Paper ML06

A Real-World View of Cataract

Walter Cedeño, Janssen Pharmaceuticals, Spring House, PA, USA

Jeffrey J. Headd, Janssen Pharmaceuticals, Spring House, PA, USA

Jun Morimura, Janssen Pharmaceuticals, Titusville, NJ, USA

ABSTRACT

Cataract, an eye disease characterized by the clouding of the lens, currently affects over 25 million people in US. This number is projected to double by the year 2050 according to the NIH Eye Institute. To further elucidate drivers of cataract progression, we present an analysis of Real-World Data (RWD), which has gained wide acceptance in the pharmaceutical sector as a tool to understand disease progression, patient journey, unmet patient needs, and other areas. In this work, we apply methods in unsupervised machine learning, including K-Means, Mini-Batch K-Means, Agglomerative Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Interactive Clustering, and K-Modes techniques to Electronic Health Records (EHR) to explore the nature of variation within cataract patients. We identify several distinct segments within our cataract patient cohort and identify important features through detailed cluster analysis that drive the observed segmentation of patients.

Introduction

Cataract is a leading cause of vision loss and is responsible for 51% of world blindness[1]. Cataract, an eye disease characterized by the clouding of the lens in the eye, currently affects over 25 million people in US. By age 75, 50% of Caucasians have cataract. The percentage goes up to 70% by age 80 compared to 53% African Americans and 61% Hispanic Americans[2]. The number of cataract patients is projected to double by the year 2050 according to the NIH Eye Institute[2]. Most cases of cataracts are related to aging. The current standard treatment is surgical replacement of the lens which is very successful in restoring sight. Smoking, UV light exposure, diabetes, drinking, steroids, and family history puts you at a higher risk for cataracts[3]. Additional research includes the comparison of intraocular lenses used following cataract surgery using data collected retrospectively from ophthalmology clinics[4]. In another study, real-world data (RWD)

was used to assess the prevalence of astigmatism after cataract surgery[5].

The initial work presented here leverages Optum® de-identified Electronic Health Record (EHR) dataset, a multi-dimensional database containing information on outpatient visits, diagnostic procedures, medications, laboratory results, hospitalizations, clinical notes and patient outcomes. The EHR database includes representation of 80M patients with at least 7M patients from each US Census region. We query the real-world data to capture three races, Caucasian, African American, and Asian, as well as Hispanic ethnicity. Optum EHR contains data from more than 700 Hospitals and 7000 Clinics; treating more than 103 million patients receiving care in the United States. Using real world data for eye condition studies and related research has previously been successfully applied to develop machine learning models. For example, metrics associated with eye health and morphology, including retinal nerve fiber layer thickness and visual field, have been used to develop a Glaucoma classifier[6,7].

We apply clustering methods to a dataset of cataract patient journeys identified in the Optum EHR data in order to find patterns associated with cataract patients and identify important variables that could be used to segment patients into distinct groups. Prior research using real-world data has shown a link between UV light exposure and cortical cataract[8]. We hypothesize that events in a patient's medical history may also be linked to the onset and progression of cataract. Patient segmentation is the first step towards uncovering these potential relationships in real-world data. Segmentation of cataract patient journeys into distinct groups may reveal variation in the observed patient population, which in turn may inspire new directions of research and the development of targeted diagnostics, interventions, and treatments for segments of the cataract patient population.

Clustering is one of the fundamental unsupervised learning techniques in machine learning. As such it does not require any training to be applied or labels in the data to distinguish between different categories a priori. In its most basic form, it works by using a "similarity" metric to gauge distance between data points to group similar points into the same cluster and separate clusters from each other. Some techniques require the number of clusters or other related parameters to be provided a priori to determine membership in a cluster that optimizes the intra and inter cluster distance between data points. A brief description of the clustering techniques used in this work is presented in the Materials and Methods section below.

