Identifying individuals at-risk of developing oesophageal adenocarcinoma through symptom, risk factor and salivary biomarker analysis

David George Graham

A thesis submitted for the degree of Doctor of Philosophy (Ph.D.)

2020

Gastroenterological Intervention Centre

Division of Surgery and Interventional Science

University College London

Statement of originality

I, David George Graham, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in my thesis. In particular I would like to acknowledge the contribution of Mr Saif Khan and Ms Rachel Wellman in their assistance in the laboratory work, Dr Rifat Hamoudi for his in contribution in the bioinformatics analysis and Professor Avi Rosenfeld for his Artificial Intelligence analysis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Acknowledgements

I would like to express my immense gratitude to Professor Laurence Lovat who not only provided me with the fantastic opportunity to undertake this research degree but has also offered great support throughout my career.

I am extremely grateful to Professor Rifat Hamoudi who demonstrated great patience and provided fascinating insights throughout the years we worked together. I look forward to continuing our research collaboration in the future.

I would like to thank Mr Saif Khan for all his help with the laboratory work. I greatly enjoyed working with him and appreciate the time he took to teach me. I wish him great success in the future.

Finally and most importantly I would like to acknowledge and thank my incredible family. My wife, Emma, who never ceases to amaze me with how truly exceptional she is and how much love, support and encouragement she gives us all. And to my three boys Sam, Josh and Max who I am so proud of and who provide Emma and I with so much joy and happiness.

Abstract

Background: Oesophageal adenocarcinoma (OAC) carries a grave prognosis. Existing early detection strategies are flawed predominately because of reliance upon symptoms known to occur late when the disease is often incurable. Detection of individuals with Barrett's Oesophagus (BO), a known pre-malignant condition, is problematic and the vast majority will not develop OAC.

Aim: To explore novel methods of identifying patients with or at risk of OAC through machine learning (ML) techniques and biomarker identification.

Materials and Methods: Initial work utilised novel ML on two existing patient symptom and risk factor questionnaire datasets. Additionally, targeted expression analysis was performed to establish whether transcriptomic biomarkers were present in blood and saliva of affected patients. Optimal RNA extraction techniques and saliva collection strategies for sufficient quality and quantity RNA were determined. Whole mRNA sequencing was performed on patient salivary RNA to identify biomarkers for future assessment. Epigenetic analysis was performed on salivary DNA to identify biomarkers. ML techniques analysed these data to derive a risk prediction tool.

Results: ML techniques on questionnaire data produced satisfactory sensitivity (90%), but accuracy not appropriate for population screening (AUC 0.77). Blood and saliva extraction and collection methods were established and samples found to contain biomarkers. Targeted transcriptomic expression analysis demonstrated 12 / 22 tested genes were significantly aberrantly expressed in patients. 5 genes, combined with 6 questionnaire data-points, identified those with or at risk of OAC 93% sensitivity, AUC 0.88. Whole mRNA sequencing identified a further 134 genes implicated in OAC pathogenesis requiring future testing. Epigenetic analysis found 25 differentially methylated regions, when combined, identified those with or at risk of OAC to 99.9% accuracy.

Conclusion: Utilisation of salivary biomarkers is a potentially effective means to identify individuals with or at risk of OAC. Further work exploring transcriptomic and epigenetic data established in this thesis should be performed.

Impact Statement

There has been little improvement in the grave prognosis of oesophageal cancer despite the advances in endoscopy, surgery and oncology. Key to this are the issues surrounding the difficulty in detecting this cancer in its early stage due to the late presentation of symptoms. It is therefore of paramount importance to develop a means to identify individuals with or at-risk of developing oesophageal cancer without a reliance on symptoms.

The work in this thesis explores using novel machine learning techniques to analyse symptom and risk factor data derived from patient questionnaires and also for biomarker analysis obtained from salivary RNA and DNA to identify these individuals. This work has the potential to significantly alter the grave prognosis of oesophageal cancer through early detection. It could lead to a change in how we screen for this and potentially other diseases and result in less invasive tests such as endoscopy being performed. This would mean less unnecessary investigations being performed and through earlier stage of detection fewer operations and fewer patients requiring oncological intervention creating a significant reduction in healthcare resource usage. Should we detect patients at an earlier stage then their treatment could consist of endoscopic therapy which would lead to improved quality of life when compared to those who undergo surgery. It is not implausible that biomarkers for other cancers or diseases could be found in patient saliva and thus one could potentially envisage this simple test being used to screen the population for a vast range of disease. At present the predictive tool created in this thesis for the early detection of oesophageal cancer has been patented by University College London (patent number: WO2017137427, https://patents.google.com/patent/WO2017137427A1/en) and could potentially be developed commercially. The work has been presented at national and international conferences. It has led to further funding through research grants from the Rosetrees Trust and CORE Digestive Disorders Foundation and the creation of the SPIT study (Saliva to Predict Disease Risk) which is currently recruiting patients from 11 sites around the UK.

Publications Associated with this Thesis

Sullivan R, Heavey S, Graham DG et al. (2020) An optimised saliva collection method to produce high-yield, high-quality RNA for translational research. PLoS ONE 15(3): e0229791.

Rosenfeld A, Graham D et al. Development and validation of a risk prediction model to diagnose Barrett's oesophagus (MARK-BE): a case-control machine learning approach. Lancet Digit Health. 2020 Jan 1; 2(1): E37–E48.

Pollit V, Graham D et al. A cost-effective analysis of endoscopic eradication therapy for the management of dysplasia arising in patients with Barrett's oesophagus in the United Kingdom. Curr Med Res Opin. 2019 May;35(5):805-815

Graham D et al. Risk of lymph node metastases in patients with T1b oesophageal adenocarcinoma. A retrospective single centre experience. World J Gastroenterol. 2018 Nov 7;24(41): 4698-4707

Puccio I et al. Immunohistochemical assessment of Survivin and BCL3 expression as potential biomarkers for NF-kB activation in the Barrett's metaplasia-dysplasia adenocarcinoma sequence. Int J Exp Path. Feb 2018

Graham D et al. Monitoring the premalignant potential of Barrett's oesophagus' Frontline Gastroenterology 2016;**7:**316-322.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AGA	American Gastroenterological Association
AI	Artificial Intelligence
BO	Barrett’s oesophagus
BMI	Body Mass Index
BO	Barrett’s oesophagus
bp	Base-pair
BSG	British Society of Gastroenterology
cDNA	Complementary DNA
cfDNA	Circulating free DNA
CpG	Cytosine-Guanine site
Ct	Cycle threshold
CTC	Circulating tumour cells
DNA	Deoxyribose nucleic acid
DMR	Differentially methylated region
EMR	Endoscopic mucosal resection
ePCR	Emulsion PCR
FDR	False discovery rate
FFPE	Formalin fixed paraffin embedded
GC	Guanine-cytosine
GORD	Gastro-oesophageal reflux disease
HGD	High grade dysplasia
IMCa	Intra-mucosal cancer
ISP	Ion sphere particles
LGD	Low grade dysplasia
miRNA	Micro-RNA
mRNA	Messenger ribonucleic acid
NDBO	Non-dysplastic Barrett’s Oesophagus
NFW	Nuclease free water
NICE	National Institute for Health and Care Excellence
NTC	No template control
OAC	Oesophageal adenocarcinoma
PCR	Polymerase chain reaction
qRT PCR	Quantitative real-time polymerase chain reaction
RFA	Radiofrequency ablation
RIN	RNA integrity number
RNA	Ribonucleic acid
RNA-seq	Whole mRNA sequencing
SCC	Squamous cell carcinoma
Tm	Melting temperature

Chapter 1

Oesophageal cancer and the existing early detection strategies

Chapter 1 - Oesophageal cancer and the existing early detection strategies.

1.1 The epidemiology and prognosis of oesophageal cancer

The incidence of oesophageal adenocarcinoma (OAC) has risen significantly across all Western populations since the 1970's. This rising incidence has been so significant that it has been reported in literature and media outlets as a 'quiet epidemic'[4, 5].In the UK there has been a 43% rise in oesophageal cancer since 1970 [5].There has been an interesting shift in the epidemiology of oesophageal cancer. Squamous cell carcinoma of the oesophagus (SCC) remains the most common histological type of oesophageal cancer worldwide, accounting for 87% of cases, but in Western populations OAC is now the predominant subtype [6]. Epidemiological studies suggest that this steep rise in OAC incidence appears to have plateaued although there is conflicting literature suggesting that although the rise has slowed, from 8% per year to 2% per year, there remains a statistically significant increase [5-8].

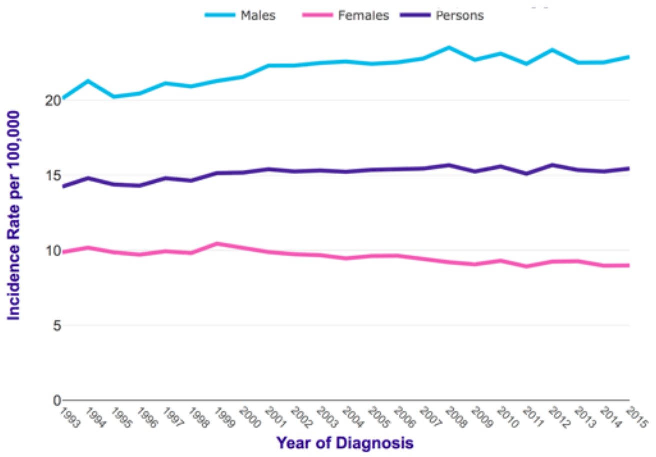


Figure 1: UK oesophageal cancer incidence rates 1993 – 2015 [5]

One must remember that although OAC incidence is rising, it remains a relatively uncommon cancer. Overall, oesophageal cancer is the 13th most common cancer in the UK[5]. Importantly, however, despite its relative uncommon occurrence, it is the 6th most common cause of cancer death [5].This is largely due to the combination of the bleak prognosis of late stage disease and the stage at which oesophageal cancer is diagnosed. *Schlansky et al* demonstrated that upon diagnosis 89% of oesophageal tumours were either at T3 or T4 stage [9]. Although it has been improving over the last 20 years, the overall 5-year survival remains only 16.5% [5].Thus the vast majority who develop oesophageal cancer will die of the disease. Not only are the costs of oesophageal cancer great in terms of lives lost but also in terms of cost of NHS resource. A report published by Cancer Research UK in 2014 demonstrated the significant costs associated with treatment for late stage cancers in comparison to that of early stage therapy and the significant savings that could be made should cancer be diagnosed at an earlier stage [10].

1.2 Barrett's oesophagus and oesophageal adenocarcinoma

Barrett's oesophagus (BO) is defined, according to the British Society of Gastroenterology (BSG), as; *"An oesophagus in which any portion of the normal distal squamous epithelial lining has been replaced by metaplastic columnar epithelium, which is clearly visible endoscopically (≥1 cm) above the gastro-oesophageal junction and confirmed histopathologically from oesophageal biopsies".* A British surgeon in 1950 named Norman Barrett described the columnar-lined oesophagus that now bears his name although the lesion had been described 50 years earlier by a pathologist named Tileston who noted the similarities between the mucosa of the stomach and that of the diseased oesophagus. Barrett maintained for many years that the condition was congenital, due to a short squamous oesophagus, although he did recognise a link between its presence and the presence of a hiatal hernia and severe oesophagitis [11]. It was not until the mid-1970's that it became widely accepted that BO was associated with severe gastro-oesophageal reflux disease (GORD) [12, 13].The importance of BO is now widely recognised. BO is the only known precursor lesion to OAC and it is thought that the majority of all OAC arise from BO [14, 15].

The prevalence of BO within the population is difficult to determine as we do not presently offer universal endoscopic screening. There are two European studies that have attempted to ascertain the prevalence of BO by offering an endoscopy to an unselected adult population. These studies performed an endoscopic examination on over 1000 patients each and found the prevalence to be 1.3% and 1.6% [16]. However, there is concern that both these studies have a selection bias due to the low participation rate. In other work attempting to ascertain the prevalence of BO the estimates have ranged from 0.4% to 25% [17]. Given the issues in knowing the true prevalence of BO, it is difficult therefore to accurately state what risk BO incurs to an individual. Although we state that the majority of OAC arises in BO, the vast majority of OAC occurs in individuals with no previous diagnosis of BO [18]. Irrespective of the true prevalence and risk, it is fair to say that although BO may be present in the majority of OAC cases, the majority of BO does not progress to OAC. At present we determine an individual’s risk of their BO progressing to OAC based on the histopathological findings of either metaplasia or dysplasia.

1.2.1 Metaplasia-Dysplasia-Carcinoma Sequence

OAC arising from BO is thought to occur in a stepwise progression, recognised histologically, named the metaplasia-dysplasia-carcinoma sequence [19]. Metaplasia originates from the Greek language and translates to a “change in form”. This change in form is classically thought to be in response to chronic tissue damage resulting in regeneration. In BO the tissue damage is predominantly thought to be an association with the consequences of acid-reflux in GORD[16]. Metaplasia can also occur due to a selective overgrowth of the minor cell types originally contained in an organ, through the differentiation of stem cells or by a switch occurring in pre-existing differentiated cells [20]. In BO the normal squamous epithelial lining of the oesophagus is replaced by a mosaic of columnar epithelia with or without goblet cells[21]. A type of metaplasia known as intestinal metaplasia has been demonstrated in a multitude of studies to be biologically unstable and carry a significant risk of progression to OAC [16]. In response to this the American Gastroenterological Association (AGA) state that the presence of intestinal metaplasia is required for the diagnosis of BO as it is the ‘only type of columnar oesophageal epithelium that clearly predisposes to malignancy’[22]. However, this view is controversial and is at odds with the BSG with both societies recognising that there is a paucity of evidence on the risk associated with other forms of metaplasia [16, 22]. This issue was recently addressed by *Lavery et al* who demonstrated that cardia type metaplasia without goblet cells in BO carried malignant potential, a finding which is in keeping with other published work [21, 23, 24]. Thus, whilst the type of metaplasia associated with OAC in BO remains controversial, it is fair to say that the presence of metaplasia is regarded as the first step in the progression towards OAC. Patients with metaplastic BO (otherwise known as non-dysplastic Barrett's oesophagus, NDBO) are regarded as having an increased risk of OAC and as such undergo lifetime surveillance endoscopies in the UK and worldwide (discussed in 1.3.2). However, it is important to remember that despite this recognised increased risk the actual risk of progression in these individuals remains low. In three large cohort studies, a Dutch study with 42,207 patients, a Danish study with 11,028 patients, and a Northern Irish study with 8522 patients, the annual risk of progression ranged from 0.12% to 0.4% [25].

At present, the only robust, routinely used, clinical marker of progression towards OAC is the presence of dysplasia found on histological analysis of biopsy sampling of BO. According to the revised Vienna classification, dysplasia is divided into two groups, low-grade and high-grade dysplasia, although one should note the existence of those determined as indefinite for dysplasia in which the morphological features are blurred and the differentiation between dysplasia and inflammation is difficult [16, 26]. Despite dysplasia being our most robust marker there remain issues with its use, not least the difficulties faced by a histopathologist in making the diagnosis. A high level of intra-observer and inter-observer variability has been demonstrated in many studies thus suggesting that a more robust, consistent method of recognising those who have progressed, or are at risk of progressing, is required [26-29]. It is worth noting also that BO is a heterogenous disease with a range of progression from NDBO to HGD seen on surgical OAC specimens. Therefore, representative sampling is key.

Low grade dysplasia (LGD) is characterised by the glandular architecture being relatively preserved with the diagnosis made on the basis of cytological atypia. One of the key features is the evidence of a loss of "surface maturation" where cytological atypia seen in deeper glands moves into the surface epithelium.

Other key features include elongated, enlarged and hyperchromatic nuclei, mild pleomorphism, mucin depletion, mild loss of polarity, nuclear crowding, and stratification of nuclei up to three-quarters of the height of the cell, but not touching the luminal surface [16, 26].The diagnosis of LGD can be challenging for a histopathologist. This was demonstrated by *Duits et al* who performed a large retrospective analysis of patients diagnosed with LGD in which only 27% of the cases were confirmed upon consensus review of the biopsies by expert histopathologists. Importantly, this study also demonstrated that those with a confirmed diagnosis of LGD had an annual progression rate to high-grade dysplasia (HGD) or OAC of 9.1%. In those who were down-staged to NDBO or indefinite for dysplasia the progression rate fell to 0.6% and 0.9%. This demonstrated the increased malignant potential of LGD whereas as previously there had been debate over its natural history with progression rates in other studies varying from 0.6% to 13.4% thought to be due to the issues of inter-observer variability [30, 31]. In response to this study and other published work, including studies on therapeutic intervention in LGD, the BSG updated their guidelines to reflect the established significant risk of progression of LGD and recommended those with a confirmed LGD diagnosis (on two occasions) be offered therapy (discussed in 1.2.2).

In contrast, whilst the exact rates of progression have been debated, there has been little debate over the significant risks of progression of HGD to OAC. In HGD there are marked architectural changes alongside further nuclear atypia. These changes include a papilliary or villous surface alongside branching, complex budding or back to back crowding arrangements [16]. In addition to this, intraluminal papillae, bridges, and cribriform patterns are also seen as well as atypical mitoses, together with mucin depletion and a loss of nuclear polarity[16]. As mentioned previously, the diagnosis of HGD may also be problematic, to a lesser extent than with LGD, with the continued observer variability, although agreement between pathologists rises to as high as 80% [27-29]. There is agreement that the risks of progression of HGD to OAC are significant. Those risks range from 16-59% at 5 years according to studies with the actual risk thought to be more in the region of 50% as per *Buttar et al and Reid et al* [32-34]. In addition to the risks associated with HGD it is also reported that up to 12-75% of those diagnosed with HGD are found to have occult adenocarcinoma in surgical resection specimens [35, 36], although it should be noted that these statistics were obtained through analysis of older studies prior to the advances in endoscopic imaging and resection techniques. It is reasonable to assume that these data would be improved now. As a consequence of this risk, it has long been accepted that patients with HGD should be offered therapy. In 2010 the UK National Institute of Clinical Excellence (NICE) published guidelines approving the use of radiofrequency ablation (RFA) and endoscopic mucosal resection (EMR) in treating HGD in BO [37].Since 2013 endoscopic therapy for HGD and early non-invasive cancer (intramucosal cancer, IMCa) has been recommended by the BSG[16]. This was in response to the excellent results from studies using endoscopic resection and ablative therapies in BO, not least the US and UK HALO Registries, and the significant morbidity and mortality associated with oesophagectomy. The 2012 UK National Oesophago-Gastric Cancer Audit found that the intraoperative mortality for all patients undergoing an oesophagectomy was 2-4%, although it should be noted that in those having surgical intervention for HGD the mortality was less, although still 1%. Additionally, 40% of patients faced not insignificant morbidity with the normal quality of life returning only after nine months [38]. Although still significant, these surgical statistics are an improvement on the 10% mortality patients faced in the 1990’s [39]. As a consequence of the associated morbidity and mortality with this operation, there was a drive to find a less invasive answer to treating those with high-risk pre-malignant HGD and early cancers. The promising data associated with endoscopic therapies resulted in a change in the treatment algorithm for patients with dysplastic BO and intra-mucosal cancer (IMCa) with endoscopic resection and ablative techniques becoming first line therapies [16]. Additionally, it is generally accepted that in selected patients endoscopic therapy can be used for those who have a cancer that has only reached the first part of the submucosa (known as Sm1 disease) [40]. In these individuals, who have no other adverse pathological features, the risk of lymph node metastases is comparable to the 30-day mortality of an oesophagectomy. In those in which the lesion extends deeper than SM1, or have other adverse pathological features, the risk of lymph-node metastases does rise accordingly and surgery is recommended, if appropriate[41-44].

1.2.2 Therapeutic interventions in Barrett's oesophagus

In 2006 *Dunkin et al* published work demonstrating complete removal of the oesophageal epithelium, without damage to deeper structures, using RFA on patients prior to them undergoing an oesophagectomy [45]. This work initiated further high-quality trials exploring the use of RFA in the treatment of dysplastic BO. The AIM dysplasia trial was the first multi-centre, randomised-control study looking at the use of RFA for BO versus a sham procedure. In this study, at 12 months, 90% of patients with LGD and 81% of patients with HGD had complete eradication of dysplasia compared to 22% and 19% of the patients receiving the sham procedure, respectively. Following on from the AIM dysplasia trial, further large-scale studies were performed in the US, across Europe, and within the UK with the eradication of dysplasia in these studies ranging from 86% to 96% [46-48].

There is little doubt that RFA in BO is an effective treatment. However, over time our skills to both recognise and treat lesions have also improved, helped by evolving endoscopic technologies. As such we have seen a paradigm shift in the therapeutic management of dysplastic BO and IMCa. The UK HALO Registry was founded by Professor Lovat in 2008 with the intention of collecting real-life data (rather than data collected within the confines of a clinical trial) on patients undergoing therapy for their BO. The first results were published in 2013 [48]. In 2014 *Haidry et al* demonstrated that over time our practice had altered to include EMR of any visible lesions prior to RFA. This shift in management has resulted in improved outcomes for patients [38]. In those treated between 2008-2010 the rates for the clearance of dysplasia were 77% compared to 92% in those treated between 2011-2013 [38]. This paper identified a significantly higher number of EMRs performed prior to RFA in the 2011-2013 group, although it may be that some of this improved dysplasia clearance can also be attributed to improving skills in delivering RFA. It is now accepted practice that resection of all visible lesions should be performed prior to RFA to achieve the optimal outcomes.

Importantly, also, endoscopic therapy is safe and durable. The AIM dysplasia trial demonstrated that 98% of patients were free of dysplasia at 3 years [49]. The UK HALO registry showed that 94% of 270 patients were dysplasia free at

19 months and 92% free in a paper looking at 6-year outcomes[38, 48]. *Haidry et al* also demonstrated that the cancer progression rate in the individuals receiving endoscopic therapy was 2-4%, which is similar to that demonstrated by *Orman et al* [38, 50]. Additionally, it should be noted that endoscopic therapy is a day-case procedure with low complication rates. Large scale studies have shown that the rates of perforation are 0.5% in the hands of experienced physicians and significant bleeding occurred in 1.2% [51, 52]. Stricturing of the oesophagus is problematic in both RFA (6%) and EMR (up to 41-88% in circumferential EMR's) [51, 52]. Importantly, NICE demonstrated in their 2010 analysis that these therapies are cost-effective [37]. Recently, *Filby et al* analysed endoscopic therapy in HGD and modelled it to be cost-effective and *Pollit et al* suggested that endoscopic therapy was cost-effective in all types of dysplasia [53, 54].

1.2.3 Screening for Barrett's oesophagus

Given the risks of BO and the time taken to move from BO to OAC, it has long been discussed whether either the whole population or only selected at-risk groups within the population should be screened. This remains a controversial issue to this day. The first aspect to the debate is trying to ascertain the exact prevalence of BO within the population. This is problematic for many reasons, not least because at least 25% of individuals with BO are thought to be asymptomatic [55]. As stated earlier, two European studies have estimated the prevalence of BO to be around 1.3%-1.6% [16]. Secondly, it should be noted that the incidence of OAC is too low for it to be classified as a major health problem and as such the risk of dying from cancer even in those with BO is also low [17]. Thirdly, the cost of screening the whole population would be vast, also because those who are found to have BO would then be entered into a surveillance programme in which they would continue to have endoscopic surveillance incurring more cost. The benefits of these surveillance programmes are also dubious (discussed in 1.4). Consequently, the BSG guidelines state that it is neither feasible nor justified to screen an unselected adult population [16]. These guidelines do state that screening can be considered for certain high-risk groups linked to key risk factors associated with BO and OAC, such as male sex, age over 50 years, white race, obesity, and GORD. When considering screening these groups, one should also remember the unwanted consequences such as patient anxiety, unnecessary follow up examinations caused by false positives, and the potential difficulties in obtaining life insurance and other types of insurances. Given that endoscopic examination is the only current diagnostic technology available to the population at present, screening does not appear to be practical or cost-effective.

Research into answering these issues with devices such as transnasal endoscopy and the Cytosponge are showing some promise [16]. Cytosponge in particular has attracted attention as a minimally invasive cell sampling device that is swallowed by the patient and then pulled back through the oesophagus. This is further discussed in section 1.4.3.

1.3 Risk factors for oesophageal adenocarcinoma and Barrett’s oesophagus

Clearly, if we aim to improve the bleak prognosis of OAC we need to identify individuals with OAC at an earlier, curable stage, ideally when the minimally invasive, low-risk, effective therapies that already exist can be applied. This task would be aided by identifying those in the population who are at risk of OAC through the development of BO who could be closely monitored and treated using the same endoscopic therapies when appropriate. The identification of these individuals is difficult as the vast majority are asymptomatic or have symptoms indistinguishable from benign disease [9]. As such we need to find novel ways to identify these people. In order to do this, we need to understand what puts an individual BO patient at risk. This is a complex interplay between our genetics (discussed in Chapter 2), diseases we develop such as GORD or obesity, and environmental exposures such as cigarette smoking. This interplay is nicely highlighted by work demonstrating that a polymorphism in the gene encoding the insulin-like growth factor receptor and obesity increases an individual’s risk of oesophageal cancer. However, those with the polymorphism alone were at no increased risk [56]. As such, in this section we explore some of the key risk factors associated with BO and OAC. The genetics, epigenetics and transcriptomics involved in BO and OAC are discussed in chapter 2, but clearly play a role alongside the below outlined risk factors.

1.3.1 Gender

An individual's gender is known to impact on many aspects of disease, and cancer is no exception. Generally speaking women are more commonly affected by autoimmune diseases and in men the incidence of cancer and infections is higher. In addition, cancer survival is poorer in men [57]. The reasons for this are not entirely clear. BO occurs at a ratio of 2:1 in men to women and almost 70% of cases of OAC in 2013 were in men [5, 25]. The general perception of this increased cancer risk in men is due to men engaging in more high-risk behaviours than women. For example, a greater proportion of men smoke or are obese. However, although these factors certainly play a role in the increased risk they do not completely explain the phenomenon. There are some known X chromosome-linked genetic mutations that explain some of the increased risk in men. There are other genetic mutations seen only in men that gain increase cancer risk. There is also an array of proposed physiological differences that may contribute to sex discrepancy in various cancers. Included within this are the hormonal differences, in particular oestrogens and androgens, that also play a role in cancer susceptibility. For example, higher oestrogen levels are thought to be protective in gastric and liver cancer [57, 58]. In addition, there is published literature, mainly in mice, showing significant differences in gene expression and protein production in men and women in genes linked to many diseases including cancer. Furthermore, women appear to produce higher innate and adaptive immune response than men and these immune differences may explain why certain diseases have a sex discrepancy [57].Finally, there is also thought to be a sex difference in the tissue response to carcinogens which is thought to underlie, at least in part, why there is a significant rise in the incidence of lung cancer in women [59].

