Visualizing Nationwide Variation in Medicare Part D Prescribing Patterns

Alexander Rosenberg[1,2,3], Christopher Fucile[2,3], Robert J. White[3,4], Melissa Trayhan[3,4], Samir Farooq[3,4], Lisa A. Nelson[5], Caroline M. Quill[3,6], Samuel J. Weisenthal[3,4], Kristen Bush[3,4], and Martin S. Zand[3,4,6,7]

[1]Department of Microbiology, University of Alabama at Birmingham, BBRB 276, 1720 2nd Ave. South, Birmingham AL 35294-2170
[2]Informatics Institute, University of Alabama at Birmingham, 1720 2nd Avenue South, Birmingham AL 35294-3412
[3]Rochester Center for Health Informatics at the University of Rochester Medical Center, 265 Crittendon - Rm 1.207, Rochester, NY 14642
[4]Clinical and Translational Science Institute, University of Rochester, 265 Crittenden Blvd, Rochester, NY 14642
[5]Department Pharmacy, University of Rochester Medical Center, 601 Elmwood Ave, Rochester, NY 14642
[6]Department of Medicine, Division of Pulmonary and Critical Care Medicine, University of Rochester Medical Center, 601 Elmwood Ave, Rochester, NY 14642
[7]Department of Medicine, Division of Nephrology, University of Rochester Medical Center, 601 Elmwood Ave - Box 675, Rochester, NY 14642

Corresponding author:
Martin S. Zand[7]

Email address: martin_zand@urmc.rochester.edu

ABSTRACT

Objective To characterize the regional and national variation in prescribing patterns in the Medicare Part D program using machine learning and dimensional reduction visualization methods.
Methods Using publicly available Medicare Part D claims data, we identified regional and national provider prescribing profile variation with unsupervised clustering and t-distributed stochastic neighbor embedding (t-SNE) dimensional reduction techniques. Additionally, we examined differences between regionally representative prescribing patterns for major metropolitan areas.
Results Distributions of prescribing volume and medication diversity were highly skewed among over 800,000 Medicare Part D providers, and medical specialties had characteristic prescribing patterns. Although the number of Medicare providers in each state was highly correlated with the number of Medicare Part D enrollees, some states were enriched for providers with >10,000 prescription claims annually. Hierarchical clustering and t-SNE dimension-reduction visualization of drug- or drug-class prescribing patterns revealed that providers cluster strongly based on specialty and sub-specialty, with large regional variations in prescribing patterns. Major metropolitan areas had distinct prescribing patterns that tended to group by major geographical divisions.
Conclusions There is substantial prescribing variation among providers in Medicare Part D both between and within specialties. Large regional variations in prescribing patterns, particularly among major metropolitan areas, were also seen. Unsupervised clustering and t-SNE dimension-reduction are an effective means to examine variation in provider prescribing patterns, including substantial regional and medical specialty variation.

Keywords:
Medicare,Prescribing, Machine learning, t-SNE, Healthcare variation

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3470v1 | CC BY 4.0 Open Access | rec: 18 Dec 2017, publ: 18 Dec 2017

INTRODUCTION

Pharmaceutical spending accounts for 5-25% of total medical care expenditures in Europe, and 16% of all Medicare expenditures in the United States. Variation in prescribing patterns is common, even within groups of providers with a similar scope of practice and patient mix. Prescribing variation may be due to a combination of provider preferences, patient case-mix, deviation from practice guidelines, insurance formulary restrictions, and occasionally fraud [1–5]. Understanding the patterns of prescribing variation is critical to improving healthcare delivery. Visualizing prescribing variation in ways that accurately reflect underlying data structure can be challenging. Good data visualization can provide a ”big picture” of complex data, especially variation and quantitative changes in large and complex data sets. In this manuscript, we apply machine learning and non-linear visualization methods Medicare Part D provider prescribing pattern data, revealing significant provider variation at the local, regional and national levels, even when controlled for provider specialty and medication volumes.

Prescription claims data captures the volume, diversity and cost of medications prescribed by individual providers. For example, the 2013 Medicare Part D prescribing pattern data set consists of 1,049,381 providers and 3,449 prescription drugs [6]. Because the claims are linked to thousands of individual provider treatment decisions, their patterns can also provide an objective measure of how medical care is actually delivered, especially with a large data set. A list of the types and volumes of possible medications prescribed by an individual provider quantifies a pattern of medical practice. In machine learning (ML), this list is termed a feature vector, and can be used to cluster providers with similar prescribing patterns. Cluster membership can then be compared to other, independent characteristics such as geographic location, medical specialty, patient case mix or outcomes. ML is very efficient for analyzing data with hundreds or thousands of features, particularly when the gold-standard or ground-truth for cluster membership is unknown (e.g. how providers should be grouped).

Visualizations that accurately reflect feature variation in high dimensional data (i.e. with a large number of features) are extremely useful for data exploration, inference and decision making [7, 8]. Standard visualization methods use classical multidimensional scaling [9], linear transformations that project multidimensional data into two or three dimensions, while preserving the relative distances between data points. Principal Components Analysis (PCA) [10] is one such method. When applied to high dimensional data, however, PCA and other linear transformation methods most often result in dense visualizations that do little to clarify the underlying data structure or to support decision making. In contrast, non-linear visualization methods [11–13] are often used to better visualize large, high dimensional, data sets. Recently, van der Maaten and colleagues developed t-distributed stochastic neighbor embedding (t-SNE) [14], which balances cluster display at the local and global levels. Given the large number of medications or medication classes that providers can prescribe, t-SNE is ideally suited to visualizing prescribing patterns variation for large numbers of providers over thousands of medications or a few hundred medication classes.

Regional variation in health services delivery has been well described [15–21]. In contrast, little is known about regional *patterns* of prescription drug utilization beyond focused studies of prescribing patterns for antibiotics [1], chemotherapy [22], cholinesterase therapy [23], psychiatric medications [24], and statin cholesterol lowering agents [25]. Such patterns have been found to reflect the nature and complexity of health status of patient populations [26, 27], patient socioeconomic factors [28–31], provider preferences with self-reinforcing regional influences [32–34], social network influence (i.e. ”prescriber contagion”) [35], and composition of specialties and Medicare formulary [34]. Variation of regional prescribing practices has important implications for behavioral, economic, and healthcare outcomes [2, 36–38]. To our knowledge, however, there are currently no published analyses that examine and visualize geographic variations in prescribing patterns at a national level, irrespective of provider specialty or medications.

The focus of this work is twofold. First, t-SNE to visualize the prescribing patterns of Medicare Part D providers based on the volume and types of medication claims, and agglomerative clustering to validate groupings of providers identified by t-SNE. Second, we identify regional prescribing pattern differences among Medicare Part D providers across specialties, and examine variations in the distribution of prescribing patterns across medical specialties, states, and geographic regions in the United States.

METHODS

Medicare Part D data

Medicare Part D 2013 provider prescribing data were downloaded directly from the Center for Medicare Services (CMS) [6]. A provider refers to any individual who is licensed to prescribe medications and appears in the data set. The data was packaged as three files: 1) a table of providers and their associated annotations, including their unique national provider identifier (NPI), address, summary statistics on numbers of claims, costs, etc., 2) a table of drugs and their associated annotations including flags for whether they are narcotics, DEA schedule II or III, or categorized as Beers (medications to avoid in older adults [39]), as well as summary statistics (e.g. numbers of claims, costs, etc.), and 3) a table of NPI, drug (both brand and generic names, which taken together are unique) and the number of claims, duration of prescription, and cost for each provider-drug combination. This third file represents a bipartite graph specifying connections between disjoint sets of nodes (i.e. providers and drugs) that are linked by a corresponding measure (e.g. number of claims). To comply with data privacy requirements, values in the provider-by-drug matrix less than 11 were set to 0 by CMS prior to data release [40]. All formatted data were imported into Matlab R2016a (Mathworks, Natick MA) or Mathematica 11.1 (Wolfram, Champaign IL) for further analysis and visualization.

For analysis, clustering and visualization, a feature vector was created for each provider $\Omega_i = \{\alpha_{i,1}, \alpha_{i,2}...\alpha_{i,m}\}$ where i is the provider number and $\alpha_{i,j}$ is the number of Medicare outpatient prescription claims for drug α_j attributed to provider i. The total number of providers is designated by n, and the total number of individual drugs by m. A restriction of the data set, implemented by CMS to ensure non-identifiability of Medicare recipients, is that if $\alpha_{i,j} \leq 11$, then $\alpha_{i,j} = 0$. With this constraint, the summary number of claims associated with a particular provider (or drug) in the CMS data set may not be exactly equivalent to the sum of the provider-by-drug matrix. Thus, there are 1,049,381 providers and 3,449 drugs in the data set, there are only 808,020 providers with $\geq$11 claims for at least one drug. Similarly, there were 2,892 drugs with $\geq$11 claims from at least one provider.

Supporting data sources

Supplemental Figure 1 shows a schema of the prescribing patterns data set along with other data sources used for this study. All data sources used in this work are publicly available. The number of Medicare Part D participants by state were obtained from CMS public use files (boxes 1, 2, and 3) [41]. For part of our analysis, we consider providers within 52 metropolitan areas with a population $\geq$1,000,000 by the July 2012 Core-Based Statistical Areas (CBSAs) estimate [42]. We link CBSAs to counties and their Federal Information Processing Standards (FIPS) code, using a look-up table from the National Bureau of Economic Research (box 8) [43]. We linked providers to their FIPS county codes using a table from the U.S. Department of Housing and Urban Development website (box 5) [44]. Finally, we obtained population estimates of Medicare Part D participants by county from the Kaiser Family Foundation website [45], where we consider both Medicare Advantage and the Prescription Drug Plan (box 7) enrollees. To group individual drugs into broader categories for analysis, we used the National Drug File from the Veterans Administration [46] followed by further, minor manual aggregation to result in 198 drug categories (box 4). Data regarding providers who have been excluded from participation in Medicare, Medicaid, and all other Federal health care programs was obtained from the Department of Health and Human Services, Office of the Inspector General, List of Excluded Individuals/Entities (LEIE) website [47]. This list provides data, including NPI numbers, of individuals and entities currently excluded from participation in Medicare, Medicaid, and all other Federal health care programs. Individuals and entities who have been reinstated no longer appear in this data set.

Dimensional Reduction, Machine Learning, and Statistical Methods

Providers with similar prescribing patterns were identified by agglomerative clustering implemented in Wolfram Mathematica with Ward's minimum variance criteria, which minimizes the total within-cluster variance [48], for determining cluster membership and number. Clusters were additionally grouped by provider geographical region, state, and medical specialty. To visualize clusters of providers based on their prescribing patterns, we used the fast t-distributed stochastic neighbor embedding (t-SNE) dimension reduction method of van der Maaten and Hinton [14]. Given the size of the data set, with $> 10^5$ providers, we used the fast Barnes-Hut implementation of t-SNE in Matlab [49], with 50 initial dimensions, a perplexity of 40, and $theta$ = 0.5. The algorithm performed 300-1,500 iterations per run and we selected the result with the minimum t-SNE cost function (error rate) [14] among 1,000 runs. Dimensional reduction to visualize the CBSA groupings CBSAs was accomplished using classical multidimensional scaling [9] implemented in Matlab using a CBSA-CBSA distance matrix with one minus correlation as the metric. Comparisons of the differences in proportion of provider fractions between geographic regions was performed using the Mann-Whitney U test.

RESULTS

Volume and Diversity of Medicare Prescriptions

We first examined the overall statistical distributions of prescribing volume and diversity (Figure 1). Overall, only a small fraction of the total Medicare Part D outpatient unique medications were prescribed by more than 5% of providers. Figures 1A and 1C show the frequencies of the 2,892 individual drugs prescribed by both percentage of providers and overall number of claims, respectively. Only 165 unique drugs (5.7%) were prescribed by at least 5% of the providers (Figure 1A). Similarly, only 197 unique drugs (6.8%) had more than one million claims across all providers (Figure 1C). To examine the patterns in terms of type of medication, reduce the effect of formulary restrictions or brand name versus generic medications, we collapsed the unique drug features into 197 categories (Figure 1B and 1D) resulting in distributions that were less skewed, with 72 drug classes (36.5%) prescribed by ≥5% of the providers, and 83 classes (42.1%) surpassing one million claims across all providers.

