APCD Sample Dataset

Introduction

  The Office of Health Care Statistics (OHCS) created a sample dataset from Utah's All Payer Claims Database (APCD) to give users with limited resources the ability to explore these valuable data. Our goal is to ensure that the sample is representative of our population, and the second goal is to make the dataset small enough to run on a computer rather than a server.

  A multitier stratification sample was designed to ensure representativeness. Three strata were selected: 3M Clinical Risk Groupings (CRGs), gender, and age[1]. 3M CRGs use longitudinal claims data to assign patients to a single, mutually exclusive, severity-adjusted risk group. The CRG stratum samples people based on health status and associated risk proportionally to the population. Gender is nested in each CRG strata and age strata are nested each gender strata. This sampling methodology produces an accurate representation of the population's risk, as well as gender and age within specific risk groups.

  Geography was not included in the sampling stratification because the limited dataset does not contain county data. The APCD limited datasets include Utah's small health area and more information about small areas can be found on IBS[2][3] However, stratifying by small area greatly increases the number of strata in the sample, from 8,297 to 160,624 due to the large number of small areas. Many of these strata would have an extremely small sample size, which increases the probability of a non-representative sample for that stratum. Sampling for small area would also make the dataset more cumbersome by increasing the number of sampled individuals from 262,229 to 374,627. The APCD Sample Validation section demonstrates that the sample has a location distribution similar to the overall population.

  Prior to sampling, it was estimated the sample dataset will be approximately 3.4 GB. The true file size may be larger but we anticipate a computer with 8 GB of RAM should be able analyze this dataset.

Steps for Creating the APCD Sample

1. Upload the entire population's eligibility records for a given year into SAS with all the strata indicators: Person_ID, CRG[4], gender, and age group.
2. Upload file containing sample size for each stratum to SAS.
   a. This file should contain the variables CRG, gender, age group, and group size.
      i. Group size was calculated by taking a 10% sample of each stratum and then rounding up.
3. Sort the population and sample size datasets by the strata used in sampling method
4. Use **PROC SurveySELECT** to generate sample

---

[1] Age was grouped into five categories (0-17, 18-39, 40-49, 50-64, and 65+) based on a recommendation from Medicaid for an unrelated project.
[2] http://health.utah.gov/hda/manual/PcLDSManual.pdf
[3] http://health.utah.gov/opha/IBIShelp/sarea/UtahSmallAreaInfo.pdf
[4] The CRG used is the most detailed CRG. This means it has the severity level attached to the base CRG

      a. Use the following options: ***Data =*** (population dataset), ***SEED=***(Large prime number), ***OUT =*** (sample dataset name), ***sampsize =*** (sample size dataset);
      b. Specify the strata using: ***strata*** (strata variables separated with spaces);
      c. Specify the person indicator using: ***ID*** (person ID);
5. Export the sample dataset.
6. Use the person ID to find the eligibility information, medical claims, and pharmacy claims for each person in the sample for the designated year.

APCD Sample Validation

      The sample's geographic representativeness was validated by comparing the geographic distribution of the sample with the population. CRG, gender, and age grouping are controlled for in the sample design so no validation is needed for these groups. The geographic representativeness was tested using patient's county of residence (or other geographic indicator if needed) and a chi-square test of independence. The chi-square test compares the distribution of the population and sample and checks to see if they are independent of each other; meaning given a certain distribution is it more likely to be in the population or sample or equally likely. The null hypothesis is that the two distributions are equal/independent with the alternative being that the distributions are not equal. The results in Table 1 show there is no evidence to suggest the geographic distribution of the sample is statistically different than the geographic distribution in the population. Or, in other words, the sample is geographically representative of the population.

Table 1. Chi-square Test of Geographic Representativeness

Number of Individuals in Population	Number of Individuals in Sample	Chi-Square	DF for Chi-Square	p-value for Chi-Square
2,577,260	262,229	28.11	29[*]	0.512

[*]*Degrees of freedom are n-1 where n is 29 Utah counties plus the unknown/other category.*

      The geographic representativeness within each type of strata was also calculated beginning with gender (see Table 2). These results suggest that the population and the sample have a similar geographic distribution within gender with males in the sample fitting the population distribution a little more closely than the females.