1.3.2 Age

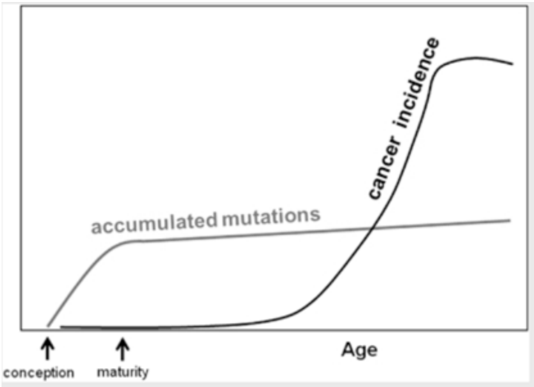


Figure 2: Comparing time courses of the onset of malignancy and accumulation of mutations in mice [3]

It has long been known that increasing age is a leading risk factor for many cancers and their pre-cancerous lesions, including OAC and BO. The reasons for this remain unclear. Specifically related to OAC and BO, the incidence of BO increases with age with the average age at diagnosis being 60 years-old whereas the average age of OAC diagnosis is 68 [5, 17]. The general view is that the development of cancer is dependent on the sequential accumulation of epigenetic alterations and oncogenic genetic mutations. The incidence of these mutations increases with age. *Armitage and Doll* published work reporting that an individual needs to develop 6-7 age-dependent, oncogenic mutations in order for cancer to develop [3]. This time-dependent view has been challenged. For example, *De Gregori* observed that the cancer incidence increases quickly in later life, but the accumulation of oncogenic mutations occurs at a maximal rate during puberty (see figure 2). Therefore, one would expect to observe more cancers occurring at a younger age. *De Gregori* further points out that oncogenic mutations are present in large numbers in healthy cells at an incidence that is greater than the incidence of cancer within these cells. He concludes that it is not necessarily the incidence of mutations that is key, although clearly mutations are required, but the deterioration of the mechanisms in place for the body to combat these mutations[3]. Recently *Xu et al* published strong evidence for the role of the epigenome in the age-related increased cancer incidence. Their work looking at 27,000 age specific methylation sites in 1000 women that was then compared to additional datasets and to seven different cancer types from “The Cancer Genome Atlas”, found that 70-90% of the age-related sites they identified showed significantly increased methylation in all seven of the cancer types. This suggests that age-related methylation changes play a significant role in expression of certain genes leading to oncogenesis [60]. It is fair to say that the mechanisms behind the age-related increased cancer incidence remain debated, although it is undoubtedly a key risk factor for many cancers.

1.3.3 Race

There is huge disparity in the incidence of oesophageal cancer depending on ethnicity. Worldwide oesophageal SCC is the most common histological subtype, seen mainly in Africa and what has been dubbed the 'Asian oesophageal cancer belt' (an area extending from Eastern Turkey, through Iraq, Iran, and the southern part of the former Soviet Union, including Kazakhstan, Turkmenistan, Uzbekistan, Tajikistan, to Mongolia and Western/Northern China). Whereas in Western populations, as discussed earlier, OAC is the most common subtype with an prominent rise in recent decades. Focusing on BO and OAC within the UK, somewhat unsurprisingly the ethnic variations in incidence of BO are echoed in OAC. The incidence of both diseases is significantly higher in white men and women. For OAC, age-standardised rates in the UK are 13.9-14.4 per 100,000 and 5.5-5.7 per 100,000 in white men and women, respectively. For black males the rates are 6-10.2 per 100,000 and for black females 2.5-4.2 per 100,000. The rates of OAC in Asian women is almost identical to that of black women and in Asian men the rate is 3.6-6.1 per 100,000 [5]. Again, the reason for these variations is not clear. Certainly, there are behavioural differences as well as contributing disease incidences, such as obesity and GORD, seen between different populations around the world and ethnic groups within a country that are likely to be a factor. This is demonstrated in the sharp rise of OAC incidence in China which may be linked to the increasing tobacco consumption or Western diet [61]. There are also certainly likely to be genetic factors involved, with genes linked to increased risk of both oesophageal SCC and OAC identified in genome wide association studies within the populations in which these respective cancers have high incidence [61]. However, there is little in the way of comparative studies between the various high-risk populations to determine what genetic reasons there may be behind differential cancer incidence across populations [61].

1.3.4 Gastro-oesophageal reflux disease

As discussed earlier, Norman Barrett who originally described BO, initially thought that this was a congenital abnormality. Despite the observation of the increased incidence of factors such as oesophagitis and hiatus hernia, the debate over its origin continued for many years until 1970 when *Bremnar et al* demonstrated the development of columnar-lined mucosa in an experimental model of GORD [62]. It is now accepted that BO is a complication of GORD. BO is found in approximately 10-15% of individuals undergoing an endoscopy for GORD and the odds ratio for BO in patients with symptoms of GORD for 1-5 years is 3.0 which rises to 6.4 in those who have had symptoms for over 10 years [17]. Similarly, the presence of at least weekly GORD symptoms has an odds ratio of 7.7 for the development of OAC. Clearly, therefore, the severity and longevity of GORD is a major risk factor for development of BO and OAC which is why, when discussing the possibility of screening for these conditions, GORD symptoms are often highlighted [17]. Acute mucosal injury alone is not sufficient to induce the development of BO as acute damage is often followed by squamous regeneration. Moreover, a chronic abnormal environment is required for the columnar mucosa to develop, thought to be a protective change in response to chronic reflux. Hence the incidence of BO and OAC increases with the duration of GORD. Support for this need for chronic exposure is demonstrated with the observation that increased oesophageal acid exposure is seen in over 90% of individuals with BO, the proximal progression of BO over time, and the induction of BO in experimental models [62].

1.3.5 Obesity

The detrimental effects of the worldwide obesity epidemic on population health and healthcare resources is a significant problem. In the UK, the prevalence of obesity has doubled in the last 25 years. Currently 25% of the adult population is obese (as defined by a Body Mass Index >30) with these numbers predicted to rise to 40% of adults and 25% of children by 2030 [5, 63]. The association between obesity and the development of diseases such as T2DM, heart disease, pulmonary disorders, and some cancers (including gastrointestinal, hepatobiliary, kidney and breast) is well documented [5, 64, 65]. It is thought that more than 5% of all cancers are attributed to obesity with obesity associated with the pathogenesis of 39% of OAC cases. Overall obese patients with cancer have a worse prognosis than those of normal weight [66]. The incidence rates of OAC have risen 65% in men and 14% in women in the UK since the 1970’s[5]. This is in parallel with the obesity epidemic [56, 64]. It is stated that there is a 2 to 3-fold increased risk of BO and OAC in obese individuals, but some studies have suggested up to a 10-fold increased risk [56, 67].

Initially the link between obesity and BO and OAC was thought to be due to the mechanical alterations that occur in obesity that predispose an individual to GORD. Firstly, it was noted that there is a higher prevalence of hiatal hernias in obese individuals and secondly there is an increase in intra-abdominal pressure in obesity which displaces the lower oesophageal sphincter. The combination of these factors was thought to increases the likelihood of GORD and thus cause the increased risk of BO and OAC [68]. It is now accepted, however, that this view has been too simplistic. There have been several high-quality, large-scale studies exploring the link between obesity and GORD with mixed results. Some have shown a modest association at most, whereas others have demonstrated no significant association [69]. Consequently, focus has shifted to the endocrine and molecular changes that occur in obesity that place an individual at increased risk not just for BO and OAC, but for the development of numerous obesity-associated diseases.

Obesity causes a derangement of the cellular and molecular mediators of immunity and inflammation resulting in a chronic low-grade inflammatory response [65]. Consequently, inflammatory markers such as TNF-α, C-reactive protein and IL-6 are increased in obese individuals, although not to the same extent as observed in classic inflammatory conditions [70]. Adipose tissue is an active secretory organ involved not just in appetite and energy expenditure but also endocrine and reproductive systems, bone metabolism and inflammation and immunity [70]. It is a major source of chronic inflammation [71]. Inflammation observed in adipose tissue is likely to serve as a feedback signal locally in adipose tissue and systemically for energy expenditure. In adipose tissue, inflammation inhibits adipocyte expansion and differentiation, changes adipocyte endocrine function and induces extracellular matrix remodelling[72]. The local response is translated into a systemic response through cytokines and free acids released from adipose tissue. Hypoxia in adipose tissue induces the secretion of pro-inflammatory adipocytokines (e.g. leptin) and cytokines (e.g. TNFα, IL-6, IL=8, IL-10, IL-1) that promote angiogenesis and upregulate key inflammatory pathways. The products of these pathways are responsible for the transcription of genes that mediate proliferation, invasion, angiogenesis, survival, and metastasis [66]. This chronic low-grade inflammation has also been linked to the pathogenesis of many diseases including insulin resistance, hyperglycaemia, and the development of T2DM, as well as the cause of cardiovascular disease due to dyslipidaemia and hypertension [70, 73]. Insulin resistance, T2DM, and metabolic syndrome have been demonstrated to be significant risk factors for cancer development thought to be associated with the pro-proliferative properties of insulin. High levels of adipose cytokines have deleterious effects on glucose homeostasis leading to chronic hyper-insulinaemia and insulin resistance. This is further worsened by the deregulation of the inflammatory pathways producing increased nitric oxide levels and consequently increasing insulin resistance and further worsening inflammation [66].

Additionally, adipose tissue is infiltrated by macrophages with the number of macrophages found directly correlating to the adiposity of human subjects. These macrophages are known to express growth factors, cytokines, chemokines and proteolytic enzymes involved in the regulation of tumour growth, angiogenesis, invasion and metastatic spread [70, 74]. Adipose macrophages may act similarly to tumour-associated macrophages found in the tumour stroma [74]. Activated adipose macrophages affect adipose tissue function, increasing insulin resistance, thereby enhancing the mitogenic effects of insulin, and increasing inflammatory cytokine production [66].

1.3.6 Smoking

The relationship between smoking and the incidence of BO and OAC go hand in hand. There is an increased risk of BO in those who smoke and this risk increases with greater pack-years [75]. Similarly, the risk of OAC is 85-96% higher in people who have smoked compared to “never-smokers” and, again, this risk increases with higher pack-years. OAC risk is 2.5 times higher in people who smoke over 20 cigarettes per day compared with those who have never-smoked. In regards to smoking cessation OAC risk is 29% lower in ex-smokers who quit over 10 years previously compared to current smokers, however, these ex-smokers remain at considerably higher risk than those who have never smoked (72% higher risk) [5].

The mechanisms behind the increased risk associated with smoking are multifactorial and not fully understood. Many of the carcinogenic contents of tobacco products have been linked to gastrointestinal cancers. There are approximately 60 carcinogens in cigarette smoke that have been demonstrated to be carcinogenic in humans or animals. Of particular note of these are the N-nitroso compounds that are a potent carcinogen in oesophageal cancer [76, 77]. Additionally, it has been demonstrated that cigarette smoke causes the creation of genotoxic DNA adducts that are central to carcinogenesis. A well-documented DNA mutation linked to cigarette smoke is mutation to the p53 gene that is known to be strongly linked with the development of OAC [1, 77].

1.3.7 Family history

Although it is thought that the majority of BO and OAC are sporadic, there is evidence for its clustering within families. The Wellcome Trust performed genome wide association studies to determine the heritability of BO and OAC and published results in 2012 and 2015. Within this work they identified more than 1000 variants contributing to the BO phenotype and estimated a heritability of 9.9%. The Barrett's and Esophageal Adenocarcinoma Consortium published work in 2013 estimating the heritability to be 35% [78]. The work within this field continues, although the requirement for large-scale studies limits its progress.

1.4 The existing early detection strategies for oesophageal adenocarcinoma

The consequences of the grave prognosis of OAC leads to the huge impact it has on worldwide morbidity and mortality as well as healthcare finances. OAC, like all cancers, has a significantly better survival if treated in its earlier stages. As such it is logical that attempts are being made to detect it earlier on. Within the UK there are three strategies currently being adopted that are either general to all cancers or specific to OAC alone. These are discussed in the section below.

1.4.1 The "two-week wait" expedited referral pathway

Reacting to UK cancer mortality rates amongst the highest in Europe and long waiting lists, the two-week wait referral system was introduced at the end of the 20th century [79, 80]. This guaranteed that institutions would face financial penalty, unless patients with symptoms suspected to be consistent with a cancer diagnosis were seen by a specialist within two weeks and receive a diagnosis and commence treatment within further time restrictions. There was little scientific foundation to its introduction and was seen by some as a political public relations exercise rather a measure to improve the standards of care in the UK [81]. Consequently, hospitals have been flooded with two-week wait referrals and the system is poorly adhered to within primary care. Analysis of the two-week wait system since its introduction has given weight to those who were sceptical. A wealth of evidence demonstrates that there is a low yield of cancer diagnosis through the two-week wait pathway across all cancers. Specific to the gastrointestinal tract only 9% of two-week wait referrals for colorectal cancer and 6% for oesophageal cancer yielded a cancer diagnosis [82, 83]. It is important to also note that this referral system does not even capture the majority of cancers with approximately two-thirds of cancers diagnosed via an alternate route [80, 82]. Thus, it is felt that a disproportionate amount of healthcare resource is invested in supplying demand for a system that appears to be of little benefit. Similarly, there is little evidence that the two-week wait referral system leads to diagnosis at an earlier stage and, in fact, data demonstrates that it offers no improvement in survival benefit [80, 82, 84]. Specific to OAC analysis of survival outcomes show that patients whose cancer is diagnosed following referral for routine gastroscopy have a better prognosis than those referred through the two-week wait pathway, although referral bias may be an issue in this analysis [79, 85].

As already discussed, symptoms occur late in OAC. Symptoms of early OAC or its pre-malignant form (BO) are often either absent or indistinguishable from benign disease. As such, it is illogical to think that the two-week wait referral system, based upon symptoms, will be an effective means for early diagnosis in OAC. A vast amount of healthcare resource is invested into the two-week wait system with data demonstrating that referral rates are increasing year upon year [86].In lieu of any evidence supporting its clinical benefit, this is now clearly a flawed approach.

1.4.2 Public Health England “Be Clear on Cancer” campaign

In 2011 Public Health England, working in partnership with the Department of Health, NHS England, and Cancer Research UK, launched the “Be Clear on Cancer” campaign. The idea behind this was to improve early detection of specific cancers through raising the public’s awareness of key symptoms and encouraging the public to seek medical attention should they suffer from any one of these. To date this campaign has addressed bowel, breast, ovarian, lung, kidney, bladder, and oesophago-gastric cancers. The campaign in relation to oesophago-gastric cancers was launched in 2015 with specific focus on heartburn. Once again, this is an attempt to improve early detection of OAC through symptoms. In our opinion, this approach is flawed. There is little data at present for the success of this campaign. However, the pilot campaign noted a significant increase in referrals through the two-week wait pathway following the campaign which was confirmed in one study. This study also demonstrated no improvement in survival or stage of diagnosis [87, 88].

1.4.3 Barrett’s oesophagus surveillance programmes and the emergence of Cytosponge

Several learned societies have issued guidelines on the endoscopic surveillance of BO including the BSG. The BSG advocates surveillance for patients with known BO [16]. As such, people with BO undergo regular endoscopic observation throughout their lives with the aim of identifying progression to dysplasia or OAC at an early, curable stage. Patients with NDBO therefore undergo endoscopic examination every 2-5 years, depending on the length of the BO segment, with close inspection followed by targeted biopsies of any areas of concern and then random four quadrant biopsies every 2cm. It is fair to say, however, that the quality of endoscopic examination and adherence to these guidelines varies greatly throughout institutions within the UK and globally, affecting standards and outcomes of BO surveillance [89]. Thus, the merits of surveillance for BO is not without its controversies.

Retrospective analysis of data on patients undergoing Barrett's surveillance have demonstrated that those undergoing surveillance who developed OAC were found to have an earlier stage of disease and improved survival. This is supported by other studies demonstrating similar findings, however, all of these are vulnerable to both lead and length time bias [17, 90]. In fact, an expert panel in the US concluded that there are no controlled studies to support surveillance. Similarly, a UK governmental review concluded that surveillance may do more harm than good and there are significant gaps in available evidence supporting that surveillance reduces morbidity and mortality from OAC [91]. Population-based studies have demonstrated that overall mortality rate in patients with BO is similar to that of the general population and that OAC is in fact an uncommon cause of death in these patients [91]. It is also of note that less than 10% of OAC occurs in patients with a known diagnosis of BO and that 90% of individuals with BO die of unrelated causes. Additionally, it has been found that 93% of OAC are not detected in current screening strategies and instead present with advanced, incurable disease [92]. Regular surveillance of individuals with BO is therefore unlikely to impact on global prognosis [93]. The role of surveillance for BO is being explored in the ongoing BOSS study in which 3400 patients have been randomised to routine surveillance versus at need endoscopy determined by patient symptoms or concerns. The study will aim to determine if there is an overall survival benefit in either group alongside also providing data on cost effectiveness, cancer-specific survival time, time to OAC diagnosis and stage of diagnosis, morbidity and mortality related to any interventions and frequency of endoscopy. This study will hopefully shed further light on the role of endoscopic surveillance in BO [94].

Cytoponge

In the search for better ways to screen the population to improve detection of BO and OAC the Cytosponge has been developed. This is a minimally invasive non-endoscopic cell collection device which is a 30-mm polyurethane sponge, contained within a capsule attached to a string. The capsule is swallowed and dissolves within the stomach after 3–5 min. The sponge is retrieved by pulling on the string, thus collecting cells on its return. The test has been shown to be acceptable for patients and the cells obtained undergo immunohistochemical labelling for Trefoil Factor 3 (TFF3), which is a biomarker known to be linked to BO as it labels goblet cells in intestinal metaplasia. This biomarker was ascertained from a gene expression study designed to distinguish between BO cells and those from the gastric cardia, squamous oesophagus, and oropharynx, which are serially sampled by the Cytosponge as it is retrieved. *Ross-Innes et al* reported a sensitivity of 79.9% and specificity of 92.4% for detecting BO using this device. The sensitivity increased to 87.2% with two passes of the Cytosponge which obviously incurs acceptability issues [95].

The BEST2 study explored the use of the Cytosponge as a means to risk stratify patients known to have BO in order to determine those at low risk of having progressed and thus not requiring a surveillance endoscopy. This was achieved by combining the biomarkers (P53 abnormality, glandular atypia, and AurKA staining) with clinical variables (patient age, length of BO and obesity) and using logistic regression to determine an individual’s risk. Whilst this study was successful in identifying these low-risk patients it did not demonstrate an ability to detect high-risk individuals [95].

The BEST3 study outcomes have recently been published. This study aimed to detect individuals within the community through primary care with BO by comparing the existing management of patients with reflux symptoms, identified as those on reflux medications, with offering these individuals a Cytosponge examination and consequent endoscopy if they are found to be TFF3 positive. The “usual” management was defined as giving lifestyle advice, commencing anti-reflux medication if required and offering an endoscopy if clinically indicated. The primary outcome was the identification of patients with BO at 12 months between these two groups and the study lasted 2 years. Interestingly of the 6834 patients who were randomised to the Cytosponge arm only 1750 attempted to undergo the procedure and 5% (n=96) of these failed to swallow the capsule. 12% of these patients (n=202) had to undergo a repeat procedure due to insufficient sampling. From the 5084 who did not undergo the Cytosponge procedure, 310 failed the screening interview and the other 4774 either declined the procedure or did not respond. Although not fully addressed in the study this is a concerningly high number of people who did not successfully undergo the procedure (75%). For those that did undergo the

Cytosponge procedure 2% (140/1654) were found to have BO and in the comparison group this was <1% (13/6388). 9 individuals were found to have dysplasia or early OAC in the Cytosponge group and none were found to have dysplasia in the "usual" management group. However, it should be noted that the follow up for this study was only 12 months and it is not stated how many of the 6388 patients in the "usual" management group underwent an endoscopy and therefore had the opportunity for BO or OAC to be diagnosed (as this can only be diagnosed at endoscopy). Given that they state <1% were found to have BO and this was 13 patients it would suggest that only a small percentage of these 6388 patients actually had an endoscopy. Clearly longer follow up is needed for this group and is a significant limiting factor [96]. Whilst this study shows promise as a possible option in screening the population for BO and OAC it also, in my opinion, raises concerns on how acceptable the procedure is and thus how successful it's uptake will be.

1.5 Chapter summary

Although the incidence of OAC is low, its poor prognosis means the vast majority who do develop the disease will die from it. As such, OAC contributes significantly to overall cancer deaths in the UK, particularly amongst men. It is established that BO represents the only known pre-cancerous lesion in the progression towards OAC and that the majority of OAC arises from within BO. The majority of those who develop OAC are, however, unaware they have BO. It has also been demonstrated that by the time individuals develop symptoms and signs of OAC their cancer is often advanced and incurable. It is therefore reasonable to hypothesize that identifying individuals with BO within the population might aid in the earlier detection of OAC, potentially at an earlier stage of development.

Endoscopic therapies for the treatment of dysplastic BO and early OAC (those confined to the mucosa or the first part of the submucosa) have been demonstrated to be effective in ridding individuals of these pre-cancerous and early cancerous lesions. The published long-term data demonstrate that these therapies are safe and durable with individuals undergoing treatment having significantly less risk of developing OAC. Therefore, we already have in our therapeutic armoury effective measures to alter the incurred risks of dysplastic BO and effectively treat early OAC.

The means to identify high-risk individuals remains problematic. At present the only means of identification we have is an invasive and expensive endoscopic procedure. Population screening using this means is impractical. There are also significant issues with the endoscopic surveillance of those with BO with programmes seemingly providing no reduction in cancer risk. We therefore need a low-cost, non-invasive means to accurately identify individuals with BO or early OAC within the general population and perhaps even better risk stratify those with BO on surveillance programmes to minimise the need for endoscopic surveillance. This is, perhaps, where Cytosponge fits in but patient acceptability is a concern. By addressing this key issue of accurate identification of those with or at risk of OAC we will significantly alter the bleak prognosis of OAC.

Chapter 2

The genome, transcriptome and epigenome. Focusing on oesophageal cancer

Chapter 2 – The genome, transcriptome and epigenome. Focusing on oesophageal cancer

In 1990 the largest ever collaborative scientific project, The Human Genome Project, was launched with the aim of mapping all the genes in the human genome. Thirteen years after its initiation and at a cost of approximately $3 billion it was declared complete. At this announcement President Clinton claimed that it would 'revolutionise the diagnosis, prevention and treatment of most, if not all, human diseases'. There is no doubt that The Human Genome Project has had its successes, but this prediction has so far failed to materialise. Through this project hundreds of genetic variants identified have been associated with human disease and traits and valuable insights have been gained on their complex genetic architecture. But these variants seem to confer little in the way of incremental risk to an individual and only a fraction of the genetic basis for disease has been identified [97]. This is highlighted by *Paynter et al* in which 101 genetic variants were tested to determine whether they could provide a predictive risk score for cardiovascular disease. The conclusion of this study was that the variants had little value in predicting disease, whereas a family history taken by the physician remained a useful predictive tool [98]. Additionally, the Human Genome Project has highlighted how similar the human genetic architecture is to lower animals such as the earthworm and that the human genome has around 25,000 genes - instead of the anticipated 120,000 genes - suggesting that the mechanisms of regulation of gene expression are key in understanding the molecular basis of disease. This has emphasised the vital and complex roles played by the epigenome and transcriptome in disease development. Thus, rather than answering the uncertainties of diagnosis, prevention and treatment of disease, perhaps the Human Genome Project has raised more questions.

This analysis paints a too simplistic picture of the impact of the Human Genome Project. An obvious success was the significant improvement in genetic sequencing technologies that now allow scientists to sequence an entire genome within hours and at a fraction of the previous cost. This has led to huge advances in scientific research and biological understanding in areas such as neurodevelopmental disorders, mitochondrial disease or other unknown disease in children in which a disease causing genetic mutation can be found in 19-33% of cases [99]. Additionally, the medical oncology field now aims to provide personalised medicine in which an individual’s tumour can be sequenced to identify the best therapies. None of these advances would have been possible without first knowing the reference genome that was outlined within The Human Genome Project. This project also triggered the birth of other vital branches of science including proteomics, bioinformatics and medical genomics. Areas such as these allow us to continue to develop our understanding of disease and enable us to strive to revolutionise the diagnosis, prevention and treatment of disease.

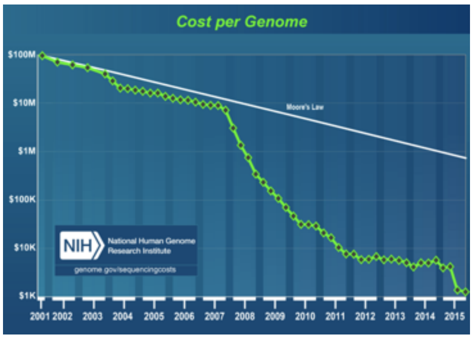


Figure 3: Decreasing cost of genetic sequencing [2]

Within this chapter a thorough review of the literature on the genomics, epigenetics and transcriptomics of BO and OAC will be presented.

2.1 Genetic heterogeneity and clonality in tumours and pre-malignant lesions

There is known to be a great deal of heterogeneity within tumours and pre-malignant lesions across most cancers. This heterogeneity refers to the coexistence of different biological, morphological, phenotypic and genotypic profiles, between tumours known as inter-tumour heterogeneity and within tumours (intra-tumor heterogeneity). There is also spatial heterogeneity between the primary cancer and metastases) and temporal heterogeneity which occurs during the course of disease progression. Finally, the tumor microenvironment which is the complex ecosystem in which cancer cells interact with non-cancerous cells also represents an additional source of intra-tumor heterogeneity. Tumours and pre-malignant lesions are composed of clones that are distinguished on the basis of a number of features including genetic alterations. As discussed by *Turajlic et al,* whilst the evolution of cancer and pre-malignant lesions is an extremely complex series of events it arises from a relatively simple, underlying evolutionary processes of mutation, genetic drift and selection, involving a large number of interacting agents [100]. The clones that make up the tumour mass are those that have mutations which have demonstrated a survival advantage and can proliferate best. *Swanton* demonstrated that branched evolution occurs in tumours and leads to spatial and temporal intra-tumour heterogeneity [101].

In cancer, genetic information is most frequently obtained through a biopsy despite the many known limitations of this approach. The first issue one faces is in the acquisition of the tissue that often involves an invasive procedure that carry risks to the patient and are expensive to perform. In addition to this, tumours can often be inaccessible, or the procedure obtains insufficient samples. These issues are often seen in cancers such as small cell lung cancer, where up to 31% are inaccessible, or in pancreatic tumours where sampling can often be problematic particularly with small tumours where the diagnostic yield can be as low as 40% [102, 103]. Furthermore, should tissue be obtained, preservation issues, such as using formalin, can cause genetic changes leading to false positive results [102]. As highlighted by *Gerlinger et al,* there is a vast amount of tumour heterogeneity and a biopsy only identifies the minority of the genetic aberrations seen within the whole tumour [104]. Thus, biopsy sampling can suffer from spatial bias and mislead the physician. Similarly, a biopsy can only provide information on the tumour at the time of sampling. In clinical practice, a tumour is often only biopsied once, at the start to confirm diagnosis, and as such analysis of this biopsy provides information on the tumour only at that given point. Tumours are known to be dynamic, especially after drug therapy, and as such basing our decisions on past findings that may be inaccurate is inadequate [102].Clearly, therefore, there are spatial and temporal bias issues with relying on biopsies to lead our decision-making and from a population screening perspective this approach is likely to be inappropriate. It has been demonstrated that chromosomal instability, which generates intra-tumour heterogeneity, is often implicated in poor cancer outcomes both in regard to biomarker identification and resistance to therapy [104]. One can therefore appreciate that alternatives to a biopsy led strategy would be appropriate from a diagnostic and therapeutic point of view and this is where the role of “liquid biopsies” are key.