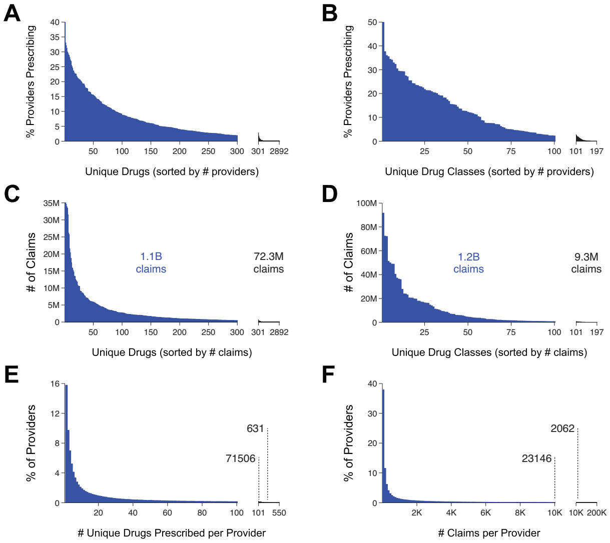


Figure 1. Overall features of 2013 Medicare Part D prescribing patterns data set. A. Distribution of percentage of providers prescribing each of 2,892 unique drugs, sorted by percentage of providers prescribing. B. Same as A except for 197 unique drug classes. C. Distribution of number of claims for each of 2,892 unique drugs, sorted by number of claims. Note that the unique drug order is not necessarily the same as in A. D. Same as B except for 197 unique drug classes. E. Distribution of drug prescription diversity across all providers sorted by number of unique claims. Numbers of providers prescribing more than 100 and 300 unique drugs are annotated on plot. F. Distribution of number of claims across all providers sorted by claims per provider. Number of providers making more than 10,000 and 25,000 claims are annotated on plot.

We next analyzed provider prescription diversity, defined as the number of different drugs prescribed by each provider. Figure 1E shows the distribution of prescription diversity across all providers. The majority (70.3%) of providers prescribe ≤25 unique drugs reimbursed by Medicare. However, this is a long-tailed distribution, with 71,506 providers prescribing ≥100, and 631 providers ≥300 unique drugs. Importantly, most providers have few Medicare drug claims (Figure 1F). This may be due to many Medicare enrollees having multiple types of prescription coverage (e.g. Medicare and Veteran Administration), with many of their prescriptions filled outside of Medicare Part D [50]. Because the data set was limited to Medicare Part D claims, we could not assess for this factor. We again observed a long

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3470v1 | CC BY 4.0 Open Access | rec: 18 Dec 2017, publ: 18 Dec 2017

tail of providers associated with ≥100,000 claims. There were 2,062 high-volume prescribing providers (HV) with ≥25,000 claims utilizing 1,954 of the 2,892 available drugs. This group of 0.2% of providers were responsible for 3.59% Medicare Part D drug costs in 2013. Compared with the standard volume prescribing providers (SV; n=805,958), the small subset of HV (n=2,062) was heavily skewed towards general practice ($p<0.001$): 89% of HV providers were categorized as internal medicine, family medicine or general practice (SV = 25.8%), and 3% were geriatric medicine (SV = 0.2%).

Regional Prescribing Volumes and Patterns

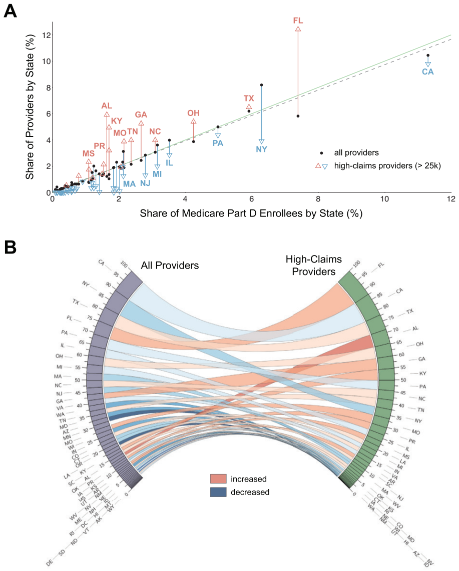


Figure 2. Distribution of Medicare Part D providers across states. A. Share of providers by state (as a percentage of the total number of providers) plotted against share of Medicare Part D enrollees by state (as a percentage of the total number of enrollees nationwide) are shown by black circles and fit to a line (gray dashed line); green line is slope of one. A similar plot based on a data subset of high-claims providers (¿ 25,000 claims resulting in 2062 providers) is shown superimposed as open triangles colored by their relation to the corresponding data from the full data set. Some states are annotated. B. Comparison of the provider composition by state for the full data set (left) and the high-claims data set (right). Ribbons connecting the two join corresponding states.

We next examined the degree to which prescribing volumes correlated with the regional distribution of Medicare Part D prescription benefit enrollees. On a state-by-state basis, the number of Medicare Part D providers was highly correlated with the corresponding number of Medicare Part D enrollees (Figure 2A, R^2=0.950). However, this relationship was not statistically significant (R^2=0.697) for providers with >25,000 claims. There were substantial deviations for several states. For Florida and New York, these deviations might partly be explained by the ratio of providers to enrollees for the elderly or enrolled populations: fewer in Florida or more in New York, such that Medicare drug prescribing was more/less concentrated among those providers. In contrast, several states with a proportional number of providers and enrollees had a high share of high-claims providers (e.g. Georgia).

Figure 2B compares the ranking of all providers versus high-claims providers, with ribbons joining corresponding states. Alabama, Georgia, and Kentucky, for example, have a larger share of high-prescribing Medicare providers as compared to all providers, whereas New York, Massachusetts, New Jersey, Arizona, and Oregon have a smaller share. In contrast to the relatively similar ratios of Medicare providers per enrollee across states, the distribution of high-prescribing providers varies regionally (Table 1). The East and West South Central States had statistically significantly higher proportions of high prescribing providers compared to the New England, Middle Atlantic, West North Central, and Mountain Federal Standard Census Regions. In addition, the Middle Atlantic regions had a significantly higher proportion of HV providers than the Mountain and New England regions (see Table 1 for p values).

We also examined the relationship between prescribing diversity and prescription claim volume. In

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3470v1 | CC BY 4.0 Open Access | rec: 18 Dec 2017, publ: 18 Dec 2017

Table 1. Differences in high-prescribing provider fractions by geographic region

Region	States	n =	Median	Min	Max	PAC	ESC	WSC	MTN	NE	SATL	WNC	ENC
Pacific	AK, CA, HI, OR, WA	5	0.19	0.0000	9.8000								
East South Central	AL, KY, MS, TN	4	4.47	2.3300	5.9200	0.1113							
West South Central	AR, LA, OK, TX	4	1.94	1.1200	6.5000	0.1113	0.1939						
Mountain	AZ, CO, ID, MT, NM, NV, UT, WY	8	0.10	0.0000	0.4800	0.5059	0.0049	0.0049					
New England	CT, MA, ME, NH, RI, VT	6	0.29	0.0000	1.2600	0.8535	0.0075	0.0139	0.5130				
South Atlantic	DC, DE, FL, GA, MD, NC, SC, VA, WV	9	1.41	0.0000	12.4200	0.3485	0.1218	0.4869	0.0470	0.1218			
West North Central	IA, KS, MN, MO, ND, NE, SD	7	0.34	0.0000	3.8800	0.9340	0.0099	0.0279	0.5161	0.8253	0.1314		
East North Central	IL, IN, MI, OH, WI	5	1.99	0.3400	5.3800	0.1437	0.1113	0.9025	0.0042	0.0222	0.4222	0.0481	
Middle Atlantic	NJ, NY, PA	3	3.93	1.3100	4.3600	0.1360	0.3768	0.5959	0.0104	0.0138	0.4588	0.0206	0.5510

p-values calculated by the Mann-Whitney U test, with significant values shown in red.
A map of the U.S. Census regions can be found in Supplemental Figure 2.

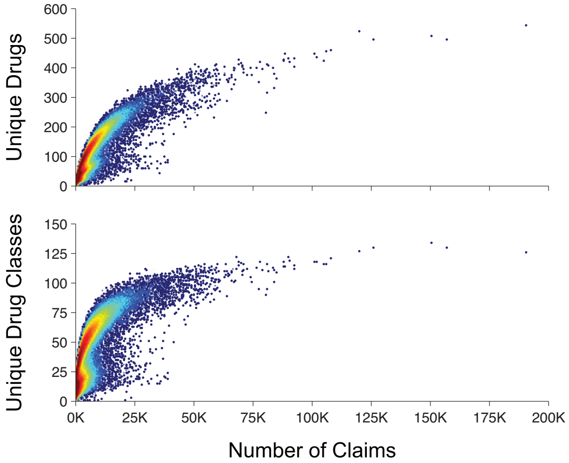


Figure 3. Comparison of prescribing diversity and prescribing volume. Density/scatter plot indicating the number of unique drugs (top) drug classes (bottom) prescribed (diversity; y-axis), number of claims (volume; x-axis) and number of providers (bin height coded as color. Bins that have a single provider are indicated by a blue dot.

general, the diversity of drugs prescribed by any individual provider increased with the number of claims for individual drugs or drug classes (Figure 3). HV providers are characterized by high prescribing diversity, with a few outliers in terms of prescribing volume. For example, only 10 Medicare providers accounted for approximately 12% of all 2013 Medicare Part D zoster vaccine claims, each with ≥10,000 claims accounting for over $30 million in claims. The reason(s) these specific such outliers could not be determined based on the available data; there was no specific pattern of geographic distribution, urban versus rural practice location, or medical specialty observed.

Provider Prescribing Patterns Highly Correlate With Provider Specialty

We next compared t-SNE visualization of provider groupings with principal components analysis (PCA). Figure 4 shows the projection of provider densities resulting from t-SNE applied to providers with ≥1,000 claims (n = 227,573) and using a feature vector of corresponding drugs (n = 2,791; Figure 4A) or drug classes (n = 195; Figure 4B), where claim volumes in Ω_i were initially normalized by total claims per provider. Note the density of the PCA projection, with the very high density areas obscuring finer variations in prescribing patterns. t-SNE has an advantage over PCA for visualizing this type of data because the embedding is not skewed by a few dominant features and t-SNE can reveal more subtleties in the similarities of provider groups [14]. In both provider-by-drug and provider-by-class t-SNE projections, there is one dominant grouping of Internal, Family, Geriatric, and General Medicine providers. The density within this grouping is relatively uniform, although there are some areas of higher density reflecting subgroupings of providers with similar prescribing patterns.

