and DNMNTR files.

It also turns out I was able to find some files from AHRQ in the archive that contain v36 of APR-DRG, so I can implement that now too, or implement both v36 and v38 to see if there is a difference. This should get me much closer to the final version.

Let’s start by looking at ICU visits in the time period... [dev01.sql]

We're looking at 3025 hospitals, quite a bit higher than the paper. So I still don't think they looked at ICU visits but heart failure (Because this wouldn't change v36 vs v38)

All of this said, I haven't filtered ICD 9v10 yet and v36 as far as I can tell, requires icd-10 so if there is a filtering step it hasn't happened yet.

Pandas told me there was an issue with my 2015 APRDRG (duplicate data) so I will investigate what's going on with that.

11265668 rows, oh that means I certainly made a mistake... that's as many as there are rows in the dataset so I didn't filter the dataset based on ICD9/10 code... which is why so many GRC11's (invalid)

Yep, I didn't add a filter in my where clause. That's something that isn't the end of the world. So, how many APRDRG194s did I get... None. Actually Zero. That's unexpected... equally unexpected was how I confidently believed 149 was 194...

Well okay 129878 in 2015... and the same with distinct, so no duplicates with this DRG, so that's great. Now to bring everything togeather. my first stab at that is dev02

So again, the target is...

Paper 430397 cases 270616 ICU visits 1374 Hospitals

Initially I have 523097 rows...

when I filter this to those with more than 100 CHF I get 1708 hospitals

When I look at hospitals with more than 100 ICU + CHF I get 872 hospitals

When I look at hospitals with more than 100 ICU visits I get 492207 visits initially. Again 1702 hospitals.

Before I go and do the v36 vs v38 stuff let me break down exactly what their inclusion criteria were.

October 1, 2015, through September 30, 2016 (Check)

claims for hospitals paid under the IPPS hmmmm Not check.

I wonder if they're filtering by pps\_ind if pps\_ind

grouped claims under Version 36 of the All-Patient Refined Diagnosis-Related Group (APR-DRG) classification system (Check)

Cases identified as transfers (discharge disposition = 2) were excluded for subsequent analysis (Check)

and flagged cases in which patients died (discharge disposition = 20) were excluded for subsequent analysis. Ambiguous... I'll try filtering out deaths potentially

The analysis was further restricted to hospitals having at least 100 ICU cases within the claims data. (Check, though, again, it's ambiguous if they mean ICU cases or ICU+CHF cases)

2 things to change,

1, I am going to be using v36 rather than v38 from here on out.

2, I am going to experiment with looking at pps\_ind

I am also going to be looking at simplified code to figure out the cohort rather than trying to get all the variables I need for everything. dev04 is my first stab at that.

492,203 cases is where I'm starting:

there are 17837 not in PPS

there are 14782 deaths

there are 3057 orgs

there are 1702 orgs with more than 100 Heart failure visits and more than 100 ICU visits.

there are 872 orgs with more than 100 heart failure ICU visits.

of my claims, 252892 are ICU claims 239315 are not.

My count of ICU claims is lower than their count, and my count of non-icu claims is like... double.

They describe their restriction as... "To review variation in patterns of ICU designation across hospitals, the analysis was restricted to a single base DRG—APR-DRG 194 CHF—that has a known high level of ICU utilization. The analysis was further restricted to hospitals having at least 100 ICU cases within the claims data. "

To me this implies they first restricted by DRG, and then restricted by ICU cases in that dataset...

To try again:

Step 1 Exclusion Criteria

1 Claims from October 1 2015 to September 30 2016 (where CLM\_THRU\_DT>20151000 and CLM\_THRU\_DT<20161000)

2 Discharge Disposition = 2 discarded (PTNT\_DSCHRG\_STUS\_CD!=2)

3 DRG=194 ( drg='194.0')

This is dev05.

523090 is the number of cases.

Now I need to restrict it to hospitals with atleast 100 ICU visits. (dev06)

same number of cases, so that's good.

523090 cases, 878 hospitals, 254690 ICU cases.

and if I just look at 100+ HF it's 1708 hospitals.

What if they didn't do what they said they did. The closest I've been to their numbers is using the 2016 dataset. What does this look like if I use the 2016 dataset (but correctly)

I no longer believe what they did was "correct" or if it was correct then I don't understand it.

Let’s look at the 2016 FY date range, using the 2016 Dataset. Maybe they used the 2016 dataset incorrectly- they believed it was the fiscal year rather than it being the calendar year.

Looks promising initially 393216 claims 1395 hospitals filtering on heart failures. Is the ratio of ICU to non ICU close? No. The ratio isn't close at all. 191451.

dev07

492019 is where I'm starting.

I really need to be focused on is that hospital number. So, before I do everything else... I should be able to get the hospital number from just ICU data.

