1 Introduction

1.1 Tobacco Related Health Issues

The harms associated with tobacco use are well recognised. Tobacco kills up to half its users who do not quit and more than 8 million people per year, including an estimated 1.3 million non-smokers due to second hand smoke [1]. Smoking causes cancer, heart and lung disease, stroke, type 2 diabetes, and harmful reproductive effects [2]. There is a growing body of evidence suggesting a causal relationship between smoking and mental health issues [3]. Clearly, such negative impacts on patient health due to tobacco use are undesirable, just as they are avoidable. For these reason, tobacco usage is of great concern to health professionals. The World Health Organization asserts that surveillance is key for addressing the tobacco epidemic, as tracking tobacco usage indicates how to shape policy [1].

1.2 Self-Reported Smoking Status

Current surveillance relies on self-reported smoking data. That is, a patient's smoking history is recorded by them personally recalling and reporting. It is a convenient and cost-effective way of collecting smoking statistics. There are two main types of smoking data used to measure tobacco exposure: smoking status and smoking pack-years. Smoking status is label based on the history and habits of tobacco use. Indiviudals are binned into never smokers, ex smokers and current smokers. Smoking pack-years is a calculated score that tries to quantify tobacco use. It is calculated as the number of packs of cigarettes smoked per day multiplied by years of smoking [4]. For example, one pack-year is one pack per day for one year, or half a pack per day for two years. Therefore, smoking pack-years quantifies both the degree of exposure and duration of exposure equally.

Self-reported smoking data has several limitations. Relying on individuals recounting information can introduce bias. Self-reported smoking data is prone to inaccuracy due to stigma, recall bias and a lack of information on second-hand exposure [5]. That is, the social pressure to deny partaking in stigmatised behaviours, forgetting details and information, and not being aware of sources of second-hand exposure can all influence the results of self-reported smoking data. A method of using objective evidence to determine smoking history could overcome these issues. On the other hand, the inaccuracy of self-reported smoking data can differ between population groups. For example, studies suggest that teens are more likely to provide false responses in smoking surveys [5]. Moreover, tobacco consumption differs between social groups, with smoking more prevalent in low-education and low-socio-economic groups [6].

To this end, developing diagnostic tests to collect smoking data that do not share the biases of self-reported methods are of interest for improving the monitoring of health. One such approach is the use of epigenetic biomarkers.

1.3 Epigenetics

Epi- is a Greek prefix meaning upon or on. So, epigenetics is the study of factors on top of or upon genetics. Specifically, it is the study of how enviromental factors and behaviours affect and modify your genetics and their expression. We consider one type of epigenetic modification: DNA methylation.

1.3.1 DNA Methylation

At a high level, DNA is a sequence of letters that provide genetic instructions. Like a human reading a book, strings of these letters are converted into information that tells cells how to function. More precisely, these letters are one of four nucleotide bases: adenine (A), cytosine (C), guanine (G) and thymine (T). To form the sequence, these bases are attatched to a deoxyribose sugar and connected by a phosphate molecule, called the sugar-phosphate backbone. Of relevance is the phosphate molecule, specifically for when a cytosine is directly followed by a guanine in the sequence. A phosphate bonding a cytosine and a guanine (called a CpG site) creates a chemical structure which allows methyl groups to attatch. This process is an epigenetic modification called DNA methylation (DNAm). As a biomarker, we measure methylation at a CpG site as a float between 0 and 1, measuring the percentage of methylation at that site.

While the genetic sequence of DNA is stable, methylation is not. It is a dynamic state that depends on factors such has behaviours and enviromental exposure **(add citation?)**. Exposure to such factors increases methylation of CpG sites, while sufficent lack of exposure causes methylation to decrease over time **(add citation?)**. As previously mentioned, DNA methylation affects the expression of genes. Methylation at a CpG site can silence the expression of the gene that site is located in **(add citation?)**. More methylation at a site leads to stronger silencing of that gene. Moreover, DNA methylation is not random. There is strong correlation between methylation of specific sites with specific factors **(add citation?)**. This means that DNA methylation of CpG sites can be used as a biomarker indicative of the factors that caused it, while also describing changes in cellular function. Therefore, DNA methylation is a biomarker not only useful for reporting on enviromental exposures, but also predicting future health outcomes or risks. Examples of this include **list some topics of epigenetics-only papers (add citation?)**. Furthermore, DNA methylation is not self-reported, and therefore overcomes the biases associated with self-reported data.