Liquid biopsies are further discussed in Chapter 5. Within the rest of this chapter, I address the heterogenous genomic, epigenetic and transcriptomic landscape of BO and OAC.

2.2 Overview of the genomics of Barrett’s oesophagus and oesophageal adenocarcinoma

The ease with which the oesophagus can be sampled and the known existence of the pre-malignant states of OAC (NDBO and dysplastic BO) has meant that this cancer has been frequently studied in an attempt to determine the molecular basis behind cancer pathogenesis. Despite intense work, however, the exact genetic relationship between BO and OAC is poorly understood. The majority of BO cases appear to be sporadic although two genome wide association studies, the Wellcome Trust Case Control consortium and the Barrett’s and Esophageal Adenocarcinoma CONsortium (BEACON), have reported its heritability to be 9.9% and 35%, respectively. Further work is required in this area but this would suggest that the genetic susceptibility to BO is not insignificant [78].

The pathogenesis of BO remains poorly understood. It is well established that GORD plays a significant role and it is thought that the reflux of bile acids may also trigger oxidative DNA damage and cell death through an increase in reactive oxygen species. The continuous cycles of injury and repair associated with the chronic inflammatory state caused by acid and bile reflux causes alterations in gene expression. CDX1 and CDX2 are caudal related homeobox transcription factors that play a key role in the regulation of intestine specific gene expression and the differentiation of the intestinal epithelium. It is thought that GORD induces the expression of the CDX genes through BMP4 and possible EGFR activation leading to these being found to be expressed in BO, but not in the normal oesophagus. This is evidenced by a reduction in methylation in the CDX gene promoter. However, it should be noted that attempts to recreate this morphological change towards metaplasia in *in vivo* and *in vitro* studies by inducing the ectopic expression of the CDX genes has not yielded metaplastic conversion, although the validity of such models remains in question. Other genes implicated in the pathogenesis of BO are those from the HOXB family. HOXB 5, 6 and 7 have been demonstrated to be upregulated in BO and their ectopic expression has induced intestinal differentiation markers, including MUC2, in normal oesophageal cells. Additionally, the Notch and Wnt signalling pathways have been implicated in the pathogenesis of BO. Their role is in the maintenance of stem cells and differentiation in the intestine. Further studies are required to elucidate these pathways further [105-107].

Once BO is established the inflammatory environment produces cytokines and reactive oxygen and nitrogen species that sustain the DNA damage through a resistance to apoptotic cell death. This is achieved through activation of inflammatory pathways including the NF-κB and IL-6/STAT3 pathways [108]. Studies have demonstrated that BO is highly mutated, even in non-dysplastic biopsy material. Interestingly, frequent mutations found in NDBO are also noted to be present in OAC. Given that few people with NDBO will progress to OAC, it is not clear what functional role, if any, these mutations play [105-107, 109, 110]. This was highlighted by *Weaver et al* who had expected to find a stepwise progression of mutations towards OAC, but instead found that the complex mutational landscape was present at the same frequencies in NDBO as it was in HGD and OAC. In fact, it was only mutations in SMAD4 and TP53 that accurately defined the boundaries between NDBO, HGD and OAC. SMAD4 mutations were only found in 13% of OAC cases although it was only seen in OAC whereas as mutations to TP53 were seen in 72% of cases of HGD, 69% of cases of OAC, but only 2.5% of

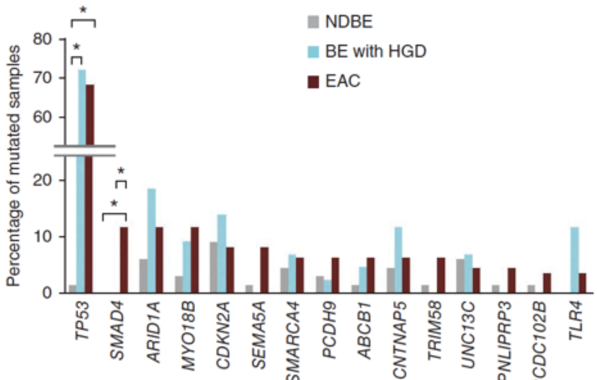


Figure 4: Recurrently mutated genes found in BO and OAC [1]

cases with NDBO [1]. Figure 4 from the *Weaver et al* paper demonstrates the recurrently mutated genes found in their study and nicely highlights that the mutated genes seen in NDBO are frequently also seen in HGD and OAC. Interestingly, although the mutational frequency between NDBO, HGD and OAC varied little, *Ross-Innes et al* observed that the frequency of copy number aberrations increased towards the development of OAC [110]. These copy number changes, known as aneuploidy, have also been investigated by *Li et al* who demonstrated the copy-number profile of patients with NDBO who did not progress to OAC remained static, whereas the percentage of the genome harbouring somatic chromosomal alterations increased rapidly from 0 and 50% at baseline to approaching 100% within 48 months of the cancer diagnosis in those who did progress [111].

Maley et al proposed that a mutation that confers a selective advantage to a cell sweeps across the Barrett's segment and is present in the majority of cells within BO. Additional advantageous mutations, most commonly the loss of TP53, accumulate and can expand across the Barrett's segment with cancer developing as a result of an accumulation of mutations [112]. Others however, propose a more heterogeneous model in which multiple independent clones arise and their associated genetic aberrations expand according to the competitive advantage of the clone [110]. Certainly, the most frequently studied genes in regard to progression of NDBO towards OAC are the tumour suppressor genes TP53 and CDKN2A. CDKN2A is a cyclin-dependent kinase inhibitor that acts as a tumour suppressors through inhibitory regulation of the cell cycle. Germline mutations in CDKN2A are implicated in familial cancer syndromes and somatic alterations have been detected in a wide variety of cancers. The CDKN2A allele has been noted to have lost expression, usually through methylation of its promoter, in 85% of NDBO cases and mutations are thought to occur through oxidative damage. This oxidative damage causes a loss of heterozygosity which has been observed in dysplastic BO. It is notable that the CDKN2A locus is also responsible for encoding p14ARF that prevents degradation of TP53. Thus, studies have focused on the role of the CDKN2A locus in the progression of BO [107, 113].

The role of TP53 in the regulation of the cell cycle and the triggering of apoptosis following genomic damage is well established. Oncogenic stress triggers TP53 to cause cell cycle arrest and DNA repair. Should the DNA damage be irreparable TP53 will induce apoptosis. Thus, its role in protecting DNA is vital and as such TP53 mutation and loss of function play a key role in the development of cancer. The deletion of one allele in the short arm of Chromosome 17 and a mutation causing inactivation of the other allele are common genetic abnormalities found in numerous cancers. In fact, mutations of TP53 are the most prevalent genetic lesion seen in human cancer. This includes the development of OAC from BO [114, 115].

Studies exploring the genomic landscape of OAC demonstrate it to be highly mutated cancer with a mutation burden of up to 10 single nucleotide variations per megabase. Additionally, it has been demonstrated to be a heterogeneous cancer with only a few genes being recurrently mutated. Studies commonly report mutations to tumour suppressors such as TP53, SMAD4 and CDKN2A. Additionally, members of the SWI/SNF chromatin remodelling complex such as MYO18B, SEMA5A, SMARCA4 and ARID1A have also been implicated. These also exert a tumour suppressor function through their role in regulating cell growth and division, DNA repair and chromosome segregation. Their mutation results in the favouring of self-sufficiency in cell growth and escape from growth-regulatory cell signals that play a role in the pathogenesis of cancer. The proto-oncogene ERBB2 (also called HER2) has also been identified as being recurrently mutated in OAC which exerts its effects through causing cell proliferation and preventing apoptosis. Thus, loss of regulation leads to uncontrolled cell growth. Finally, receptor tyrosine kinases, in particular KRAS, EGFR, IGF1R and VEGF-A, have also been demonstrated to be recurrently altered which results in an interruption in the regulation of normal cellular processes [110, 111, 116].

Finally, *Gharahkhani et al* performed a meta-analysis of all genome-wide association studies in relation to BO and OAC. This work included 6167 patients with BO, 4112 patients with OAC, and 17,159 representative control patients. This meta-analysis identified 16 independent risk loci for the development of BO and OAC. All the risk loci identified for BO were also linked with the development of OAC. This again highlights the highly mutated environment of non-dysplastic BO and questions the role these mutations play in BO progression. There was one risk locus on chromosome 3q27, near HTR3C and ABCC5 that was specific to OAC. Of the 16 risk loci outlined within this paper 7 of these had been previously identified to be linked to BO and OAC development. MHC, FOXF1, GDF7 and TBX5 have been associated with the development of BO, and CRTC1, FOXP1 and ALDH1A2 have been associated with the development of both BO and OAC. Of note, BARX1 that has been previously identified to be associated with the development of BO and OAC did not meet the threshold of significance for this study, but was still strongly associated. This study identified 9 new risk loci associated with BO and OAC these were linked to the following genes; SATB2, HTR3C, TPPP/CEP72, KHDRBS2/MTRNR2L9, CFTR, MSRA, LINC00208/BLK, TMOD1 and LPA. The most strongly linked of these was associated with CFTR which is of interest as it is known that patients with cystic fibrosis have an increased incidence of GORD and this may suggest a common pathophysiological process [117].

The existing literature suggests that the options to identify key genomic biomarkers for early detection of the at-risk individual with BO are limited. The majority of individuals with NDBO will not progress to OAC and identification of individuals with NDBO may well lead to many patients being over-diagnosed and subjected to invasive surveillance with little evidence that it improves survival. Thus, it is important to identify individuals who progress to dysplastic BO in order to truly impact on the prognosis of this disease without placing further, unmanageable strain, on healthcare resource. The manner of which cancer develops also plays a significant role in the early detection of cancer. For many years a linear model of progression was postulated with cancer occurring due to an accumulation of molecular abnormalities including single point mutations and structural aberrations. A window of opportunity for identification of the at-risk individual may therefore exist provided the key steps along that linear progression can be identified. Modern genomic technologies now suggest a branched somatic genome evolution in OAC. Additionally, neoplastic evolution can be accelerated by increased mutation rates or punctuated by catastrophic chromosomal events that can occur in a few cell divisions. This, therefore, impacts on the possibility that genomic events might provide windows of opportunity to identify the at-risk individual[118].

2.3 Overview of the epigenetics of Barrett's oesophagus and oesophageal adenocarcinoma

The initial theory was that cancer was a genetic disease caused by a change in the DNA sequence, an accumulation of mutations and eventual cancer pathogenesis. It is now clear that the epigenome plays a significant role in the development of cancer. The epigenome regulates gene expression via the promoter region of DNA through three processes; DNA methylation, histone modification, and post-transcriptional gene regulation by micro ribonucleic acid, also known as microRNAs (or miRNA). Importantly, it is apparent that these epigenetic alterations are heritable between cell lineages [119].

DNA methylation is the most well-studied epigenetic regulator. The regulation of genes through methylation plays a vital role in genome integrity, genomic imprinting, transcriptional regulation, and developmental processes. DNA methylation usually occurs at the 5' position of a cytosine ring within a CpG dinucleotide. CpG dinucleotides are distributed unevenly throughout the human genome, however, the areas in which they are densely concentrated are called CpG islands. CpG islands are often located at the start of a gene with 50-60% of gene promoters located within CpG islands. The methylation of a gene silences it as it denies access for transcription factors and chromatin associated proteins as well as recruiting methyl-CpG binding domain proteins that are associated with histone modification [119].

A histone is the chief component of chromatin and acts as a spool around which DNA winds. It plays a key role in gene regulation, which is performed through post-translational modifications of histone tails which alter the structure of chromatin and thus affecting the transcriptional status of the gene within that locus. Genome-wide studies have demonstrated that histone modifications lead to an “open” or “closed” structure resulting in activation or repression of gene expression[119]. Histone modification may therefore play a role in carcinogenesis.

The third epigenetic regulatory process of gene expression is through miRNAs. These exert their affect through post-translational control of messenger RNA (mRNA) translation into proteins. They can also regulate gene expression through causing histone modification and DNA methylation. Approximately 1000 miRNAs have been computationally predicted, each targeting multiple protein-coding transcripts. It is thought they regulate translation rate in more than 60% of protein coding genes. miRNAs play a key role in normal physiology and their mis-expression has been linked to numerous diseases, including cancer.

There have been numerous studies on the epigenome in BO and OAC. Importantly it has been demonstrated that aberrant DNA methylation has been shown to occur early in the metaplasia – dysplasia – carcinoma sequence. Additionally, it has been demonstrated that methylation status of multiple genes is a powerful biomarker for risk prediction and a strong predictor of survival and of recurrence [120, 121]. Many studies have focused on the hypermethylation and consequent inactivation of tumour suppressor genes. Work by *Agarwal et al* demonstrated that areas of hypomethylation were strongly associated with the progression of BO to OAC. This, therefore, suggests that with hypomethylation the activation of growth promoting genes plays a significant role in BO progression. Their work highlighted 3 genes with which hypomethylation was associated; IGF1R, TLX3 and JUN-D. IGF1R is a tyrosine kinase receptor functioning as an anti-apoptotic agent promoting cell survival. It has been demonstrated to be highly over-expressed in malignant tissues. TLX3 is a DNA-binding nuclear transcription factor and has been associated with poor prognosis in leukaemia. Finally, JUN-D is a proto-oncogene that protects cells from TP53-associated senescence and apoptosis [122].

Environmental, behavioural and demographic factors influence epigenetic state and an individual’s behaviours, or those of their predecessors, through epigenetic inheritance, can influence the epigenome and thus susceptibility to cancer [121]. *Kaz et al* demonstrated that obesity and tobacco smoking affected the methylation status of patients with BO, dysplastic BO, and OAC. In regard to obesity it was identified that overweight BO patients differentially methylated promoter regions [121]. Cases of HGD and OAC were hypermethylated in all functional areas when patients with high body mass index (BMI) were compared to those with a low/normal BMI. When pathways were mapped according to the methylation status it was noted that it mapped to insulin growth factor pathways, pro-inflammatory pathways (IL1B), and pathways that directly effect TP53 (NCI-PID). All of these have been linked to cancer pathogenesis. A similar picture was seen with tobacco smoking in which significant differences were seen in those with BO, HGD and OAC. Interestingly, again when these methylation changes were mapped they linked to genes known to be linked to cancer pathogenesis. This included GFI1 that is a transcriptional repressor involved in the regulation of TP53 activity and Notch signalling [121].

The following genes have been demonstrated to be frequently hypermethylated in BO outlined in *Agarwal et al*[120];

AKAP12. This maps to chromosome 6q24-25.2 that is involved in cell signalling, cell adhesion, mitogenesis and differentiation and possesses tumour suppressor activity. It has been shown to be frequently deleted in cancer. A study by *Jin et al* using 259 oesophageal tissues demonstrated a hypermethylation frequency in AKAP12 of 38.9% in BO and 52.5% in patients with dysplastic BO and OAC compared to 0% in normal oesophageal tissue [123].

APC. This is a known tumour suppressor gene involved in colorectal cancer progression and defects are also thought to mediate chromosomal instability. Methylation of the APC promoter has been demonstrated in 0-25% of normal oesophageal tissues (NB it should be noted that higher levels of APC methylation have been seen in normal tissue from OAC cases), 50-95% of NDBO cases, and 61-100% of dysplastic BO cases. Another study demonstrated that methylation of the APC promoter was a better predictor of survival and tumour recurrence than age or stage. Hypermethylated APC DNA has also been observed in plasma with one study demonstrating its presence in 25% of patients with OAC.

CDH13. This maps to chromosome 16q24 which, again, has been demonstrated to undergo deletion in cancers including OAC. In another study by *Jin et al* using 259 oesophageal biopsy specimens, hypermethylation of CDH13 was seen in 0% of normal oesophagi, 70% of those with BO, 77.5% of dysplastic BO, and 76.1% of those with OAC.

DAPK1. This is a mediator of gamma-interferon induced programmed cell death and as such acts as a tumour suppressor. Work by *Kuester et al* demonstrated hypermethylation of the DAPK promoter in 20% of normal oesophagi, 50% of those with BO, 53% of those with dysplastic BO, and 60% of those with OAC. The consequent loss of protein expression as a result of this hypermethylation was associated with advanced depth of tumour invasion and advanced tumour staging.

GPX and GST. These gene families are involved in antioxidative systems that catalyse the conjugation of exogenous and endogenous chemicals, including gastric and bile acids, and thus protect cells against oxidative damage. *Peng et al* analysed the promoter regions of 23 genes in the GST and GPX families on 37 normal, 11 NDBO, 11 dysplastic BO and 100 OAC samples. In this study they found the range of methylation for normal oesophageal tissue to be 0-8.1%, 12.5-90% for NDBO, 37.5-87.5% of dysplastic BO and 15-69.1% of OAC cases. This echoed work by *Lee et al* who demonstrated hypermethylation in GPx3 promoter in 16.7% of normal oesophagi, 81.1% of NDBO, 82.2% dysplastic BO, and 88% of OAC cases.

MGMT. This gene maps to chromosome 10q26 and is involved in DNA repair, in particular protecting cells against G to A mutations. Multiple studies have explored the involvement of this gene in the pathogenesis of OAC. Methylation of MGMT in normal oesophageal tissue ranges from 1.8-54.8%, 25-88.9% of NDBO, 71.4-100% of dysplastic BO, and 23.4-78.7% of OAC cases. Hypermethylation of MGMT has also been shown to correlate with advanced disease stage.

NELL1. This locus frequently shows a loss of heterozygosity in cancers including OAC. NELL1 encodes for a protein involved in signalling molecules controlling cell growth and differentiation. The exact role of NELL1 is not known however it has been demonstrated that overexpression apoptosis. NELL1 hypermethylation has been demonstrated to correlate with the length of BO and also with survival in patients with OAC. Hypermethylation of NELL1 was shown to have a frequency of 0% in normal oesophagi, 41.7% in NDBO, 52.5% in dysplastic BO, and 47.8% in OAC.

REPRIMO / RPRM. This gene encodes for a protein that is controlled by TP53. As such it is involved in the regulation of the TP53 mediated cell cycle. Methylation frequency of REPRIMO has been found to be 0% in normal oesophageal tissues, 36% of NDBO, 63.6% of dysplastic BO, and 62.6% of OAC cases. Promoter methylation of REPRIMO was also associated with length of BO with those with long segment BO having significantly increased methylation.

CDKN2A. As discussed previously this is a tumour suppressor gene that is commonly linked with cancer pathogenesis via the cell cycle. The most common alterations in the CDKN2A gene are mutations, loss of heterozygosity, and hypermethylation which leads to a loss of regulation of cell proliferation and causes genomic instability. Multiple studies have reported on methylation status of CDKN2A with the range of methylation frequency being 0-24.4% in normal oesophageal tissues, 7.4-27.5% of NDBO, 22.2-55% of dysplastic BO, and 39-59.5% of OAC cases. It has also been shown that CDKN2A is significantly more methylated in patients who have progressed to OAC from NDBO when compared to those who have not.

SFRP. These are tumour suppressor genes that modulate the Wnt pathway. Five SFRP genes have been identified with three of them found to be consistently methylated in BO and OAC. In normal oesophageal tissues the SFRP genes were methylated in 0-13% of cases, but 73-89% of cases of BO and 73-93% of cases of OAC. This was further evidenced by a lack of SFRP mRNA and protein expression being seen in cases of BO and OAC.

SOCS. These genes are involved in cytokine signalling and in particular are cytokine-inducible negative regulators of cytokine signalling. It has been observed that in normal oesophageal tissues there is no methylation of the SOCS genes, whereas in BO SOCS-3 was found in be methylated in 13% of cases and in 74% of dysplastic BO cases. This supports the theory that SOCS genes are involved in progression towards OAC.

SST. Within the GI tract somatostatin regulates endocrine and exocrine secretion, modulates motor activity and is the primary inhibitor of gastric acid secretion. SST gene promoter methylation has been noted to increase as tissues progress towards OAC. In normal oesophageal tissue hypermethylation was noted in 9% whereas it was seen in 70% of BO case, 71.4% of HGD cases and 71.6% of OAC cases. It has also been noted that hypermethylation correlates with the length of BO with more frequent hypermethylation occurring in longer segments.

TAC1. This gene has been mapped to chromosome 7q21-22 which has been identified in multiple studies to undergo a loss of heterozygosity in cancer including OAC. The role of TAC1 in carcinogenesis is not fully understood although part of its function is to encode for substance P which has been demonstrated to have proliferative and anti-apoptotic effects. *Jin et al* demonstrated a significant rise in the frequency of TAC1 hypermethylation as tissues progress towards OAC. Hypermethylation was seen in 7.5% of normal tissues, 55.6% of BO cases, 57.5% of dysplastic BO cases and 61.2% of cases of OAC. Again, hypermethylation was seen more frequently in longer segment BO. Interestingly, circulating hypermethylated TAC1 DNA was identified in patients with OAC.

TIMP3. Again, the silencing of this gene due to loss of heterozygosity or hypermethylation has been frequently observed in cancer. It acts as a tumour suppressor gene as it can inhibit tumour growth and angiogenesis as well as playing a role in apoptosis. This has been reported on in multiple studies with hypermethylation occurring in 0-19.3% of normal tissues, 54.1-87.5% in BO, 71.4-77.8% of dysplastic BO, and 19.5-86.3% of OAC cases.

WIF1. This gene plays a role in carcinogenesis through its function as an antagonist to the Wnt signalling pathway. The Wnt signalling pathway has been strongly linked to the pathogenesis of cancer and the overexpression of its signalling components and the downregulation of its antagonists plays a vital role. The hypermethylation of WIF1 has been demonstrated to range from 0% in normal oesophagus to 51.6% in BO and 83.3% in OAC. WIF1 hypermethylation was also more frequently seen in patients who had progressed from BO to OAC than those who had not.

Panels of epigenetic markers have been used on endoscopic biopsies in attempt to accurately detect patients at high-risk of progressing from BO to OAC. An example of this is where a combination of APC, TIMP3 and TERT promoter hypermethylation was observed in 91-100% of cases of progressors, whereas it was only seen in 17-36% of non-progressors [120]. Additionally, a retrospective, double blind trial by *Jin et al* used 8 genes and studied 145 non-progressors and 50 progressors. This panel demonstrated an area under the curve (AUC) of 0.84 meaning it had higher predictive sensitivity and specificity than clinical features[124].A study by *Alvi et al* found that four genes could distinguish between NDBO and dysplastic BO / OAC with an AUC of 0.988 [125]. Epigenetic panels have also been used on plasma for the detection of OAC. APC and DAPK were studied with 61% of patients with OAC having detectable levels of methylated promoter DNA which was also associated with worse prognosis [120].

Finally, it is also important to recognise the work exploring miRNA in BO and OAC as these provide a potentially alternate means of risk stratification and early detection. miRNA-196a has been identified as a potential biomarker in OAC and a study using a panel of 4 miRNA including this found that expression of these four miRNAs was significantly higher in patients who progressed from NDBO to OAC than those who did not. Additionally, studies have identified miRNA-21 as being implicated in the carcinogenesis of numerous cancers including OAC and it has been shown to be upregulated in a progressive manner through the metaplasia-dysplasia-carcinoma sequence [126]. Work by *Slaby et al* compared samples of 24 normal oesophageal mucosa, 33 NDBO cases, 21 LGD cases, 6 HGD cases and 35 cases of OAC. This identified four diagnostic miRNA that could distinguish between normal and BO cases (NDBO and dysplasia) with an AUC of 0.971, between normal and OAC cases with an

AUC of 1.0, between NDBO and dysplastic BO with an AUC of 0.856 and between NDBO and dysplastic BO / OAC with an AUC of 0.886 [127]. Finally, *Drahos et al* found, in a study looking at 150 OAC cases and 148 BO cases, 46 distinct miRNAs that significantly increased in OAC when compared to BO. Of note in this study, when cases of early stage OAC were looked at (T1b and T2 cancers) 35 miRNAs could distinguish between BO and early stage OAC. This provides particular promise in the pursuit of the early detection of OAC [128]. It is important to note, however, that whilst mRNA directly relates to the protein and thus phenotype, miRNA is part of a complex regulatory network involving multiple target mRNAs.

2.4 Overview of the transcriptomics of Barrett's oesophagus and oesophageal adenocarcinoma

Much of the work on the pathogenesis of BO and OAC has focused on the genomics behind the diseases. As discussed earlier it is apparent that the BO and OAC, like all cancers and their pre-cancerous conditions, have a highly heterogeneous genetic landscape making the search for the key mutations in the pathogenesis of OAC from BO a difficult task. Cancers and their pre-cancerous conditions clearly accumulate genetic changes but not all of these will drive tumour progression. It is also becoming clear that cancers are not homogenous cells undergoing transformation by themselves but are instead complex biological systems of intertwined interactions and signalling within their microenvironment. Within progression of BO it is thought that the cells accumulate genetic and epigenetic alterations that are influenced by surrounding cells, external factors and their microenvironment [129]. Thus, alterations in gene expression highlighted by over or under expression through the metaplasia-dysplasia-carcinoma sequence provides invaluable insights into disease progression. Importantly it is vital to note that alterations to gene expression are not only caused by mutations to a gene but also through changes to gene regulation that can occur at the epigenetic, transcriptional, post-transcriptional, translational or post-translational levels. Advances in biotechnology, in particular the development of complementary DNA (cDNA) microarray technology and the accompanying bioinformatics, have allowed for gene expression profiles of cancers and their pre-cancerous conditions to be performed. These allow in-depth study of the pathways associated with cancer pathogenesis and progression [130, 131].

Greenawalt et al performed gene expression profiling on biopsy samples using cDNA microarrays from 25 patients with NDBO, 38 with OAC and compared these to 39 normal patients. Following normalisation and background correction, unsupervised hierarchical clustering separated the tissue types with BO and OAC tissues being closely related. They summarise that the BO and OAC samples had overexpression of genes involved in tissue development, specifically keratinisation, intercellular junctions, calcium-ion binding and endopeptidase activity. The BO tissues were noted to have overexpression of genes associated with digestion and alcohol metabolism. In keeping with known literature genes in the MUC and TFF families were noted to be overexpressed in BO. In the OAC tissues, overexpression of genes associated with response to external or biotic stimuli, immune and inflammatory responses, collagen catabolism, and proteolysis were noted [131].

Hyland et al performed gene expression profiling on matched normal oesophagus, normal cardia and NDBO tissue from 40 patients using Affymetrix microarray chip. In keeping with known data, they also observed genes such as the TFF family (1, 2 and 3), MUC5AC, and genes involved in keratinisation were up-regulated in BO. In keeping with known data, this work noted significant increased expression of genes within the HOX family and the activation of their downstream intestinal markers in the BO tissues. HOX genes have been demonstrated to be involved in pathogenesis of cancer through coding for proteins that regulate transcription factors during development. This includes regulating apoptosis, receptor signalling, differentiation and angiogenesis. When performing pathway-based analysis *Hyland et al* demonstrated the involvement of TGF-β signalling pathway, showing a possible loss of TGF-β signalling when comparing BO with normal oesophagus. Alterations in TGF-β signalling have been demonstrated previously to be involved in the metaplasia-dysplasia-carcinoma sequence. This paper also reported alterations in the Notch signalling pathway, which is also linked to the metaplasia-dysplasia-carcinoma sequence. They noted 19 genes (of 32) to be down-regulated in BO when compared to normal oesophagus and in contrast 23 genes to be up-regulated when compared to normal cardia. *Hyland et al* hypothesised that the role of these two pathways in BO may be through disrupting the ability of the cells to differentiate or maintain their differentiated state [132].