Table 2. Chi-square Test of Geographic Representativeness within Gender Strata

Gender[*]	Number of Individuals in Population	Number of Individuals in Sample	Chi-Square	DF for Chi-Square	p-value for Chi-Square
F	1,311,580	133,380	30.31	29	0.399
M	1,265,567	128,818	23.70	29	0.744

[*]*Due to small number, individuals with unknown gender did not have geographic representativeness tested although they are included in the sample.*

The geographic representativeness held for age group as well (Table 3). All age groups were well above the 0.05 p-value needed to reject the null hypothesis. The age group that was closest to this threshold was the 50-64 year old group.

Table 3. Chi-square Test of Geographic Representativeness within Age Group Strata

Age group	Number of Individuals in Population	Number of Individuals in Sample	Chi-Square	DF for Chi-Square	p-value for Chi-Square
0-17	850,725	85,879	24.84	29	0.686
18-39	840,889	84,989	25.89	29	0.631
40-49	276,753	28,587	27.58	29	0.541
50-64	373,032	38,262	35.71	29	0.182
65+	235,861	24,512	32.40	29	0.303

The chi-square test was handled differently for the CRGs due to the fact many strata are small and the chi-square test requires at least five expected observations in each cell.[5] Geographic areas were formed from the counties to increase cell sizes. The geographic groups are Davis-Morgan, Eastern Utah, Salt Lake, Southern Utah, Utah County, Weber and surrounding counties, Western Utah, and an unknown category.[6] The combined population and sample size had to contain at least 1,000 individuals to exclude small CRG strata[7]. Of the 1,054 CRG strata, only 164 met the threshold.

Table 4 shows only two of the CRGs have a statistically different sample geographic distribution from the population at a .05 alpha level. Four other CRGs show signs of having a sample geographic distribution that is statistically different than the population, but are not significant at a .05 alpha level. The two that are statistically different are 61442, Diabetes and Hypertension Level – 2, and 61902, Two Other Dominant Chronic Diseases Level – 2.

Table 4. Chi-square Test of Geographic Representativeness within CRG Strata

CRG	Number of Individuals in Population	Number of Individuals in Sample	Chi-Square	DF for Chi-Square	p-value for Chi-Square
61442	9,821	987	16.42	7	0.022**
61902	1,150	119	15.34	7	0.032**
62913	1,743	179	13.96	7	0.052*
61611	1,043	109	13.08	7	0.07*
30801	4,985	504	12.54	7	0.084*
50141	2,920	296	12.02	7	0.1*
62013	1,752	181	11.29	7	0.126

[5] Meaning that the sample population should have an expectation of at least five individuals in each county/geographic region for each CRG.
[6] The county to area grouping is mapped as follows Davis and Morgan into **Davis-Morgan**, Carbon, Daggett, Duchesne, Emery, Grand, Summit, Uintah, and Wasatch into **Eastern**, Salt Lake into **Salt Lake**, Beaver, Garfield, Iron, Kane, Piute, San Juan, Washington, Wayne into **Southern**, Utah into **Utah**, Box Elder, Cache, Rich, and Weber into **Weber**, Juab, Millard, Sanpete, Sevier, and Tooele into **Western**, unknown or outside of Utah into **unknown**.
[7] 1,000 was used because the smallest area, Western, had about 4% of the state population. A 10% sample of 1000 would be about 100 people or an expected 4 for the Western cell. Depending on the CRG the Western area may have more or less than 4% of that population, however a lower bar of 1,000 seemed appropriate for our purposes.