1374 hospitals... I've been using ORG\_NPI\_NUM to define hospital... but maybe that's not correct. Maybe they're filtering on CLM\_FAC\_TYPE\_CD=1.

dev08

How many ICU encounters do we have in 2015-2016.

803289

3994528 in 2015

984584 Oct-Dec

4037761 in 2016

3025857 Jan-Sep

4010441 in FY 2016

4155 orgs in FY 2016

Orgs >100 ICU FY 2016

3082 orgs in FY2016

Excluding Non-Hospital

3082 orgs in FY2016

Exclusing deaths and transfers

2949 orgs in FY2016

So either something very flawed has happened with data selection, or, they did the APR-DRG filter before this step.

If I look at the most likely flawed case it comes out to 1857 hospitals... The issue is that in their figures they say "a Source: Medicare Financial Year 2017 claims data." and "aSource: Medicare FY2017 claims data. FY2016; 1374 IPPS hospitals; 430 397 CHF admissions; 270 616 (62.9%) ICU admissions."

Maybe there is some definition of IPPS hospital that I can't figure out. I think you could define IPPS with the CLP\_PPS\_IND\_CD variable as hospitals that don't participate in PPS would have that blank.

I do not believe there is any possible way they're looking at the number of hospitals before looking at heart failure.

So, let’s look at our heart failure cohort for 2015-2017

drg 194

2015 = 259762

Oct-Dec also 259762, that makes sense.

2016 = 532981 (This implies that heart failure is not evenly distributed)

Jan-Sep 293216

552978 cases in FY16 with no filters.

4912 orgs in 2015

5592 in 2016

5590 in FY 2016

HF > 100 is 1708 FY2016

HF > 100 is 858 2015

HF > 100 is 1733 2016

HF > 100 is 1395 2016 jan-sep (294248 cases)

So what do I want to figure out...

HF 2017 569150 cases

HF >100 2017 1785 hospitals

HF FY 2017 564832 cases

HF >100 FY 2017 1799 hospitals

2017 jan-sept 1483 hospitals

2016 oct-dec 276 hospitals

ICU >100 2017 3051 hospitals

ICU >100 FY2017 3055 hospitals

2017 jan-sept 2862 hospitals

2016 oct-dec 2003 hospitals

HFICU>100 2017 930 hospitals

HFICU>100 FY2017 926 hospitals

100ICU and 100HF

2017 1779 hospitals 477183 cases

2017FY 1774 hospitals 472773

2017 jan-sept 1479 hospitals 330269 cases

2016 oct-dec 276 hospitals

2016 1715 hospitals

2016 jan-sept 1385 hospitals 292870 cases

2015 oct-dec

2016FY 1702 hospitals 426408 cases

So, here is more or less where I'm at,

I know I'm not looking at clm\_Fac\_type\_cd, PTNT\_DSCHRG\_STUS!=2 or CLM\_PPS\_IND\_CD.. I've come this far, I might as well.

FY2017 1703 Hospitals 444285 cases 229473 icu cases

FY2016 1638 Hospitals 399297 cases

What I believe is that they seem to have some different way of determining hospitals than I do and that is likely causing everything else to cascade OR they did something wrong. The initial way that they describe their dataset still bothers me because they reference the medicare dataset by FY, which is not how the medicare datasets are organized. They retain hospitals that paid under IPPS, ignored patients with discharge 2 and 20, and looked only at hospitals, which are the 4 extra variables I filtered on. And they only looked at hospitals with... atleast 100ICU cases within the claims data... which the defined earlier as only containing heart failure visits. Which to me implies the HFICU case (the most restrictive case) So let’s explore that most restrictive case again. Hospitals with 100 heart failure ICU claims per year.

HFICU>100 and CLM\_FAC\_TYPE\_CD=1 and PTNT\_DSCHRG\_STUS\_CD!=2 and PTNT\_DSCHRG\_STUS\_CD!=20 and CLM\_PPS\_IND\_CD !=''

2016 FY 829 hospitals 172910 icu visits

2016 840 hospitals 177963 icu visits

2017 FY 883 hospitals 193759 icu visits

2017 903 hospitals 196710 icu visits

2017+2016q4 1053 hospitals 260002 icu visits

So, what does this tell me... What I believe is the most likely thing if they have some other way of figuring out what hospital the charge was in. I'm able to get close to their ICU count, I'm able to get close to their claim count, I'm just not able to get close to their hospital count unless I do things that I believe would be incorrect. Yeah, that's really all I can figure out. Their methods are extremely clear, they call out almost everything... except for how they determine what a hospital is, and maybe my assumption of ORG\_NPI is incorrect... it just feels like the only "hospital identifier" that I get.

So my org NPIs are certainly linking to corporations. Not hospitals. What npis do I get...