14 studies exploring gene expression profiling in BO were reviewed by *Wang et al*. Six of these studies made their raw data publicly available and, of those, the authors selected 3 studies to analyse the data and a further single study to validate their findings. In the three studies used for analysis there were 75 normal controls, 56 patients with NDBO and 43 with OAC. This work identified 40 genes that were differentially expressed in BO from normal controls. These were validated with immunohistochemistry. *Wang et al* found that the BMP/TGFβ pathway and transcription factors such as CDX1 and CDX2 were linked to the development of BO. The genes involved included those in the MUC, TFF, and FOX families that have been previously linked to the development of BO [133].

The ability to differentiate patients with NDBO from those with dysplasia is of key importance as it would allow the identification of individuals at high risk of progression to OAC rather than those with a low risk. *Sabo et al* obtained microarray data from formalin-fixed paraffin embedded (FFPE) samples of patients with HGD and NDBO and then selected the genes with the greatest differential expression between the two categories (16 upregulated and 11 downregulated) to confirm these findings on 17 new samples (8 NDBO and 9 HGD) using real-time polymerase chain reaction (PCR). The results of this work were in keeping with previously published work on altered gene expression in the progression to OAC. Notably, the expression of calgranulin B (S100A9) which has a role in chemotaxis and proliferation, was noted to be consistently upregulated in HGD. Additionally, ADAMTS12 and pleckstrin homology-like family domain were also seen to be upregulated in HGD as reported in other work. MUC5AC and TFF1 that have been well described to be upregulated in NDBO was noted to be down-regulated in HGD[134]. *Murao et al* had similar aims of identifying the high-risk individuals when comparing the expression profile of biopsies (and brushings) from the Barrett's segment of 9 patients with OAC and 50 with NDBO. Their most notable finding was that CD55 (DAF), linked to numerous cancers and involved in the innate immune response, was over-expressed in the brushings of those with OAC [135].

Finally, *Visser et al* performed a systemic review of all studies, between 2000 and 2015, exploring gene expression profiles in patients with either OAC or SCC of the oesophagus with the aim of identifying prognostic gene signatures. They analysed the results of 22 studies and interestingly identified that the findings were heterogeneous with only 12 of all the genes reported within these studies being identified in more than one study. Pertinently there were 6 studies looking at OAC alone. In 5 of these studies gene signatures were identified that associated with survival. This included one study producing a 4 gene signature that, if present, reduced 5-year survival from 58 to 14%. Another study produced a separate 4 gene signature, whilst a third found a 2 gene signature of SPARC and SPP1 over-expression being linked to poor outcome. Two studies produced large gene signatures of 165 and 59 genes, respectively, which were associated with better survival. Whilst this work demonstrates the heterogeneous landscape of OAC and these gene signatures require validation in large studies, it does demonstrate the potential for the development of gene signatures that might accurately identify individuals with OAC (and potentially BO) as well as predict risk, progression, and prognosis. Similar work to this has been validated in both breast and colon cancer [136]. It should be noted that unsupervised clustering is not effective at deconvoluting heterogenous signatures and thus can struggle to identify key pathways and specific molecules. A way to resolve this issue is to use Gene Set Enrichment Analysis that has been used to identify potential biomarkers in BO [137].

The role of miRNAs is an area of interest in cancer and their role in the diagnosis of oesophageal cancer has been studied. miRNAs are small noncoding RNAs that play a role in the regulation of the translation of genes in many physiological and pathological processes including cancer development. Aberrant expression of miRNAs has been demonstrated in many types of cancer and been utilised in the diagnosis and prognostication of patients[138]. *Jia et al* reported on miR-25 being overexpressed in SCC tissue when compared to normal tissue. miR-150 has been demonstrated to be under-expressed in oesophageal SCC tissue when compared to normal which is thought to cause the overexpression of Glioma-associated oncogene homolog 1 and consequently a loss of cell cycle regulation through Cyclin D1[139]. Much of the work in relation to oesophageal cancer and miRNAs have focussed on SCC and the utilisation of tissue. However, *Zhang et al* demonstrated that serum concentrations of 4 miRNAs (overexpression of miR-25-3p and miR-151a-3p and under-expression of miR-100-5p and miR-375) were significantly different in patients with OAC when compared to normal controls[140].

The canonical view put forward by early studies on cancer tissues promoted the view that the development of cancer occurred through a stepwise acquisition and accumulation of genetic alterations [141]. However, this view is now challenged through the recognition that individual tumours show great heterogeneity in their patterns of genetic alterations, epigenetic changes and gene expression even within single histological groups [142]. Additionally, it has been demonstrated that malignant phenotypes can be maintained solely by a sub-population of cells with stem cell properties [143].Finally, it has also been shown that early stage cancer has a similar gene expression profile to metastatic cancers and, interestingly, in regards to OAC it has been shown that gene expression profiles, particularly in stromal gene expression associated with tumour growth, is found to be similar both in BO and OAC[144]. Therefore, one can hypothesise that diagnostic tests based purely on genomics may be challenging as epigenetics and transcriptomics provide a more complete and dynamic view of a tumour or disease phenotype. The epigenome can inform on the environmental changes that affect disease, whilst the transcriptome is closely related to the proteins and thus the disease phenotype. Genomic biomarkers tend to provide an accurate diagnosis for inherited diseases rather than complex diseases such as cancer and autoimmune conditions. Consequently, a diagnostic test focussed only on genetic mutations may fail to identify individuals at an early or ideally high-risk pre-malignant stage.

2.5 Chapter summary

It is clear that the genomic, epigenomic and transcriptomic landscape of BO and OAC are complex and heterogeneous with this landscape not only affected by interactions with each other but also through the influence of external factors and the microenvironment. Whilst consistent markers have been found to identify individuals with BO such as TFF3, this does not necessarily improve the prognosis of OAC as the vast majority of these individuals will not progress and the identification of those that will remains highly problematic. This potentially leads to the subjection of more individuals to invasive testing at the cost of great healthcare resource with little in the way of improved patient outcomes. Clearly, however, further work in these fields holds the key to early identification of individuals with or at risk of developing OAC and as such improving its grave prognosis.

Chapter 3

Hypothesis and aims of this thesis

Chapter 3 - Hypothesis and aims of this thesis

The hypothesis for this thesis is;

Individuals with or at-risk of OAC can be accurately identified through analysis of questionnaire data and / or liquid biopsies.

In summary the specific aims of this thesis are;

1. Can individuals with or at-risk of OAC be accurately identified through analysis of demographic, symptom and risk factor questionnaire data?

2. Can sufficient quality and quantity of RNA and DNA be extracted from blood and / or saliva to be used in downstream analysis for biomarker discovery that could potentially identify the at-risk individual?

3. Can saliva potentially be used as a liquid biopsy for population screening?
    a. What are the optimum collection and storage procedures to provide the best quality and quantity sample?
    b. What biomarkers can be detected utilising salivary DNA and RNA from epigenetic and transcriptomic analysis?
    c. Can the discovered biomarkers be utilised with or without clinical questionnaire data to accurately identify the at-risk individual and potentially create a population screening tool?

The central aim of this thesis is to develop a novel non-invasive means to identify individuals who already have or are at risk of developing OAC. For this work the at-risk individual will be defined as those with the precursor lesion BO either in its low-risk form, NDBO, or its high-risk form of HGD. Patients with LGD were excluded from the analysis due to the complexities of its diagnosis outlined in Chapter 1.

This will be approached in differing manners acknowledging that a population screening tool will need to be safe and acceptable. Chapter 4 explores the use of novel artificial intelligence (AI) analysis of patient demographic, risk factor and symptom data (collected via a patient questionnaire) to determine whether this alone can identify at-risk individuals. The AI analysis was performed by Professor Avi Rosenfeld at Jerusalem College of Technology. AI analytical techniques were developed on an already existing data-set collected as part of a previous study on OAC. Following this work, this thesis explores whether additional data collected through a specifically designed questionnaire, with enhanced data points, can improve our ability to identify individuals with or at risk of OAC.

This thesis then explores whether the identification of at-risk individuals can be improved by either integrating transcriptomic and / or epigenetic data with the questionnaire data or whether this biomarker data alone is sufficient to identify at-risk individuals. In Chapter 5, focus is on whether sufficient RNA can be extracted from blood and / or saliva for downstream analysis and biomarker discovery. Optimisation of extraction techniques are undertaken and preliminary biomarker work is performed using matched blood and saliva samples for targeted expression analysis.

Taking the salivary diagnostic work further, Chapter 6 explores the use of saliva as a tissue for population screening. Further method optimisation work is performed to determine the ideal collection and storage of salivary samples. Following this biomarker discovery is performed on a large cohort of patients using salivary RNA for targeted expression analysis and whole mRNA sequencing is performed on a smaller cohort of patients to identify future potential biomarkers that require validation. Additionally, DNA is extracted from the saliva samples and submitted for epigenetic analysis using the Illumina EPIC array.

Chapter 7 concludes and summarises the key findings of the work presented as well as outlines future work required.

Chapter 4

Identification of the at-risk individual through analysis of questionnaire data using machine learning

Chapter 4 - Identification of the at-risk individual through analysis of questionnaire data using machine learning

4.1 Introduction. Using machine learning to analyse questionnaire data

4.1.1 Overview of artificial intelligence in medicine

Artificial intelligence (AI) is the science and engineering behind the creation of intelligent machines and / or intelligent computer programs. Machine learning (ML) is a subset of AI and is a branch of data mining that applies mathematical models to generate computerised algorithms. These can create novel prediction models. ML involves a computer 'learning' important features of a dataset to enable predictions about other, unseen, data. This can be particularly useful to create predictive models about which subjects have a disease.

The medical field has lagged behind using AI although there is an increasing clinical footprint. There are complex issues behind this. Firstly, physicians had felt uncomfortable with the risks of medical error and, although in some areas the clinical risks may be no higher when using AI, are generally reluctant to embrace the "black-box nature" of an automated system. A cofounding factor in this was that there were early instances in which the AI systems were unable to perform as well as the physician. An example was the use of AI to analyse clinical and demographic data and 477 electrocardiogram (ECG) biomarkers from 33,144 women to predict mortality in post-menopausal women. Whilst the results of this were encouraging, it was not as effective as already existing risk prediction scores [145]. This is not always the case, however, AI improved the detection of ovarian and breast cancers using ultrasonography [146, 147]. It has also aided in the classification of prognosis in melanoma, to predict susceptibility for cerebrovascular disease, risk of recurrence of breast cancer, and diagnosis of thyroid disease with outcomes as good as, if not better, than specialists within those fields[148-150]. Within the field of gastroenterology there is now an increasingly large AI footprint, particularly within endoscopy. AI-assisted endoscopy with recognition and characterisation of lesions is now becoming increasingly common and part of an endoscopists every day work [151].

A second obstacle is the data that is provided to the AI programs. If the inputs are limited, known as “supervised learning”, then their ability to perform is equally limited. An example of this is the AI automated interpretation of ECG’s in which it has been given a limited number of diagnoses and patterns to recognise and consequently provides only a limited interpretation. In this scenario, the program can only perform as well as a physician. In many applications of AI in medicine the data supplied has been limited to what is known. There are, however, examples where AI systems were developed without constraints. For example, *Beck et al* published work aiming to improve the identification of high-risk breast cancer cases by analysing pathological specimens called “C-Path”. In this work, instead of using the historical markers of risk such as atypical nuclei, the AI system used image processing to identify 6642 predictors that, when tested on 2 independent test sets, were superior to results achieved by pathologists. These results were significantly associated with 5-year survival beyond any established clinical or molecular markers [152, 153]. Examples like this are currently rare. The opportunity and appetite to collect the required huge amounts of unbiased data from a large cohort of individuals is costly, particularly when there is a possibility that little will be learned from doing so. We therefore frequently see new algorithms being applied to old data.

Examples like the above suggest that AI may have a role in providing support for under-resourced areas of medicine such as decreasing the burden on histopathologists who are facing increasing numbers of cases to review as a result of the increasing numbers of diagnostic tests being performed. A study by *Bejnordi et al* demonstrated that an AI model was able to detect breast cancer better than 11 pathologists if the pathologists were allowed only 1 minute to review the slides. The mean of the 5 best performing algorithms had an AUC of 0.966 with the pathologists having a mean AUC of 0.810. It should be noted, however, that when the pathologists had no time constraints they performed as well as the AI model (AUC 0.960) and the pathologists were able to find difficult to detect cases more often[154]. Another example of AI being used to challenge the current accepted means of service delivery was with a recent study exploring the decision making of a multidisciplinary team in 638 breast cancer patients. In this study the AI model (Watson for Oncology) gave recommendations that were highly concordant (93%) with that of the 15-member team of physicians [155].

AI is of course vulnerable to being affected by small subsets of data that differ from the general pattern. This can lead to the creation of algorithms that are inaccurate when tested in larger cohorts. However, combining this analysis with molecular biology creates "anchors" that prevents data being swayed. The field of molecular biology is an area in which AI has huge potential. One study using AI on mRNA data from 1208 kidney transplant biopsies was able to predict allograft loss better than current histological methods[155]. The use of AI for accurate diagnosis and prognostication is becoming more frequent in the field of clinical oncology. *Cheng et al* have published and won the Breast Cancer Prognosis Challenge for their work on building predictive models using genomic and clinical data in breast cancer. They also published work on so-called attractor metagenes in which they identified, using unsupervised learning of genomic data, clusters of genes required for cancer progression in multiple tumour types. They found that by combining both genomic and clinical data they were able to build predictive models that out-performed any known system [153].

Clearly AI offers great potential within the field of medicine. It is hoped that the utilisation of big data will allow for better disease surveillance, earlier detection, improved diagnosis, uncover new treatments and bring the prospect of personalised medicine into reality. Additionally, it is hoped that AI will obviate the repetitive tasks that strain healthcare resources such as documentation and the review of straightforward histology or imaging. However, with this potential there are also fears that AI will disrupt the physician-patient relationship removing the crucial role of emotional intelligence, empathy and human judgement to guide and advise.

There are many branches of AI. However in this work, through a collaboration with Professor Avi Rosenfeld, Professor of Data Mining at the Jerusalem College of Technology, we focus on machine learning. Machine learning involves using intelligent computer programs and algorithms to collate, aggregate, and analyse large data sets that were traditionally too large for interpretation and analysis to reveal patterns, trends, and associations. The computer program operates not by following explicitly programmed instructions but by building models using the algorithms created from the data inputs. This allows the program to iteratively learn as it receives more data inputs. Professor Rosenfeld utilised Bayesian machine learning for this work in which the algorithm was developed around a set of variables and their conditional dependencies e.g. what is the probability of outcome A, if variable B occurs. In medicine this can be applied as; what is the likelihood of developing OAC if a patient smokes tobacco. The model can then be built exploring a multitude of variables and in various combinations, referred to as a Bayesian Network.

Machine learning is applied for the analysis of heterogeneous data sets and as such is often used in the field of medicine when analysing questionnaire or molecular biology data. There are three categories of machine learning; supervised, unsupervised and reinforcement learning. Supervised learning is most commonly used in medicine as it analyses the data when the outcome is known. With unsupervised learning the data is uncategorised and the AI model looks to find hidden patterns of groups within the data. Finally, reinforcement learning uses simulated data and as such is rarely used in medicine. Within this work Professor Rosenfeld utilised supervised learning to identify specific trends across the data. In general terms, this process involves an algorithm being applied to pre-processed data (the cleaning of data, removal of incomplete data and addressing any missing information) that has undergone transformation that includes accurate scaling, decomposition (splitting attributes into constituent parts) and / or aggregation (combining related attributes into a single feature). These steps maximise learning from a training set and allow for the creation of algorithms that can be applied to the testing set. The AI model then undergoes cross-validation by splitting the data into subsets and testing the model a number of times on each subset. The prediction error is then averaged over the number of trials on each subset to obtain the total effectiveness of the AI model. Once complete statistical methods are utilised to provide parameters such as sensitivity and accuracy and should this be acceptable the algorithm can be applied to further data sets.

Alongside machine learning, the use of an AI technique called “feature selection” in analysing bioinformatics data has become an essential tool. It allows the analysis of a large volume of data which may feature numerous irrelevant features. This approach avoids the pitfalls of over-fitting the data, provides faster and cost-effective models and generates results that provide a deeper insight into processes [156].Generally speaking there are three types of selection techniques: filter, wrapper and embedded. For our work, Professor Avi Rosenfeld focused on using the multivariate filter techniques alongside a novel filter technique created by himself in collaboration with our team, called MIAT [157]. MIAT is an acronym for “Minority Interesting Attribute Threshold” and is unique in that it is designed to identify significant data that is strongly correlated to a diagnosis that only occurs in a minority of instances. By identifying those thresholds, it can adaptively identify subgroups from a big dataset that correlate to clinical variables.

Several screening tools for BO have been evaluated with variable degrees of success [158]. In addition, the annual incidence of OAC in patients with BE may be as low as 0.1% [159]. The merits of current surveillance approaches for BO therefore remain controversial. The minimally invasive Cytosponge may add an important triaging step as it can be administered in general practice, is acceptable to patients, and has a sensitivity of 87% for patients with >3cm of BE with a specificity of 92% [160, 161]. An important question that then arises is which patients to screen with this test.

The obvious groups to target would be those with symptoms of disease as well as those with known risk factors. The main factors include age, sex, reflux symptoms, obesity, cigarette smoking, and family history [162]. Ethnic background has also been suggested [163]. Further risk factors for oesophageal cancer include ingestion of red meat and pickled vegetables, although the latter are probably only relevant in squamous cell cancer, whereas processed meat may be important in OAC [164]. Anticholinergic drugs have been shown to significantly increase risk of OAC [165]. We have previously tried to identify patients at risk by analysing these factors, with relatively poor success [166].

We hypothesised that machine learning techniques may yield better and more reproducible discrimination between patients with and without BO compared to the statistical models used previously. Previous works often did not validate their results and those that did found large reductions in model accuracy between the training and validation cohorts [161, 163, 167]. Additionally, most studies focused on only a few symptoms, making comparison to previous work difficult. For example, individual studies have been published reporting risk factors associated with BE including older age [168, 169] , male gender [156, 168, 170], Caucasian race [171-173], gastro-oesophageal reflux disease (GORD) [169, 174, 175], smoking [176-178], and central obesity [178, 179]. To our knowledge, no studies exist which considered all of these factors together, particularly whilst considering cofounding effects. In this current study we evaluate two extensive independent datasets to develop the model with an additional independent validation set.

4.2 Aims of chapter

The aim of this chapter is to establish whether using machine learning techniques to analyse medical data enhances our ability to identify individuals with Barrett's oesophagus when compared to traditional medical statistical analytical techniques. The targeted objective of this chapter is:

1. Compare machine learning analysis to traditional medical statistical analysis by applying techniques to an identical dataset to determine whether superior results are achieved by artificial intelligence.

This work has now been published [180]

4.3 Methods

4.3.1 Study design and Participants

In this prospective study, machine learning risk prediction in Barrett's oesophagus (MARK-BE), we collected data from two case-control studies done in the UK to construct training, testing, and external validation datasets. We collected data on patients with Barrett's oesophagus and controls, both as defined in the inclusion criteria of the studies. All patients with a diagnosis of dysplastic Barrett's oesophagus or oesophageal adenocarcinoma were included in the Barrett's oesophagus group and those with ultra-short segment Barrett's oesophagus (Prague classification of less than C1Mx or C0M3) were removed from the analysis completely to create a clear distinction between the groups.

BEST2 (ISRCTN 12730505) was a case-control study undertaken nationwide in 14 UK hospitals, with patients recruited between 2011–14, that compared the accuracy of the Cytosponge-trefoil factor 3 test for the detection of BO with endoscopy and biopsy as the reference standard [95, 181]. BO was defined as endoscopically visible columnar-lined oesophagus (Prague classification > C1 or M3), with histopathological evidence of intestinal metaplasia on at least one biopsy sample. Controls were symptomatic patients without BO referred for routine endoscopy. Of 1299 patients, 880 (67·7%) had BO, 40 (3%) had invasive oesophageal adenocarcinoma, and 419 (32·3%) were controls. In parallel to assessing the accuracy of the Cytosponge test, patients were asked to complete a questionnaire giving details of 40 symptoms and risk factors of their condition to analyse whether these symptoms and risk factors could be used to stratify patients by risk, such as we have done previously [182]. Questionnaire data were collected from all 1299 participants. For the current study, we randomly split this large dataset (6:4) using a computer algorithm into a training dataset (n=776) and a testing dataset (n=523). We split the dataset using this ratio to allow sufficient training data to quantify the model's complexity, while maintaining adequate data to validate the model.

BOOST (ISRCTN 58235785) was a case-control study undertaken in four European hospitals (two in the UK in London and Nottingham, one in Leuven, Belgium, and one in Madrid, Spain), with patients recruited between 2013–15, that used enhanced endoscopic techniques to target high-risk lesions that occur in patients with BO [168]. Clinical and demographic data were collected. Controls were patients referred by their primary care physician with suspected oesophageal cancer who had neither BO nor oesophageal adenocarcinoma and were analogous to those in BEST2. Although BOOST was a multicentre study, questionnaires were only collected from 398 patients at a single site, University College London Hospital, London, UK. 197 (50%) of 398 participants who completed questionnaires were controls and 24 (6%) of 398 had oesophageal adenocarcinoma. Patients were asked to complete a questionnaire similar to that in BEST2. This questionnaire was designed from the outset to include the same questions as in the BEST2 questionnaire so that the cohort could be used as a validation dataset for a symptom-based algorithm that was to be generated from the BEST2 dataset in line with TRIPOD guidelines [169]. However, some extra questions were included relating to food intake, anxiety, and depression. We used this dataset as the external validation dataset.

The primary outcome of both studies was a diagnosis of BO, which was ascertained by histopathologists who were masked to predictor variables.

For BEST2, symptoms of gastro-oesophageal reflux disease (GORD) were collected with a questionnaire adapted from the GERD Impact Scale together with the GORD questionnaire [158]. BOOST also included the hospital anxiety and depression scale. The total number of variables reported in BEST2 was 40 and in BOOST was 204. In both studies, data were collected on paper case-report forms and transferred into electronic databases by investigators.

4.3.2 Data Handling and Machine Learning approaches

We imputed missing data for nominal and numerical features with the modes and means of the training data. Here we describe how predictors were handled, and the workflow is shown in the figure below.

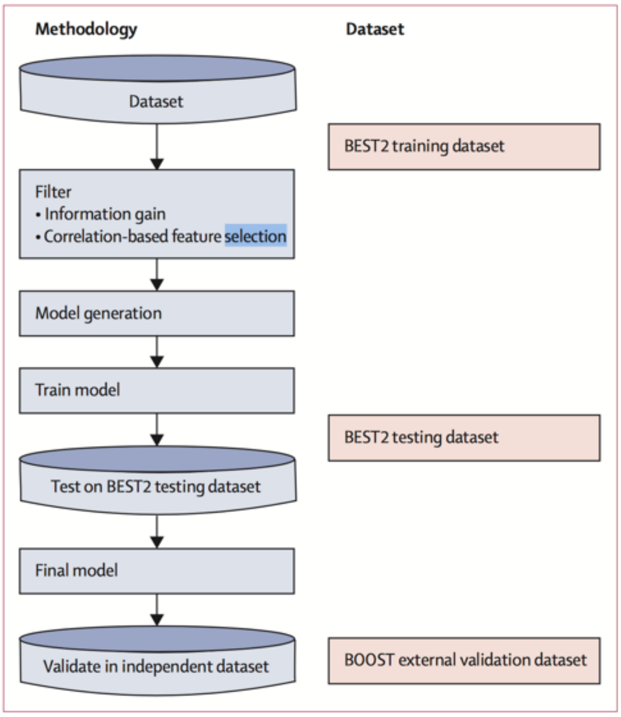

Methodology
Dataset
Dataset
BEST2 training dataset
Filter
• Information gain
• Correlation-based feature selection
Model generation
Train model
BEST2 testing dataset
Test on BEST2 testing dataset
Final model
Validate in independent dataset
BOOST external validation dataset

Figure 5: Machine learning workflow for data processing and model development

We used feature analysis to process data and identify important predictors. For the training dataset, we analysed data using two accepted feature selection filters: information gain and correlation-based feature selection. Information gain is a machine learning univariate filter that compares each feature separately and its correlation with the class. Features are chosen on the basis of how much each one discriminates between the groups being investigated; in our case, BO versus no BO. Correlation-based feature selection filtering is a multivariable filter that specifically considers features' correlation to each other and removes redundant features that are highly correlated. The final set of features is then used to generate the analysis model.

Both information gain and correlation-based feature selection are filter feature selection methods and thus have the advantage of being fast, scalable, and independent of the classifier [156]. Independence from the classifier is crucial to our study because it allows us to understand which features are being selected by the algorithm and their medical importance. As made clear by Nie and colleagues [170], filters that are independent of the classifier enable improved interpretability. They should also lead to more stable algorithms than conventional statistical approaches, such as backward logistic regression, because they minimise data overfitting. Similar to our previous work [182], we initially identified *k* features that had at least a minimal correlation to BO. We then plotted the change in mean area under the receiver operator curve (AUC) for prediction of BO using between 1 and *k* features.

We identified the smallest number of independent features in the BEST2 training dataset to create our model. The smaller the set of predictors, the more stable and robust the model, which minimises the risks of overfitting the data.

Once our features were defined, we considered five different machine learning methods: logistic regression, a decision tree based on the Gini measure of quality, a naive Bayes classifier assuming a Gaussian distribution, a support vector machine using the radial bias function kernel, and a random forest classifier using ten trees. These five algorithms were chosen for comparison because they are well-accepted machine learning methods in medical applications [171-176]. The relative strengths and weaknesses of learning algorithms remain a major research topic but in principle, when training data are restricted, simpler models usually perform better because they will generalise more reliably e.g. linear and logistic regression models. Random forest and decision trees usually perform better when training data are abundant and a complex interaction exists between features. Support vector machines can be extremely robust if the number of predictive features is very large compared with the number of training examples, a situation in which overfitting often occurs. Naive Bayes should be preferred over logistic regression if data are sparse but one is confident of the modelling assumptions [177]. We also considered using deep neural networks, but given the lack of dimensionality of our data, these models are substantially less accurate and interpretable [178, 179, 183]. Although we considered several options for building a supervised prediction model, unless otherwise specified, we present the results from a logistic regression prediction model.

We developed a prediction model using 90% of the BEST2 training dataset to train and 10% to internally test the model (Figure 8). This process was repeated ten times. We used the mean AUC to determine which model performed best, which was then tested with the BEST2 testing dataset. Finally, we validated the model on the BOOST external validation dataset. For the AUC calculation, we set the sensitivity of the model to 90%, because we deemed this sensitivity to be clinically relevant.