CRG	Number of Individuals in Population	Number of Individuals in Sample	Chi-Square	DF for Chi-Square	p-value for Chi-Square
35111	1,628	167	11.27	7	0.127
20800	6,207	624	11.05	7	0.136
32741	3,876	392	10.98	7	0.14
20770	3,291	332	10.66	7	0.154
53511	2,461	251	10.61	7	0.156
61603	904	96	10.57	7	0.158
61436	1,690	173	10.43	7	0.166
62613	1,189	122	10.24	7	0.175
61413	1,092	113	10.15	7	0.18
57472	961	98	9.89	7	0.195
62701	7,767	782	9.66	7	0.209
53912	1,579	161	9.57	7	0.214
62402	949	99	9.54	7	0.216
34451	18,712	1,875	9.46	7	0.221
37542	3,107	315	9.43	7	0.224
33571	4,445	449	9.12	7	0.244
20810	1,496	154	9.07	7	0.248
10040	3,307	333	9.04	7	0.25
61454	1,047	108	8.94	7	0.257
53513	1,792	183	8.68	7	0.276
20840	1,828	188	8.67	7	0.277
61416	983	103	8.51	7	0.29
10000	990,607	99,068	8.37	7	0.301
57431	4,735	479	8.28	7	0.309
36981	1,771	181	8.14	7	0.32
56652	1,063	112	8.03	7	0.33
51923	1,447	150	7.98	7	0.334
62412	2,508	254	7.9	7	0.342
54791	1,111	116	7.86	7	0.345
61443	8,364	841	7.84	7	0.347
33562	1,261	131	7.6	7	0.369
10170	9,637	968	7.47	7	0.382
62512	2,211	226	7.44	7	0.385
62011	1,279	131	7.4	7	0.389
57491	8,858	890	7.39	7	0.39
57511	968	101	7.37	7	0.391
62931	1,294	134	7.36	7	0.392
54411	3,036	309	7.16	7	0.412
57432	1,737	178	7.13	7	0.415
20780	3,093	315	7.11	7	0.418
62411	1,108	117	7.04	7	0.424
40001	27,183	2,724	6.94	7	0.435
40003	23,130	2,317	6.87	7	0.442
34462	2,445	250	6.87	7	0.443

CRG	Number of Individuals in Population	Number of Individuals in Sample	Chi-Square	DF for Chi-Square	p-value for Chi-Square
56651	2,567	262	6.82	7	0.448
61435	3,651	370	6.75	7	0.455
34461	23,902	2,394	6.69	7	0.462
10030	12,477	1,251	6.66	7	0.465
62421	1,467	151	6.66	7	0.465
53541	1,939	199	6.64	7	0.468
70713	973	103	6.57	7	0.475
20400	3,219	327	6.49	7	0.484
53481	1,055	111	6.49	7	0.484
20720	12,135	1,219	6.38	7	0.496
37552	17,481	1,752	6.24	7	0.512
33581	15,530	1,559	6.22	7	0.514
51912	1,170	123	6.06	7	0.532
51911	1,880	193	5.97	7	0.543
10130	13,426	1,347	5.96	7	0.544
51921	61,506	6,155	5.75	7	0.57
61412	1,222	127	5.74	7	0.57
51922	28,631	2,868	5.65	7	0.581
61713	946	99	5.47	7	0.603
20730	11,187	1,123	5.43	7	0.607
40004	9,039	908	5.4	7	0.611
34452	2,798	285	5.33	7	0.619
34091	982	103	5.32	7	0.62
62602	1,959	199	5.31	7	0.623
33561	3,118	317	5.29	7	0.625
62012	3,451	350	5.21	7	0.634
62702	8,288	834	5.18	7	0.638
62703	9,786	984	5.17	7	0.639
53521	2,580	262	5.15	7	0.642
61712	1,256	131	5.06	7	0.652
37551	15,370	1,541	5.01	7	0.659
54792	1,192	122	4.99	7	0.661
61432	2,790	282	4.95	7	0.666
31421	2,117	216	4.81	7	0.683
40002	4,527	458	4.79	7	0.686
20870	1,383	143	4.78	7	0.687
61613	1,886	193	4.78	7	0.687
20100	3,104	316	4.76	7	0.689
52661	1,873	191	4.75	7	0.69
57852	1,295	135	4.71	7	0.695
10140	6,768	679	4.65	7	0.703
20300	32,836	3,287	4.64	7	0.704
51872	1,318	136	4.64	7	0.704
62413	1,436	147	4.54	7	0.716