The t-SNE groupings are highly correlated with provider specialty and subspecialty (Figure 5). These

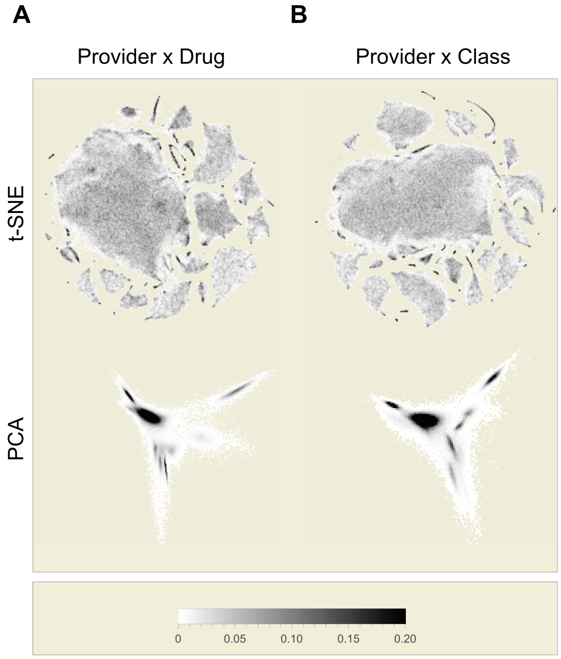


Figure 4. **Low-dimension embedding of providers using t-SNE and PCA.** 2-D density plots in low dimensional space created using t-SNE or of 227,573 Medicare Part D providers, each with ≥1000 prescription claims in 2013 organized by (A) the 227,573 x 2,791 drug claims matrix or (B) the 227,573 x 195 drug class claims matrix. Number-of-claims data per provider by drug or drug class is scaled by the total claims per provider to express the prescribing pattern as a composition prior to t-SNE.

plots, based on the provider-by-drug matrix and cross-referenced with provider specialty from the NPPES database, highlight that some specialties have single, distinct dominant clusters (e.g. Dermatology, Endocrinology, Nephrology) whereas others can have multiple clusters or sub-clusters that may reflect divisions within a specialty (e.g. Gastroenterology, Urology). Furthermore, related specialties can be spatially resolved in this fashion, for example, Cardiology and Cardiac Electrophysiology, as well as Ophthalmology and Optometry.

Variations in Provider Prescribing Patterns

We next used t-SNE visualize the diversity of prescribing across many different provider cluster regions, including generic and branded medication formulations (Figure 6) using the full provider-drug matrix. Ten random providers were chosen from 20 regions of the low-dimension tSNE visualization (Figure 6, labeled A-T), which mapped to 47 different agglomerative clusters. Corresponding compositional prescribing patterns from these providers are shown as a heat map where the columns are 200 individual providers and the rows are drug names. To to allow legibility, the union of the top eight prescribed drugs in each of the 20 clusters is displayed, resulting in 111 medications. Color corresponds to the percentage of claims for medications prescribed by a particular provider over the provider's total Medicare Part D medication claims. Both medication diversity (number of unique drugs prescribed) and claim volume are shown above the heat map.

Note that location variation on the embedding corresponds to different prescribing patterns, as indicated by the heat map. For example, clusters E and P both are dominated by Urology (see Figure 5), but E is characterized by large proportions of claims for tamsulosin and finasteride whereas P is mainly tamsulosin (though more subtle differences may not be evident in this figure based on the filtering described above). Cluster L is largely Ophthalmologists, consistent with high proportions of latanoprost and to a lesser extent, timolol maleate, Lumigan, Alphagan and similar drugs. The area K providers are concentrated for Allergists that prescribe high proportions of fluticasone proprionate and montelukast sodium. Clusters G and S are dominated by Neurologists, yet the prescribing patterns in these two groups have substantially different patterns. Cluster S has providers prescribing large amounts of carbidopa-levodopa, ropinirole, amantidine, azilect and similar medications whereas cluster G is more biased towards

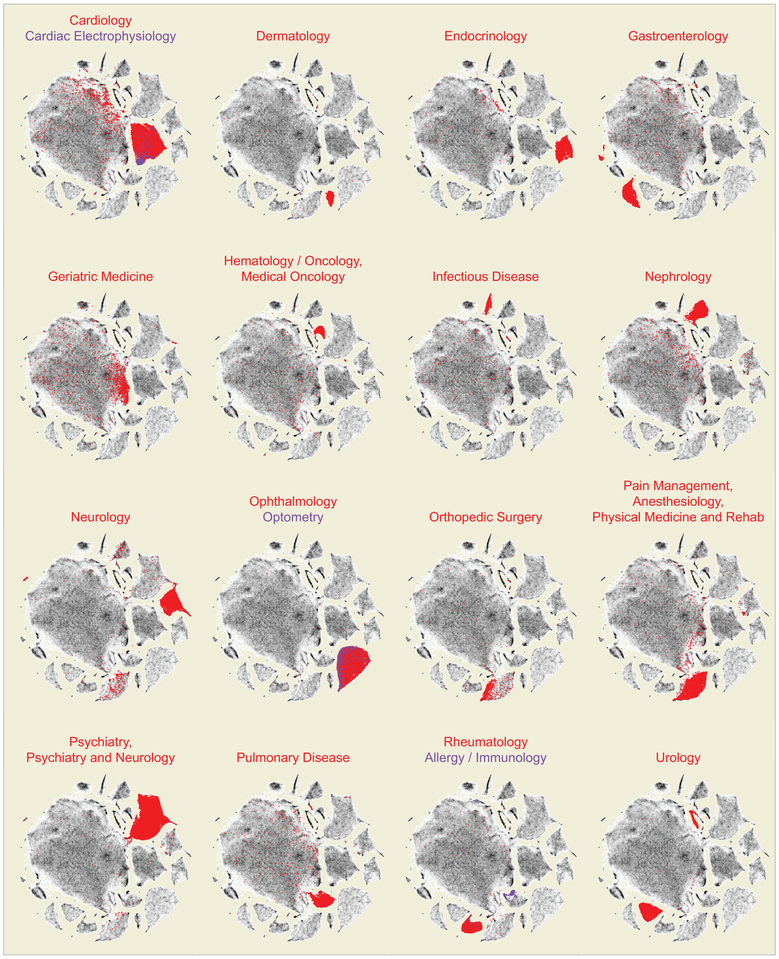


Figure 5. Array of t-SNE plots each highlighting providers of a specific specialty. Each 2-D density plot is the same as shown in Figure 4A. Annotation of different indicated provider specialties reveals their collocation by prescribing pattern.

drugs like levetiracetam, lamotrigine, vimpat, topiramate, namenda and donepezil, suggesting that the S providers see more Parkinson's disease patients and that the G providers deal more with epilepsy and Alzheimer's disease. In other cases, regional variation may strongly influence prescribing patterns. For example, cluster A is dominated by providers in Puerto Rico. These results demonstrate the utility of using t-SNE to visualize variation of prescribing patterns that highly correlate with formal provider clusters.

Visualizing Prescribing Volume and Medication Distribution Patterns

t-SNE plots can also be annotated by the prescribing proportions for individual drugs as shown in Figure 7. Here, for eight drugs typically prescribed for cardiovascular-related conditions, the percentage of claims for individual providers relative to their total number of claims are coded by color. Note that these are visible as high proportions within the tSNE region corresponding to Cardiology (see Figure 5). Even within the tSNE Cardiology region, high prescription rates of these drugs are associated with different provider groupings (see for example, atorvastatin, cloidogrel, and warfarin). These groupings may reflect differences in provider scope of practice, patient populations, Medicare formularies, or provider prescribing preferences.

In a similar fashion, the dimension-reduced space can be annotated by claim volume as shown in Supplemental Figure 3. In this figure, each point is color coded by claim volume. There is slight gradient of claim volume in the large, central General Medicine/Internal Medicine/Family Practice region with several small densities of extremely high prescribing volume providers (e.g. $\geq$10,000 claims). Claim volume also correlated with drug diversity (see Figure 3), so volume will be somewhat conflated with

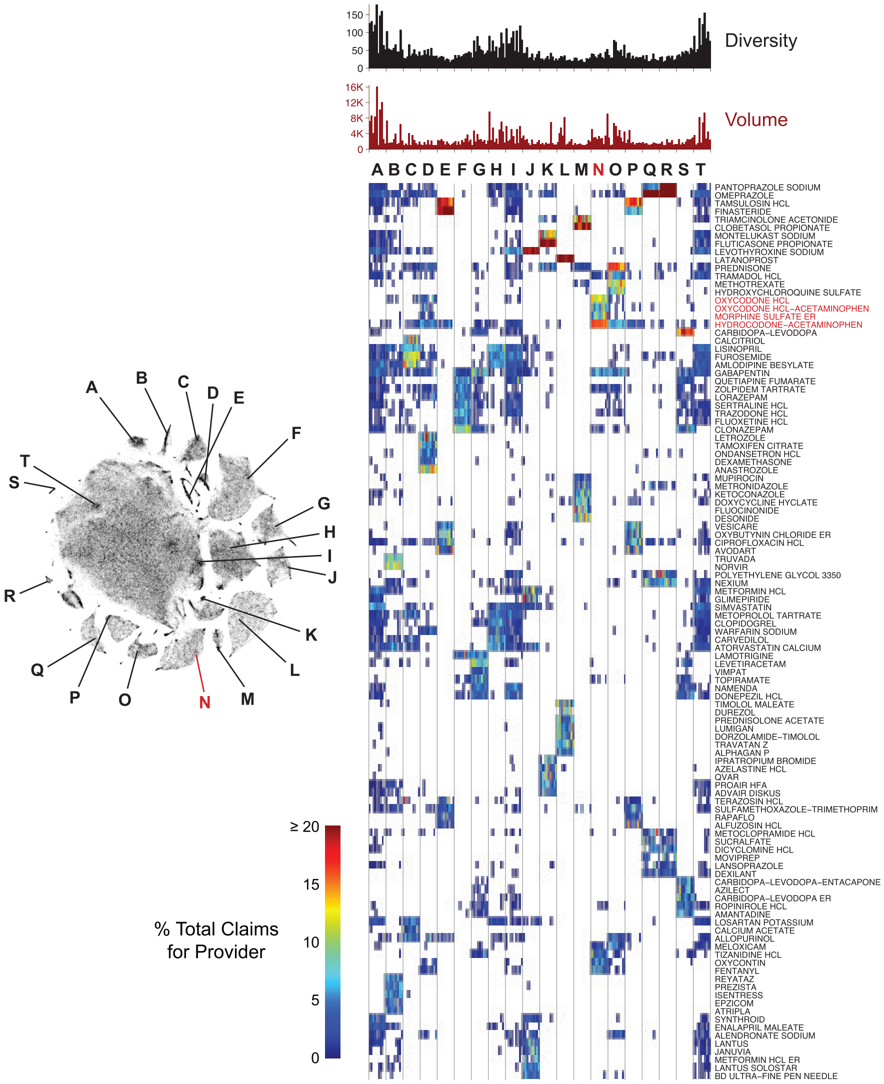


Figure 6. Representative prescribing patterns corresponding to different regions of t-SNE plot. Left: t-SNE plot as shown in Figure 4A with 20 different regions labeled as A through T. Right: Heat map showing prescribing patterns. Columns are individual providers, 10 randomly selected from each of the 20 regions. Each row represents a drug. To allow labeling, drugs shown are the union of the top eight most frequently prescribed in each region. Color corresponds to the percent of claims for a particular drug made by a provider relative to their total claims, with white denoting no claims. Prescribing volume (total claims) and diversity (number of unique drugs prescribed) are shown above the heat map as bar graphs. Note region N, which is enriched for providers with a high volume of opioid analgesic claims.

prescribing pattern and will affect position in the low-dimensional embedding. However, plots highlighting single drugs (Figure 7) suggest that the variation across the large tSNE region correlate well with the prescribing patterns of individual providers.

Figure 8 shows the specialist-annotated embeddings based on medication class (see Figure 4B). As with the embeddings based on individual medications, specialists are enriched in the smaller clusters surrounding the main cluster. Figure 9 shows this embedding annotated for prescription proportion of six cardiology-related drug *classes* (similar to Figure 7). Even when considering classes instead of individual drugs, which eliminates clustering differences due to separately considering different formulations of the same drug (i.e. generic and brand name), there are clearly large variations in prescription patterns within the cardiology cluster (see for example, anticoagulents, calcium channel blockers, and platelet aggregation inhibitors).