Materials and Methods

Procedure

We used Real World Data from Optum® de-identified Electronic Health Record dataset, containing patient records from 2007 to 2019. The following data tables were used from Optum:

1. Patient – patient demographic information
2. Diagnosis – patient diagnosis information
3. Labs – patient laboratory measurements
4. Observations – observation notes for patients
5. Medication_administrations – Rx administered to patients in a hospital setting
6. Prescriptions_written – Rx prescribed to patients

We identified a cohort of cataract patients with a recorded birth year in the Optum database using the following ICD codes:

I. ICD9
   a. 366.0* - Infantile juvenile and presenile cataract
   b. 366.1* - Senile cataract
   c. 366.3* - Cataract secondary to ocular disorders
   d. 366.8* - Other cataract
   e. 366.9* - Unspecified cataract

II. ICD10
   a. H25* - Age-related cataract
   b. H26.0* - Infantile and juvenile cataract
   c. H26.2* - Complicated cataract
   d. H26.8* - Other specified cataract
   e. H26.9* - Unspecified cataract

For our study, only patients with birth year recorded and age between 30 and 79 at the time of their first diagnosis of cataract are considered. These ranges were selected to account for about 85% of the adult cataract patients. Additional mapping between ICD9 and ICD10 codes was done to identify a common name. Drug class names were obtained from the Medi-Span[9] table using GPI codes in the data. The dates associated with diagnosis, labs, observations, medications, or prescriptions events were collected from the available information. Initial tests that included lab and observation events without extracting true measurements exhibited ambiguous results, therefore we only evaluate the use of diagnosis and prescriptions occurrences for this study. We also examined four different cohorts based on race (Caucasian, African American, and Asian which are not Hispanics) and ethnicity (Hispanic).

Cataract Cohorts

All available patients diagnosed with cataract who met our inclusion/exclusion criteria were considered for inclusion in one of the four cohorts; Caucasian, African American, Asian, and Hispanic. A summary of the cohort sizes is shown below:

- All Cataract Patients: 2,270,943
- Caucasian (not Hispanic): 1,719,067 total, limited to 100,000 of patients
- African American (not Hispanic): 189,738 total, limited to 100,000 of patients
- Asian (not Hispanic): 44,191
- Hispanic (any race): 35,577

All Asian and Hispanic patients with cataract were included in their respective cohort and considered for the study. Caucasian and African American cohorts were limited to 100,000 to reduce execution time given the large number of cataract patients identified. The inclusion of patients in these cohorts was assigned at random prior to feature selection and data preparation.

Feature selection and dataset preparation

The initial set of features in the data are those patient events for diagnosis (prefix flag_dg_) and prescriptions (prefix flag_rx_) as described above. For each event included, a Boolean value indicating the occurrence of the event (shown in dataset with prefix followed by the event name and value of 1) was created. Demographic information including age, gender, race (Hispanic cohort only), USA census location division and region, education level of region, and average income of region was added as well for all cohorts. The number of features in each cohort is driven by the diversity of medical events observed for the included patients. The initial dataset had the following number of rows and features:

- Caucasian cohort: 100,000 x 2,974
- African American cohort: 100,000 x 2,959
- Asian cohort: 44,191 x 2,774
- Hispanic cohort: 35,577 x 2,826

For all techniques tested in this study, features were transformed into appropriate data types for each respective methodology (e.g., for K Modes, numeric features were transformed into categorical values). Additional feature engineering included:

1. Remove rows with patient age below 30 or above 79.
2. Remove rows with undefined gender.
3. Patients with no records of a given diagnosis or prescription (absence of flag_* value) were set to 0 (no presence of diagnosis or prescription) in the corresponding flag_ feature.
4. Remove features with near zero variance based on the ratio of the most frequent value over the second most frequent value being greater than 95/5.
5. Dummy encode all demographic categorical features and treat missing values as an additional category.
6. Center and scale numeric features age, average income, and percent college educated using Dataiku Data Science Studio (DSS) standard average rescaling and fill missing values with median.
7. For K-Modes, discretize numeric features to treat them as categorical features.

