Project: Capstone Project 1: Milestone Report

Glaucoma

Problem Statement:

Glaucoma is a general term for a number of eye conditions that progressively damage the optic nerve, consequently causing vision loss. Diagnosis of glaucoma is difficult and complex[1], but is often associated with elevated intra-ocular pressure, optic nerve damage, and reduction in visual acuity and visual field. Vision loss from glaucoma is permanent, but progression may be slowed or halted through early diagnosis and treatment.

With their small collection of retinal data Cambia Health is looking to develop a machine learning algorithm in order to reduce medical practitioner time of diagnosing patients with Glaucoma.

Dataset:

For this project I will use data gathered from the Harvard Dataverse collected and uploaded by Ungsoo Kim2. The retinal image dataset consists of 788 'normal control' images, 289 'early stage' images, and 467 'advanced stage' images. The images are already preprocessed, (scaled to 800 pixels and cropped so the nerve endings are of 240 pixels) thus ready to be used for machine learning tasks. However, this dataset will present itself particularly challenging as the number of retinal images needed for this task are less than desirable; thus representing a real-world problem of its own.

For analysis of describing the Glaucoma problem I used the most recent survey of the Behavioral Risk Factor Surveillance System (BRFSS) that is publicly released collected responses from participants about being diagnosed with Glaucoma (2010). I also used multiple years of the BRFSS survey to create a trend graph in order to get an idea of the prevalence fluctuation from one year to the next. Each survey reflects over 400,000 nationwide responses from participants and then is weighted in order to adjust the data to reflect state, or nationwide results. The Behavioral Risk Factor Surveillance System (BRFSS) is a nationwide telephone survey collected by the CDC and freely available online at their host website.

Data Wrangling:

Neither the image dataset nor the telephone survey required much data cleaning. For the telephone survey I had to create a calculated variable in order to flag the responses I was interested in investigating for Glaucoma; this included properly investigating and coding NaN values and distinguishing them from "Refuse" to respond and "Don't Know" options. Furthermore, workings with weighted analyses are also particularly tricky to do in Python since available statistical APIs are poorly documented on how to work with complex weighting schemes so there was some trial and error involved in order to return the correct or near correct values. There were additional difficulties working with multi-year surveys since typically not all BRFSS surveys always have consistent question design; requiring specialized attention to detail to make sure the responses match.

For the image dataset there wasn't any data wrangling involved aside from correctly specifying locations to store the images. The image set was provided in the form of a zip file that had the 'normal' control images, 'early stage', and 'advanced stage' images. It is not typically a good practice to have such a small dataset to be used for image classification; since I was dealing with a small set rather than trying to classify into three distinct categories I chose to simply classify to 'Glaucoma' or 'Not Glaucoma'. My reasoning is statistically speaking the greater number of units we try to classify will increase the error for

[1] Glaucoma diagnosis is based on clinical assessment of intraocular pressure, visual fields, and cup to disk ratio.

each additional level of classification; therefore I chose to keep the classification binary to avoid such problems since the dataset was small anyways. This required moving images from the ‘Advanced’ and ‘Early Stage’ Glaucoma into a separate folder marked as ‘Cases’ and the ‘Normal Controls’ to simply ‘Controls’. After separating the image files into two distinct folders in the appropriate folder locations I wrote a class object and a number of functions that randomized the images, performed image augmentation, and split them into appropriate train-test splits (70%, 10%, 20%) into training, validation, testing, and augmentation folders[2]. Lastly, I created a Dockerfile so I could work on training my algorithm in a closed development environment that can be potentially utilized by Amazon AWS Sagemaker.

Exploratory Analysis:

Summary:

Glaucoma is not a common problem; but it is a costly and crippling disease particularly for individuals and the family of those diagnosed with it. Compared to the general population few adults are diagnosed with Glaucoma each year thus prevalence data for those diagnosed with Glaucoma can be difficult to come by. Luckily, there are a key number of national surveys and clinical studies that can help to understand the scope of the chronic condition. Unfortunately, these sources of data are not collected in a systematic fashion so the findings that can be obtained are limited.