Because the AUC measurements might have restricted accuracy for imbalanced datasets, we calculated precision recall and log loss to show the stability of the derived model. We calculated and present extended metrics for the machine learning application for the training model when applied to the BEST2 testing dataset, and for the BEST2 training model after external validation on the BOOST dataset. We present accuracy, which is the ratio of the correctly labelled participants to the whole dataset; recall, which is equivalent to sensitivity (of all the people with BO, how many could we correctly predict?); precision, which is equivalent to positive predictive value (how many of those labelled with BO actually have it?) measured at the highest point on the receiver operator curve; and the F-measure, which is the harmonic mean (average) of the precision and recall.

The input datasets included obvious biases, such as different sex prevalences in the BO and control groups and duration of symptoms. Patients with BO are known to have a higher prevalence of long-term gastro-oesophageal reflux disease [161, 164]. Additionally, controls presented with new symptoms whereas those with BO were mostly in surveillance programmes. We reconstructed the datasets so that race, sex ratios, and age profiles were similar across all datasets. We also removed all features relating to symptom duration. We then repeated all machine learning with this reconstructed dataset to build a new risk prediction panel. The risk prediction panel was tested on both the BEST2 testing dataset and the BOOST independent validation dataset with the actual diagnoses withheld. Once the panel had predicted the diagnoses, the results were compared with the true diagnoses and the accuracy of the model was then calculated.

4.3.3 Statistical analysis

This Article is reported in alignment with TRIPOD guidelines [169]. No generally accepted approaches exist to estimate sample size requirements for derivation and validation studies of risk prediction models. We used all available data to maximise the power and generalisability of our results. Model reliability was enhanced by our use of an external validation cohort.

We present discrete variables as numbers and percentages and continuous variables as mean (SD). We calculated *p* values for the association of each factor with presence and absence of BO using Student's *t* test or the χ2 method. We calculated AUCs by generating a univariate logistic regression model using only that feature.

We present the ranked features from the training dataset using regression coefficients of the association of each feature in the final prediction model. We present the risk of BO associated with each feature using odds ratios and 95% CIs.

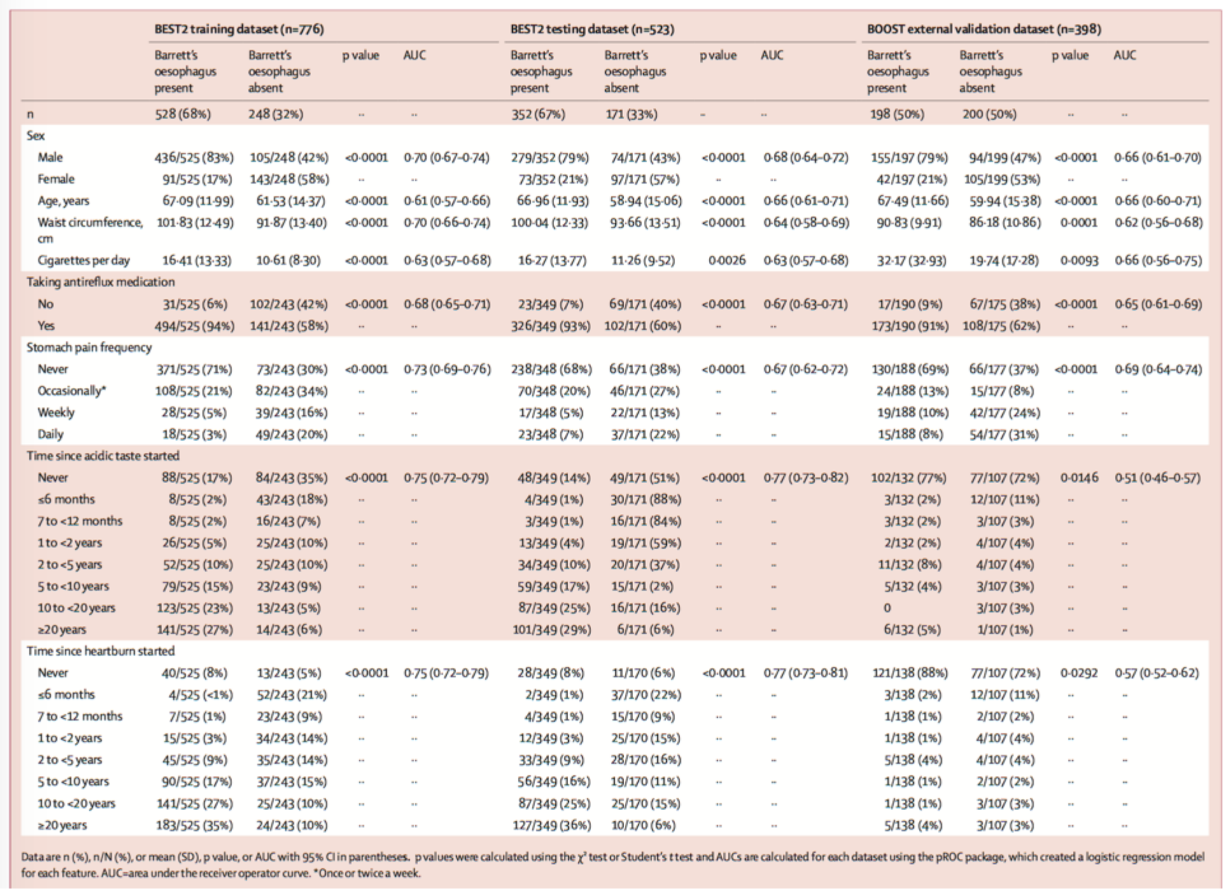
We did all analyses using the RWeka, cvAUC and pROC packages in R (version 3.6.1).

4.3.4 Role of funding sources

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

4.4 Results

Demographic and symptom characteristics for all three datasets are shown in Figure 6 below:



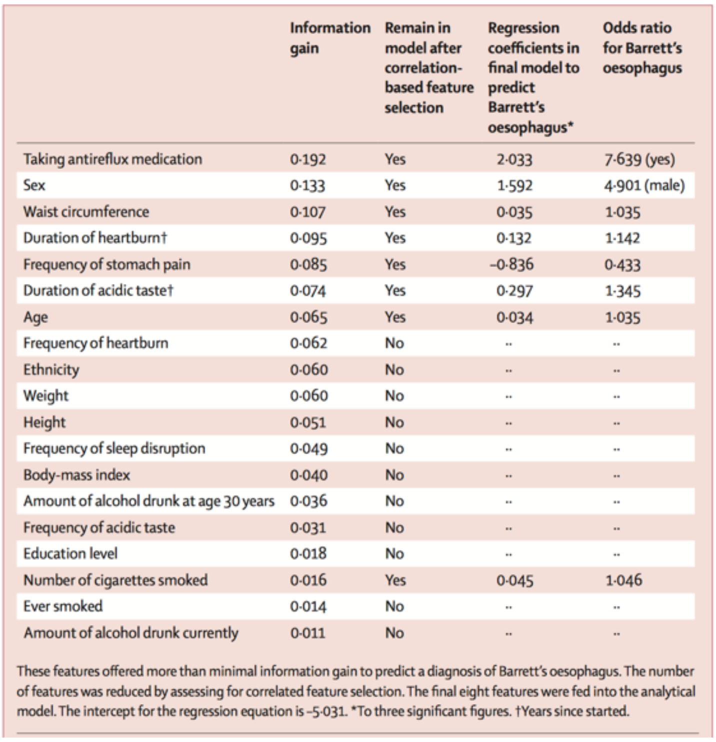
	BEST2 training dataset (n=776)				BEST2 testing dataset (n=523)				BOOST external validation dataset (n=398)			
	Barrett's oesophagus present	Barrett's oesophagus absent	p value	AUC	Barrett's oesophagus present	Barrett's oesophagus absent	p value	AUC	Barrett's oesophagus present	Barrett's oesophagus absent	p value	AUC
n	528 (68%)	248 (32%)	··	··	352 (67%)	171 (33%)	-	··	198 (50%)	200 (50%)	··	··
Sex												
Male	436/525 (83%)	105/248 (42%)	<0·0001	0·70 (0·67–0·74)	279/352 (79%)	74/171 (43%)	<0·0001	0·68 (0·64–0·72)	155/197 (79%)	94/199 (47%)	<0·0001	0·66 (0·61–0·70)
Female	91/525 (17%)	143/248 (58%)	··	··	73/352 (21%)	97/171 (57%)	··	··	42/197 (21%)	105/199 (53%)	··	··
Age, years	67·09 (11·99)	61·53 (14·37)	<0·0001	0·61 (0·57–0·66)	66·96 (11·93)	58·94 (15·06)	<0·0001	0·66 (0·61–0·71)	67·49 (11·66)	59·94 (15·38)	<0·0001	0·66 (0·60–0·71)
Waist circumference, cm	101·83 (12·49)	91·87 (13·40)	<0·0001	0·70 (0·66–0·74)	100·04 (12·33)	93·66 (13·51)	<0·0001	0·64 (0·58–0·69)	90·83 (9·91)	86·18 (10·86)	0·0001	0·62 (0·56–0·68)
Cigarettes per day	16·41 (13·33)	10·61 (8·30)	<0·0001	0·63 (0·57–0·68)	16·27 (13·77)	11·26 (9·52)	0·0026	0·63 (0·57–0·68)	32·17 (32·93)	19·74 (17·28)	0·0093	0·66 (0·56–0·75)
Taking antireflux medication												
No	31/525 (6%)	102/243 (42%)	<0·0001	0·68 (0·65–0·71)	23/349 (7%)	69/171 (40%)	<0·0001	0·67 (0·63–0·71)	17/190 (9%)	67/175 (38%)	<0·0001	0·65 (0·61–0·69)
Yes	494/525 (94%)	141/243 (58%)	··	··	326/349 (93%)	102/171 (60%)	··	··	173/190 (91%)	108/175 (62%)	··	··
Stomach pain frequency												
Never	371/525 (71%)	73/243 (30%)	<0·0001	0·73 (0·69–0·76)	238/348 (68%)	66/171 (38%)	<0·0001	0·67 (0·62–0·72)	130/188 (69%)	66/177 (37%)	<0·0001	0·69 (0·64–0·74)
Occasionally*	108/525 (21%)	82/243 (34%)	··	··	70/348 (20%)	46/171 (27%)	··	··	24/188 (13%)	15/177 (8%)	··	··
Weekly	28/525 (5%)	39/243 (16%)	··	··	17/348 (5%)	22/171 (13%)	··	··	19/188 (10%)	42/177 (24%)	··	··
Daily	18/525 (3%)	49/243 (20%)	··	··	23/348 (7%)	37/171 (22%)	··	··	15/188 (8%)	54/177 (31%)	··	··
Time since acidic taste started												
Never	88/525 (17%)	84/243 (35%)	<0·0001	0·75 (0·72–0·79)	48/349 (14%)	49/171 (51%)	<0·0001	0·77 (0·73–0·82)	102/132 (77%)	77/107 (72%)	0·0146	0·51 (0·46–0·57)
≤6 months	8/525 (2%)	43/243 (18%)	··	··	4/349 (1%)	30/171 (88%)	··	··	3/132 (2%)	12/107 (11%)	··	··
7 to <12 months	8/525 (2%)	16/243 (7%)	··	··	3/349 (1%)	16/171 (84%)	··	··	3/132 (2%)	3/107 (3%)	··	··
1 to <2 years	26/525 (5%)	25/243 (10%)	··	··	13/349 (4%)	19/171 (59%)	··	··	2/132 (2%)	4/107 (4%)	··	··
2 to <5 years	52/525 (10%)	25/243 (10%)	··	··	34/349 (10%)	20/171 (37%)	··	··	11/132 (8%)	4/107 (4%)	··	··
5 to <10 years	79/525 (15%)	23/243 (9%)	··	··	59/349 (17%)	15/171 (2%)	··	··	5/132 (4%)	3/107 (3%)	··	··
10 to <20 years	123/525 (23%)	13/243 (5%)	··	··	87/349 (25%)	16/171 (16%)	··	··	0	3/107 (3%)	··	··
≥20 years	141/525 (27%)	14/243 (6%)	··	··	101/349 (29%)	6/171 (6%)	··	··	6/132 (5%)	1/107 (1%)	··	··
Time since heartburn started												
Never	40/525 (8%)	13/243 (5%)	<0·0001	0·75 (0·72–0·79)	28/349 (8%)	11/170 (6%)	<0·0001	0·77 (0·73–0·81)	121/138 (88%)	77/107 (72%)	0·0292	0·57 (0·52–0·62)
≤6 months	4/525 (<1%)	52/243 (21%)	··	··	2/349 (1%)	37/170 (22%)	··	··	3/138 (2%)	12/107 (11%)	··	··
7 to <12 months	7/525 (1%)	23/243 (9%)	··	··	4/349 (1%)	15/170 (9%)	··	··	1/138 (1%)	2/107 (2%)	··	··
1 to <2 years	15/525 (3%)	34/243 (14%)	··	··	12/349 (3%)	25/170 (15%)	··	··	1/138 (1%)	4/107 (4%)	··	··
2 to <5 years	45/525 (9%)	35/243 (14%)	··	··	33/349 (9%)	28/170 (16%)	··	··	5/138 (4%)	4/107 (4%)	··	··
5 to <10 years	90/525 (17%)	37/243 (15%)	··	··	56/349 (16%)	19/170 (11%)	··	··	1/138 (1%)	2/107 (2%)	··	··
10 to <20 years	141/525 (27%)	25/243 (10%)	··	··	87/349 (25%)	25/170 (15%)	··	··	1/138 (1%)	3/107 (3%)	··	··
≥20 years	183/525 (35%)	24/243 (10%)	··	··	127/349 (36%)	10/170 (6%)	··	··	5/138 (4%)	3/107 (3%)	··	··

Data are n (%), n/N (%), or mean (SD), p value, or AUC with 95% CI in parentheses. p values were calculated using the χ^2 test or Student's t test and AUCs are calculated for each dataset using the pROC package, which created a logistic regression model for each feature. AUC=area under the receiver operator curve. *Once or twice a week.

Figure 6: Demographics and symptom characteristics of the three datasets by the presence or absence of Barrett's

BO patients were generally older than those without BO and were also more likely to be male and smokers, had more central obesity, took more anti-reflux medication, and had less frequent stomach pain. Additionally, those with BO had experienced acidic taste and heartburn for significantly longer than those without.

In the case-control BEST2 training dataset, all cases had a confirmed diagnosis of BO. We selected features with a non-negligible information gain. In line with previous work [184], we used a threshold of 0·01 (i.e. above a negligible zero value) to select features that would positively affect the model. Features with a weaker correlation to disease were removed. A total of 19 features were selected (Figure 7).



	Information gain	Remain in model after correlation-based feature selection	Regression coefficients in final model to predict Barrett's oesophagus*	Odds ratio for Barrett's oesophagus
Taking antireflux medication	0·192	Yes	2·033	7·639 (yes)
Sex	0·133	Yes	1·592	4·901 (male)
Waist circumference	0·107	Yes	0·035	1·035
Duration of heartburn†	0·095	Yes	0·132	1·142
Frequency of stomach pain	0·085	Yes	–0·836	0·433
Duration of acidic taste†	0·074	Yes	0·297	1·345
Age	0·065	Yes	0·034	1·035
Frequency of heartburn	0·062	No	..	..
Ethnicity	0·060	No	..	..
Weight	0·060	No	..	..
Height	0·051	No	..	..
Frequency of sleep disruption	0·049	No	..	..
Body-mass index	0·040	No	..	..
Amount of alcohol drunk at age 30 years	0·036	No	..	..
Frequency of acidic taste	0·031	No	..	..
Education level	0·018	No	..	..
Number of cigarettes smoked	0·016	Yes	0·045	1·046
Ever smoked	0·014	No	..	..
Amount of alcohol drunk currently	0·011	No	..	..

These features offered more than minimal information gain to predict a diagnosis of Barrett's oesophagus. The number of features was reduced by assessing for correlated feature selection. The final eight features were fed into the analytical model. The intercept for the regression equation is –5·031. *To three significant figures. †Years since started.

Figure 7: Ranked features in the BEST2 training

We sorted these features from highest to lowest information gain correlation with BO and considered subsets with the top *k* features ranging between 1 and 24. We selected the eight features with the highest information gain and found no significant increase in the AUC (*p* value of the moving average of the next 10 points compared with the original values being 0·7; Figure 8 below).

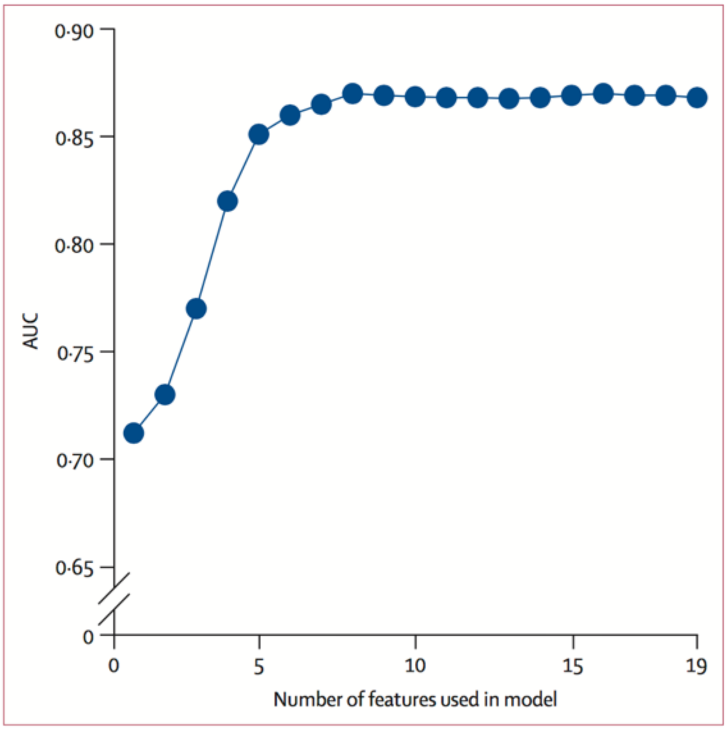
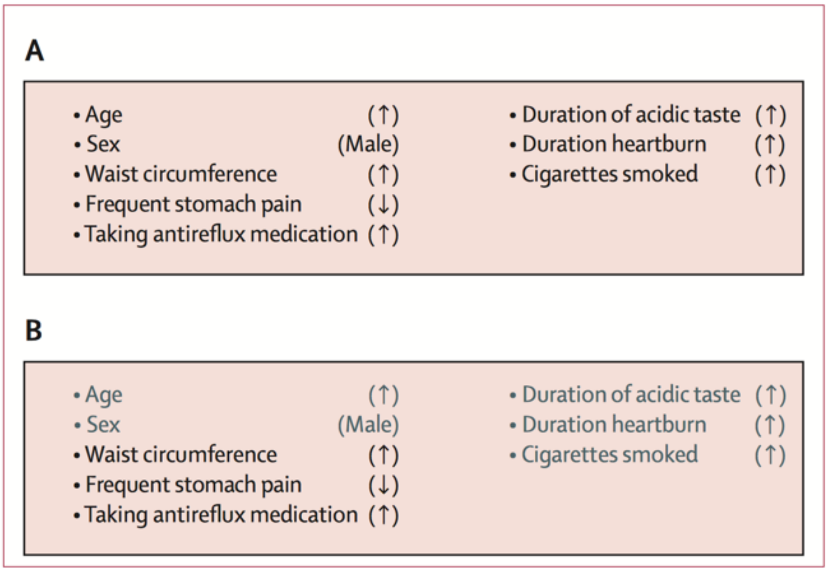


Figure 8: Performance of the model using the BEST2 training dataset Increasing the number of features strengthens the model to a plateau point that is reached around eight features. The model AUC remains unaffected when up to a total of 19 features are added. AUC=area under the receiver operator curve.

This finding is consistent with the concept that adding features, even those with strong correlation to BO (Figure 7), does not necessarily improve model performance.

We developed multivariable models using correlation-based feature selection based on all 24 common features. Correlation-based feature selection selected eight features as independent predictors of BO (age, sex, waist circumference, stomach pain, taking anti-reflux medication, duration of heartburn, duration of acidic taste in the mouth, and smoking; Figure 9A below).



A

- Age (↑)
- Sex (Male)
- Waist circumference (↑)
- Frequent stomach pain (↓)
- Taking antireflux medication (↑)
- Duration of acidic taste (↑)
- Duration heartburn (↑)
- Cigarettes smoked (↑)

B

- Age (↑)
- Sex (Male)
- Waist circumference (↑)
- Frequent stomach pain (↓)
- Taking antireflux medication (↑)
- Duration of acidic taste (↑)
- Duration heartburn (↑)
- Cigarettes smoked (↑)

Figure 9: Risk prediction model panels for Barrett's oesophagus
(A) The eight features selected by correlation-based feature selection for the BEST2 training dataset, and the direction of association with presence of Barrett's oesophagus. (B) Of the eight features identified, those that are still associated using the correlation-based feature selection model using the reconstructed datasets, excluding potential age, sex, race and symptom duration biases, are shown in black, with those no longer associated in grey. Arrows show the direction of association, with an arrow pointing up indicating an increased likelihood of Barrett's

These features were not the same as the top eight features identified with information gain analysis (Figure 7).

The prediction model was based on the features selected via the correlation-based feature selection analysis (Figure 9A). Once we had our small panel of features, we tested the five different machine learning methods and found that logistic regression yielded the best median AUC, and so we elected to use this model (Figure 10).

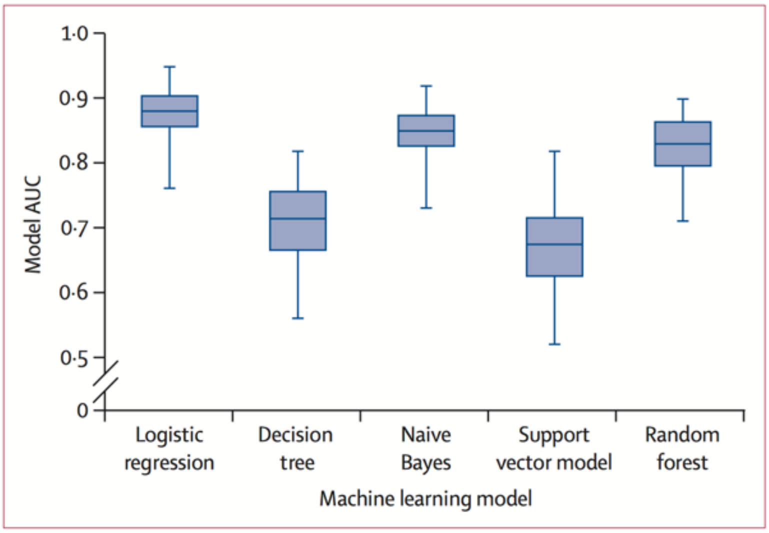


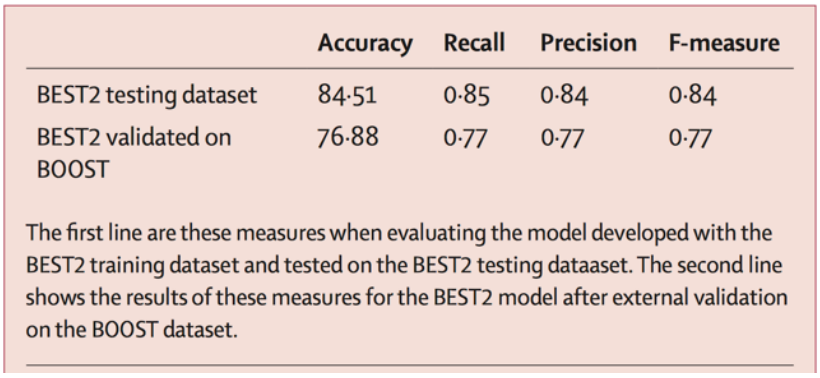
Figure 10: Comparison of model's AUC with different machine learning classification algorithms
Box plots show AUCs and 95% CIs. AUCs when using the BEST2 training dataset with 13 features. AUC=areas under the receiver operator curve.

Furthermore, this model is most readily understandable to a medical audience making it easy to convert into a usable tool for clinical practice.

We used the testing dataset to provide an upper estimate of the model's predictive ability (Figure 11). Using the BEST2 testing dataset, the AUC was 0·87 (95% CI 0·84–0·90) and, for a sensitivity arbitrarily set at 90%, the specificity was 68%.

We validated this model using the BEST2 testing dataset. The model reproduced well, with an AUC of 0·86 (95% CI 0·83–0·89), with sensitivity set at 90%, and specificity of 65%. The model was finally tested on the independent validation BOOST dataset. Here the model achieved an AUC of 0·81 (95% CI 0·74–0·84), with a set sensitivity of 90%, and a specificity of 58%. This three-stage development process led to a stable, reproducible model.

For completeness, we also present the accuracy, recall, precision, and F-measure results of the training model applied to the BEST2 training dataset and the external validation model on the BOOST dataset (Figure 11). The results were in a relatively narrow range (e.g. accuracy 76·88–84·51% and F-measure 0·77–0·84) with the lowest values being recorded when validating the BEST2 model on the BOOST data. These results are consistent with the AUC results.

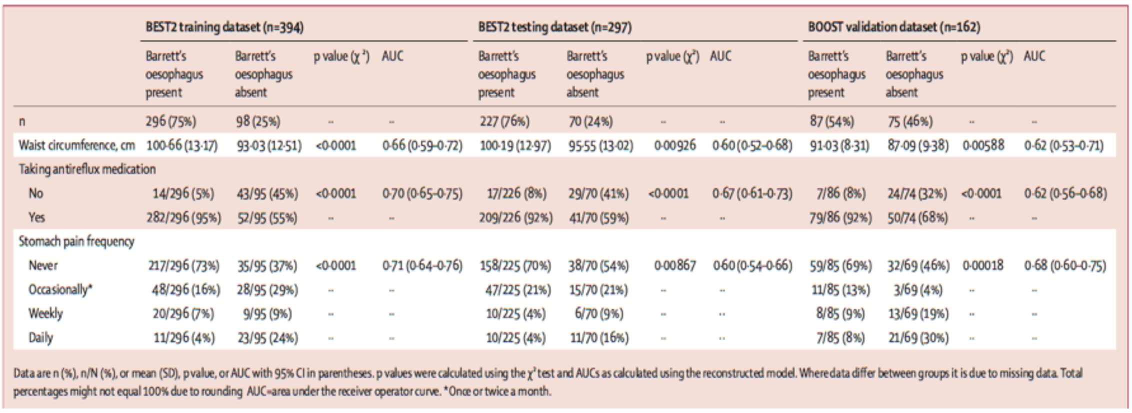


	Accuracy	Recall	Precision	F-measure
BEST2 testing dataset	84·51	0·85	0·84	0·84
BEST2 validated on BOOST	76·88	0·77	0·77	0·77

The first line are these measures when evaluating the model developed with the BEST2 training dataset and tested on the BEST2 testing dataaset. The second line shows the results of these measures for the BEST2 model after external validation on the BOOST dataset.