The resulting dataset contained the following number of rows and features;

- Caucasian cohort (CA): 84,479 x 380
- African American cohort (AA): 85,405 x 362
- Asian cohort (AS): 40,289 x 277
- Hispanic cohort (HI): 31,667 x 357

Clustering Techniques

1. K-Means – Partitions the data space with the goal to minimizes intra-cluster variances using squared Euclidean distances. May converge to local optimum.
2. Mini-Batch K-Means – A variation of K-Means that uses a subset of randomly sampled “mini-batches” in each iteration to reduce the convergence time of K-Means.
3. Agglomerative Clustering – A form of hierarchical clustering that starts with each data point being its own cluster and then successfully merging clusters according to a chosen metric.
4. DBSCAN – Density-based non-parametric clustering that aims to group together data points which are closely together.
5. Interactive Clustering – A two-stage algorithm which first groups data points into small clusters using K-Means and then uses Agglomerative clustering for building a hierarchy between the smaller clusters.
6. K-Modes – A variation of K-Means suitable for clustering categorical variables. It uses matching categories to define the clusters in contrast to K-Means which uses Euclidean distance.

Clustering model evaluation criteria

We applied six clustering algorithms described previously to the cohort datasets. Additionally, we considered the application of Gaussian Mixture, as well as Spectral Clustering. Gaussian Mixture was not considered due to the overwhelming number of categorical data present. Spectral Clustering was not used due to the running time and Isolation Forest was not used as it focused on anomaly detection. The remaining algorithms K-Means, Mini-Batch K-Means, Agglomerative Clustering, DBSCAN, Interactive Clustering, and K-Modes were used with the following parameters:

1. K-Means
    a. Number of clusters: 5, 10, 15, 20, 25, 30, 35, 40
2. Mini-Batch K-Means
    a. Number of clusters: 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40
3. Agglomerative Clustering
    a. Number of clusters: 5, 10, 15, 20, 25, 30, 35, 40
4. DBSCAN
    a. Epsilon: 0.1, 0.25, 0.5, 0.75, 1, 2
    b. Min. sample ratio: 0.01
5. Interactive Clustering
    a. Number of Pre-clusters: 100
    b. Number of clusters: 2, 4, 5, 25
    c. Max Iterations: 75
6. K-Modes
    a. Number of clusters: 2, 3, 4, 5, 6, 7, 8, 9, 10

The Silhouette width metric was used for comparing algorithmic results. This metric measure how well each patient belongs to its cluster compared to other clusters. The values range between -1 to 1 with values near 1 indicating well matched to its own cluster and poorly matched to neighboring clusters. Initially, we applied principle component analysis (PCA) to reduce the dimensionality of the tested datasets. The results from using clustering techniques on the PCA derived datasets were inconclusive, with most cases identifying no segmentation of patients into multiple distinct clusters. Potentially this could be due to non-linear relationship between features, and techniques such as Kernel PCA could be more appropriate. As such, the results that follow are based on all clustering applied to all features included in the datasets. An outlier cluster was created to place patients that are far from identified clusters by pre-clustering with many clusters and selecting small clusters with less than 100 as outliers.

Results

The results of our clustering analysis were obtained using an iterative approach. The first run included all techniques described in the Materials and Methods section, with the exception of K-Modes, by varying the number of clusters (where applicable) between 5 and 40 for every interval of 5. K-Modes was introduced after the initial results to evaluate a technique designed for categorical values. Figure 1 shows the results of this run. DBSCAN did not yield meaningful clusters for any tested parameterization and is therefore excluded from our reported results. As shown in Figure 1, the highest silhouette values were obtained for 5 clusters for most techniques with Mini-Batch K-Means and Interactive Clustering having the best overall results.

Figure 1: Performance of clustering techniques from 5 to 40 clusters.

Table 1: Best two K values for each cohort.