According to an assessment of the economic burdens associated with Glaucoma, an estimated 3% of the global population over 40 years of age currently has Glaucoma; the majority of whom are undiagnosed. The direct medical costs for those diagnosed with Glaucoma include ocular hypertensive medications, physician and hospital visits, and glaucoma-related procedures. The Indirect costs include lost productivity, days missed from work, and the cost borne by caregivers such as family and friends. Direct cost estimates for approximately 2 million US citizens are 2.9 billion dollars each year for those whom have been diagnosed, to say the least for those whom are undiagnosed. For those receiving treatment for glaucoma the estimated average annual incremental cost is 137 dollars per patient per year; a cost that only increases with severity of disease.

I investigated the overall prevalence of glaucoma using the most recent estimates I could find from the Behavioral Risk Factor Surveillance Survey (BRFSS). By doing this task I can start to build a case against glaucoma and helping to build awareness and context around the disease. I also investigate multiple year trends from different years of the BRFSS in order to gain some insight on the stability of the chronic condition. Lastly, I finish with a simple meta-analysis of literature review provided by the CDC that shows the difficulty in assessing the prevalence from multiple data sources.

Conclusions:

From my brief investigation of the data I found that the overall prevalence to be anywhere between 4.8 and 6 percent in the United States among adults 40 years or older. Interestingly, the nationwide prevalence among US adults is nearly double that of the global population. This could possibly be attributed to a greater overall prevalence of chronic diseases such as diabetes; however, another possible explanation is access to vision care in the overall population leading to a greater degree of Glaucoma diagnoses. A third and most likely explanation is simple statistical variation among small sample sizes, as can be illustrated by the estimates of the *meta-analysis* section. Regardless in the difference in overall population estimates, assuming there are approximately 154911104 adults over 40 in the United States, an estimated 4647333.12 to 143415641.9 adults per year suffer from glaucoma. Furthermore, I also

[2] *Image augmentation is the act of creating new images by adding rotation, flipping images, adding distortions, and introducing noise in order to increase the dataset size and introduce statistical noise and thereby increasing generalizability of the trained model.*

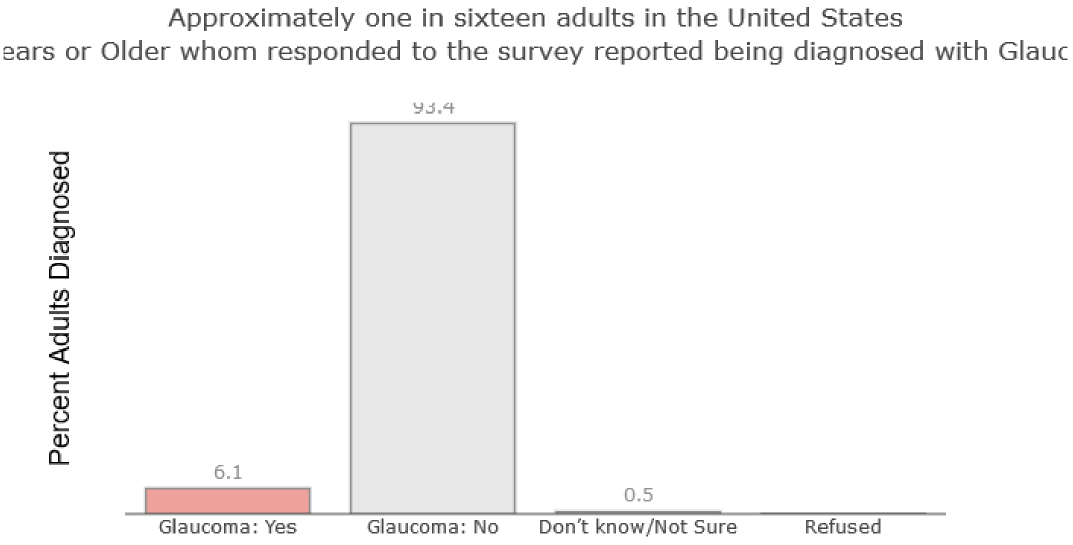
concur with the background research that finding an accurate prevalence source for glaucoma to be difficult because few research studies glaucoma at the cross-sectional or national level. If I didn't already have a different project path in mind, a good follow-up to the materials presented here would be to investigate the association between glaucoma and different risk factors collected from the BRFSS such as race, sex, socio-economic status, tobacco use, hypertension, obesity, and heart-disease. Judging from the results shown thus far however; I do believe this task could be difficult considering the small sample sizes presented in the BRFSS data.