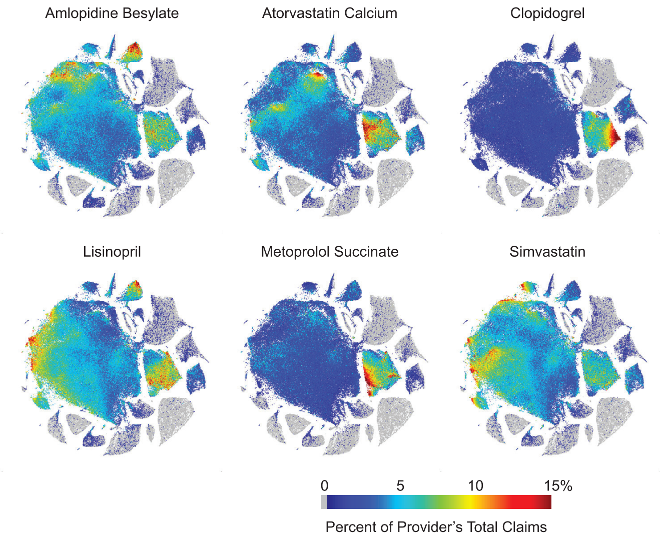


Figure 7. Array of t-SNE plots of providers annotated for fraction of claims for each of eight heart/circulation related drugs. The color for each provider corresponds to the percentage of claims for the indicated drug relative to the provider's total claims. Gray is 0%, the maximum scale (red) is 15%.

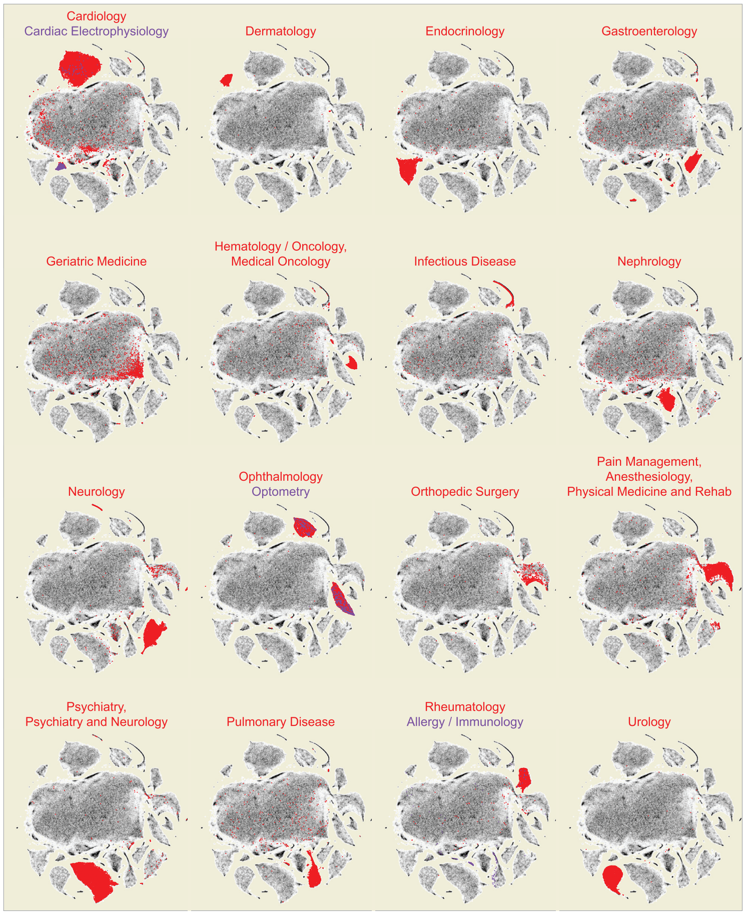


Figure 8. Array of t-SNE plots each highlighting providers of a specific specialty. Same as Figure 5 except t-SNE plots are based on provider by drug class matrices (as shown by Figure 4B).

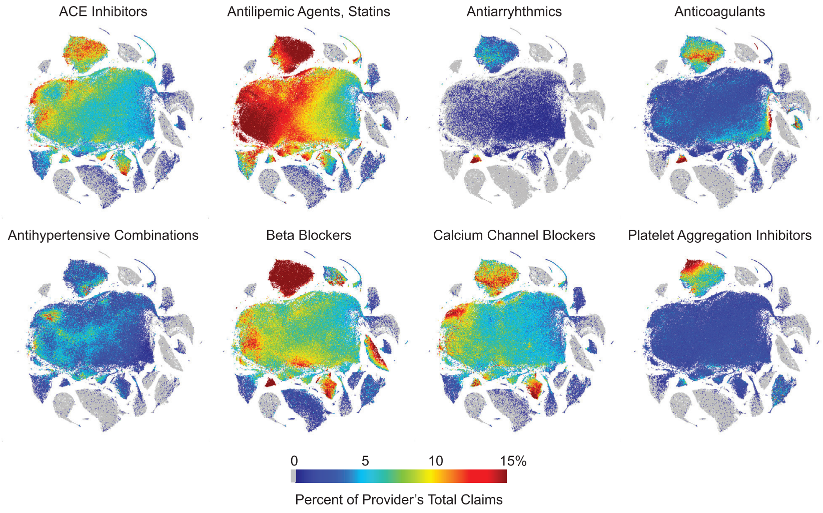


Figure 9. Array of t-SNE plots of providers annotated for fractions of claims for each of six cardiac drug classes Drug classes include all medications (generic and brand name) collapsed into the indicated class. The color for each provider corresponds to the percentage of claims for the indicated drug relative to the provider's total claims. Gray is 0%, the maximum scale (red) is 15%.

Hierarchical Clustering of Provider Prescribing Patters

To more rigorously identify provider subspecialty association with t-SNE heatmap regions, we performed unsupervised hierarchical cluster analysis with a centroid linkage method on the provider-by-drug-class matrix. We identified 605 provider clusters using agglomerative clustering with Ward's minimum intercluster variance linkage minimization (Supplemental Figure 5, and Figure 10A). The dominant provider subspecialty classification within a cluster, taken from the NPPES data, was used to map each of the 605 sub-clusters to provider sub-specialties. 91% of the clusters had a dominant provider specialty identity corresponding to ≥30% of the providers within the cluster (Figure 10B). When mapped to US Federal Regions (Figure 10C), clusters also reflected regional variation in prescribing patterns. For example, within the t-SNE projection, we highlighted sub-clusters of providers identified as Family Medicine and then divided by Federal Region. This combination of clustering and t-SNE visualization made visible striking regional variations in regional medication prescribing volumes and patterns within Family Practice.

We also examined whether these providers who appeared in smaller outlier clusters represented Medicare prescribing fraud [51]. We found 397 providers that were also present in the 2017 List of Excluded Individuals/Entities, indicating that these providers had been barred from billing Medicare within 4 years after the 2013 Medicare Prescribing Patterns data set was released. The most frequent medical specialties in this group included Internal Medicine (27%), Family Practice (23%), and Psychiatry (10%). Excluded individuals were present in 87 of the 605 sub-clusters. No cluster with ¡18 individuals contained excluded providers, and no single cluster had ≥6% excluded providers. No consistent fraud associated pattern was found using data regarding prescription volume, diversity, or medications prescribed.

Regional variation in prescribing patterns

Given the variation in prescribing patterns observed within the Internal Medicine-Family Practice-General Medicine cluster, we next performed an in depth characterization of regional differences in prescribing patterns over all sub-specialties by census region 2. For these visualizations, we used heat maps of provider density within the tSNE embedding. Figure 11 shows how the prescribing patterns of providers with ≥1,000 Medicare Part D claims are clustered within each census region, as compared to a non-overlapping random sample from the entire data set. This type of visualization accounts for equivalent sample sizes, but not variation in the proportion of Medicare Part D provider types (e.g. Family Practice versus Nephrology) between the random and regional samples. For example, the East North Central region has a much higher percentage of Neurologists compared with the East South Central region. Differences in provider and population density, and thus prescribing patterns and volumes, may also contribute to

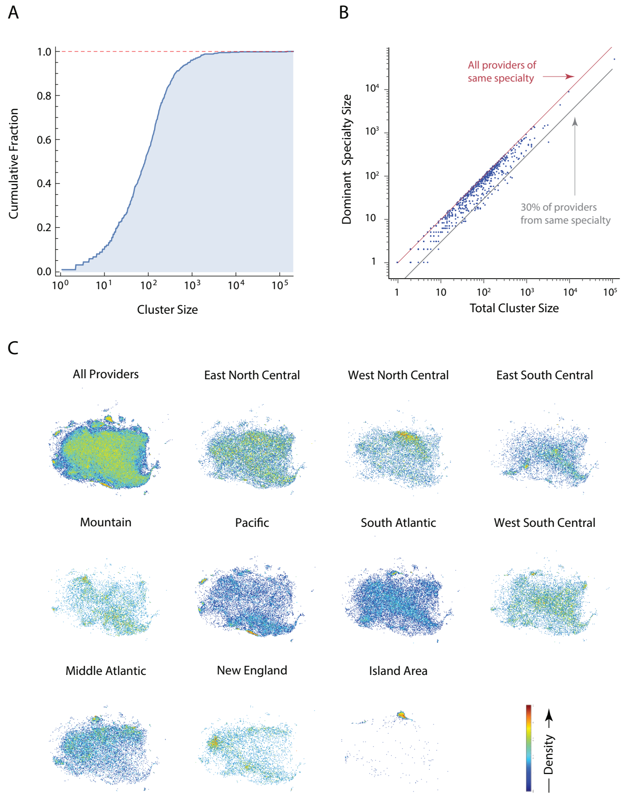


Figure 10. Unsupervised hierarchical clustering by drug class Provider clusters obtained by hierarchical agglomerative clustering using a Euclidean distance measure and centroid criteria. A) Cumulative distribution of provider size over 605 clusters. B) Provider specialties within each cluster were tallied and the number of providers in the dominant specialty plotted against cluster size. The lines indicates where 100% (red), or 30% (gray) of providers in the cluster are the same medical specialty. C) tSNE visualization of provider prescribing pattern variation for Family Medicine providers by United States Federal Region. Each plot represents a 2D density histogram.

regional variations in Medicare part D prescription costs.

To further explore regional variations in prescribing patterns while diminishing the impact of population density, we selected 52 metropolitan areas (core-based statistical areas, CBSA) with populations greater than one million. Among the large metropolitan areas, there were large regional differences in terms of proportion of Medicare Part D enrollees of the total population, as shown in Supplemental Figure 6, ranging from 4.6% (Washington DC) to just under 15.7% (Pittsburgh). These results were not statistically correlated to overall population of the respective CBSAs.

Dimension-reduction with t-SNE visualizations also revealed regional variation in prescribing patterns across CBSAs. To characterize prescribing profiles in CBSAs, we selected the 532 drugs with over 100,000 claims for all states. A 52 CBSA by 532 drug number-of-claims matrix was computed and each row was divided by the number of Medicare Part D enrollees in the corresponding CBSA, expressing the data as drug claims per enrollee. Figure 12A shows the first two coordinates of the resulting multidimensional scaling based on pairwise CBSA-CBSA distances $d_{i,j} = 1 - r_{i,j}$, where $r_{i,j}$ is the Pearson product-moment correlation coefficient for the CBSA pair i and j. The red dots near the center of the plot are the result of multi-dimensional scaling following random permutation of the CBSA provider memberships (preserving the relative numbers of providers per CBSA) used as a reference against which to interpret the dispersion of the real data. Although the data do not segregate into distinct clusters in this dimension, there are apparent regional variations, notably, that most of the southern CBSAs appear on the left half of the plot, reflecting similar regional prescribing profiles within the southern CBSAs.