Cohort	Technique	K	Silhouette
AA	MB K-Means	2	0.242
AA	Interactive	5	0.198
AS	MB K-Means	2	0.237
AS	Interactive	4	0.200
CA	MB K-Means	2	0.214
CA	KModes	3	0.170
HI	MB K-Means	2	0.221
HI	KModes	3	0.180

To further optimize our clustering results, additional runs were executed for K cluster values between 2 and 10 for Mini-Batch K-Means. We also tried K-Modes for the same number of clusters as we hypothesized its suitability for datasets with a high percentage of categorical data would yield improved clustering results. Interactive Clustering, due to its long run time, was limited to k values of 2, 4, and 5 clusters which were the best results obtained by the other techniques. The results are shown in Figure 2. The best and the second-best silhouette results for different K values for each cohort are shown in Table 1. The best result was found for K=2 clusters by Mini-Batch K-Means for every cohort. The average silhouette width for AA, AS, CA, and HI cohorts was 0.242, 0.237, 0.214, and 0.221, respectively. The silhouette values are consistent with optimal clustering results observed in other published RWD-based clustering anlyses[10]. The second-best result for different value of K was found by K-Modes for CA and HI cohorts and by Interactive Clustering for AA and AS cohorts. In this case, for CA and HI cohorts a value of K=3 clusters was found in contrast to AA and AS cohorts with values of K=5 and K=4 respectively.

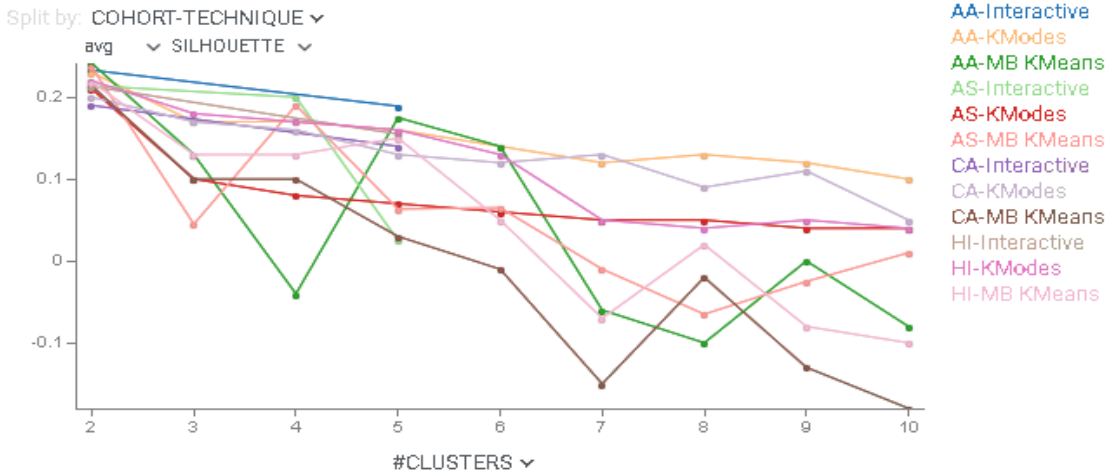

Figure 2: Performance of best techniques from 2 to 10 clusters.

Following the identification of our optimal clustering result of K=2 clusters for every cohort, shown in Table 1, we examined the features that contribute most significantly to segmenting patients into their respective clusters. To do this a simple random forest classifier model was used in DSS to determine the important features for classifying the datasets into the output classes found by Mini-Batch K-Means. The random forest classifier from scikit-learn version 0.22.1 was used with 100 trees, no depth restriction, and minimum leaf nodes of 1. Table 2 shows five common features that were found to be among the ten most important features for the optimal clustering results in all cohorts. In this table we have highlighted (bold values) for each cohort the most important common feature. For the AA, CA, and HI cohorts the most important common feature is the prescription of heparins and heparinoid like agents and matches the most important feature for each cohort. Only the AS cohort has 5 ht3 receptor antagonists drug class prescription as the most important common feature and analgesics prescriptions (3.10%, not shown) as the most important feature for the cohort.

Table 2: Importance of five common features found among the top 10 in optimal clustering results, K=2, for every cohort.

