

Data Sciences &
Data Management
computing the future



A Real-World View of Cataract

Walter Cedeño, Jeffrey Headd, & Jun Morimura

Agenda

Problem Statement & Overview

Real-World Data & Cohorts

Data Transformation & Clustering

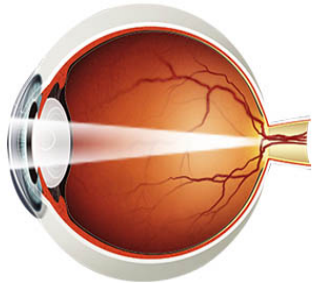
Results & Discussion

Acknowledgements

What is Cataracts?

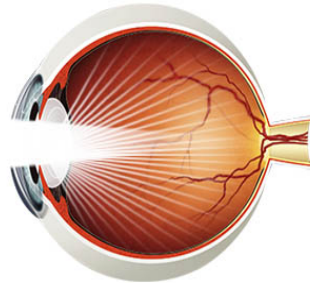
A medical condition affecting the eye that causes clouding of the lens. Most cataracts are related to aging.

Normal Eye



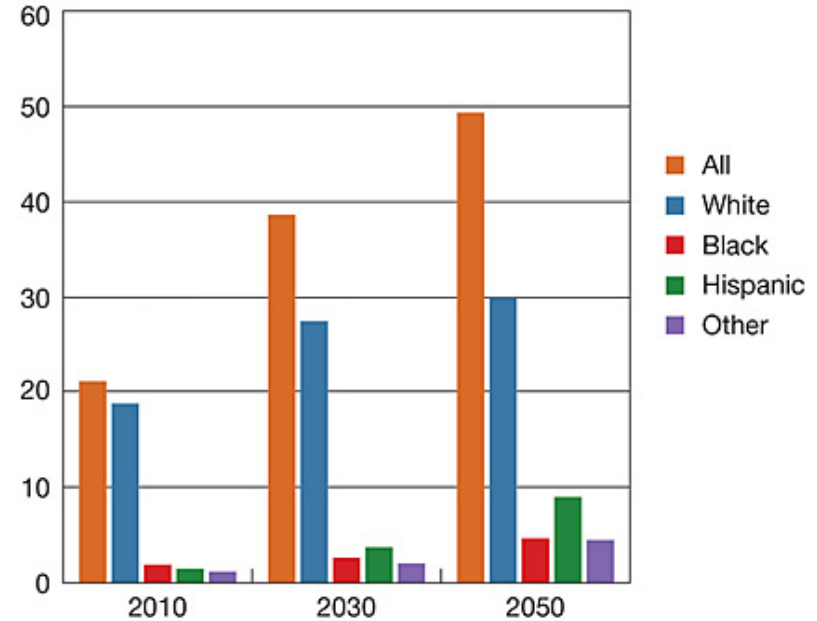
A healthy lens allows for all parts of the retina to receive the image.

Cataract Eye



A cloudy lens scatters light, causing an image that's out of focus and hazy.

Projections for Cataract
in 2030 and 2050 (in millions)



Problem Statement



- Through analysis of Real-World Data, can we identify patient journey patterns that segment Cataract patients into distinct groups?

- What are the important variables that segment patients into these distinct groups?

Overview



- Real-World Data
 - The data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.
- Real-World Evidence
 - The clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD.

Identify Real-World Data

Optum® de-identified Electronic Health Record dataset, containing patient records from 2007 to 2019:

- Patient*
 - Diagnosis*
 - Labs
 - Observations
 - Medication_administrations*
 - Prescriptions_written*
- ICD9
 - 366.0*, 366.1*, 366.3*, 366.8*, 366.9*
 - ICD10
 - H25*, H26.0*, H26.2*, H26.8*, H26.9*

Define Cohorts

- Patients with birth year recorded
- Age between 30 and 79 at the time of their first diagnosis
- Gender is recorded
- Available patient records for cohorts:
 - All Cataract Patients: 2,270,943
 - Caucasian (not Hispanic): 1,719,067*
 - African American (not Hispanic): 189,738*
 - Asian (not Hispanic): 44,191
 - Hispanic (any race): 35,577

Feature Selection

- Demographic
- Diagnosis
- Prescriptions
 - Drug class names using Medi-Span
- Cohorts (rows x columns) after feature selection:
 - 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