Figure 11: Extended metrics for evaluating the machine learning application, by dataset

We repeated our analyses using reconstructed databases to remove potential biases. Reconstructing the cohorts reduced the BEST2 training dataset from 776 to 394 patients; the BEST2 testing dataset from 523 to 297 patients; and the BOOST external validation dataset from 398 to 162 patients (Figure 12).



	BEST2 training dataset (n=394)				BEST2 testing dataset (n=297)				BOOST validation dataset (n=162)			
	Barrett's oesophagus present	Barrett's oesophagus absent	p value (χ^2)	AUC	Barrett's oesophagus present	Barrett's oesophagus absent	p value (χ^2)	AUC	Barrett's oesophagus present	Barrett's oesophagus absent	p value (χ^2)	AUC
n	296 (75%)	98 (25%)	··	··	227 (76%)	70 (24%)	··	··	87 (54%)	75 (46%)	··	··
Waist circumference, cm	100·66 (13·17)	93·03 (12·51)	<0·0001	0·66 (0·59–0·72)	100·19 (12·97)	95·55 (13·02)	0·00926	0·60 (0·52–0·68)	91·03 (8·31)	87·09 (9·38)	0·00588	0·62 (0·53–0·71)
Taking antireflux medication												
No	14/296 (5%)	43/95 (45%)	<0·0001	0·70 (0·65–0·75)	17/226 (8%)	29/70 (41%)	<0·0001	0·67 (0·61–0·73)	7/86 (8%)	24/74 (32%)	<0·0001	0·62 (0·56–0·68)
Yes	282/296 (95%)	52/95 (55%)	··	··	209/226 (92%)	41/70 (59%)	··	··	79/86 (92%)	50/74 (68%)	··	··
Stomach pain frequency												
Never	217/296 (73%)	35/95 (37%)	<0·0001	0·71 (0·64–0·76)	158/225 (70%)	38/70 (54%)	0·00867	0·60 (0·54–0·66)	59/85 (69%)	32/69 (46%)	0·00018	0·68 (0·60–0·75)
Occasionally*	48/296 (16%)	28/95 (29%)	··	··	47/225 (21%)	15/70 (21%)	··	··	11/85 (13%)	3/69 (4%)	··	··
Weekly	20/296 (7%)	9/95 (9%)	··	··	10/225 (4%)	6/70 (9%)	··	··	8/85 (9%)	13/69 (19%)	··	··
Daily	11/296 (4%)	23/95 (24%)	··	··	10/225 (4%)	11/70 (16%)	··	··	7/85 (8%)	21/69 (30%)	··	··

Data are n (%), n/N (%), or mean (SD), p value, or AUC with 95% CI in parentheses. p values were calculated using the χ^2 test and AUCs as calculated using the reconstructed model. Where data differ between groups it is due to missing data. Total percentages might not equal 100% due to rounding AUC=area under the receiver operator curve. *Once or twice a month.

Figure 12: Demographic and symptom characteristics of reconstructed datasets, by presence or absence of Barrett's Oesophagus

We used the same workflow to create a new model. We determined the new correlation-based feature selection variables (Figure 9B). The same features remain apart from age, sex, and symptom duration. No new features entered the correlation-based feature selection analysis. As for the initial analyses, we selected features with non-negligible information gain, and selected a total of seven features. We then built multivariable models based on correlation-based feature selection on these seven features. Three were selected as independent predictors of Barrett’s oesophagus (waist circumference, frequent stomach pain, and taking anti-reflux medication; Figure 9B). The overall accuracies are lower than the original eight features, but a clear difference remains between patients with and without BO. The initial model had an AUC of 0·84 (95% CI 0·79–0·88; sensitivity 90%, specificity 43%), which decreased to 0·78 (95% CI 0·72–0·84; sensitivity 90%, specificity 41%) after testing internally, and to 0·77 (0·64–0·81; sensitivity 90%, specificity 37%) after external validation.

Most features identified through both iterations of the model are readily understandable such as age, male sex, longer duration of symptoms, taking anti-reflux medications, and central obesity (i.e. waist circumference). However, the feature of lower frequency of stomach pain appears counterintuitive.

4.5 Discussion

We have shown that a panel with eight features, including detailed stomach and chest symptoms, can predict the presence of BO with high sensitivity and specificity in a case-control population. The currently used system for identifying patients with BO, or those at risk of oesophageal adenocarcinoma, is flawed because it is based on symptoms that trigger expensive and unpleasant invasive tests. Simple triaging of individuals might be possible on the basis of predictive panels that include variables that are widely available or easy to obtain. Work on the QResearch database has shown the usefulness of this approach to predict oesophageal cancer [185]. This approach is slowly being incorporated into general practice but it has not yet been robustly confirmed to detect the premalignant phenotype of BO, potentially because BO is frequently asymptomatic and takes many years to develop into cancer. Nevertheless, this condition needs to be recognised because of the success of early intervention in preventing oesophageal adenocarcinoma with its dire prognosis [186].

In our study, we specifically did not include patients with ultrashort Barrett's oesophagus (i.e. Prague classification of <C1 or <M3). Differences exist between UK and US guidelines on follow-up for this low-risk group and our aim was to create a prediction tool that avoided this ambiguity. Although the methods we used are generally applicable and should be considered for prediction of other diseases, we focused on BO as an example of how a tool could be used by primary-care physicians to better target people for formal screening. Patient age and sex, together with medication and smoking history, are routinely captured in primary care systems. Additionally asking about duration of heartburn and acidic taste, frequency of stomach pain, and measuring waist circumference should be simple for physicians. Alternatively, a patient could do a self-assessment using a web-based app and generate a personalised BO risk profile. Precise cut-offs between patients and controls will need to be defined once this risk prediction panel is tested prospectively in a primary care population in which the prevalence of BO is lower than in our cohorts. For a particular AUC, the sensitivity chosen for use in clinical practice can be altered depending on the clinical question. If triaging for cancer in symptomatic individuals would require a sensitivity of 95% or greater, missing a diagnosis of BO might not be so critical, and a sensitivity of even lower than 90% might be adequate. Indeed, machine learning might offer a way to create accurate predictive panels to pre-screen for many other diseases and could be tuned to achieve the desired sensitivity depending on the importance of the disease in question.

Reflux duration is strongly correlated with cancer risk and is longer in patients with BO. In our panel, use of anti-reflux medicines was a strong BO predictor. Metabolic obesity characteristically presents with truncal obesity and is also a risk factor for BO [187], which explains why our model predicted patients with BO to have greater waist circumference. Waist circumference is not routinely collected, but is an easy measurement to collect, particularly for patients who wish to self-triage. A clear correlation exists between waist circumference and body mass index (BMI), which is routinely collected. Our method identified the most important independent predictors of BO. In routine practice, replacing waist circumference with BMI might be more practical but the model would then need to be reworked. Another finding that initially appears counterintuitive is the negative correlation between BO and frequency of stomach pain; however, on further investigation this correlation makes sense. Most patients with oesophageal adenocarcinoma are not identified before cancer develops despite many of them having BO. Indeed, 40% of patients with oesophageal adenocarcinoma have not previously had symptomatic reflux and many probably had BO [188]. Therefore, BO has been hypothesised to not be associated with severity of reflux symptoms [189], which fits with the model determined from our data.

Our panel of features differs from the QResearch database work for oesophageal cancer [190]. The QResearch panel includes dysphagia, appetite loss, weight loss, and anaemia as predictors for cancer and does not include duration of symptoms or central obesity data. These differences reflect the different realities of BO and oesophageal adenocarcinoma.

Previous works have identified risk factor panels, including multiple biomarkers, such as leptin and interleukin levels, or data from genome-wide association studies, which are not easily available, and others included only a few symptoms [158, 191]. For those in which the risk factor panels were larger, several key differences exist between our analyses and these previous works. We confirmed the importance of older age [161, 191, 192], male sex [159, 160, 162], gastro- oesophageal reflux disease [158, 159, 161, 164, 192, 193], smoking [165, 166, 191, 192], and central obesity [191, 192, 194]; however, we found that many of these risk factors were cross-correlated in our data analysis. We overcame the challenge of panels failing external validation through a combination of univariate and multivariable feature selection techniques that yielded a stable panel. The results are better than previous panels with sensitivities of 70–80% and specificities of 50–60% or AUCs of 0·7 or lower [158, 159, 191, 192]. By contrast, our panel validates between completely different datasets with an AUC of at least 0·81 when only considering eight risk factors. This predictive panel of risk factors might be adequate to be used as a BO triaging tool in clinical practice.

Three recent studies support our risk prediction panel. Xie and colleagues followed-up 63000 patients for 20 years in Norway for risk of developing oesophageal adenocarcinoma and they constructed a model based on a very similar risk panel to ours [195]. Their data were taken from a patient cohort without the level of symptom granularity we achieved by using data from cohorts in which patients were interviewed. The AUC of their model to identify 15-year risk of oesophageal adenocarcinoma was 0·84 (95% CI 0·76–0·91) but it did not attempt to identify patients with BO. Similarly, Kunzmann and colleagues examined 355 034 individuals from the UK Biobank for risk of developing oesophageal adenocarcinoma [195]. Their panel including age, sex, smoking, BMI, and history of oesophageal conditions or treatments and they identified individuals who would later develop oesophageal adenocarcinoma with an AUC of 0·80 (95% CI 0·77–0·82). Once again, their study did not specifically aim to identify BO, although the features are remarkably similar to those we identified, suggesting that many patients they identified might have undiagnosed BO [196]. We found one study that targeted sporadic BO alone that was undertaken in a small Australian cohort in which their choice of risk factors was determined by complex deduction; however, this approach did lead to a tool with an AUC of 0·82 (95% CI 0·78–0·87) [197]. This tool was later validated in an independent dataset [197, 198]. One additional feature of that model was hypertension, which was not identified as an independently important feature in our model even though we queried for it, raising the question of the stability of their model.

Because our aim was to create a tool for pre-screening, we intentionally used the BEST2 and BOOST datasets, which had a higher incidence of BO than the general population. Generally, an open challenge to machine learning is how to properly identify important so-called minority categories, such as BO. Because BO is relatively uncommon, with a prevalence as low as 2% found in Mexico [199], one could create extremely accurate models by assuming no individuals have BO. In the BEST2 and BOOST datasets, this issue was mitigated by use of a targeted collection of suspected at-risk individuals, which led to a distribution of BO that is much higher than that in the general population. Several methods exist to computationally rebalance the data beyond or in addition to this approach. The most common approach is under-sampling, whereby existing records belonging to a prevalent category are intentionally removed to create a different ratio between the classes. Here, the relatively high number of BO patients could be adjusted by randomly removing some of the patients. Alternatively, oversampling could be used, whereby control individuals without BO are added to generate a new balance between the target patients. One popular example of this approach is the synthetic minority oversampling technique [200], which synthetically adds artificial cases to the minority class. Another approach would be to apply a ratio of controls to known cases to train the model with a prevalence that more closely aligns with the real-world setting.

The advantage of using datasets that inherently have higher distributions of patients with BO is that our data are non-synthetic and thus more likely to be effective as a screening tool; although, one could argue that undertaking this study in a cohort with a prevalence of BO that is similar to that of the general population might yield different results. However, further studies are needed to confirm this hypothesis and to study any potential effects of false positives or false negatives generated in a real-world setting. To this end, we propose that our algorithm should be applied to the data generated from the BEST3 study, which is a pragmatic, multisite, cluster-randomised controlled trial set in primary care centres in England, UK, where the prevalence of BO is representative of the general UK population and in which the same questions have been asked as in BEST2 and BOOST. We are also undertaking another prospective study (ISRCTN 11921553) to test this hypothesis independently in a second population that more closely aligns with the general population prevalence of the disease.

The methods used to apply the machine learning analysis also present a challenge. Many researchers carry out both univariate and multivariable analyses of each dataset independently, which often leads to selecting similar features in both datasets. We have previously used this approach ourselves. We made very small changes in our definitions of BO (i.e. with or without intestinal metaplasia), each of which was associated with different risk factors being important in the ensuing algorithms. These differences stem from a lack of so-called stability in the features that each model independently selected [182], too many features, even those with relatively high prediction value, often reduces the model's power.

One current solution to both these challenges is effective feature selection. We approached this challenge by identifying which features add information. This approach is called information gain, a univariate approach. In our previous work, we used a threshold of 0·1 within χ2 with one degree of freedom to select eight features in the dataset [182]. An advantage to using feature selection to determine important features is that they are based on a filter approach to selection, which is undertaken without any connection to a specific learning algorithm. Similarly, no human bias is involved. We incorporated this approach as one step in our current analysis.

Our results show stability across the BEST2 and BOOST datasets. Although each of these datasets was collected independently, their collection methodologies and definitions were similar enough for effective comparison. This study shows that such analyses are possible if stable features are identified that are not influenced by random artefacts in the data collection process [201].

We considered using other multivariate feature selection algorithms including least absolute shrinkage and section operator (LASSO) [202]. LASSO is one type of feature selection that is embedded in logistic regression because its feature analysis is inherently linked to this machine learning method. It has a similar limitation to the support vector machine recursive feature selection (RFE-SVM) approach [203]. Both approaches are limited to only one algorithm, in the case of RFE-SVM, the support vector machine algorithm that we also considered. Because we aimed to consider a variety of machine learning methods, we preferred using information gain and correlation-based feature selection, which are filter methods and can be used without any connection to a specific machine learning prediction model, thus facilitating improved medical understanding [156, 183].

We also considered correlations between features, which often exist in medical datasets. We used the multivariable correlation-based feature selection algorithm to do this. We reasoned that features selected by correlation-based feature selection should be more stable than other approaches. This hypothesis is borne out by the high AUC of the predictive model and its stability against the independent validation cohort.

Having created our dataset, we considered possible biases and sought to minimise these by reconstructing the cohorts to avoid any age, sex, or race bias; however, we found that our model remains robust.

The risk prediction panels we generated are easy to use in practice. Theoretically, people could enter their symptoms into a smartphone app and receive an immediate risk factor analysis. These data could then be uploaded to a central database (e.g. in the cloud) that would be updated after that person sees their medical professional.

Our study had several limitations. Because both datasets were collected from at-risk individuals, the dataset was enriched for BO patients. Additionally, patients attending for symptom assessment are more symptomatic than those undergoing surveillance endoscopy. Nevertheless, all BO patients undergoing surveillance would have presented initially with symptoms. Notably, many individuals with BO have no symptoms and so this risk prediction panel is unlikely to work for these people. Nonetheless, given the robustness of the models generated, the predictive panel produced here could be of benefit to rapidly triage symptomatic patients for minimally invasive screening tools, such as the Cytosponge test, because many symptomatic individuals currently undergo no testing at all [204].

Further prospective data collection is needed using a cohort study design in a primary care setting where BO prevalence will be much lower to confirm the validity of our findings and to establish the final best risk prediction model parameters.

4.6 Chapter conclusions

The work within this chapter demonstrates that demographic, symptom and risk factor data may be of use in identifying individuals with BO who are at risk of OAC. Additionally, the encouraging results obtained using AI analysis over standard medical statistical analysis suggests that this is an area deserving of further exploration. However, although the results presented are encouraging, they may not be sufficiently successful on their own to be used as a screening tool and the reliance on self-reporting of medical data is vulnerable to bias which will damage the ability of the predictive system. The AUC of 0.77 (following external validation) demonstrates that using this method alone would result in many undergoing unnecessary testing. As such, whilst this work would suggest this data can be used to identify at risk individuals, we should not rely on it alone. More robust, consistent means are required.

Chapter 5

Biomarker discovery using different tissue types

Chapter 5 - Biomarker discovery using different tissue types

5.1 Introduction

The work outlined in Chapter 4 demonstrates some promise in regard to the ability to identify those with or at risk of OAC through AI analysis of questionnaire data. However, as discussed in Chapter 4 there are two key issues with the reliance on this form of data. Firstly, there are the issues in regards to the consistency and accuracy of responses obtained through questionnaire data and the impact this has on outcomes [205]. Secondly, the specificity obtained through this analysis was low and thus, should it be used for population screening, this would result in large numbers of normal individuals being subjected to invasive and expensive investigations. Clearly we need to create a more robust system that offers identification with higher specificity. It is at this juncture that the possible addition of genomic, epigenetic, and transcriptomic data becomes appealing.

5.1.1 Liquid biopsies

At present, genetic information is obtained in some diseases, particularly within the field of oncology, to aid in the management of the disease. There is little use of transcriptomics or epigenetics currently in clinical practice [206, 207]. However, as discussed in Chapter 2 the epigenetic and expression profiles of tumours and other disease states provides invaluable data on the phenotype that genomic information alone cannot. Clearly basing diagnosis or clinical decision making on genomics alone has limitations. Furthermore, as discussed in section 2.2 a biopsy led approach to diagnosis and therapeutic management of pre-malignant and malignant lesions is flawed and this is where liquid biopsies can provide benefit.

In 1948 Mandel and Metais published work on circulating free DNA (cfDNA)[208]. Unbeknownst to them, this discovery was the first step to the development of liquid biopsies, an area of research that has received a lot of attention of late due to the push for personalised medicine. It has, however, taken many years for this research to begin to be seen in clinical practice.

Liquid biopsies provide a non-invasive means of disease detection and monitoring and have advanced greatly in recent years aided by the evolution of sensitive technologies that can detect and analyse cfDNA. There are issues with the collection of this informative genetic material with variation in quantity and quality seen in differing storage methods, extraction processes including delays in time of extraction, and natural variation amongst individuals even whether they have exercised prior to sample collection. However, the advancing technologies are addressing these issues which is allowing for the analysis of even lower quantity and quality genetic material [102]. It is now possible to sequence a tumour's entire genome from peripheral blood [209]. The non-invasive, low cost means by which a liquid biopsy can be obtained also allows for the potential for them to be used not just as a means of disease detection, but also a means to longitudinally monitor an individual including identification of potential relapses or recurrences.

The question arises as to the origin of the genetic information that is obtained in liquid biopsies. There are three discussed, and not mutually exclusive, mechanisms by which DNA and RNA can enter the bloodstream from a lesion. The majority of work focusses on CTC DNA with little work exploring the use of cfRNA or CTC RNA. The mechanisms are outlined below:

Circulating tumour cells (CTC) are shed into the blood stream or lymphatics from a primary tumour and may act as a seed for the tumour to metastasise. They were first described by Thomas Ashworth in 1869 who reported them to be "*cells identical to those of the cancer itself*". CTCs are present, at varying but generally low frequencies, in many different types of metastatic tumour, but are rarely found in pre-malignant diseases [210]. Their frequency is so low that *Zhe et al* estimated that among the cells to have detached from the primary tumour only 0.01% can form metastases [211]. Due to their low frequency of occurrence, it is necessary to isolate the CTCs from other cells prior to analysis. The Food and Drug Administration have approved the use of CellSearch in the United States, a means to isolate and analyse CTCs, where it has been used as a prognostic indicator for breast, lung, colorectal, and prostate cancer. However, it should be noted that CTCs are fragile and degrade quickly. Thus collected samples need to be processed, when using CellSearch, within 96 hours[102].

cfDNA is primarily thought to occur due to the release of nuclear and mitochondrial DNA from cells that have undergone apoptosis or necrosis. In normal physiological conditions the role of phagocytes is to clear this cellular debris and thus healthy individuals will have low levels of cfDNA. However, in certain circumstances the phagocytes are unable to clear this debris sufficiently and as such the level found in the blood rises. This is found in inflammatory conditions, following exercise and in relation to tumour mass [212]. cfDNA has also been demonstrated to be present in pre-malignant conditions. For example, *Perrone et al* demonstrated that cfDNA was present in those with early colorectal cancer as well as those with pre-malignant disease. However, it should be noted that in this study the ability to predict those with cancerous or premalignant lesions using cfDNA quantification was found only to be significant for those with early cancers [213]. cfDNA tends to be fragmented, approximately 150-180 base pairs in length, and has a high prevalence of tumour associated mutations. When cfDNA and CTC's were compared for their mutation detection, there was a higher abundance of the mutation found in the cfDNA of the same patient [102]. Another advantage of cfDNA is that it can be analysed from bio-banked fluids, it is not as fragile as CTC DNA. However, there are limitations to the use of cfDNA. It should be noted that cfDNA can provide genetic and epigenetic information in regards to a tumour, but cannot inform on the transcriptome or proteome [102].

The establishment of **exosomes** as a means by which to extract and analyse DNA, RNA and protein in the field of research has grown significantly. Exosomes are actively released extracellular vesicles that are found in biofluids such as serum, plasma, saliva and urine. Their function remains debated although it is thought they have a role in intercellular communication and specifically in cancer contribute to micro-metastasis, angiogenesis and immune modulation[102, 214]. A key advantage of using exosomes is that they are stable carriers of DNA and RNA. Studies have reported finding tumour specific mutations, using both RNA and DNA, years after their collection and storage in a freezer [102]. Although the size of an exosome means that a whole transcriptome cannot be identified inside, due to their abundance most of the transcriptome can be identified [215]. RNA derived from exosomes has been used to profile not only cancer, but also many other diseases including inflammatory, neurodegenerative, cardiovascular and metabolic diseases [102].

The use of liquid biopsies has generated a lot of excitement and has been discussed frequently in the media. In relation to cancer there has been published work exploring the use of liquid biopsies in the following fields:

Bettegowda et al performed work using liquid biopsies as a prognostic indicator in colorectal cancer. Their work found that the detection of KRAS-mutant cell-free circulating tumour DNA in those with a KRAS mutant colorectal tumour was associated with a decreased two-year survival with a sensitivity of 87.2% and specificity of 99.2% [216].In breast cancer it has been reported that persistent tumour-associated microsatellite DNA alterations, detected in the blood using PCR following mastectomy, has been associated with tumours with vascular invasion, metastasis to more than 3 lymph-nodes and higher histological grade. And similarly, in breast cancer, the presence of tumour-associated genetic aberrations and loss of heterozygosity found in the blood correlated to overall survival. Overall, there have been studies demonstrating a statistically significant correlation between disease stage and tumour-associated genetic aberrations in the blood in numerous cancers including prostate, ovarian, breast, pancreatic, colorectal, lung, and oral cancer [212].

The use of liquid biopsy to identify those with a relapse or recurrence of cancer is a promising field. *Diehl et al* reported that by monitoring for tumour specific aberrations, including KRAS, TP53 and APC, they were able to identify those with recurrence of colorectal cancer with 100% sensitivity. This work has recently been followed up by*Tie et al* who report the ability to identify those at high-risk of disease recurrence following resection of their colorectal cancer using circulating tumour DNA [212, 217]. Similar work exploring this field has been done in other cancers, including work on ovarian cancer that has recently been reported in the media [218].

Current publications show clearly that liquid biopsies have great potential to diagnose cancers at an earlier stage. This is key to improve the prognosis of any cancer. Exciting work by *Ilie et al* found that the detection of CTCs in patients with chronic obstructive pulmonary disease (COPD) without clinically detectable lung cancer lead to an early diagnosis of the disease [216].

An area in which there has been a lot of research using liquid biopsies is in the response to treatment and monitoring acquired resistance. As discussed earlier, at present we are reliant on biopsies to guide targeted therapy for cancers, which is prone to spatial or temporal bias. Work has been published demonstrating that liquid biopsies can be used to monitor an individual's HER2 status in breast cancer following treatment, the response to anti-EGFR therapy via KRAS mutations in colorectal cancer, and the response to therapy in lung cancer [212]. Furthermore, liquid biopsies have been shown to be able to demonstrate the changing landscape of a tumour following therapy and in particular identify resistant clones. This work has been done in lung, colorectal, melanoma and gastrointestinal stromal tumours with the aim to identify a means to continue to target therapy at a tumour using novel combinations of drugs [212].

The use of liquid biopsies is clearly exciting with an abundance of research being performed in this field. However, whilst the collection of blood can certainly be described as minimally invasive it is still invasive and the collection of blood still requires trained personnel, specialised equipment for collection and storage, and a degree of discomfort for an individual. This does not take into account that the prevalence of needle phobia is reported to be 10%, in American adults, which was thought to be underestimate [219]. Therefore, the question arises as to whether saliva offers a more acceptable option.

5.1.2 Salivary diagnostics

The use of saliva as a diagnostic tissue dates as far back as the early 20th century when Kirk and Michaels identified protein biomarkers in saliva for rheumatism and gout [220]. However, it is not until recent years in which the field of salivary diagnostics has generated more interest particularly in the diagnosis of disease. Saliva has been described as the 'mirror of the body'

[221]. Whilst saliva is predominantly made up of aqueous solution, the other constituents include a variety of enzymes, hormones, antibodies and growth factors that enter the saliva from the blood. The porous, tightly packed capillaries that surround the salivary glands allow constituents to enter the saliva through spaces between cells by transcellular (passive intracellular diffusion and active transport) or paracellular routes (extracellular ultrafiltration). Importantly, exosomes circulating in the blood have also been identified in saliva. In fact, saliva has been shown to contain identical components to blood and is thus able to reflect the physiological state of the body similar to blood [222]. The main hurdle for the use of saliva as a diagnostic fluid has been that the informative constituents of saliva are at a low level and thus difficult to obtain and analyse. However, with the advances in technology this is becoming less of a barrier. As such we are seeing the potential for saliva to be used as a clinically useful tool, providing insights into prognosis, diagnosis and the monitoring and management of patients. Increasing bodies of work are utilising saliva and are detecting useful biomarkers in fields such as cardiovascular disease, renal disease, diabetes and infections, although more work is required in these fields [223].

Specific to cancer, emerging work is finding both DNA and RNA signatures associated with cancer risk in saliva. Mutations to TP53 and other cancer related genes have been reported to be found in the saliva of breast cancer patients. Similarly, CA125 has been found in the saliva of those with ovarian cancer [223]. Other work has identified salivary biomarkers linked to oral, oesophageal, and pancreatic cancer [224-226]. The latter of these studies is of particular interest as it demonstrates the potential for salivary biomarkers to detect individuals with resectable pancreatic cancer with high specificity and sensitivity [226]. Similar to OAC, pancreatic cancer is often diagnosed late when the majority of patients are incurable.

The use of saliva as a diagnostic tissue carries great potential but is still in its infancy. However, certainly, the potential use for liquid biopsies in the early detection of OAC could significantly impact on its grave prognosis. Similarly, one would want to address the potential for liquid biopsies to also identify those with BO who are at increased risk of OAC. The ability to identify these individuals, particularly those with dysplasia within their BO who are at highest risk, without the need of an endoscopy, could significantly alter the prognosis as well as address the vast sums of healthcare resource that is wasted performing invasive and unpleasant investigations on healthy individuals.

5.2 Aims of chapter

The aim of this chapter is to determine whether transcriptomic biomarkers, linked to those with or at risk of OAC, can be found in blood and / or saliva. The targeted aims are:

1. Can sufficient quality and quantity RNA be extracted from blood and saliva samples for downstream analysis?

2. Does blood and / or saliva contain potentially useful biomarkers for the detection of those with or at risk of OAC?

3. Is the RNA expression data comparable between the two tissue types and thus do the biomarkers found in saliva mirror those found in blood?

5.3 Objective 1 – Establishing RNA extraction methods from blood and saliva

In this preliminary work the focus was establishing the methods for sufficient quality and quantity RNA to be extracted from blood and saliva samples that could potentially be used for downstream analysis.