Feature Name	AA	AS	CA	HI
heparins and heparinoid like agents prescriptions	4.00%	2.14%	3.72%	3.35%
5 ht3 receptor antagonists drug class prescription	2.87%	3.07%	2.91%	2.39%
cough diagnosis	2.07%	2.62%	2.85%	2.39%
opioid combinations drug class prescription	2.25%	2.71%	2.68%	3.21%
somnolence stupor and coma diagnosis	2.17%	2.31%	2.50%	2.73%

There were many common features among the important variables for every cohort. Only the order of importance differed among the cohorts. For every cohort, the results identified between 20 to 25 top features with 1% or more importance for segmenting the patients into two clusters. A visual inspection revealed the important features are all associated with prescriptions and diagnosis events. No demographic related features were among the top features. For the AS cohort the top features are about half for both prescription and diagnosis events. For the other cohorts between 70-75% of the top features are associated with prescription events compared to 25-30% diagnosis events.

Table 3: Relative importance (RI) of the ten most important features for both clusters, low and high EHR events in each cohort.

AFRICAN AMERICAN			
LOW EHR		HIGH EHR	
FEATURE	RI	FEATURE	RI
rx heparins and heparinoid like agents	83.03	rx opioid agonists	143.76
rx carbohydrates	75.86	rx sodium	129.66
rx diabetic other	71.99	rx 5 ht3 receptor antagonists	126.75
rx stimulant laxatives	69.83	rx analgesics other	123.96
rx 5 ht3 receptor antagonists	69.55	dg cough	118.51
rx magnesium	69.51	rx opioid combinations	118.41
rx potassium	69.30	rx heparins and heparinoid like agents	112.23
rx analgesics other	68.24	dg somnolence stupor and coma	109.26
rx surfactant laxatives	67.30	rx proton pump inhibitors	106.61
rx laxatives miscellaneous	66.53	dg other and unspecified soft tissue disorders	105.92
ASIAN			
LOW EHR		HIGH EHR	
FEATURE	RI	FEATURE	RI
rx heparins and heparinoid like agents	42.45	dg cough	74.44
rx potassium	37.47	rx opioid agonists	69.34
rx carbohydrates	37.25	dg somnolence stupor and coma	66.76
rx 5 ht3 receptor antagonists	37.22	rx proton pump inhibitors	61.70
rx stimulant laxatives	37.03	rx 5 ht3 receptor antagonists	61.58
rx opioid combinations	36.92	dg abnormalities of gait and mobility	60.59
rx analgesics other	36.60	rx opioid combinations	59.99
rx surfactant laxatives	35.73	dg other and unspecified soft tissue disorders	59.26
rx opioid agonists	34.20	rx analgesics other	58.58
rx diabetic other	34.14	rx sodium	55.94

CAUCASIAN			
LOW EHR		HIGH EHR	
FEATURE	RI	FEATURE	RI
rx heparins and heparinoid like agents	88.08	rx opioid agonists	140.64
rx stimulant laxatives	77.55	rx 5 ht3 receptor antagonists	125.92
rx potassium	77.33	rx sodium	118.70
rx 5 ht3 receptor antagonists	74.00	rx opioid combinations	116.47
rx saline laxatives	73.63	rx analgesics other	111.97
rx surfactant laxatives	73.14	dg cough	111.85
rx analgesics other	72.17	rx glucocorticosteroids	109.54
rx opioid combinations	71.88	dg somnolence stupor and coma	105.84
rx loop diuretics	68.95	dg encounter for other postprocedural aftercare	104.77
rx cephalosporins 1st generation	68.15	rx non barbiturate hypnotics	98.72
HISPANIC			
LOW EHR		HIGH EHR	
FEATURE	RI	FEATURE	RI
rx heparins and heparinoid like agents	48.67	rx opioid agonists	79.71
rx stimulant laxatives	42.82	rx opioid combinations	74.99
rx opioid combinations	42.37	rx 5 ht3 receptor antagonists	74.27
rx potassium	41.79	rx sodium	70.25
rx carbohydrates	41.47	dg cough	69.83
rx 5 ht3 receptor antagonists	41.04	rx proton pump inhibitors	65.94
rx surfactant laxatives	39.39	rx analgesics other	65.13
rx magnesium	39.27	dg other and unspecified soft tissue disorders	63.41
rx analgesics other	39.20	rx sympathomimetics	61.87
rx proton pump inhibitors	39.00	dg somnolence stupor and coma	61.59