Data Preparation

- Set diagnosis or prescription flag to 1 or 0.
- Remove features with near zero variance.
- Dummy encode categorical features and treat missing values as a category.
- Center and scale numeric features and fill missing values with median.
- For K-Modes, discretize numeric features
- Cohort (rows x columns) after data prep:
 - 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

Hierarchical

- Agglomerative
- Interactive

Centroid-based

- K-Means, Mini-Batch K-Means, K-Modes
- Gaussian Mixture Models

Graph-based

- Spectral

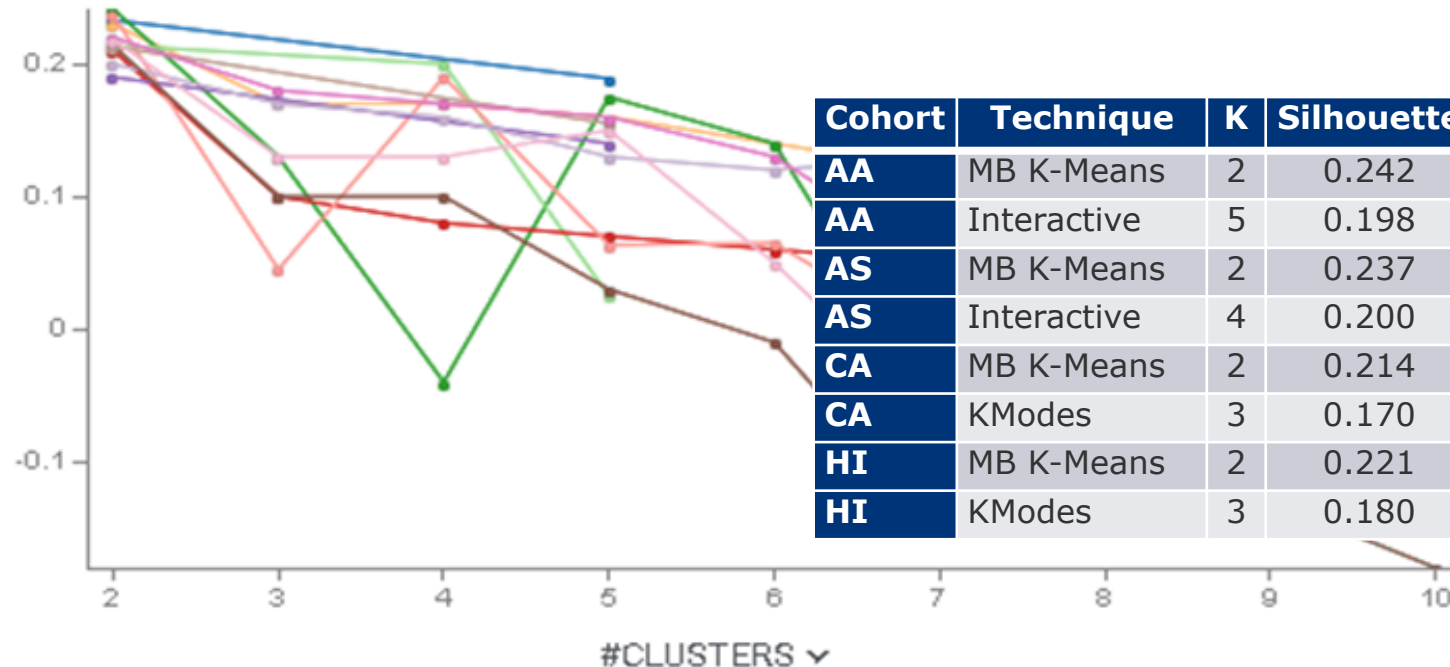
Density-based

- DBSCAN

Performance of Best Techniques from 2 to 10 Clusters

Split by: COHORT-TECHNIQUE ▾

avg ▾ SILHOUETTE ▾



AA-Interactive
 AA-KModes
 AA-MB KMeans
 AS-Interactive
 AS-KModes
 AS-MB KMeans
 CA-Interactive
 CA-KModes
 CA-MB KMeans
 HI-Interactive
 HI-KModes
 HI-MB KMeans