Single Year Prevalence

For best viewing of my results and interactive graphs follow this link:
https://nbviewer.jupyter.org/github/pmleffers/Glaucoma/blob/3d07c76b765dde9c627960bddd442970ca532130/Glaucoma%20Analysis.ipynb

BRFSS 2010: Self-reported diagnosis with Glaucoma.

	Percent	Frequency	Weight_freq
Glaucoma: Yes	6.061600	1478.0	4.815537e+05
Glaucoma: No	93.401140	22774.0	1.077470e+07
Don't know/Not Sure	0.516753	126.0	4.687352e+04
Refused	0.020506	5.0	3.439010e+03
BLANK	1849.956937	451075.0	0.000000e+00
Total	1949.956937	475458.0	1.130657e+07



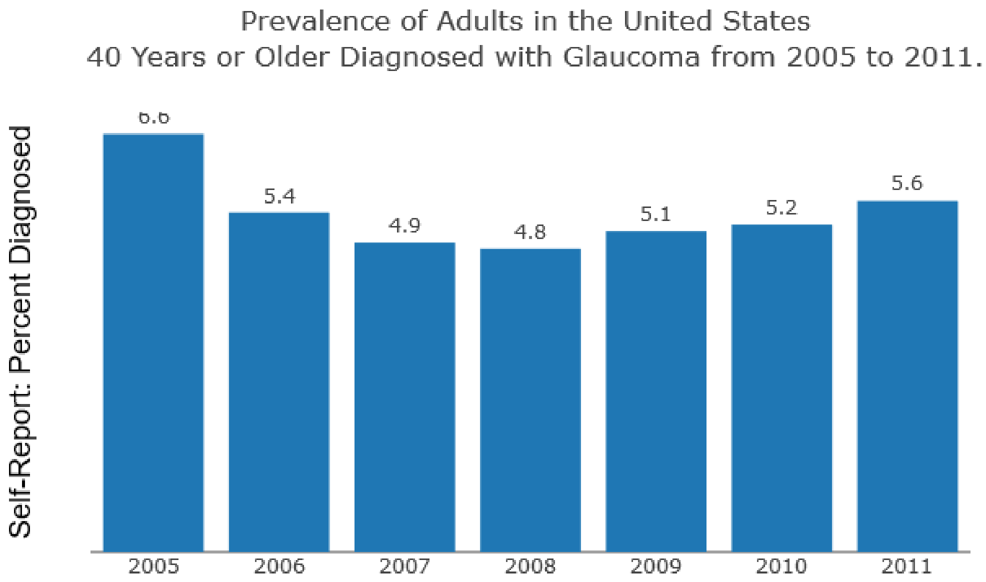
Explanation: In the year 2010 among those whom were 40 years of age that received the vision module and responded to the question on the survey, six percent or one-in-sixteen adults responded as having been diagnosed with Glaucoma.

Multi- Year Prevalence

For best viewing of my results and interactive graphs follow this link:
https://nbviewer.jupyter.org/github/pmleffers/Glaucoma/blob/3d07c76b765dde9c627960bddd442970ca532130/Glaucoma%20Analysis.ipynb