A closer look at both clusters in every cohort shows that one of them contained between 70-79% of the patients. This cluster contained a high proportion of patients within a below average number of EHR events in their data. We refer to this cluster as the LOW EHR cluster. In contrast, the second cohort with the least number of patients is largely comprised of patients with the most EHR events in their data. We refer to this cluster as the HIGH EHR cluster. Table 3 shows the ten most important features for both LOW EHR and HIGH EHR clusters in every cohort. The features are shown in descending order of relative importance (RI) which is calculated as 100 * abs(CM - GM) / GM, where CM is the cluster mean value and GM the cohort mean value. The LOW EHR cluster have top features with CM << GM for all cohorts due to presence of patients with lower than normal medical events. Conversely, the HI EHR cluster have top features with CM >> GM for all cohorts due to the presence of patients with higher than normal. We can see the five common features identified in Table 2 in many of the clusters as they play an important role in the classification. The RI in HIGH EHR clusters are higher than those in the LOW EHR in every cohort because in general we have less patients with a high number of medical events, making the GM value lower. The top feature, heparins and heparinoid like agents prescription, is the same in the LOW EHR clusters for every cohort. Except for the AS cohort, opioid agonists prescription is the top feature for HIGH EHR clusters in every cohort. Many of the same important features are also shared between the same cluster type across all cohorts, but only HIGH EHR clusters have diagnosis features within the top ten. Further analysis is necessary to compare these values against non-cataract patient cohorts to determine any statistical differences between them. The presence of a significant number of cataract patients with high number of medical events provides the opportunity for future studies.

A visual inspection of demographic values revealed a similar profile across both clusters for all cohorts and between them. One notable exception is the percent of patients above 70 years of age in the HIGH EHR cluster. The HIGH EHR cluster in every cohort was between 9-31% higher when compared to the percent of patients above 70 years of age in the cohort. Most patients were diagnosed with cataract between the ages of 60 to 70 in all cohorts. The percentage of male and female in each cluster for every cohort was within 1% of the percentage in the cohort. A more

detailed analysis of cluster variation will serve as the basis of a future study.

Discussion

The initial characterization of patients with cataracts explored in this study through unsupervised machine learning techniques yielded preliminary results that may be used for a simple segmentation of patients into two distinct clusters. Similar important features are included across all the cohorts for the cluster containing many EHR events. A small set of 20-25 features was found to provide the necessary information for segmenting the patients into two clusters across all cohorts. Only minor differences were found between the top features across all cohorts with prescription events being the most common in most cohorts. Across our observed results, demographic features did not contribute significantly in segmenting cataract patients. To assess whether the presence of any of these features is enriched in cataract patients (or a subset thereof) will require a secondary analysis where patients with an elevated number of touch points with the healthcare system including patients with and without cataract diagnosis. Follow-up studies leveraging both supervised and unsupervised machine learning techniques will follow.

It is also important to acknowledge limitations to the clustering algorithms tested due to the presence of a high number of features that affect the number of clusters found. As the number of dimensions increase, the search space increases exponentially making the data sparse and appearing dissimilar. Euclidean distance breaks down due to distance to the origin growing as the square root to the number of dimensions[11]. We took steps to reduce the number of features, but it is difficult to determine conclusively that our mitigations are enough. In addition, the diversity of feature data types, numeric, Boolean, and categorical required different type of transformations to allow techniques designed for numeric values to be used. Similarly, K-Modes required features to be transformed into categorical values. Techniques, like K-Prototypes could be used in the future to keep a mix of numeric and categorical data for the analysis. Other techniques based on graph relationships with cluster shape independence, Deep Learning and Fuzzy clustering should be evaluated as well.