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

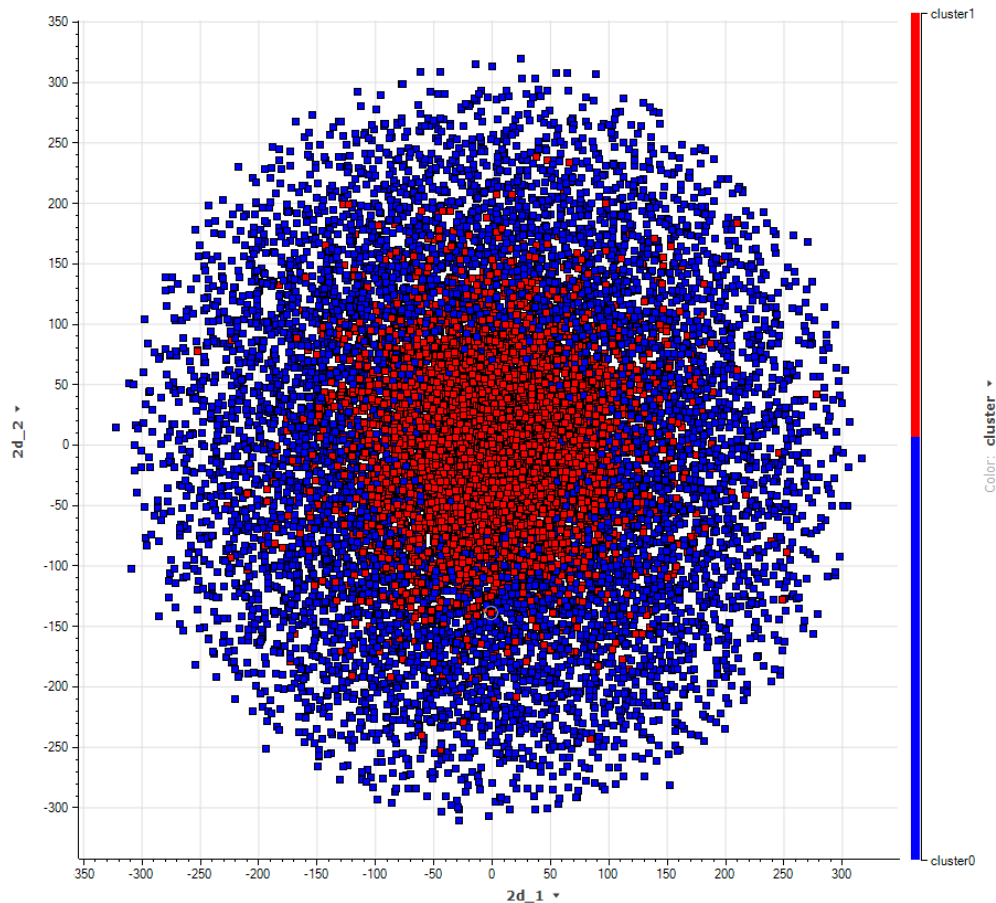
Feature Name	AA	AS	CA	HI
heparins and heparinoid like agents prescriptions	4.00%	2.14%	3.72%	3.35%
5 ht3 receptor antagonists drug class prescription	2.87%	3.07%	2.91%	2.39%
cough diagnosis	2.07%	2.62%	2.85%	2.39%
opioid combinations drug class prescription	2.25%	2.71%	2.68%	3.21%
somnolence stupor and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

Results Highlights

- For every cohort, 20 to 25 top features with 1% or more importance for segmenting the patients into two clusters.
- AS cohort top features: 50% prescription and 50% diagnosis events
- AA, CA, HI top features: 70-75% prescription and 25-30% diagnosis events.
- Two clusters with similar cardinality for all cohorts
 - LOW EHR: 70-79% of the patients
 - HIGH EHR: 21-30% of the patients
- Relative importance of top features in HIGH EHR >> LOW EHR

Cluster Embedding of Hispanic Cohort

- Stochastic Proximity Embedding
 - Self-organizing algorithm
 - Scales linearly
- Map 357d -> 2d
- Low EHR cluster in red
- High EHR cluster in blue



Summary & Next Steps

- Summary

- Segmentation of Cataract patient journeys found two distinct clusters.
- A small set of 20-25 features provided the necessary information for segmenting patients into two clusters across all cohorts tested.

- Next Steps

- Examine the journeys of patients with elevated number of healthcare events for both Cataract and non-Cataract patients.
- Additional analysis of the identified important features to understand their significance in the segmentation of cataract patients.
- Exploration of additional RWD sources (EHR, claims, non-healthcare data).
- Test additional supervised and unsupervised techniques.

Acknowledgements

Jeffrey Headd & Jun Morimura for their collaboration and support

Janssen Business Technology Data Sciences team

Janssen R&D for resources and general RWD information