5.3.1 Objective 1 – RNA extraction from salivary samples utilising spin columns. Methods

The extraction of RNA from patient samples is notoriously problematic due to the fragility of RNA. As such when embarking upon this work the optimisation of the extraction process was going to be key to the downstream results. It was decided initially that the ideal approach to RNA extraction was to utilise spin columns as they provide a quick and simple means of nucleic acid extraction. The focus was to determine how much patient sample was required to provide sufficient RNA, what ratio of phenol / chloroform used for sample lysis provided optimal results and then whether there was a notable difference in outcome depending on the make of spin column used. Prior to use on patient samples, 3 healthy volunteers within the research department provided whole saliva samples for this process. I anticipated that the extraction of RNA from saliva would be most challenging, due to the known degradation, and as such the initial optimisation process focused purely on salivary RNA extraction. Each volunteer was asked to provide a 3x1ml samples of saliva on multiple occasions, each 2-4 days apart. The collection process for the saliva is discussed in chapter 6. The samples were immediately stored in a -80$^{\circ C}$ freezer and were only freeze-thawed once for extraction. On the day of extraction samples were thawed on ice.

In detail, the three variables initially tested were:

1 - Quantity of patient sample

Using a minimal amount of patient sample to obtain RNA is ideal as it allows for the rest of the sample to be stored and used for further testing and validation. Similarly, I found that people can struggle to provide more than 1ml of saliva, particularly after fasting. Thus, the quantities of patient saliva sample initially tested were 200µl, 400µl and 600µl.

2 - Phenol / chloroform ratio

The column-based methods require the sample to be lysed and RNA isolated prior to purification using the columns. This initial step is done using a phenol / chloroform-based method. Following a literature review of RNA extraction using columns on varying tissue types there were multiple varying ratios of phenol to chloroform used for this initial step. The two most common ones identified were 5:1 (i.e. 1ml of phenol to 200µl chloroform) and 4:1 (i.e. 1.2ml of phenol to 300µl chloroform). Thus, these two ratios were tested on the aforementioned differing patient sample quantities.

3 - Column type and centrifugation method

Although there are multiple spin column-based RNA extraction kits available, I decided to test two that had previously been used by Prof. Rifat Hamoudi for RNA extraction in different tissue types from saliva. These were the Qiagen RNeasy (Qiagen, Hilden, Germany) and the RNA Purification and Clean Up (Cambridge Bioscience, Cambridge, UK). The methods used for these steps were as per the manufacturers protocol, however, both the protocols supplied state that centrifugation can be performed at room temperature. Given that I was aware of the fragility of RNA I decided to test both centrifugation at room temperature and at $4^{\circ C}$ which a literature search suggested had been performed for RNA extraction in different tissue types.

For each variable 6 samples were tested, two from each volunteer. An overview of the variables tested is as follows:

Phenol / chloroform ratio	5:1 (1ml Phenol : 200µl chloroform)			4:1 (1.2ml Phenol : 300µl chloroform)		
	Sample quantity					
Column	200µl	400µl	600µl	200µl	400µl	600µl
Qiagen	6 samples	6 samples	6 samples	6 samples	6 samples	6 samples
Qiagen $4^{\circ C}$	6 samples	6 samples	6 samples	6 samples	6 samples	6 samples
Cambridge	6 samples	6 samples	6 samples	6 samples	6 samples	6 samples
Cambridge $4^{\circ C}$	6 samples	6 samples	6 samples	6 samples	6 samples	6 samples

Table 1: Samples for column type and centrifugation method quality control

Turbo™ DNase (Thermo Fisher Scientific, Waltham, USA):

Once successful extraction was confirmed via the NanoDrop (Thermo Fisher Scientific, Waltham, USA), discussed below, the samples underwent the vital removal of traces genomic DNA contamination using Turbo DNase (Thermo Fisher Scientific, Waltham, USA). This is necessary to ensure that there is no genomic DNA within the sample that can affect downstream quantitative real-time PCR (qRT-PCR). For the work within this objective this step is also vital in order to ensure our method outcomes are not influenced by unwanted DNA. We elected to use Turbo DNase (Thermo Fisher Scientific, Waltham, USA) rather than a column-based treatment as this has been demonstrated to provide a more effective process.

The 10x Turbo DNase Buffer (Thermo Fisher Scientific, Waltham, USA) and the Turbo DNase (Thermo Fisher Scientific, Waltham, USA) were added to the extracted RNA sample alongside nuclease free water. This was mixed using a pipette prior to incubation at 37$^{\circ C}$ for 30 minutes. Following this the DNase Inactivation Reagent (Thermo Fisher Scientific, Waltham, USA) was added to the sample and Turbo DNase (Thermo Fisher Scientific, Waltham, USA) mixture and again mixed using a pipette and left at room temperature for 2 minutes. Following this, the mixture was centrifuged at 10,000 rpm at room temperature for 1.5 minutes and then the supernatant was transferred to a new 1.5ml Eppendorf. Again, the sample had initial quality and quantity assessment performed using NanoDrop (Thermo Fisher Scientific, Waltham, USA) to ensure extraction had been successful prior to further quality control using the Bioanalyzer (Agilent Technologies, Santa Clara, USA) and then storage in a -80$^{\circ C}$ freezer.

NanoDrop (Thermo Fisher Scientific, Waltham, USA)

For this work the NanoDrop (Thermo Fisher Scientific, Waltham, USA) alone was used for quality and quantity control to determine whether the extraction process had been successful. The NanoDrop (Thermo Fisher Scientific, Waltham, USA) provided an immediate assessment as to whether the extraction process had been successful. Samples underwent analysis on the NanoDrop 1000 (Thermo Fisher Scientific, Waltham, USA) immediately after the RNA extraction process and again after they had undergone the TurboDNase (Thermo Fisher Scientific, Waltham, USA) step.

The NanoDrop™ (Thermo Fisher Scientific, Waltham, USA) is a spectrophotometer that provides quantity and quality assessment of samples through absorbance measurements although it is generally recognised that the quantity assessment can be an overestimate often caused by sample contamination. Nucleic acids absorb at 260 nanometres (nm) and as such the NanoDrop™ (Thermo Fisher Scientific, Waltham, USA) estimates quality by providing a 260/280 ratio and a 260/230 ratio. The 260/280 ratio should be approximately 1.8 for DNA samples and 2.0 for RNA samples, the higher value for RNA representing the higher 260/280 ratio of uracil (when isolated) compared to thymine (when isolated), although it should be said that this will vary according to the composition of the nucleic acid. The 260/230 ratio is commonly around 2.0 and lower values reflect contamination of the sample in particular, with relevance to this work, with phenol that will come from the Qiazol (Qiagen, Hilden, Germany) used for RNA extraction.

5.3.2 Objective 1 – RNA extraction from salivary samples utilising spin columns. Results

The initial RNA extraction process using a spin column-based approach yielded disappointing results. Generally speaking none of the processes using 200µl of sample yielded any quantifiable RNA (as seen in table 2). Using either 400µl or 600µl of sample produced inconsistent and unreliable RNA yields with the vast majority of these samples also not producing any quantifiable RNA. Those that did yield RNA were from differing volunteers and thus one cannot attribute this to inter-person variability. The best result was a single sample from the 4:1 phenol to chloroform ratio, 400µl of patient sample and using the Qiagen RNeasy kit (Qiagen, Hilden, Germany) at room temperature in which the yield was 19.99µl/ml and the 260/280 and the 260/230 ratios were 1.84 and 1.91 respectively (demonstrated in Figure 13). However, this was the exception rather than the rule. In total, quantifiable RNA was only extracted on 21 occasions, out of a possible 144, and when this occurred it was often at low yields (5-10µl/ml) which makes its downstream use limited (demonstrated in Table 2 and Figure 13). The below table outlines when RNA was obtained using the varying methods:

Phenol / chloroform ratio	5:1 (1ml Phenol : 200µl chloroform)			4:1 (1.2ml Phenol : 300µl chloroform)		
	Sample quantity					
Column	200µl	400µl	600µl	200µl	400µl	600µl
Qiagen	0/6 *(0%)*	2/6 *(33%)*	1/6 *(16%)*	0/6 *(0%)*	3/6 *(50%)*	2/6 *(33%)*
Qiagen 4°C	0/6 *(0%)*	3/6 *(50%)*	2/6 *(33%)*	0/6 *(0%)*	1/6 *(16%)*	1/6 *(16%)*
Cambridge	0/6 *(0%)*	1/6 *(16%)*	0/6 *(0%)*	0/6 *(0%)*	2/6 *(33%)*	0/6 *(0%)*
Cambridge 4°C	0/6 *(0%)*	0/6 *(0%)*	1/6 *(16%)*	0/6 *(0%)*	0/6 *(0%)*	2/6 *(33%)*

Table 2: Results for column type and centrifugation method quality control

Given how infrequently RNA was successfully extracted it is impossible to make any real assessment of which of the variables provided a preferred method. The success of all these methods were so low that this approach could not be used reliably for the work on patient samples. Given that no RNA was extracted using 200µl of patient sample, it was felt that this was too little a sample quantity to take forward. Similarly, there is a trend to suggest that the Qiagen RNeasy kit (Qiagen, Hilden, Germany) outperformed the RNA Purification and Clean Up (Cambridge Bioscience, Cambridge, UK) given that the former obtained some quantifiable RNA on 15 out of the 21 occasions. However, again the RNA that was extracted was often not usable thus rendering these results redundant.

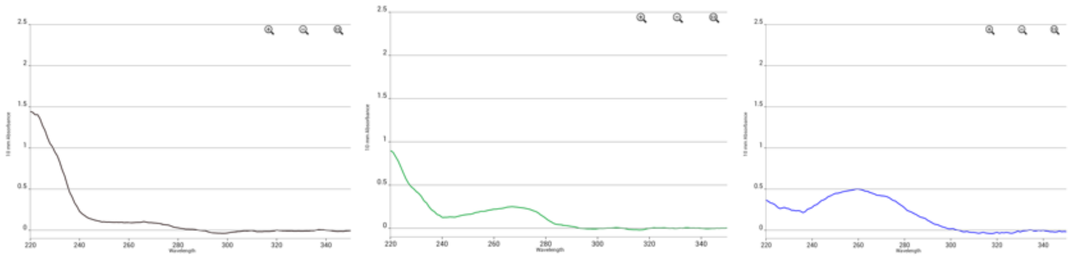


Figure 13: Nanodrop results for column method
a) *Nanodrop demonstrating no RNA yielded*

Figure 13:
b) *Nanodrop demonstrating 8.82µl/ml RNA yielded although likely over-estimate due to contamination*

Figure 13:
c) *Nanodrop demonstrating 19.99µl/ml RNA yielded. 260/280 = 1.84*
260/230 – 1.91

5.3.3 Objective 1 – RNA extraction from salivary samples utilising isopropanol and ethanol precipitation. Methods

The spin column-based approach to RNA extraction did not consistently provide sufficient quantity of RNA for downstream analysis. As such following a literature search and discussion with Prof. Rifat Hamoudi it was decided to attempt RNA extraction using isopropanol and ethanol precipitation and purification after the phenol / chloroform lysis and isolation. Given that the results from the spin column work showed that RNA could be obtained, although inconsistently, with 400µl of sample being used it was decided to keep this amount of sample as a constant when using these new methods. Again, following a literature review of RNA extraction using this method from different tissue types there were 4 variables identified that required testing:

1 – Phenol / chloroform ratio

As discussed above. Given the paucity of data from the column-based approach this was tested again using the same two ratios.

2 – Centrifugation temperature

As discussed above, the fragility of RNA meant that many protocols published in the literature centrifuged at 4°C rather than room temperature. I, therefore, tested both.

3 – Isopropanol / sample ratio

Following the initial phenol / chloroform step the upper 400µl of the aqueous phase is placed in isopropanol for purification and to increase yield. The isopropanol was kept at 4°C. Published literature uses different ratios of sample to isopropanol. The most common 2 were identified and tested. These were 1:1 and 2:1 (i.e. 800µl isopropanol to 400µl sample).

4 – Length and temperature of isopropanol incubation

Again, published literature used differing lengths of incubation period for the sample / isopropanol mix and either at room temperature or on ice. As such three time periods were tested: 15 minutes, 30 minutes, and 60 minutes with samples left on ice or at room temperature in each of these.

The final step of ethanol precipitation was consistent throughout the literature search and thus was kept constant. The ethanol was stored in a -20°C freezer. A summary table of the variables tested is as follows:

Phenol / chloroform ratio	5:1 (1ml Phenol : 200µl chloroform)		4:1 (1.2ml Phenol : 300µl chloroform)	
	Isopropanol sample ratio 1:1	Isopropanol sample ratio 2:1	Isopropanol sample ratio 1:1	Isopropanol sample ratio 2:1
Isopropanol incubation				
Isopropanol 15mins - ICE	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)
	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)
Isopropanol 15mins - RT	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)
	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)
Isopropanol 30mins - ICE	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)
	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)
Isopropanol 30mins - RT	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)
	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)
Isopropanol 60mins - ICE	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)
	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)
Isopropanol 60mins - RT	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)	4 samples (centrifuged at RT)
	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)	4 samples (centrifuged at 4°C)

Table 3: Samples for phenol / chloroform extraction method quality control

As per the methods in 5.3.1 the samples underwent the Turbo DNase (Thermo Fisher Scientific, Waltham, USA) process and quality control using NanoDrop (Thermo Fisher Scientific, Waltham, USA).

5.3.3 Objective 1 – RNA extraction from salivary samples utilising isopropanol and ethanol precipitation. Results

This method was more successful with RNA extracted on 79 out of 192 cases. The highest yield was obtained using 4:1 phenol to chloroform ratio, 1:1 isopropanol to sample ratio with the sample left to incubate for 30 minutes and all centrifugation performed at 4°C. The highest yield was 50.70µl/ml with the 260/280 and 260/230 ratios being 1.93 and 1.8 respectively. A summary of the overall results is outlined in the tables below (one table for room temperature centrifugation and one for 4°C):

CENTRIFUGATION AT ROOM TEMPERATURE

Phenol / chloroform ratio	5:1 (1ml Phenol : 200µl chloroform)		4:1 (1.2ml Phenol : 300µl chloroform)	
	Isopropanol sample ratio 1:1	Isopropanol sample ratio 2:1	Isopropanol sample ratio 1:1	Isopropanol sample ratio 2:1
Isopropanol incubation				
Isopropanol 15mins - ICE	**1/4 yield RNA** *(25%)*	**0/4 yield RNA** *(0%)*	**0/4 yielded RNA** *(0%)*	**0/4 yield RNA** *(0%)*
Isopropanol 15mins - RT	**0/4 yielded RNA** *(0%)*	**0/4 yielded RNA** *(0%)*	**0/4 yielded RNA** *(0%)*	**0/4 yielded RNA** *(0%)*
Isopropanol 30mins - ICE	**0/4 yielded RNA** *(0%)*	**1/4 yield RNA** *(25%)*	**2/4 yield RNA** *(50%)*	**1/4 yield RNA** *(25%)*
Isopropanol 30mins - RT	**2/4 yield RNA** *(50%)*	**0/4 yielded RNA** *(0%)*	**0/4 yielded RNA** *(0%)*	**1/4 yield RNA** *(25%)*
Isopropanol 60mins - ICE	**2/4 yield RNA** *(50%)*	**0/4 yielded RNA** *(0%)*	**0/4 yielded RNA** *(0%)*	**2/4 yield RNA** *(50%)*
Isopropanol 60mins - RT	**1/4 yield RNA** *(25%)*	**0/4 yielded RNA** *(0%)*	**2/4 yield RNA** *(50%)*	**1/4 yield RNA** *(25%)*

CENTRIFUGATION AT 4^{oC}

Phenol / chloroform ratio	5:1 (1ml Phenol : 200µl chloroform)		4:1 (1.2ml Phenol : 300µl chloroform)	
	Isopropanol sample ratio 1:1	Isopropanol sample ratio 2:1	Isopropanol sample ratio 1:1	Isopropanol sample ratio 2:1
Isopropanol incubation				
Isopropanol 15mins - ICE	**1/4 yield RNA** *(25%)*	**2/4 yield RNA** *(50%)*	**2/4 yield RNA** *(50%)*	**1/4 yield RNA** *(25%)*
Isopropanol 15mins - RT	**1/4 yield RNA** *(25%)*	**1/4 yield RNA** *(25%)*	**2/4 yield RNA** *(50%)*	**2/4 yield RNA** *(50%)*
Isopropanol 30mins - ICE	**2/4 yield RNA** *(50%)*	**4/4 yield RNA** *(100%)*	**3/4 yield RNA** *(75%)*	**4/4 yield RNA** *(100%)*
Isopropanol 30mins - RT	**3/4 yield RNA** *(75%)*	**3/4 yield RNA** *(75%)*	**4/4 yield RNA** *(100%)*	**4/4 yield RNA** *(100%)*
Isopropanol 60mins - ICE	**3/4 yield RNA** *(75%)*	**2/4 yield RNA** *(50%)*	**4/4 yield RNA** *(100%)*	**3/4 yield RNA** *(75%)*
Isopropanol 60mins - RT	**3/4 yield RNA** *(75%)*	**2/4 yield RNA** *(50%)*	**3/4 yield RNA** *(75%)*	**4/4 yield RNA** *(100%)*

Table 4: Results for quality control using phenol / chloroform extraction method with centrifugation at room temperature and 4^{oC}

It was necessary to look at each variable individually to determine the optimum method for the extraction of RNA from saliva.

Centrifugation temperature

There is a striking difference in overall consistency of RNA extraction and yields between the room temperature and 4°C groups. RNA extraction was successful in 16 out of 96 samples (17%) centrifuged at room temperature whereas it was successful in 63 out of 96 samples (66%) in the 4°C group. When RNA was extracted at room temperature the yields tended to be low. Within this group the mean yield of the 16 samples in which RNA was obtained was 17.97µl/ml. The mean yield of the 63 samples centrifuged at 4°C in which RNA was obtained was 38.08µl/ml. It is, therefore, obvious that the samples need to be centrifuged at 4°C, which is not surprising given the known fragility of RNA.

Given the inconsistency of RNA extraction when centrifuged at room temperature I did not use the data from this group of samples for analysis of the other variables.

Length and temperature of isopropanol incubation

The next striking observation is that samples that were incubated in isopropanol for 15 minutes, either on ice or at room temperature, only yielded RNA on 12 out of 32 occasions (38%). Whereas samples incubated in isopropanol for 30 minutes or 60 minutes, either on ice or at room temperature, yielded RNA on 51 out of 64 occasions (80%). It is clear, therefore, that a 15-minute incubation period is too short. It is also notable that those incubated for 15 minutes had a significantly lower yield, outlined in Table 10 below. However, when comparing 30 minutes to 60 minutes incubation period there is little difference. RNA was successfully yielded in 27 out of 32 (84%) occasions when incubated for 30 minutes, at either room temperature or on ice, and 24 out of 32 occasions (75%) when incubation was 60 minutes. There are no significant differences in the yields obtained although the quality of sample, according to the 260/280 and 260/230 ratio is marginally better in those incubated for 30 minutes.

INCUBATION PERIOD	No. samples	Mean. yield	260/280	260/230
15 minutes	12/32	22.15µl/ml	1.82	1.24
30 minutes	27/32	45.68µl/ml	1.94	1.42
60 minutes	24/32	46.43µl/ml	1.78	1.35

Table 5: Table demonstrating the quantity and quality of RNA extracted from saliva samples incubated for differing lengths of time in isopropanol

Placing the sample / isopropanol mix on ice for the incubation period also seemed to make little difference. Overall, when samples were incubated on ice RNA was successfully extracted on 31 out of 48 occasions whereas when incubated at room temperature RNA was successfully extracted on 32 out of 48 occasions. Removing the 15-minute incubation period group also makes no difference. RNA extraction was successful 13 out of 16 occasions in the 30-minute group on ice and 14 out of 16 in the 30-minute group at room temperature. When incubating for 60 minutes RNA was extracted in 12 out of 16 occasions in both groups. There is no significant difference in the yields obtained or quality of sample, according to the 260/280 and 260/230 ratios.

INCUBATION TEMP.	No. samples	Mean. yield	260/280	260/230
ICE	31/48	37.46μl/ml	1.90	1.37
ROOM TEMP	32/48	38.69μl/ml	1.87	1.41

Table 6: Table demonstrating the quantity and quality of RNA extracted from saliva samples incubated at different temperatures

Isopropanol / sample ratio

This variable did not appear to make any significant difference in terms of RNA yield. In the 2:1 ratio group RNA was extracted on 32 out of 48 occasions and in the 1:1 ratio group RNA was extracted on 31 out of 48 occasions with the mean yield being similar in both groups. Similarly, incubation on ice or at room temperature did not impact on the isopropanol to sample ratio. Incubation on ice at 1:1 ratio yielded RNA on 15 out of 24 whereas incubation at 1:1 ratio at room temperature yielded RNA on 16 out of 24 occasions. At a 2:1 ratio incubating on ice or at room temperature yielded RNA on 16 out of 24 occasions. Again, little difference was observed in concentration or quality.

Phenol / chloroform ratio

The phenol / chloroform ratio used did make a difference. When using a 5:1 ratio RNA was extracted on 27 out of 48 occasions with the mean yield being 31.37μl/ml. Whereas, when using a 4:1 ratio RNA was extracted on 36 out of 48 occasions with the mean yield being 44.79μl/ml.

Overall, putting these results together one can say that all samples should be centrifuged at 4°C and the ideal phenol / chloroform ratio is 4:1. 15 minutes is too short an isopropanol incubation period but there is little difference between 30 minutes and 60 minutes and the use of ice or room temperature for this period. Logistically thinking, therefore, a shorter period (30 minutes) off ice is the most straight forward and time efficient option. Finally, the ratio of sample to isopropanol also has little impact and thus again logistically thinking using less isopropanol will be more cost efficient when processing more samples. Putting this together when one looks at the group in which all samples are centrifuged at 4°C, a 4:1 phenol / chloroform ratio is used, a 1:1 isopropanol to sample ratio is used with samples incubated for 30 minutes at room temperature one notes that all the samples in this group (4/4) yielded RNA with a mean yield of

47.01µl/ml and acceptable 260/280 and 260/230 ratios were achieved. Importantly, the highest yielding sample is also found within this group (Figure 14). This, therefore, was the preferred method.

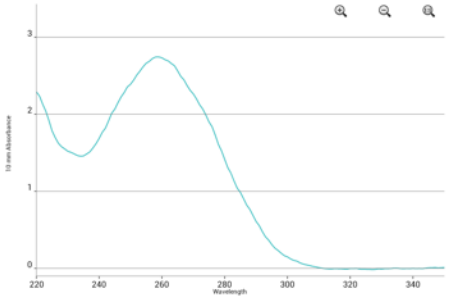


Figure 14: Nanodrop image demonstrating the highest yielding sample using the preferred method

Yield = 50.70µl/ml
260/280 = 1.93
260/230 = 1.80

5.3.4 Objective 1 – RNA extraction from salivary samples optimal method

The method utilised for the extraction of RNA from saliva samples is outlined below;

Samples were allowed to defrost on ice. Once fully thawed they were mixed using a pipette prior to 400µl of the sample being placed in a 2ml Eppendorf. 1.2ml of Qiazol (Qiagen, Hilden, Germany) was added, vortexed, and the mix left at room temperature for 5 minutes. Following this 300µl of chloroform was added to the mix, vortexed and again left for 5 minutes at room temperature. This mix was then centrifuged at 4$^{\circ C}$ for 10 minutes at 14,000 rpm. The upper 400µl of the aqueous phase was pipetted off and placed in a 1.5ml Eppendorf with an equal volume of isopropanol added to it and vortexed. Only the upper 400µl of the aqueous phase was used as it was approximately half of the aqueous phase and therefore avoided some of the contamination of the sample with DNA. This mix was left at room temperature for 30 minutes. This was then centrifuged at 4°C for 30 minutes at 14,000 rpm. The supernatant was then removed using a pipette, carefully avoiding the pellet, and 1ml of 70% ethanol was added to the Eppendorf. This was vortexed for 1 minute and centrifuged at 4$^{\circ C}$ for 7 minutes at 14,000 rpm. Again, the supernatant was carefully removed and the sample allowed to air dry for 10 minutes. Following this 32µl of Diethylpyrocarbonate (DEPC) treated water was added and mixed with a pipette. Following initial quality control assessment using the NanoDrop (Thermo Fisher Scientific, Waltham, USA) to ensure the process had been successful, the samples then underwent the Turbo DNase (Thermo Fisher Scientific, Waltham, USA) process and again quality was checked using the NanoDrop (Thermo Fisher Scientific, Waltham, USA).

5.3.5 Objective 1 – RNA extraction from blood samples. Method

Following the results of the work outlined in section 5.3.4 the preferred extraction method was then applied to blood samples. For this, the same three volunteers provided 3 x 2mls of whole blood collected in a vacutainer blood collection tube containing ethylenediaminetetraacetic acid (EDTA). These were immediately stored in a -80°C freezer. Samples were thawed on ice prior to extraction and again followed the identical Turbo DNase (Thermo Fisher Scientific, Waltham, USA) process and quality checked using the NanoDrop (Thermo Fisher Scientific, Waltham, USA). Samples were stored at -80°C for between 14-28 days.

5.3.6 Objective 1 – RNA extraction from blood samples results

The method, optimised with saliva, worked well for all 18 of the samples tested. In all samples RNA was extracted with the lowest yield being 56.32µl/ml and the highest being 258.74µl/ml. The mean yield of the 18 samples was 144.65µl/ml. The mean 260/280 and 260/230 ratios were 1.86 and 1.92 respectively. The poorest quality sample in regard to the yield and ratios was still of sufficient quantity and quality for downstream use. Given the success of this approach no further optimisation of the extraction process was performed. Examples of the NanoDrop (Thermo Fisher Scientific, Waltham, USA) images for the blood extractions are found below.

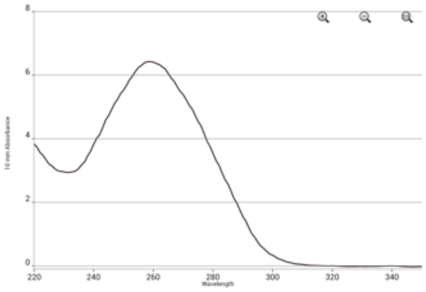


Figure 15: Nanodrop results for RNA extraction from blood
a) *Nanodrop demonstrating 258.74µl/ml of RNA yielded from blood.*
260/280 = 1.82

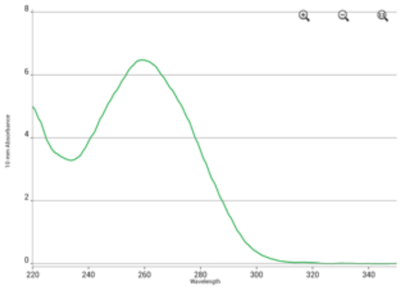


Figure 15:
b) *Nanodrop demonstrating 135.04µl/ml of RNA yielded from blood.*
260/280 = 1.83
260/230 = 1.91

As a consequence of these results the extraction method outlined in section 5.3.4 was used for both blood and saliva samples.

5.4 Objective 2 – Do patient blood and saliva samples contain useful biomarkers for the detection of those with or at risk of OAC?

Once extraction methods had been established the focus now moved towards whether these tissue types could potentially provide useful biomarkers.