Figure 12B shows an example of claims-per-enrollee of the 532 drugs for two geographically distant

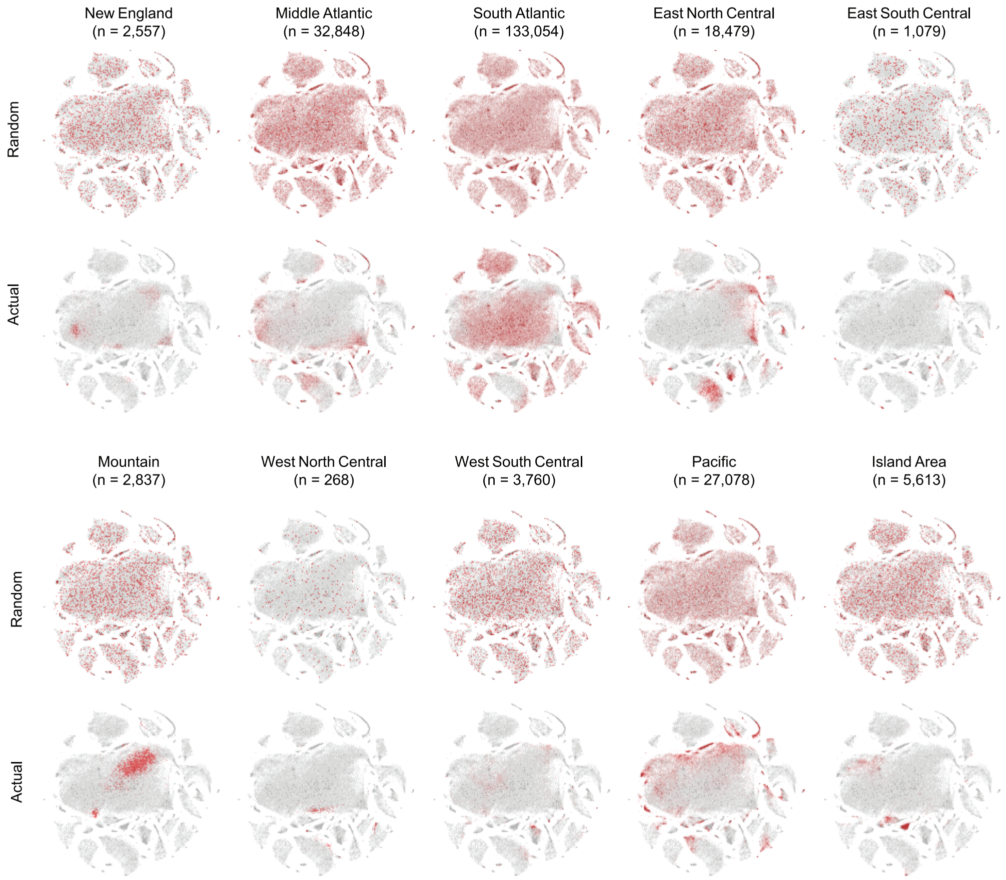


Figure 11. Distribution of provider prescribing patterns by census region. Providers with ≥1,000 claims (n=227,573) were divided into subsets by census region (lower figures within regional pairs). For comparison, a random sample of equivalent size was taken from the entire data set such that the providers in each random sub-sample did not overlap with any of the others (upper figures). This allows visual comparison of regional provider distributions with a random national sample of equivalent size.

but similarly-sized CBSAs: Rochester, NY (ROC) and Oklahoma City, OK (OKC). Although their populations are similar, they have different median household incomes and percent Medicare Part D enrollees (see Supplemental Figure 3): $43,955 and 14.1%, respectively for ROC, and $36,797 and 7.8% for OKC. The dashed lines represent 5-fold differences in claims-per-enrollee for specific drugs, with those outside the range annotated. The selected CBSAs are annotated in t-SNE density plots shown in Supplemental Figure 4A. For comparison, Figure 11C shows another pairwise visual comparison between two geographically proximate and similarly sized CBSAs: Dallas-Fort Worth, TX (DFW) and Houston, TX (IAH). If prescribing patterns reflect regional prescribing homophily or state specific Medicare Part D approved medication formularies, such pairs would be expected to have similar prescribing profiles and could be considered an internal control. In this example, the claims per enrollee are more similar between the two CBSAs. The median household incomes and percent enrolled are $47,418 and 6.6% for Dallas Fort Worth (DFW), and $44,714 and 6.3% for Houston (IAH).

Figures 12D-F show similar results based on profiles of 198 drug categories instead of drugs, which should reduce effects of regional differences in Medicare Part D formularies. These visualizations still show substantial differences in profiles of drug categories prescribed between CBSAs. Figure 12E compares the Boston, MA (BOS) and Miami, FL (MIA) CBSAs (also see t-SNE plots in Supplemental Figure 4B) and shows that there are 5 to 10-fold differences in the claims-per-enrollee within particular categories. Although similarly sized metropolitan areas, there are almost twice as many enrollees per provider in MIA than in BOS (see Supplemental Figure 6 and Figure 2). As an example, "Amphetamines and Amphetamine-Like Stimulants" generate almost 6-fold more claims per 1,000 Boston enrollees as

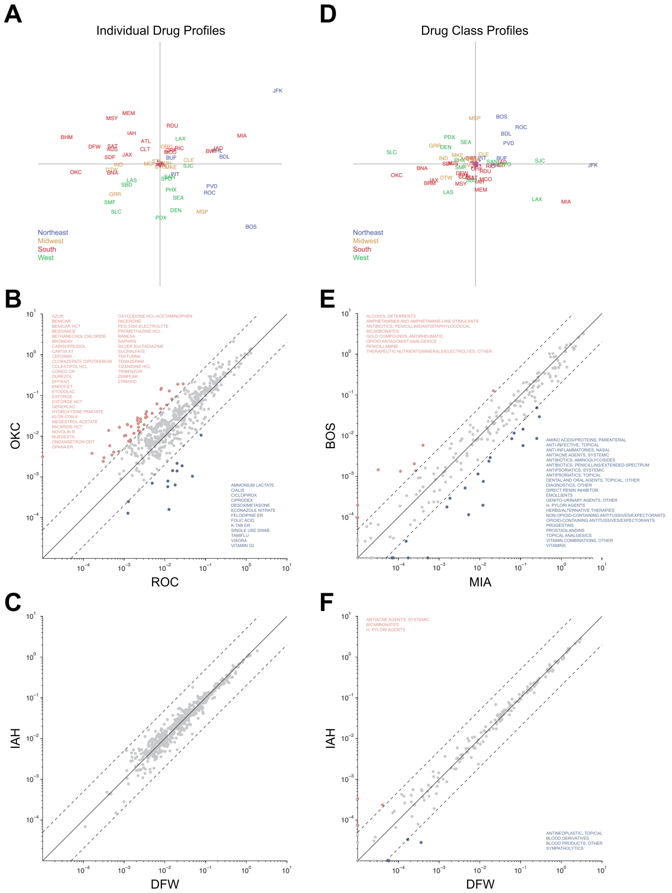


Figure 12. **Variation of prescribing pattern by core-based statistical areas.** A. Multidimensional scaling (MDS) of 52 CBSAs based on 532 drugs that have over 100,000 claims (across 50 states and Washington DC). Data were expressed as number of claims for a particular drug in a particular CBSA per number of enrollees in that CBSA. CBSAs are specified by IATA airport code. Magenta dots indicate MDS performed on a randomly permuted data sets where the data corresponding to the CBSA providers were shuffled, preserving the number of providers for each CBSA. B. Comparison of two CBSAs of similar sizes: Oklahoma City OK vs. Rochester NY. Dots represent individual drugs and axes are the number of claims per enrollee in log scale (for the respective CBSAs). Dashed lines indicate 5-fold differences in the per-enrollee numbers of claims. Drugs beyond these regions are indicated. C. Comparison of Houston TX and Dallas-Fort Worth Texas CBSAs that might be expected to have similar profiles as an internal control. D. MDS plot of 52 CBSAs based on 198 drug categories, similar to part A. E. Comparison of prescribing patterns in Boston MA and Miami FL based on drug categories. F. Houston TX vs. Dallas-Fort Worth TX based on drug categories.

compared to claims per 1,000 Miami enrollees (126.4 vs. 21.7). In contrast, “Genito-Urinary Agents, Other” generate almost 10-fold more claims per 1,000 enrollees in MIA as compared to BOS (28.9 vs. 2.9). Figure 12F shows that the Dallas-Fort Worth vs. Houston profiles are substantially more similar, with the largest differences for rarely prescribed drug categories.

DISCUSSION

Our results show that an approach combining unsupervised clustering and t-SNE dimensional reduction can be used to identify prescribing patterns in large administrative data sets. We identified a skewed distribution in provider claims volume and drug prescribing diversity, with most participating providers making relatively few claims of a small number of drug types. Previously, a number of focused studies have examined prescription diversity, mostly with respect to opioid analgesics [52–58], antibiotics [1, 59–63], psychiatric medications [64–67], and among general practitioners [31, 68–72]. One web site has made the Medicare Part D prescribing data searchable with varios filters for provider, charges, and medications [73–75]. As far as we are aware, however, this is the first high level, aggregate analysis of provider prescribing diversity and patterns on a national scale (40 million patients and over 800,000 providers) across multiple specialties, medication classes and practitioner types. This type of analysis may be used as a starting point for future work comparing national prescribing patterns, especially in countries where regional formulary composition is centrally tracked.

This work also suggests that provider prescribing volume and diversity patterns are a powerful proxy for how practitioners actually provide care, as opposed to self-identified medical specialty. This approach also enhances generation and testing of hypotheses about the root causes of such variation. For example, correlating these findings with outcomes data may enhance comparative effectiveness studies of prescribing patterns for specific diagnoses (e.g. effect of anti-hypertensive regimens with and without diuretics on blood pressure control and mortality) [76]. Similar approaches have recently been used to conduct ”virtual clinical trials”, replicating the results of randomized prospective clinical trials [77, 78]. As our results demonstrate, these methods can also be used to identify prescribing behaviors of interest (e.g. opioid prescribing) in geographically comprehensive data sets. In the future, studies coupling prescribing patterns, patient outcomes, and genomic data may aid in identification of genotype-phenotype associations and facilitate precision targeting of effective therapies to specific individual genotypes [79].

This work also highlights prescribing variation in groups of metropolitan providers with similar Medicare claims patterns. Our findings complement reports showing considerable geographic variation in both claims volume [80] and cost [4] across the United States. Potential contributing factors to such variation [29, 81–83], include suboptimal care or health services delivery inefficiencies [84, 85], and regional differences in prescriptions for branded drugs compared to generic counterparts [86–89]. Our analysis of metropolitan areas, adjusted to reduce the effect of population density, also reveals considerable residual variation in prescribing patterns, with up to ten-fold variations for both individual drugs and drug classes. Further work, incorporating more detailed data (e.g. regional Medicare formularies, provider-health system associations), are needed to determine the factors associated with such variation.

Several caveats apply to this analysis. First, we recognize that most Medicare providers have a patient population with a mix of prescription plans, and our results may not be applicable beyond the Medicare population demographics [50]. For example, only 15.5% of Medicare Part D enrollees were $\leq$65 years of age. Thus, the prescribing profiles and provider cluster memberships described here cannot be generalized to younger individuals. Approximately 50% of individuals enrolled in Medicare part D also have private or supplemental insurance for medication coverage, and thus the prescription claims captured by Medicare Part D may differ from the overall claims. This bias is somewhat mitigated by our selection of 227,000 providers with $\geq$1000 claims. In addition, there is currently no available data set integrating the medication formularies of all the Medicare plans. Thus, we are unable to judge to what extent prescribing variation is dependent on Medicare Part D plan formulary differences. Future work might explore these issues with more comprehensive US data sets, or data sets from countries with national healthcare systems where formulary information is available.

In conclusion, we have presented a focused approach for analyzing prescribing variation in a national administrative data set. The analysis highlighted regional variations in prescribing practices in the United States Medicare Part D program. The use of the t-SNE visualization algorithm enhances the analysis and visualization of variation in Medicare Part D provider prescribing patterns.

AVAILABILITY OF DATA AND MATERIAL

The datasets analyzed during the current study are all publicly available, and URLs for their download are listed in the Methods Section and references.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

This work was supported in part by research grants from National Institute of Health: National Center for Advancing Translation Science UL1 TR000042 and U54 TR001625 (MZ, CQ, RW, and SW), KL2

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3470v1 | CC BY 4.0 Open Access | rec: 18 Dec 2017, publ: 18 Dec 2017

TR001999 (CQ), TL1 TR002000 (SW, KB), R01AI098112 and R01AI069351 (MZ), from the Philip Templeton Foundation (MZ, RW, SF), and from the University of Rochester Center for Health Informatics (MZ, RW, AR, CF, SF, SW).