CRG	Number of Individuals in Population	Number of Individuals in Sample	Chi-Square	DF for Chi-Square	p-value for Chi-Square
10120	10,698	1,076	4.5	7	0.721
52411	1,050	109	4.39	7	0.734
53512	3,543	358	4.39	7	0.734
57492	3,504	355	4.39	7	0.734
54241	9,359	940	4.36	7	0.738
10060	2,157	220	4.3	7	0.744
51382	9,316	936	4.27	7	0.748
20850	12,520	1,256	4.27	7	0.749
20830	1,866	191	4.24	7	0.752
62614	1,097	114	4.14	7	0.764
62601	1,537	159	4	7	0.78
30181	3,500	355	3.91	7	0.79
37571	7,481	752	3.83	7	0.799
51331	937	98	3.83	7	0.799
10020	5,792	584	3.79	7	0.804
35251	2,036	208	3.69	7	0.815
50061	1,384	143	3.68	7	0.816
70712	1,043	110	3.64	7	0.82
54421	8,832	888	3.63	7	0.822
61444	2,551	260	3.57	7	0.828
20600	9,601	965	3.53	7	0.832
53911	1,911	198	3.44	7	0.841
61452	2,370	241	3.42	7	0.844
62934	3,087	312	3.42	7	0.844
30781	1,925	198	3.39	7	0.846
51381	13,810	1,385	3.37	7	0.849
20500	1,566	161	3.31	7	0.855
31001	998	105	3.28	7	0.858
62604	3,763	382	3.28	7	0.858
61423	1,163	120	3.21	7	0.865
10070	21,823	2,188	3.18	7	0.868
61422	1,575	161	2.97	7	0.888
62605	1,826	187	2.92	7	0.892
57481	1,222	127	2.9	7	0.894
61614	1,152	119	2.89	7	0.895
62933	5,536	559	2.89	7	0.895
54242	13,453	1,351	2.88	7	0.896
61434	5,045	508	2.84	7	0.9
61441	12,240	1,230	2.77	7	0.906
20820	3,584	363	2.74	7	0.908
33572	1,962	201	2.73	7	0.908
61433	3,682	373	2.71	7	0.91
61414	1,151	118	2.7	7	0.911
30191	13,011	1,306	2.67	7	0.913

CRG	Number of Individuals in Population	Number of Individuals in Sample	Chi-Square	DF for Chi-Square	p-value for Chi-Square
53522	936	99	2.62	7	0.918
20200	10,718	1,075	2.58	7	0.921
10150	11,160	1,122	2.54	7	0.924
30192	1,409	145	2.51	7	0.927
54243	2,132	218	2.51	7	0.927
10080	2,781	282	2.47	7	0.929
33582	1,495	154	2.41	7	0.933
10110	6,401	644	2.4	7	0.935
62704	4,701	475	2.39	7	0.935
10100	41,165	4,121	2.38	7	0.936
10010	535,785	53,587	2.32	7	0.94
37541	15,584	1,562	2.32	7	0.94
37572	964	100	2.18	7	0.949
57512	1,009	105	2.09	7	0.955
53451	961	100	2.02	7	0.959
61612	1,944	198	1.96	7	0.962
34101	2,913	296	1.82	7	0.969
62932	15,073	1,512	1.81	7	0.969
10160	5,945	600	1.77	7	0.972
62422	2,648	270	1.72	7	0.974
62603	2,820	287	1.7	7	0.975
56621	1,637	168	1.62	7	0.978
57471	2,712	277	1.48	7	0.983
70711	2,274	232	1.41	7	0.985
35101	4,284	433	1.01	7	0.995

[**]*Statistically significant at a .05 alpha level*

[*]*Statistically significant at a .1 alpha level*

Conclusion

Overall, the sample distribution represents the population distribution. However, there will be cases where the sample CRG distribution may be different from the population distribution. Caution should be used when the number of observations in any stratum is small. We encourage data requests for limited and research datasets when researchers are interested in rare events or small geographic areas.

If you have any question concerning how this sample was obtained, feel free to contact OHCS at (801)-538-7048 or by email at healthstat@utah.gov.