I get the organization, attending, operating, other, and rendering...

what are each of these things

Organizational these are health systems it seems like

attending these are individuals

operating these are individuals

other these are individuals

rendering these are individuals

How do Provider Numbers work?

provider number... for some random sample I have 4208 orgnpis and 4240 provider numbers...

400104

Positions 1-2 are a GEO-SSA Code

Positions 3-4 are category indicators

Positions 4+ are serial numbers...

Furthermore, there are hospitals with a letter in position 3 that aren't in PPS

So 12 are where

34 are what

and 56 are which...

So, I don’t use 12or56 to filter on hospital

Just 34... so I need to understand 34 but before I get into that, let’s just turn off the brain for a moment and see how it looks when we use provnum rather than orgnpi

Honestly it doesn't change much

2017 HFICU case... 903-->898

2017 HF100 case... 1799-->1704

2017 ICU 100 / HF 100 -->1697 hospitals, 451317 cases... So, it's getting closer?

if I look at the fiscal years for icu100/hf100

FY 2017 1686 hosp 446210 cases

FY 2016 1491 hosp 341481 cases

So, my number of cases is closer, forsure. but the average size of my hospital is just too large.

and the HFICU 2017 cases reduced the number of hospitals, it didn't increase it.

Let’s dig into the numbers again...

Well, like, this certainly splits off stuff but it still doesn't break it down further.

OSU has atleast 3 hospitals, but they all seem to be the same provider number and npi...

So NPI is too large

To provider number smaller but is too large

provider\_npi is in appropriate...

Maybe the identifier is in the revenue center file.

I know rev\_cntr is -which- revenue center in a hospital, but like... that does me no help as that's just a category list.

I sent out a message to CCW to ask but I suspect that they're going to tell me there isn't any other way to break it down. (There isn’t)

So. Here is my plan then... I plan to just do my best and to try and get a statistical replication, now that I no longer believe that I can actually recreate their cohort after days of trying.

So, lets start from the top

Dev 09

Step 1: The research team obtained financial year 2016 (FY2016) Medicare fee-for-service claims spanning October 1, 2015, through September 30, 2016.

year claim

2015 2757860

2016 8473922

11231782

Step 2: The team retained claims for hospitals paid under the IPPS and grouped claims under Version 36 of the All-Patient Refined Diagnosis-Related Group (APR-DRG) classification system.

year claim

2015 121415

2016 368349

489764

Step 3: Cases identified as transfers (discharge disposition = 2) and flagged cases in which patients died (discharge disposition = 20) were excluded for subsequent analysis.

year claim

2015 117671

2016 357231

474902

Step 4: To identify patient admissions reported as receiving days in ICUs, the research team flagged a claim as ICU if any claim charges were made for a revenue center code listed in Supplementary Appendix 2.

year icu claim

2015 0 58790

2016 0 178569

2015 1 58881

2016 1 178662

2015 x 117671

2016 x 357231

x 1 237543

x 0 237359

That is oddly close to a 50-50 split. This is also the first piece of evidence that we have deviated with the paper, as even before we perform the 100 ICU filter, we have fewer ICU cases than they do.

Step 5: To review variation in patterns of ICU designation across hospitals, the analysis was restricted to a single base DRG—APR-DRG 194 CHF (This was done in step 2) however because everything until this point is transitive it should not matter.

Step 6: The analysis was further restricted to hospitals having at least 100 ICU cases within the claims data.

So, this is the first decision point. I believe that they are referring to 100 icu cases within the claims data in this state. Not the initial dataset but the current dataset.

this restriction is by hospital (provider number according to CCW) and ICU cases (sum ICU=1 >= 100 group by provider\_number)

4143 providers total of which 834 providers have 100 or more HF ICU visits per year with all previous criteria

year icu claim

2015 0 13951

2016 0 43413

2015 1 43186

2016 1 131790

icu 0 57364

icu 1 174976

2015 57137

2016 175203

total cases:232340

So the final break down is...

Paper 430397 cases 270616 ICU visits 1374 Hospitals

My Cohort 232340 cases 174976 ICU visits 834 Hospitals

their % ICU is 62.9, my % ICU was 75.3

On to figure 5. Figure 5 says 2017 claims data, but their cohort said 2016, so I'm going to just continue with my existing dataset as the numbers in their cohort are the same.

For this figure I need:

This figure also suggests that deaths were not excluded (as the text above did suggest) So I will be adding them back in now. This increases my number of hospitals by 1 (865 now)

year icu (No column name)

2015 0 15519

2015 1 45419

2016 0 48303

2016 1 138670

ended up being the final figure.

So for figure 5 I need

Case Count (claim\_no, year)

LoS (los)

% died (death)

% high intensitive (dxall, prall)

And I need SOI (SOI)

and I need ICU (ICU)

and I need hospital identifier for Figure 5 (PRVDR\_NUM)

Dev 11 has the final python code needed to build the dataset

now all that's left is building the tables in R.