AUTHOR'S CONTRIBUTIONS

The project was designed and overseen by MZ and AR. Data wrangling and domain expertise were provided by CF, MT, LN, AR, MZ and SF. Statistical and machine learning analyses were performed by AR, CF, MZ, SF. Figures were produced by AR, MZ and RW. The manuscript was written by AR, MZ, RW, SW, SF, KB and LN.

ACKNOWLEDGMENTS

The authors would like to thank Rick McGuire, Orna Intrator and David Mitten for discussions related to this work.

REFERENCES

1. Zhang Y, Steinman MA, Kaplan CM. Geographic variation in outpatient antibiotic prescribing among older adults [Journal Article]. Arch Intern Med. 2012;172(19):1465–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23007171.
2. Zhang Y, Baicker K, Newhouse JP. Geographic variation in Medicare drug spending [Journal Article]. N Engl J Med. 2010;363(5):405–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20538621.
3. Stuart B, Shoemaker JS, Dai M, Davidoff AJ. Regions With Higher Medicare Part D Spending Show Better Drug Adherence, But Not Lower Medicare Costs For Two Diseases. Health Affairs. 2013;32(1):120–126.
4. Donohue JM, Morden NE, Gellad WF, Bynum JP, Zhou W, Hanlon JT, et al. Sources of regional variation in Medicare Part D drug spending. New England Journal of Medicine. 2012;366(6):530–538.
5. Chen JH, Humphreys K, Shah NH, Lembke A. Distribution of opioids by different types of medicare prescribers. JAMA internal medicine. 2016;176(2):259–261.
6. Center for Medicare Medicaid Services. Part D Prescriber Data CY 2013; 2016.
7. Tufte ER. Visual explanations: images and quantities, evidence and narrative. vol. 36. Graphics Press Cheshire, CT; 1997.
8. Lavrač N, Bohanec M, Pur A, Cestnik B, Debeljak M, Kobler A. Data mining and visualization for decision support and modeling of public health-care resources. Journal of Biomedical Informatics. 2007;40(4):438–447.
9. Zand MS, Wang J, Hilchey S. Graphical representation of proximity measures for multidimensional data: Classical and metric multidimensional scaling [Journal Article]. Mathematica Journal. 2015;17(7):1–31.
10. Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of educational psychology. 1933;24(6):417.
11. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. science. 2000;290(5500):2323–2326.
12. Sammon JW. A nonlinear mapping for data structure analysis. IEEE Transactions on computers. 1969;100(5):401–409.
13. Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. science. 2000;290(5500):2319–2323.
14. Maaten Lvd, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008;9(Nov):2579–2605.

15. Owen RR, Feng W, Thrush CR, Hudson TJ, Austen MA. Variations in prescribing practices for novel antipsychotic medications among Veterans Affairs hospitals [Journal Article]. Psychiatric Services. 2001;.

16. Baxter C, Jones R, Corr L. Time trend analysis and variations in prescribing lipid lowering drugs in general practice [Journal Article]. BMJ. 1998;317(7166):1134–1135.

17. Heins JK, Heins A, Grammas M, Costello M, Huang K, Mishra S. Disparities in analgesia and opioid prescribing practices for patients with musculoskeletal pain in the emergency department [Journal Article]. Journal of Emergency Nursing. 2006;32(3):219–224.

18. Ashworth M, Charlton J, Ballard K, Latinovic R, Gulliford M. Variations in antibiotic prescribing and consultation rates for acute respiratory infection in UK general practices 1995–2000 [Journal Article]. Br J Gen Pract. 2005;55(517):603–608.

19. Birkmeyer JD, Reames BN, McCulloch P, Carr AJ, Campbell WB, Wennberg JE. Understanding of regional variation in the use of surgery [Journal Article]. The Lancet. 2013;382(9898):1121–1129.

20. Goldberg T, Kroehl ME, Suddarth KH, Trinkley KE. Variations in Metformin Prescribing for Type 2 Diabetes [Journal Article]. The Journal of the American Board of Family Medicine. 2015;28(6):777–784.

21. Reames BN, Shubeck SP, Birkmeyer JD. Strategies for Reducing Regional Variation in the Use of Surgery A Systematic Review [Journal Article]. Annals of surgery. 2014;259(4):616.

22. Porter MP, Kerrigan MC, Donato BMK, Ramsey SD. Patterns of use of systemic chemotherapy for Medicare beneficiaries with urothelial bladder cancer. Urologic oncology. 2011 2011 May-Jun;29:252–8.

23. Fong RK, Johnson A, Gill SS. Cholinesterase inhibitors: an example of geographic variation in prescribing patterns within a drug class [Journal Article]. Int J Geriatr Psychiatry. 2015;30(2):220–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25639835.

24. Golberstein E, Rhee TG, McGuire TG. Geographic Variations in Use of Medicaid Mental Health Services [Journal Article]. Psychiatric Services. 2015;.

25. Ohlsson H, Vervloet M, van Dijk L. Practice variation in a longitudinal perspective: a multilevel analysis of the prescription of simvastatin in general practices between 2003 and 2009 [Journal Article]. Eur J Clin Pharmacol. 2011;67(12):1205–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21698375.

26. Brookes-Howell L, Hood K, Cooper L, Little P, Verheij T, Coenen S, et al. Understanding variation in primary medical care: a nine-country qualitative study of clinicians' accounts of the non-clinical factors that shape antibiotic prescribing decisions for lower respiratory tract infection [Journal Article]. BMJ Open. 2012;2(4). Available from: https://www.ncbi.nlm.nih.gov/pubmed/22918670.

27. Omar RZ, O'Sullivan C, Petersen I, Islam A, Majeed A. A model based on age, sex, and morbidity to explain variation in UK general practice prescribing: cohort study [Journal Article]. BMJ. 2008;337:a238. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18625598.

28. Davis MM, Patel MS, Halasyamani LK. Variation in estimated Medicare prescription drug plan costs and affordability for beneficiaries living in different states [Journal Article]. J Gen Intern Med. 2007;22(2):257–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17356996.

29. Forster DP, Frost CE. Use of regression analysis to explain the variation in prescribing rates and costs between family practitioner committees [Journal Article]. Br J Gen Pract. 1991;41(343):67–71. Available from: https://www.ncbi.nlm.nih.gov/pubmed/2031739.

30. Fretheim A, Oxman AD. International variation in prescribing antihypertensive drugs: its extent and possible explanations [Journal Article]. BMC Health Serv Res. 2005;5(1):21. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15762983.

31. Sorensen HT, Steffensen FH, Nielsen GL, Gron P. Variation in antibiotic prescribing costs in Danish general practice: an epidemiological pharmaco-economic analysis [Journal Article]. Int J Risk

Saf Med. 1996;8(3):243–50. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23511984.

32. Cutler D, Skinner J, Stern AD, Wennberg D. Physician beliefs and patient preferences: a new look at regional variation in health care spending. National Bureau of Economic Research; 2013.

33. Rothberg MB, Bonner AB, Rajab MH, Kim HS, Stechenberg BW, Rose DN. Effects of local variation, specialty, and beliefs on antiviral prescribing for influenza [Journal Article]. Clin Infect Dis. 2006;42(1):95–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16323098.

34. Munson J, Morden N, Goodman D, Valle L, Wennberg J. The Dartmouth atlas of Medicare prescription drug use. Lebanon, NH: The Dartmouth Institute for Health Policy and Clinical Practice; 2013.

35. Christakis NA, Fowler JH. Commentary—Contagion in Prescribing Behavior Among Networks of Doctors. Marketing Science. 2011;30(2):213–216.

36. Epstein AM. Geographic variation in Medicare spending [Journal Article]. N Engl J Med. 2010;363(1):85–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20463335.

37. Newhouse JP, Garber AM. Geographic variation in health care spending in the United States: insights from an Institute of Medicine report [Journal Article]. JAMA. 2013;310(12):1227–1228.

38. Zhang Y, Baicker K, Newhouse JP. Geographic variation in the quality of prescribing [Journal Article]. N Engl J Med. 2010;363(21):1985–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21047217.

39. Curtis LH, Østbye T, Sendersky V, Hutchison S, Dans PE, Wright A, et al. Inappropriate prescribing for elderly Americans in a large outpatient population. Archives of internal medicine. 2004;164(15):1621–1625.

40. Center for Medicare Services. Physician shared patient patterns technical requirements;. https://downloads.cms.gov/foia/physician_shared_patient_patterns_technical_requirements.pdf.

41. Center for Medicare Services. CMS 2013 Medicare Part D Statistical Supplement [webpage]; 2016. Available from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareMedicaidStatSupp/Downloads/2013PartD.zip [cited May 20, 2016].

42. United States Department of Agriculture Economic Research Service. Rural-Urban Continuum Codes [webpage]; 2016. Available from: https://www.ers.usda.gov/webdocs/DataFiles/RuralUrban_Continuum_Codes__18011/ruralurbancodes2013.xls?v=41404 [cited May 20, 2016].

43. National Bureau of Economic Research. SSA to FIPS CBSA and MSA County Crosswalk Files [webpage]; 2016. Available from: http://www.nber.org/data/cbsa-msa-fips-ssa-county-crosswalk.html [cited May 20, 2016].

44. Office of Policy Development and Research: U S Department of Housing and Urban Development. HUD USPS Zip Code Crosswalk Files [webpage]; 2016. Available from: https://www.huduser.gov/portal/datasets/usps_crosswalk.html [cited May 20, 2016].

45. Kaiser Family Foundation. Total Number of Medicare Beneficiaries Data File; 2013. Available from: http://kff.org/medicare/state-indicator/total-medicare-beneficiaries/?currentTimeframe=2&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D [cited May 4, 2017].

46. Kaiser Family Foundation. Total Number of Medicare Beneficiaries Data File; 2016. https://catalog.data.gov/dataset/va-national-drug-file-may-2015.

47. Office of the Inspector General, Department of Health and Human Services. List of Excluded Individuals/Entities (LEIE) ; 2017. https://oig.hhs.gov/exclusions/exclusions_list.asp.

48. Ward Jr JH. Hierarchical grouping to optimize an objective function. Journal of the American statistical association. 1963;58(301):236–244.

49. Van Der Maaten L. Accelerating t-SNE using tree-based algorithms. Journal of machine learning research. 2014;15(1):3221–3245.

50. Barnett JC, Vornovitsky M. Health Insurance Coverage in the United States: 2015. United States Census Bureau; 2016. P60-257. Available from: https://www.census.gov/content/dam/Census/library/publications/2016/demo/p60-257.pdf.

51. U S Department of Justice. Seventy-three charged in Southern District of Florida as part of largest national Medicare fraud takedown in history [webpage]; 2015. Available from: https://www.justice.gov/archives/opa/page/file/479006/download [cited July 2, 2018].

52. McDonald DC, Carlson K, Izrael D. Geographic variation in opioid prescribing in the U.S [Journal Article]. J Pain. 2012;13(10):988–96. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23031398.

53. McDonald DC, Carlson KE. The ecology of prescription opioid abuse in the USA: geographic variation in patients' use of multiple prescribers ("doctor shopping") [Journal Article]. Pharmacoepidemiol Drug Saf. 2014;23(12):1258–67. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25111716.

54. Curtis LH, Stoddard J, Radeva JI, Hutchison S, Dans PE, Wright A, et al. Geographic variation in the prescription of schedule II opioid analgesics among outpatients in the United States [Journal Article]. Health Serv Res. 2006;41(3 Pt 1):837–55. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16704515.

55. Paulozzi LJ, Mack KA, Hockenberry JM, Division of Unintentional Injury Prevention NCfIP, Control CDC. Vital signs: variation among States in prescribing of opioid pain relievers and benzodiazepines - United States, 2012 [Journal Article]. MMWR Morb Mortal Wkly Rep. 2014;63(26):563–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24990489.