Altogether, this motivates the use of DNA methylation data to develop methods for collecting smoking history of indiviudals.

1.3.2 DNAm Platforms

- Two types of DNAm platform, illumina 450k and illumina EPIC

- ∼28 million CpG sites in the human genome, platforms choose specific sites to ranked
- Cell type of sample matters, different cells with have different methylation. Typically whole blood is used, good general methylation signal.

1.4 Machine Learning in Epigenetics

- applications
- citations

Machine learning has already seen use in many areas of clinical epigenetics, including prognosis and diagnosis of cancer, cardiovascular diseases

- broad description of statistics produced
- classification/labels
- regression/scores
- algorithms

1. Malta et al. [7] proposed a method for assessing oncogenic dedifferentiation (cells becoming cancerous). This approach seeks to model a "stemness index" which indicates how similar a cell is to stem cell - a trait found in cancerous cells. Of relevance is the developed epigenetic approach using one-class logistic regression. The training features consisted of 219 hypermethylated CpG sites associated with stem cells. Training data only consisted of a single, positive, class: stem cells. The resulting model can then be fed non-stem cells to compare how similar they are to stem cells, i.e. cancerous cells.

2. Adorján et al. [8] proposed a method for using DNA methylation to classify cancer tissues. CpG sites were ranked using a two sample t-test, and then fed into a support vector machine. Models were evaluated using the average of 50 runs of 8-fold cross-validation. The top two CpG sites could classify leukaemia from healthy cells with 84% accuracy, while the top 60 sites achieve 94% accuracy.

3. Dogan et al. [9] proposed a method for integrated genetic and epigenetic classification of coronary heart disease. The dataset consisted of 1,545 indiviudals. An approach combining undersampling and ensemble learning was used to address class imbalance [10]. Point biserial correlation and Pearson correlation were used for feature selection, resulting in 107,799 CpG sites for training. Random Forest classifiers were then trained

1.5 Smoking Algorithms

In the context of smoking, the two most significant machine learning epigenetic algorithms use Elastic Net regression.

1.5.1 Elastic Net Regression

Elastic Net [11] is a regularised form of linear regression (and in turn, logistic regression) that includes two additional penalty terms. Given p features, we find β that produces an output:

$$\hat{y} = \sum_{i=1}^{p} \beta_i x_i$$

and minimises the function:

$$\mathcal{L}(y, \hat{y}, \beta) = ||y - \hat{y}||_2^2 + \alpha\lambda||\beta||_1 + \alpha(1-\lambda)||\beta||_2^2$$

where:

$$||\beta||_1 = \sum_{i=1}^{p} |\beta_i|$$

and:

$$||\beta||_2 = \sqrt{\sum_{i=1}^{p} \beta_i^2}$$

We can see this is simply mean-squared error, with the L_1-norm and L_2-norm included as penalisation terms on β. This in turn is a combination of two other modifications to linear regression: lasso and ridge regression [12, 13].

- effect of L_1-norm: sparse solutions
- effect of L_2-norm: grouping effect
- why this makes sense for epigenetics
- used a lot in epigenetics (biological age)

- dataset/cohort/demographic
- algorithms
- result/stat it produces
- limitations

1.5.2 DNAmPACKYRS

Lu et al. proposed DNAmPACKYRS as a DNA methylation based score for calculatign pack years. This score was originally developed as surrogate biomarker for use in the DNAm GrimAge and DNAm GrimAge v2 epigenetic clocks [14, 15]. DNAm GrimAge is a regression model for estimating mortality risk. The DNAm GrimAge score is calculated using covariates sex, age, 7 surrogate biomarkers of plasma proteins, and of relevance, the surrogate biomarker for smoking pack-years: DNAmPACKYRS. Elastic Net regression was used to train both DNAm GrimAge and the surrogate biomarkers of plasma proteins and pack-years. Training data consisted of 1731 individuals from the Framingham Heart Study [16]. The intersection of sites available on Illumina 450k and Illumina EPIC were chosen as the available CpG sites for training. This was to ensure compatability and future-proofing with new datasets. A total of 450161 CpG sites were available. 10-fold cross-validation was used for hyperparameter tuning the regularisation strength of the Elastic Net model. This resulted in the DNAmPACKYRS score which used 172 CpG sites.