Trend Graph

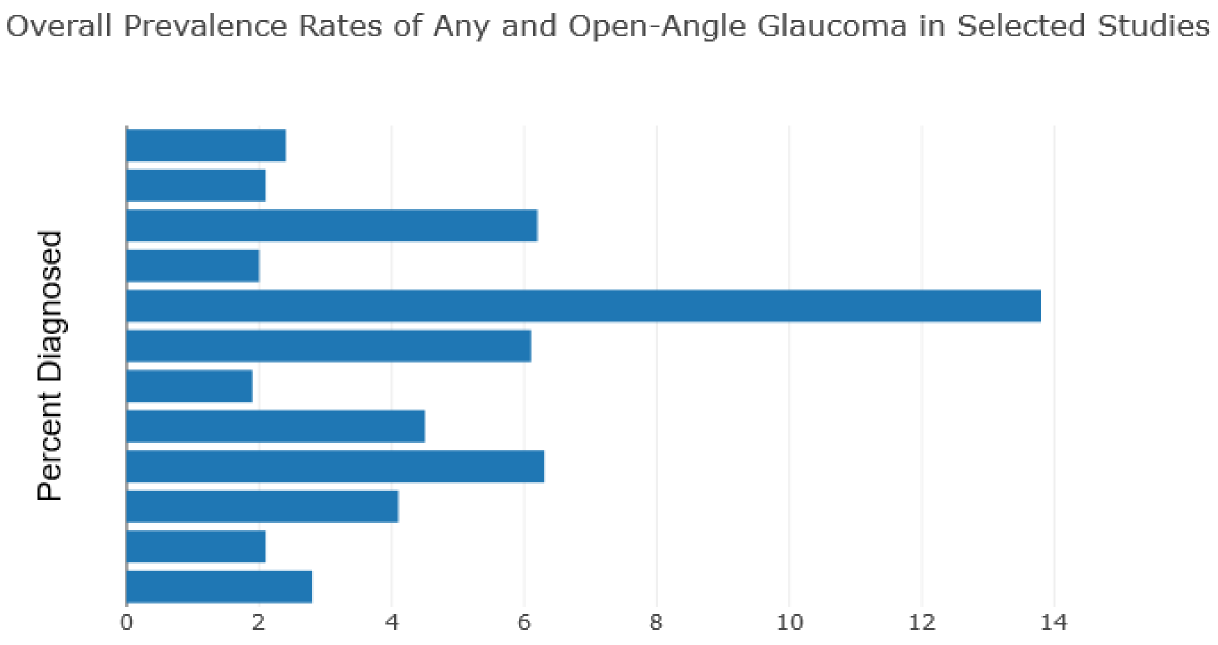
	Year	Data_Value
0	2005	6.635965
1	2006	5.389437
2	2007	4.913592
3	2008	4.818158
4	2009	5.093863
5	2010	5.199083
6	2011	5.572816



Explanation: From the years 2005 to 2011 the overall crude prevalence of Glaucoma nationwide appears to be relatively stable over time hovering between 4.8 and 6.6 percent among US adults over the age of 40.

Meta-Analysis

	Author	Title	Data_Source	Sample_Size	percent
0	Park D, Mansberger SL, et al.	Prevalence of Age-Related Macular Degeneration...	Telemedicine Screening Program	424.0	2.8
1	Gupta P, Zhao D, et al.	Prevalence of Glaucoma in the United States: T...	NHANES 2005-2008	4798.0	2.1
2	Maa AY, Evans C, et al.	Veteran Eye Disease After Eligibility Reform:...	Atlanta VA Medical Center Chart Review	658.0	4.1
3	Cassard SD, Quigley HA, Gower EW, et al.	Regional Variations and Trends in the Prevalen...	Medicare Claims	NaN	6.3
4	Kim E, Varma R.	Glaucoma in Latinos/Hispanics.	LALES	6142.0	4.5
5	EDPRG	Prevalence of Open-Angle Glaucoma Among Adults...	EDPRG	NaN	1.9
6	Mansberger SL, Romero FC, et al.	Causes of Visual Impairment and Common Eye Pro...	Northwest AIAN	288.0	6.1
7	Lee PP, Feldman ZW, Ostermann J, et al.	Longitudinal Prevalence of Major Eye Diseases	National Long-Term Care Survey	NaN	13.8
8	Quigley HA, West SK, Rodriguez J, et al.	The Prevalence of Glaucoma in a Population-Bas...	Proyecto VER	4774.0	2.0
9	Haronian E, Wheeler NC	Prevalence of Eye Disorders Among the Elderly ...	UCLA MEC	431.0	6.2
10	Klein B, Klein R, et al.	Prevalence of Glaucoma: The Beaver Dam Study	BDES	4926.0	2.1
11	Tielsch JM, Sommer A, et al.	Racial Variations in the Prevalence of Primary...	BES	5308.0	2.4



Source(s):

Varma, Rohit et al. "An assessment of the health and economic burdens of glaucoma." American journal of ophthalmology vol. 152,4 (2011): 515-22. doi:10.1016/j.ajo.2011.06.004
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3206636/

Coleman AL, Kodjebacheva G. Risk factors for glaucoma needing more attention. Open Ophthalmol J. 2009;3:38–42. Published 2009 Sep 17. doi:10.2174/1874364100903020038
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2759104/

NORC | Published Examination based Prevalence of Major Eye Disorders. JUNE 22, 2018.
http://www.norc.org/PDFs/VEHSS/EyeConditionExamLiteratureReviewVEHSS.pdf

** Population estimates collected from Census Reporter*
https://censusreporter.org/profiles/01000US-united-states/