56. Paulozzi LJ, Mack KA, Hockenberry JM. Variation among states in prescribing of opioid pain relievers and benzodiazepines–United States, 2012 [Journal Article]. J Safety Res. 2014;51:125–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25453186.

57. Tang Y, Chang CC, Lave JR, Gellad WF, Huskamp HA, Donohue JM. Patient, Physician and Organizational Influences on Variation in Antipsychotic Prescribing Behavior [Journal Article]. J Ment Health Policy Econ. 2016;19(1):45–59. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27084793.

58. Schirle L, McCabe BE. State variation in opioid and benzodiazepine prescriptions between independent and nonindependent advanced practice registered nurse prescribing states [Journal Article]. Nursing outlook. 2016;64(1):86–93.

59. Brookes-Howell L, Hood K, Cooper L, Coenen S, Little P, Verheij T, et al. Clinical influences on antibiotic prescribing decisions for lower respiratory tract infection: a nine country qualitative study of variation in care [Journal Article]. BMJ Open. 2012;2(3). Available from: https://www.ncbi.nlm.nih.gov/pubmed/22619265.

60. Steinman MA, Yang KY, Byron SC, Maselli JH, Gonzales R. Variation in outpatient antibiotic prescribing in the United States [Journal Article]. Am J Manag Care. 2009;15(12):861–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20001167.

61. Cordoba G, Siersma V, Lopez-Valcarcel B, Bjerrum L, Llor C, Aabenhus R, et al. Prescribing style and variation in antibiotic prescriptions for sore throat: cross-sectional study across six countries [Journal Article]. BMC Fam Pract. 2015;16:7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25630870.

62. Fleming-Dutra KE, Hersh AL, Shapiro DJ, Bartoces M, Enns EA, File J T M, et al. Prevalence of Inappropriate Antibiotic Prescriptions Among US Ambulatory Care Visits, 2010-2011 [Journal Article]. JAMA. 2016;315(17):1864–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27139059.

63. Williamson DA, Roos R, Verrall A, Smith A, Thomas MG. Trends, demographics and disparities in outpatient antibiotic consumption in New Zealand: a national study [Journal Article]. J Antimicrob Chemother. 2016;71(12):3593–3598. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27601293.

64. Hansen DG, Sondergaard J, Vach W, Gram LF, Rosholm JU, Kragstrup J. Antidepressant drug use in general practice: inter-practice variation and association with practice characteristics [Journal Article]. Eur J Clin Pharmacol. 2003;59(2):143–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12721774.

65. Pharoah PD, Melzer D. Variation in prescribing of hypnotics, anxiolytics and antidepressants between 61 general practices [Journal Article]. Br J Gen Pract. 1995;45(400):595–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/8554839.

66. Lund BC, Abrams TE, Bernardy NC, Alexander B, Friedman MJ. Benzodiazepine prescribing variation and clinical uncertainty in treating posttraumatic stress disorder [Journal Article]. Psychiatr Serv. 2013;64(1):21–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23070039.

67. Mayne SL, Ross ME, Song L, McCarn B, Steffes J, Liu W, et al. Variations in Mental Health Diagnosis and Prescribing Across Pediatric Primary Care Practices [Journal Article]. Pediatrics. 2016;137(5). Available from: https://www.ncbi.nlm.nih.gov/pubmed/27244791.

68. Scrivener G, Lloyd DC. Allocating census data to general practice populations: implications for study of prescribing variation at practice level [Journal Article]. BMJ. 1995;311(6998):163–5. Available from: https://www.ncbi.nlm.nih.gov/pubmed/7613429.

69. Davis P, Gribben B. Rational prescribing and interpractitioner variation. A multilevel approach [Journal Article]. Int J Technol Assess Health Care. 1995;11(3):428–42. Available from: https://www.ncbi.nlm.nih.gov/pubmed/7591545.

70. Davis PB, Yee RL, Millar J. Accounting for medical variation: the case of prescribing activity in a New Zealand general practice sample [Journal Article]. Soc Sci Med. 1994;39(3):367–74. Available from: https://www.ncbi.nlm.nih.gov/pubmed/7939853.

71. Sinnige J, Braspenning JC, Schellevis FG, Hek K, Stirbu I, Westert GP, et al. Inter-practice variation in polypharmacy prevalence amongst older patients in primary care [Journal Article]. Pharmacoepidemiol Drug Saf. 2016;Available from: http://www.ncbi.nlm.nih.gov/pubmed/27133740.

72. Tomlin AM, Gillies TD, Tilyard MW, Dovey SM. Variation in the pharmaceutical costs of New Zealand general practices: a national database linkage study [Journal Article]. J Public Health (Oxf). 2016;38(1):138–46. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25599688.

73. ProPublica. Prescriber Checkup Data; 2016.

74. Ornstein C. Government Releases Massive Trove of Data on Doctors' Prescribing Patterns; 2015.

75. ProPublica. Prescriber Checkup; 2016.

76. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Use of Electronic Medical Records for Health Outcomes Research A Literature Review. Medical Care Research and Review. 2009 DEC;66(6):611–638.

77. Tannen RL, Weiner MG, Xie D. Replicated studies of two randomized trials of angiotensin-converting enzyme inhibitors: further empiric validation of the 'prior event rate ratio' to adjust for unmeasured confounding by indication. Pharmacology and Drug Safety. 2008 JUL;17(7):671–685.

78. Tannen RL, Weiner MG, Xie D, Barnhart K. A simulation using data from a primary care practice database closely replicated the women's health initiative trial. JOURNAL OF CLINICAL EPIDEMIOLOGY. 2007 JUL;60(7):686–695.

79. Caudle KE, Gammal RS, Whirl-Carrillo M, Hoffman JM, Relling MV, Klein TE. Evidence and resources to implement pharmacogenetic knowledge for precision medicine [Journal Article]. Am J Health Syst Pharm. 2016;73(23):1977–1985. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27864205.

80. The Dartmouth Institute. The Dartmouth Atlas of Medicare Prescription Drug Use;. `http://www.dartmouthatlas.org/downloads/reports/Prescription_Drug_Atlas_101513.pdf`.

81. Jaye C, Tilyard M. A qualitative comparative investigation of variation in general practitioners' prescribing patterns [Journal Article]. Br J Gen Pract. 2002;52(478):381–6. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/12014535`.

82. Skegg K, Skegg DC, McDonald BW. Is there seasonal variation in the prescribing of antidepressants in the community? [Journal Article]. J Epidemiol Community Health. 1986;40(4):285–8. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/3655619`.

83. Johnson RE, Azevedo DJ, Kieburtz KD. Variation in individual physicians' prescribing [Journal Article]. J Ambul Care Manage. 1986;9(1):25–37. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/10275117`.

84. Kahn MG, Banade D. The impact of electronic medical records data sources on an adverse drug event quality measure. Journal of the American Medical Informatics Association. 2010 MAR;17(2):185–191.

85. Parsons A, McCullough C, Wang J, Shih S. Validity of electronic health record-derived quality measurement for performance monitoring. JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION. 2012 JUL;19(4):604–609.

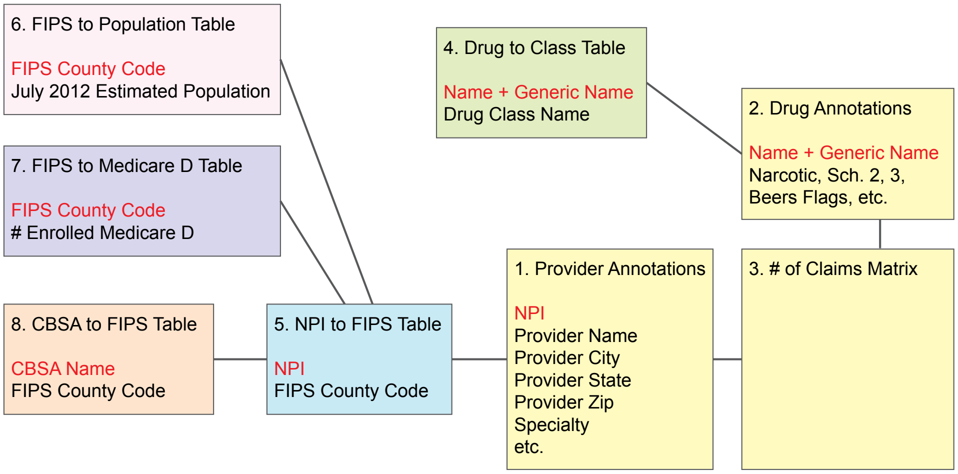
86. Newman-Casey PA, Woodward MA, Niziol LM, Lee PP, De Lott LB. Brand Medications and Medicare Part D: How Eye Care Providers' Prescribing Patterns Influence Costs [Journal Article]. Ophthalmology. 2017;Available from: `https://www.ncbi.nlm.nih.gov/pubmed/28625684`.

87. Kesselheim AS, Avorn J, Sarpatwari A. The High Cost of Prescription Drugs in the United States: Origins and Prospects for Reform [Journal Article]. JAMA. 2016;316(8):858–71. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/27552619`.

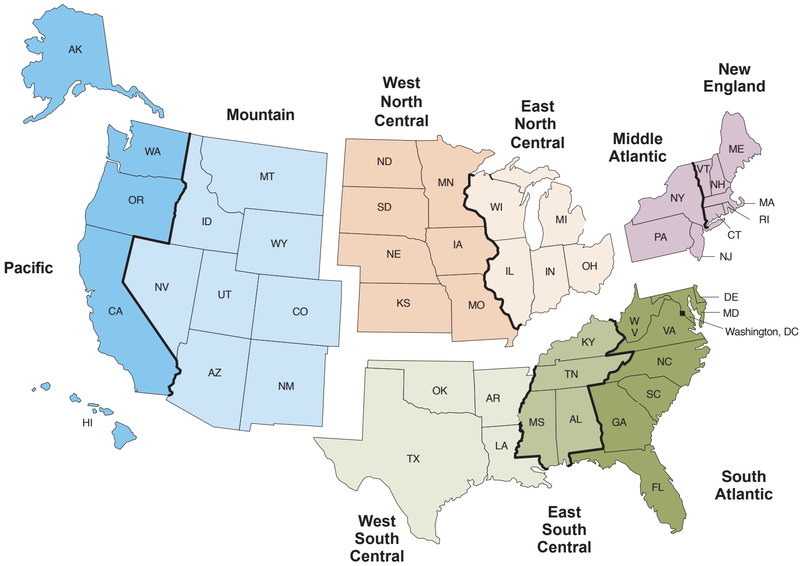
88. Manzoli L, Flacco ME, Boccia S, D'Andrea E, Panic N, Marzuillo C, et al. Generic versus brand-name drugs used in cardiovascular diseases [Journal Article]. Eur J Epidemiol. 2016;31(4):351–68. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/26620809`.

89. Corrao G, Soranna D, La Vecchia C, Catapano A, Agabiti-Rosei E, Gensini G, et al. Medication persistence and the use of generic and brand-name blood pressure-lowering agents [Journal Article]. J Hypertens. 2014;32(5):1146–53; discussion 1153. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/24569417`.