- improvement vs self-reported data
- limitation 1: poor performance of never vs ex and/or ex vs current
- limitation 2: see mCigarette

1.5.3 mCigarette

- (?) seeking to address not being able to differentiate never smokers and ex smokers
- 17865 indiviudals
- filtered CpG sites to FDR < 0.05 ($n = 18760$)
- elastic net regression
- 10-fold cross-validation, $\lambda = 0.012577$
- 1255 CpG sites used in model
- in validation cohort, AUCs of Current vs Never: 0.98, Current vs Former: 0.90, Former vs Never: 0.85
- limitation 2: choice of ground truth (sr pack-years) is potentially confusing for the model

1.6 Aim of This Work

- used smoking status label instead of pack-years
- try improve on classification of never vs ex and ex vs current

2 Method

3 Results

References

[1] World Health Organization. *Tobacco*. Accessed: 2024-11-04. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/tobacco.

[2] U.S. Department of Health and Human Services. *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. Accessed: 2024-04-24. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2014. URL: https://www.ncbi.nlm.nih.gov/books/NBK179276/pdf/Bookshelf_NBK179276.pdf.

[3] Gemma M. J. Taylor and Marcus R. Munafò. "Does smoking cause poor mental health?" In: *The Lancet Psychiatry* 6.1 (2019). Accessed: 2024-11-04, pp. 2–3. DOI: 10.1016/S2215-0366(18)30459-0. URL: https://doi.org/10.1016/S2215-0366(18)30459-0.

[4] National Cancer Institute. *Definition of Pack Year*. Accessed: 2025-04-29. 2024. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/pack-year.

[5] Myung Bae Park et al. "The Correlation of Different Cotinine Levels With Questionnaire Results: A Comparative Study for Different Measurement Methods of the Adolescent Smoking Rate in Korea". In: *Asia-Pacific Journal of Public Health* 27.5 (July 2015). Epub 2015 Jan 1, pp. 542–550. DOI: 10.1177/1010539514565447.

[6] Centers for Disease Control and Prevention. "Disparities in Adult Cigarette Smoking — United States, 2018". In: *Preventing Chronic Disease* 16 (2019). Accessed: 2024-11-04, p. 180553. DOI: 10.5888/pcd16.180553. URL: https://www.cdc.gov/pcd/issues/2019/18_0553.htm.

[7] Tathiane M Malta et al. "Machine learning identifies stemness features associated with oncogenic dedifferentiation". In: *Cell* 173.2 (2018), pp. 338–354.

[8] Péter Adorján et al. "Tumour class prediction and discovery by microarray-based DNA methylation analysis". In: *Nucleic acids research* 30.5 (2002), e21–e21.

[9] Meeshanthini V Dogan et al. "Integrated genetic and epigenetic prediction of coronary heart disease in the Framingham Heart Study". In: *PloS one* 13.1 (2018), e0190549.

[10] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. "Exploratory undersampling for class-imbalance learning". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2 (2008), pp. 539–550.

[11] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320.

[12] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[13] Arthur E Hoerl and Robert W Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67.

[14] Ake T Lu et al. "DNA methylation GrimAge strongly predicts lifespan and healthspan". In: *Aging (albany NY)* 11.2 (2019), p. 303.

[15] Ake T Lu et al. "DNA methylation GrimAge version 2". In: *Aging (Albany NY)* 14.23 (2022), p. 9484.

[16] Thomas R. Dawber, Gilcin F. Meadors, and Felix E. Moore. "Epidemiological Approaches to Heart Disease: The Framingham Study". In: *American Journal of Public Health and the Nations Health* 41.3 (1951). PMID: 14819398, pp. 279–286. DOI: 10.2105/AJPH.41.3.279. eprint: https://doi.org/10.2105/AJPH.41.3.279. URL: https://doi.org/10.2105/AJPH.41.3.279.