The presence of cataract patients with high number of events as characterized in the HIGH EHR clusters may provide an opportunity to investigate for hidden relationships between these events and the onset of cataract when combined with similar non-cataract patients. Future analysis will be conducted to objectively identify features which may be applicable to cataract patient characterization. There are also additional EHR data sources that could be leverage for similar analyses. Of potential interest is the inclusion of lab measurement information over time and observation results that may lead to a more complex dataset better suited to revealing complex variation in subsets of cataract patients.

CONCLUSION

The clustering analysis presented in this manuscript reveals the potential of real-world data for the segmentation of cataract patients in distinct subpopulations. Additional analysis of the identified important features and additional EHR sources and feature engineering techniques may lead to a better understanding of the role prescription and diagnosis events play in the segmentation of cataract patients, and whether such segmentation could drive the basis of futures diagnostics, interventions, and/or treatments. The number of features available for real-world data analysis will continue to growth over time, and improved techniques that can be

successfully applied to high dimensional spaces could lead to enhanced patient segmentation that further enables the development of solutions for cataract patients.

References

1. World Health Organization, “World Report on Vision”, Geneva: World Health Organization; 2019, License: CC BY-NC-SA 3.0 IGO, ISBN 978-92-4-151657-0, 8 October 2018.
2. National Eye Institute, “2010 U.S. age-specific prevalence rates for cataract by age and race/ethnicity”, Cataract Data and Statistics, https://www.nei.nih.gov/learn-about-eye-health/resources-for-health-educators/eye-health-data-and-statistics/cataract-data-and-statistics, Accessed May 2019.
3. National Eye Institute, “Am I at risk for cataracts?”, Cataracts, https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/cataracts, Accessed May 2019.
4. Ursell, Paul G et al. “Three-year incidence of Nd:YAG capsulotomy and posterior capsule opacification and its relationship to monofocal acrylic IOL biomaterial: a UK Real World Evidence study.” Eye (London, England) vol. 32,10 (2018): 1579-1589. doi:10.1038/s41433-018-0131-2.
5. Day, Alexander C et al. “Distribution of preoperative and postoperative astigmatism in a large population of patients undergoing cataract surgery in the UK.” The British journal of ophthalmology vol. 103,7 (2019): 993-1000. doi:10.1136/bjophthalmol-2018-312025.
6. Amazon Marketplace, “Glaucoma Detection”, PERSISTENT Glaucoma Detection, https://aws.amazon.com/marketplace/pp/prodview-bbi5wdcjq6oyy?qid=1573763067156&sr=0-42&ref_=srh_res_product_title, Accessed November 2019.
7. Kim, Seong Jae et al., “Development of machine learning models for diagnosis of glaucoma.” PloS one vol. 12,5 e0177726. 23 May. 2017, doi:10.1371/journal.pone.0177726.
8. McCarty, Catherine A, and Hugh R Taylor. “A review of the epidemiologic evidence linking ultraviolet radiation and cataracts.” Developments in ophthalmology vol. 35 (2002): 21-31. doi:10.1159/000060807
9. Clinical Drug Information, “Medi-Span Electronic Drug File (MED-File) v2”, Wolters Kluwer Clinical Drug Information, March 2016.
10. Lerner, Jason, et al., "Preoperative Behavioral Health, Opioid, and Antidepressant Utilization and Two-Year Costs After Spinal Fusion-Revelations from Cluster Analysis." Spine (2019).
11. Oskolkov, Nikolay, “How to cluster in High Dimensions: Automated way to detect the number of clusters”, Mathematical Statistics and Machine Learning For Life Sciences, https://towardsdatascience.com/how-to-cluster-in-high-dimensions-4ef693bacc6, Accessed August 2019.

Contact Information

Author Name: Walter Cedeño
Company: Janssen Pharmaceuticals
Address: 1400 McKean Rd, Spring House, PA
City / Postcode: 19477
Work Phone: 215-793-7120
Email: wcedeno AT its.jnj.com
Web: https://www.linkedin.com/in/waltercedeno/