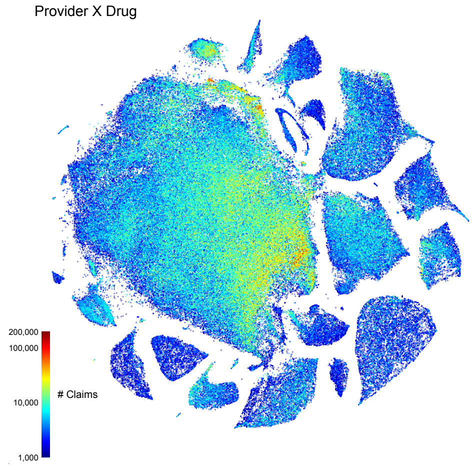
SUPPLEMENTARY FIGURES AND FILES



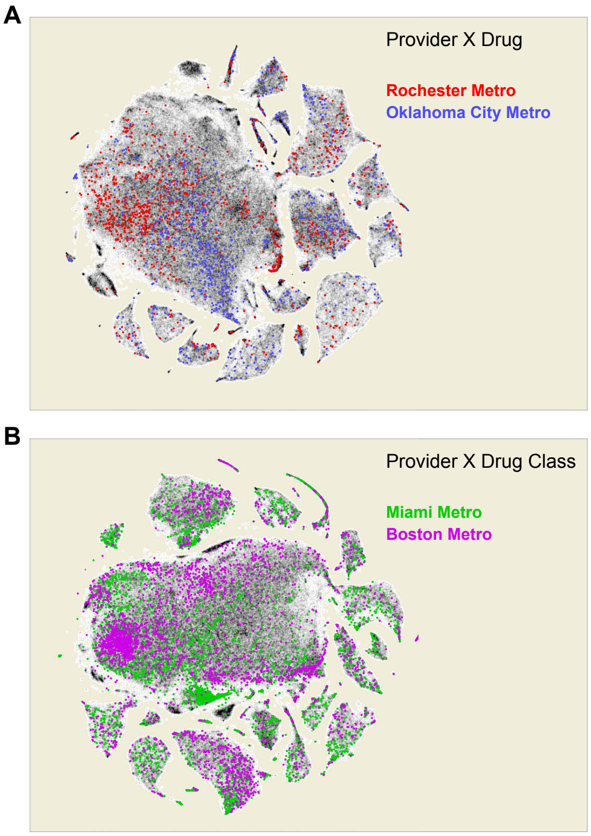
Supplementary Figure 1. Data sources used for this study. This schema depicts various sources of data and how they are related. Red font indicates a data column with unique values



Supplementary Figure 2. United States Census Regions. Map of United States Census Regions used for geographic data comparisons. Adapted from the United States Census Bureau.



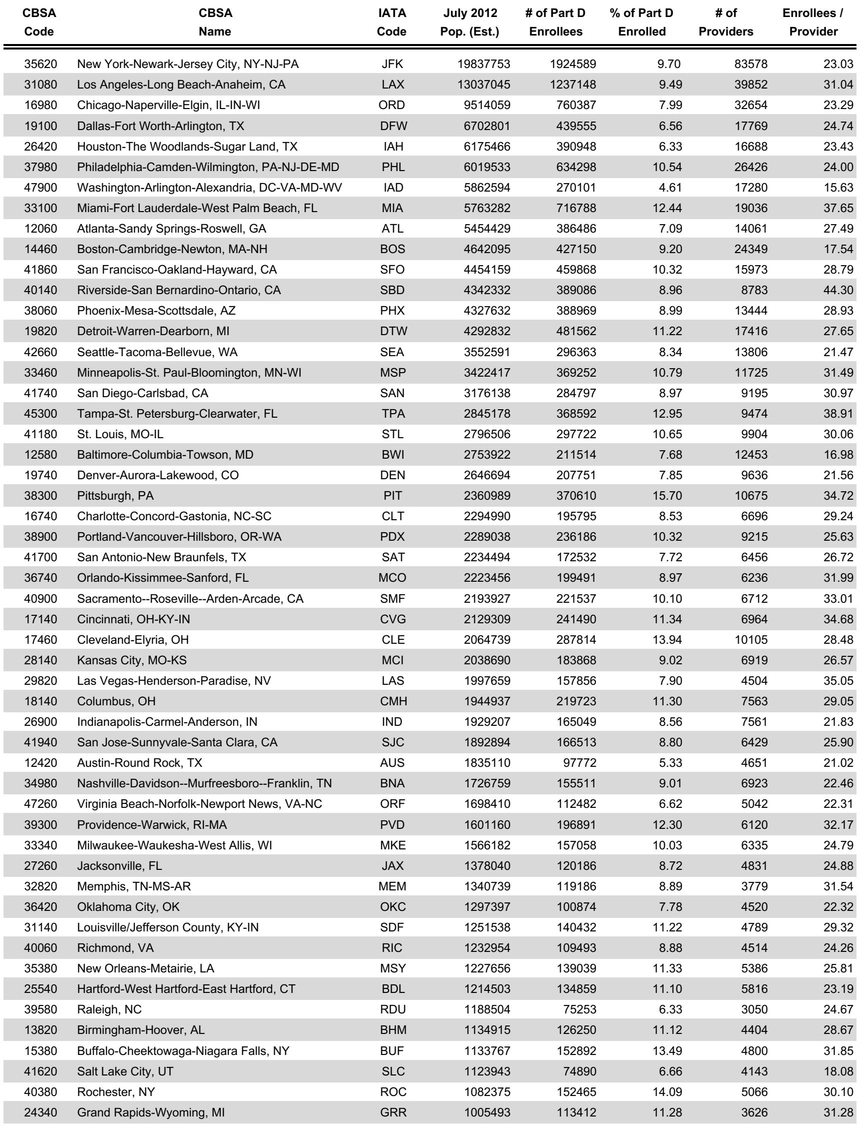
Supplementary Figure 3. t-SNE plot showing distribution of claim volume per provider. This t-SNE plot is based on the provider by drug matrix, as shown in Figure 4A. Color corresponds to the Log_{10} of claims per provider (each represented by a dot).



Supplementary Figure 4. t-SNE plots with particular CBSAs highlighted. A. t-SNE plot based on provider by drug matrix (as in Figure 4A) with providers in Rochester and Oklahoma City annotated (see Figure 11B). B. t-SNE plot based on drug class by provider matrix (as in 4B) with providers in Miami and Boston annotated (see Figure 11E).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3470v1 | CC BY 4.0 Open Access | rec: 18 Dec 2017, publ: 18 Dec 2017

Supplementary Figure 5. Hierarchical clustering. Plots of the 605 clusters identified by hierarchical clustering with linkage using Ward's minimization criteria. The background is the full tSNE projection, while each cluster is in red. This 19 page figure is available for download from *https://figshare.com/s/33aed8901d3185f92c43*



CBSA Code	CBSA Name	IATA Code	July 2012 Pop. (Est.)	# of Part D Enrollees	% of Part D Enrolled	# of Providers	Enrollees / Provider
35620	New York-Newark-Jersey City, NY-NJ-PA	JFK	19837753	1924589	9.70	83578	23.03
31080	Los Angeles-Long Beach-Anaheim, CA	LAX	13037045	1237148	9.49	39852	31.04
16980	Chicago-Naperville-Elgin, IL-IN-WI	ORD	9514059	760387	7.99	32654	23.29
19100	Dallas-Fort Worth-Arlington, TX	DFW	6702801	439555	6.56	17769	24.74
26420	Houston-The Woodlands-Sugar Land, TX	IAH	6175466	390948	6.33	16688	23.43
37980	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	PHL	6019533	634298	10.54	26426	24.00
47900	Washington-Arlington-Alexandria, DC-VA-MD-WV	IAD	5862594	270101	4.61	17280	15.63
33100	Miami-Fort Lauderdale-West Palm Beach, FL	MIA	5763282	716788	12.44	19036	37.65
12060	Atlanta-Sandy Springs-Roswell, GA	ATL	5454429	386486	7.09	14061	27.49
14460	Boston-Cambridge-Newton, MA-NH	BOS	4642095	427150	9.20	24349	17.54
41860	San Francisco-Oakland-Hayward, CA	SFO	4454159	459868	10.32	15973	28.79
40140	Riverside-San Bernardino-Ontario, CA	SBD	4342332	389086	8.96	8783	44.30
38060	Phoenix-Mesa-Scottsdale, AZ	PHX	4327632	388969	8.99	13444	28.93
19820	Detroit-Warren-Dearborn, MI	DTW	4292832	481562	11.22	17416	27.65
42660	Seattle-Tacoma-Bellevue, WA	SEA	3552591	296363	8.34	13806	21.47
33460	Minneapolis-St. Paul-Bloomington, MN-WI	MSP	3422417	369252	10.79	11725	31.49
41740	San Diego-Carlsbad, CA	SAN	3176138	284797	8.97	9195	30.97
45300	Tampa-St. Petersburg-Clearwater, FL	TPA	2845178	368592	12.95	9474	38.91
41180	St. Louis, MO-IL	STL	2796506	297722	10.65	9904	30.06
12580	Baltimore-Columbia-Towson, MD	BWI	2753922	211514	7.68	12453	16.98
19740	Denver-Aurora-Lakewood, CO	DEN	2646694	207751	7.85	9636	21.56
38300	Pittsburgh, PA	PIT	2360989	370610	15.70	10675	34.72
16740	Charlotte-Concord-Gastonia, NC-SC	CLT	2294990	195795	8.53	6696	29.24
38900	Portland-Vancouver-Hillsboro, OR-WA	PDX	2289038	236186	10.32	9215	25.63
41700	San Antonio-New Braunfels, TX	SAT	2234494	172532	7.72	6456	26.72
36740	Orlando-Kissimmee-Sanford, FL	MCO	2223456	199491	8.97	6236	31.99
40900	Sacramento--Roseville--Arden-Arcade, CA	SMF	2193927	221537	10.10	6712	33.01
17140	Cincinnati, OH-KY-IN	CVG	2129309	241490	11.34	6964	34.68
17460	Cleveland-Elyria, OH	CLE	2064739	287814	13.94	10105	28.48
28140	Kansas City, MO-KS	MCI	2038690	183868	9.02	6919	26.57
29820	Las Vegas-Henderson-Paradise, NV	LAS	1997659	157856	7.90	4504	35.05
18140	Columbus, OH	CMH	1944937	219723	11.30	7563	29.05
26900	Indianapolis-Carmel-Anderson, IN	IND	1929207	165049	8.56	7561	21.83
41940	San Jose-Sunnyvale-Santa Clara, CA	SJC	1892894	166513	8.80	6429	25.90
12420	Austin-Round Rock, TX	AUS	1835110	97772	5.33	4651	21.02
34980	Nashville-Davidson--Murfreesboro--Franklin, TN	BNA	1726759	155511	9.01	6923	22.46
47260	Virginia Beach-Norfolk-Newport News, VA-NC	ORF	1698410	112482	6.62	5042	22.31
39300	Providence-Warwick, RI-MA	PVD	1601160	196891	12.30	6120	32.17
33340	Milwaukee-Waukesha-West Allis, WI	MKE	1566182	157058	10.03	6335	24.79
27260	Jacksonville, FL	JAX	1378040	120186	8.72	4831	24.88
32820	Memphis, TN-MS-AR	MEM	1340739	119186	8.89	3779	31.54
36420	Oklahoma City, OK	OKC	1297397	100874	7.78	4520	22.32
31140	Louisville/Jefferson County, KY-IN	SDF	1251538	140432	11.22	4789	29.32
40060	Richmond, VA	RIC	1232954	109493	8.88	4514	24.26
35380	New Orleans-Metairie, LA	MSY	1227656	139039	11.33	5386	25.81
25540	Hartford-West Hartford-East Hartford, CT	BDL	1214503	134859	11.10	5816	23.19
39580	Raleigh, NC	RDU	1188504	75253	6.33	3050	24.67
13820	Birmingham-Hoover, AL	BHM	1134915	126250	11.12	4404	28.67
15380	Buffalo-Cheektowaga-Niagara Falls, NY	BUF	1133767	152892	13.49	4800	31.85
41620	Salt Lake City, UT	SLC	1123943	74890	6.66	4143	18.08
40380	Rochester, NY	ROC	1082375	152465	14.09	5066	30.10
24340	Grand Rapids-Wyoming, MI	GRR	1005493	113412	11.28	3626	31.28

Supplementary Figure 6. Characteristics of core-based statistical areas (CBSA). 52 CBSAs are listed that have July 2012 population estimates greater than 1,000,000 residents. See Methods for data sources.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3470v1 | CC BY 4.0 Open Access | rec: 18 Dec 2017, publ: 18 Dec 2017