5.4.1 Objective 2 - Patient recruitment and sample collection

The approval to collect samples for this, and other work within this project, was within the BOOST (ISRCTN 58235785) study outlined in Chapter 4. The inclusion and exclusion criteria are as follows:

Inclusion Criteria

1. Patients will be recruited from those with Barrett's oesophagus or oesophageal cancer undergoing endoscopy
2. Patients without Barrett's oesophagus attending for a clinically indicated endoscopy may be recruited as controls.
3. Patients must sign an informed consent form.

Exclusion criteria

1. Patients in whom endoscopy and biopsy is contraindicated.
2. Patients who are unable to give informed consent
3. Pregnant women
4. People under the age of 21 years.
5. People who are non-English speakers (due to potential issues in the collection of accurate questionnaire data, chapter 5).

Once enrolled patients consented for questionnaire completion and the provision of saliva, blood and biopsy samples. Individuals were recruited who were attending UCLH to undergo an upper gastrointestinal endoscopy. For the purposes of this work the patients were grouped accordingly;

Normal – Those referred for an upper gastrointestinal endoscopy based on symptoms of dyspepsia or GORD or those referred along the expedited two-week wait pathway (as discussed in Chapter 1) who have a normal upper gastrointestinal tract found on endoscopy. Mild benign pathology, such as mild oesophagitis or gastritis, was allowed to be recruited as normal. These individuals must not be found to have a cancer or any new active disease in any other investigations performed as part of this referral.

NDBO (low risk category) – Those known to UCLH for the surveillance of their NDBO. These individuals must have no history of dysplasia within their BO and no other active cancers.

HGD (high-risk category) – Those referred to UCLH for confirmation and / or treatment of the HGD within their BO. Those recruited must not have undergone previous photodynamic therapy or RFA for their HGD. EMR was permitted prior to recruitment. These individuals should also not have any other active cancers.

OAC – Those with known OAC. These individuals must not have undergone any chemo or radiotherapy prior to recruitment. They must also not have any other active cancer.

Patients eligible for recruitment received an information sheet outlining the research and their potential involvement at least 48 hours prior to attending UCLH for their endoscopic procedure. All patients were recruited prior to their endoscopy. On the day of the procedure informed consent was taken by a member of the research team and patients were offered the opportunity to ask any questions. Once consented, patients completed the enhanced questionnaire (discussed in chapter 4) and provided a 1ml saliva sample collected in 15ml Corning centrifuge tubes (Sigma-Aldrich, St Louis, USA) and 3mls of blood collected in a vacutainer blood collection tube containing ethylenediaminetetraacetic acid (EDTA). The intricacies of the saliva collection process are discussed further in Chapter 6. These samples were then immediately stored in a -80°C freezer.

Patients were placed in one of the four diagnostic groups following review of the endoscopy report and corresponding histology (if biopsies were taken) and review of the patient's medical history obtained via the hospital computer records system. Diagnoses were then confirmed by another physician following an identical review of the hospital records.

For this work the number of patients with matched blood and saliva that underwent the RNA extraction process and analysis were as follows:

Diagnosis	No. of patients recruited
Normal	7
NDBO	5
HGD	7
OAC	7

Table 7: Number of patients recruited into each diagnostic category for matched blood and saliva analysis

The power calculation for this work was performed by Dr Rifat Hamoudi who based this on whole transcriptomic data, where it has been shown that the standard deviation for detection of differentially expressed genes is 0.35 (σ = 0.35) and the delta (the measure of effect) is 1 [227, 228]. The power calculation was carried out using R statistical software version 2.14.2 with *p* = 0.01 (1% significance testing) and power of 90%. This found that minimum number of patients is n = 5.55289 suggesting we required a minimum of 6 samples in each group in order for a significant transcriptomic signature to be discovered.

5.4.2 *Objective 2 – RNA extraction, quality control and targeted expression analysis. Methods*

The extraction method outlined in 5.3.4 was used for the RNA extraction in the patient samples, however, they also underwent a further quality control assessment prior to targeted expression analysis utilising the Bioanalyzer (Agilent Technologies, Santa Clara, USA).

Bioanalyzer (Agilent Technologies, Santa Clara, USA).

This utilises chips, such as the Nanochip, to generate a reliable assessment of RNA quality and quantity. The chip comprises of sample wells, gel wells and an external standard (ladder) interconnected with a series of channels. During the preparation of the chip it is filled with a sieving polymer and fluorescent dye which is then followed by the samples and ladder into their respective wells. Once filled the chip becomes an electrical circuit with the 16-pin electrodes of the cartridge arranged to fit into the wells of the chip. The charged biomolecules, such as RNA, are driven electrophoretically by a voltage gradient through the sieving polymer matrix and separated by size. The dye intercalates with the RNA and is detected by laser-induced fluorescence. The Bioanalyzer (Agilent Technologies, Santa Clara, USA) generates data into electrophoresis like gels and electropherograms.

Quality control, using the Bioanalyzer (Agilent Technologies, Santa Clara, USA), was carried out using the RNA 6000 Nano Kit (5067-1511, Agilent, Santa Clara, CA, USA) according to the manufacturer’s instructions. 1µl of each sample and RNA ladder were aliquoted and incubated at 70°C for 2min in thermal cycler. 1µl nano dye concentrate was added to 65µl aliquot of filtered nano gel matrix and centrifuged for 10minutes at room temperature at 14,000rpm.

The RNA Nanochip was primed on chip priming station, with 9µl gel-dye remix pipetted into the well labelled “G” in the third row of the nano chip. 9µl of gel-dye mix was pipetted into the other 2 “G” wells. 5µl of the RNA 6000 nano marker was pipetted into the “ladder” well and 12 sample wells. 1µl of RNA ladder was pipetted into the ladder well and 1µl of RNA sample were pipetted into each well (Figure 7). The Nanochip was vortexed for 60seconds at 2000rpm with the chip MS 3 vortexer (*IKA*) and read on the Bioanalyzer (Agilent Technologies, Santa Clara, USA).

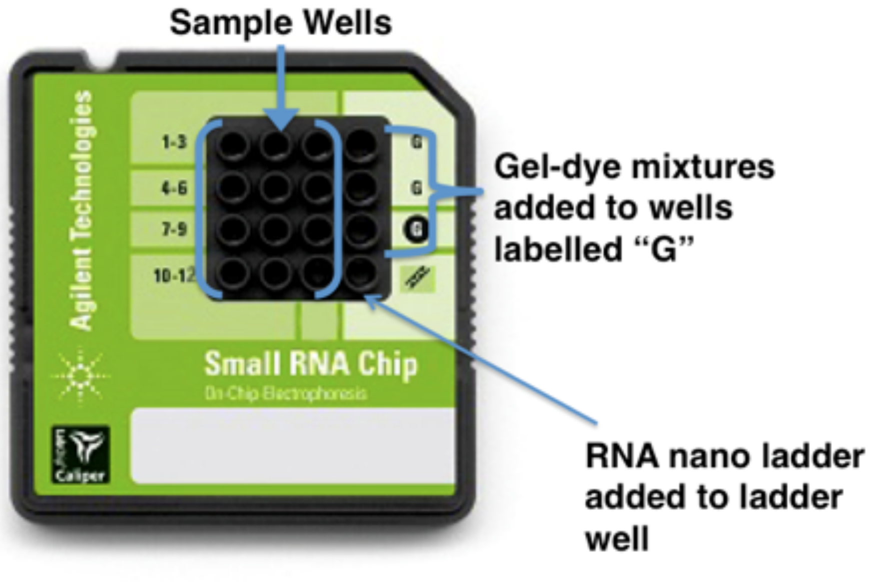


Figure 16: Pictorial representation of configuration of wells on the Agilent RNA 6000 nano chip

Expert 2100 version B.02.08 (Agilent, Santa Carla, CA, USA) was used for bioanalysis.

Figure 17 below demonstrates typical matched blood and saliva Bioanalyzer (Agilent Technologies, Santa Clara, USA) images. From this it is notable how degraded the salivary RNA sample is. The issues surrounding the degradation of salivary RNA are discussed in chapter 6. The average RNA integrity number (RIN), a measure of RNA integrity and quality, for the salivary RNA extracted was 3.85 (range 1.7 to 6) whereas the average RIN value for the blood RNA extracted was 7.8 (range 6.2 to 8.5). This is in keeping with published literature discussed in section 7.7.

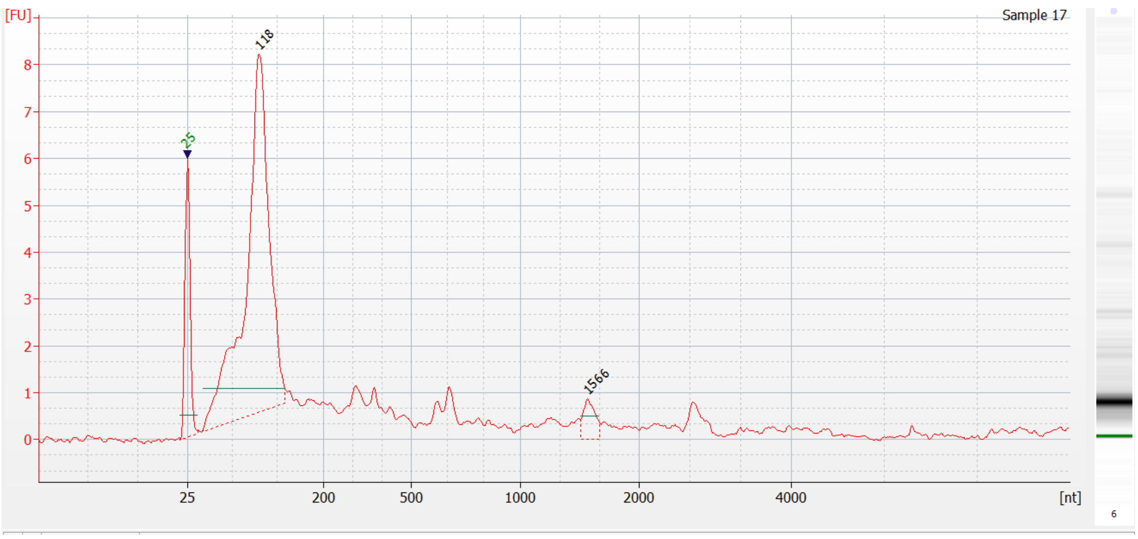


Figure 17a: Bioanalyzer (Agilent Technologies, Santa Clara, USA) image from salivary sample. RIN value 3.4

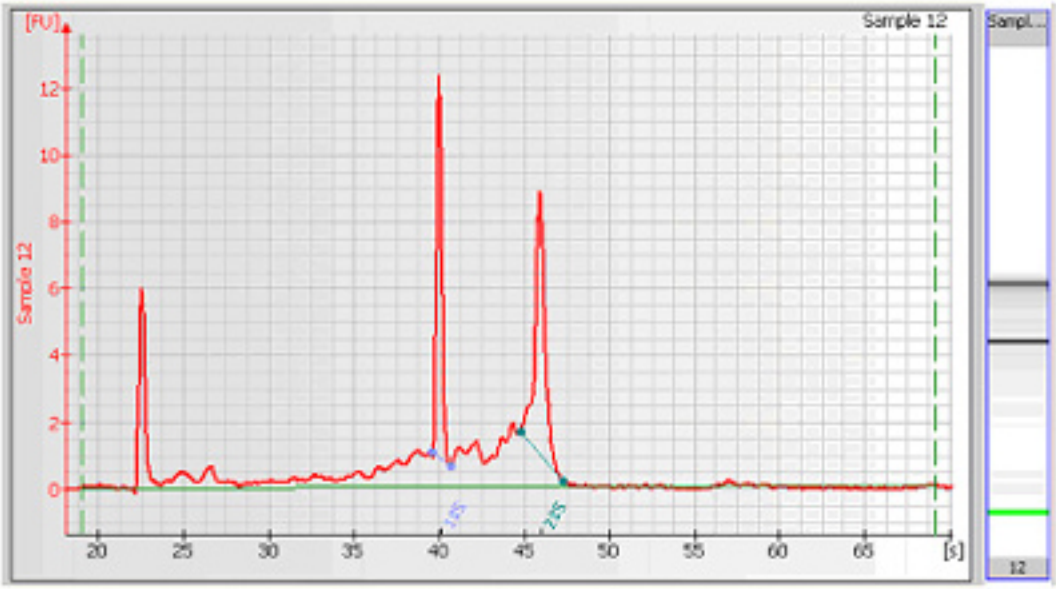


Figure 17b: Bioanalyzer (Agilent Technologies, Santa Clara, USA) image from blood sample. RIN value 7.9

Primer design and selection

The targets selected to be tested on these extracted samples were those known to be linked to cancer development or their risk factors following a thorough literature search and reanalysis of publicly available transcriptomic dataset using absolute Gene Set Enrichment Analysis according to *Hamoudi et al.* and *Rosebeck et al.* [229, 230] and the genes that are differentially expressed were checked for mutations using COSMIC database [231]. The data set were obtained from published data [232-235].

The literature review was performed by myself and Prof. Hamoudi, whilst the reanalysis of the dataset was performed by Prof. Hamoudi alone. For this preliminary work 6 targets were identified and tested which were known to be linked to OAC. However, the literature search and reanalysis of the dataset also identified additional targets which were used in further work outlined in chapter 6. These additional targets included those linked with inflammation and the associated carcinogenesis [236-238].

For this preliminary work 5 genes were selected: TP53, CDKN2a, SMAD7, TLR6 and AMY2 with 2 different exons selected for CDKN2a, making 6 targets in total.

Primers were designed by Prof. Hamoudi to be less than 150bp in length in order to cope with degraded samples [239], melting temperature (Tm) of 60°C and guanine-cytosine (GC) content of 30% or higher. The mutation targeted had around 25-30 base pair (bp) padding space, i.e. the mutation is generally in the middle of the amplicon. Primers are tested individually for specificity and to validate the correct length is obtained.

AMY2_F	TATAACTGTTCGTATTTCCCGG
AMY2_R	CTCATTAAATAGAGAAGCTAGC
CDKN2A_R80_F	CGGTGCAGCACCACCAGC
CDKN2A_R80_R	AGCGCCCGAGTGGCGGAGCTGC
CDKN2A_R58_F	ACTGCAACCTCCACCTCCCAGG
CDKN2A_R58_R	TATGGTGAAACCCCATCTCTTC
SMAD7_118_F	GTCCAGGGTCCTCCCTCCTCAG
SMAD7_118_R	TATCAGAACAGCCAATTTCCTG
TP53-6_F	CAGTTGCAAACCAGACCTCAG
TP53-6_R	CTCCTCAGCATCTTATCCGAGT
TLR6_F	AACAAGTACCACAAGCTGAAG
TLR6_R	CTCTAATGTTAGCCCAAAAGAG

cDNA synthesis

First-strand cDNA synthesis was conducted on the extracted RNA samples using the SuperScript First-Strand Synthesis System for RT-PCR (Part No.: 11904018, ThermoFisher Scientific, Waltham, MA, USA). An approximate amount of sample RNA required to obtain 200ng of RNA for each mixture was calculated. A RNA -

primer mixture was prepared with the components outlined in Table 8 with the final sample volume being 10µl in each PCR tube. Table 5 demonstrates an example of three of the samples used for this work. 18S rRNA gene specific primer (GSP) was used as an internal reference gene for normalisation and comparison of gene expression levels. Normalisation accounts for potential error in nucleic acid loading during the qPCR process[240]. 18S rRNA is a ribosomal structural gene and was chosen as it is stably expressed, regardless of experimental conditions. A No Template Control (NTC) mixture was prepared as well to detect any background signal.

RNA sample type	Concentration of eluted RNA (ng/µl)	Volume of eluted RNA (µl)	Volume of dNTP (µl)	Volume of 18S rRNA GSP (µl)	Volume of GSP anti-sense primer (µl)	Volume of nuclease free water (µl)
Saliva	34.5	5.8	1	1	1	2.2
Saliva	45.6	4.39	1	1	1	3.61
Saliva	53.8	3.72	1	1	1	4.28
NTC	N/A	N/A	1	1	1	7

Table 8: Volumes of various components in RNA/primer mixtures in first-strand cDNA synthesis

The RNA/primer mixture was incubated at 65°C for 5 minutes in the Applied Biosystems 2720 thermal cycler (Thermofisher Scientific, Inc, Waltham, MA, USA), and cooled on ice for 2 minutes thereafter. 10µl reaction mixture (Table 8) was added and samples incubated at 50°C for 50minutes, followed by 85°C for 15minutes. 2µl RNase H was added and samples were incubated at 20minutes at 37°C. Incubation times and temperatures are summarized in Figure 18.

Components of reaction master mixture	Volumes added for each reaction (µl)
10x RT Buffer	2
25nM $MgCl_2$	4
0.1M DTT	2
RNaseOUT™Recombinant Ribonuclease Inhibitor	1
SuperScript® III Reverse Transcriptase	1
Total	10

Table 9: Volumes of various components in reaction RNA/primer master mixture for first-strand cDNA synthesis

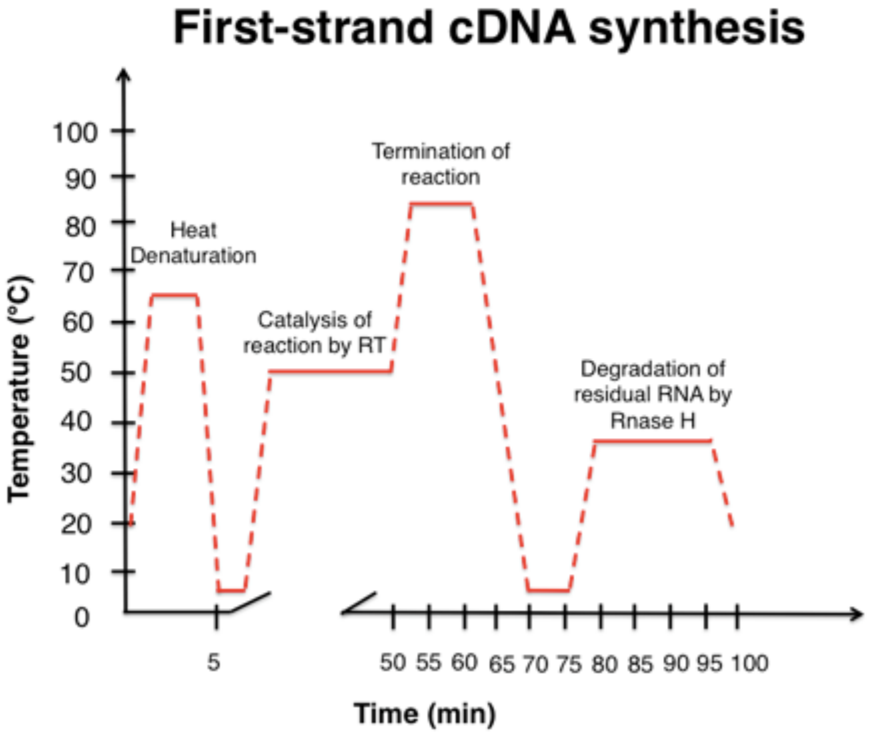


Figure 18: Illustration of steps for first-strand cDNA synthesis. The RNA/primer mixtures are first incubated at 65°C for 5min to denature double-stranded RNA. The samples were then incubated with the reaction master mixtures incubated at 50°C for 50min to allow the synthesis of cDNA strand on the RNA template. The reaction was terminated by incubating the samples at 85°C for 15min. The samples were then incubated with RNase H for 20min at 37°C to degrade any remaining RNA molecules.

qRT-PCR

9µl of the master mixture (Table 10) and 1µl of the sample cDNA was pipetted into each well of aHard-Shell96-Well PCR Plate (HSP9641, Bio-Rad, Inc., Hercules, CA, USA). Each sample was pipetted in triplicate in addition to NTC. The plate was placed in the CFX Connect Thermal Cycler (185-5201, Bio-Rad, Inc., Hercules, CA, USA) and incubated according to the protocol described in Figure 19. Melting curve analysis was performed in addition to real-time quantification of amplicons.

Components of reaction master mixture	Volumes added for each reaction (µl)
iTaq™ Universal SYBR® Green supermix by Biorad (172-5121)	5
Sense GSP	0.5
Anti-sense GSP	0.5
Nuclease free water	3
Total	9

Table 10: Volumes of various components in master mixture for qRT-PCR

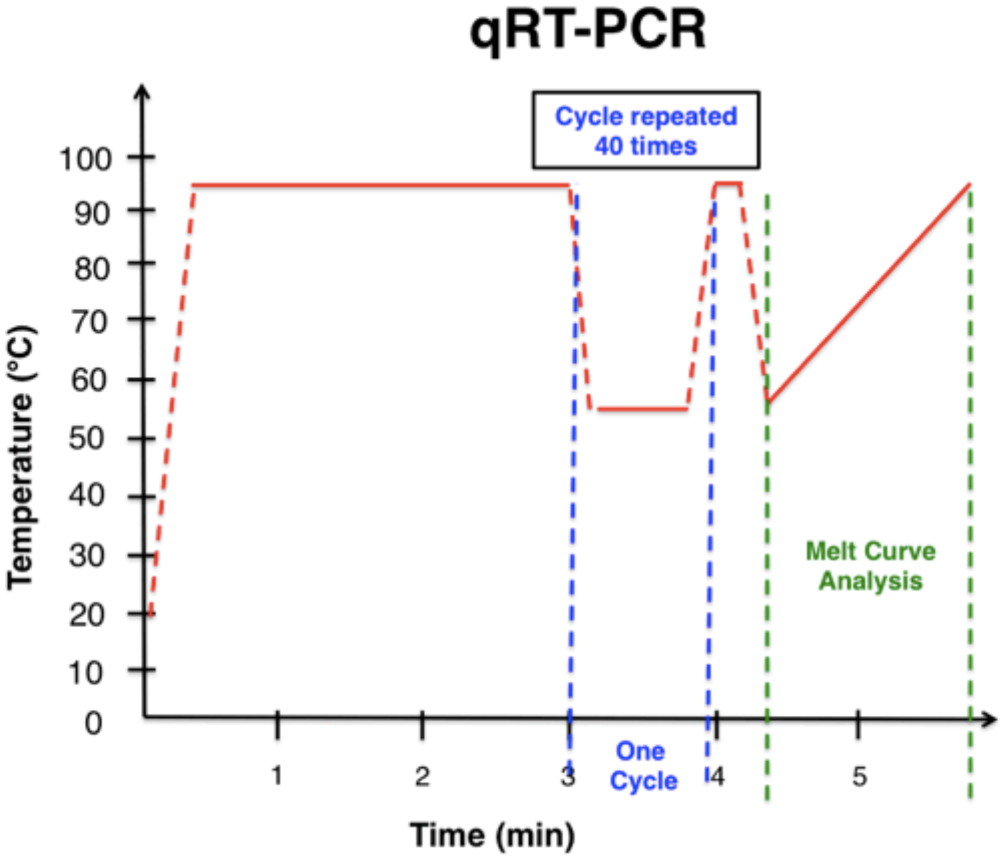


Figure 19: Protocol of qRT-PCR cycling. The plate was heated at 95°^{C} for 3minutes. One PCR cycle includes heating the plate at 95°^{C} for 10seconds, followed by heating at 55°^{C} for 45seconds. 40 cycles were carried out, followed by a melt curve analysis, in which the temperature was increased by 0.5°^{C} for 5seconds from 55°^{C} to 95°^{C}.

Statistical analysis

Statistical analysis was performed by myself and Prof. Hamoudi with Microsoft Excel 2011 and IBM SPSS Statistics 22. Relative expression was calculated by comparing the cycle threshold (Ct) from each gene with that from the 18S rRNA which is the housekeeping gene. One-way analysis of variance (ANOVA) using Bonferroni post-hoc correction was used to carry out the multiple comparisons. $p < 0.05$ taken to be statistically significant, was used to analyse qRT-PCR results. Bio-rad CFX Manager was used for curve fitting and regression equation in qPCR.

For each sample a ΔCt was calculated by comparing the Ct of each amplicon against 18S rRNA (housekeeping gene). A ΔΔCt was also calculated in order to compare the four categories of recruited patients (Normal v NDBO v HGD v Cancer) and demonstrate trends in data. The normal group are used as the category to compare against and the fold change is calculated as a ratio compared to the control.

5.4.3 Objective 2 – Targeted expression analysis results of matched blood and saliva patient RNA samples

The below table demonstrates the key demographic data of the patients analysed:

Diagnosis	No. of patients recruited	Sex	Av. Age	Av. Length of BO	Staging of cancer
Normal	7	5F, 2M	66	N/A	N/A
NDBO	5	5M	59	C3M5	N/A
HGD	7	7M	67	C7M8	N/A
OAC	7	6MF1	70	C3M6 Not recorded in 4 cases	3 x T1N0M0 T2N1M0 T2N1M1 T2N2M1 T3N1M1

Table 11: Demographic data for patients with matched blood and saliva undergoing analysis

When analysing the data on this table one initially notes that in the NDBO, HGD and OAC groups the individuals recruited are predominantly male which reflects the gender disparity of these diseases (discussed in chapter 1). However, within the normal group the ratio of female to male patients is 5:2. The average age of those recruited is similar in all four groups and again reflects the age at which one would expect patients to present with BO or OAC. The length of BO is similar in those with NDBO and OAC, although the length of BO was not recorded in 4 out of 7 cases in the OAC group and is longest in those with HGD. In all three of these categories the average length of BO would be described as a long segment and thus be thought to be of higher risk. Finally, in the OAC group the staging of cancer is evenly split with 4 patients having resectable (curable) disease and 3 with incurable disease.

Following targeted expression analysis, a ΔCt was calculated for each sample and then an average ΔCt was calculated for each diagnostic group. A ΔΔCt was then calculated using the normal patients as the reference. The table below demonstrates these results, with complete results found at the end of the chapter:

BLOOD

	ΔCt					
	TP53-6	CDKN2a_r58	SMAD7-118	AMY2	CDKN2a-80	TLR6
DIAGNOSIS						
Normal	13.67	3.68	16.6	19.86	12.9	17.88
NDBO	16.23	10.18	13.3	21.78	12.83	21.02
HGD	11.44	2.07	14.16	17.79	11.05	13.35
OAC	13.3	3.5	12.8	19.66	11.37	15.82
	ΔΔCt – Fold change					
DIAGNOSIS						
Normal	1.00	1.00	1.00	1.00	1.00	1.00
NDBO	1.19	2.77	0.80	1.10	0.99	1.18
HGD	0.84	0.56	0.85	0.90	0.86	0.75
OAC	0.97	0.95	0.77	0.99	0.88	0.88

SALIVA

	ΔCt					
	TP53-6	CDKN2a_r58	SMAD7-118	AMY2	CDKN2a-80	TLR6
DIAGNOSIS						
Normal	11.39	2.63	13.36	19.44	10	10.36
NDBO	11.81	6.32	10.48	19.95	9.61	8.08
HGD	14.04	2.96	14.62	21.52	10.01	13.24
OAC	11.39	-0.81	10	17.42	8.72	13.39
	ΔΔCt – Fold change					
DIAGNOSIS						
Normal	1.00	1.00	1.00	1.00	1.00	1.00
NDBO	1.04	2.40	0.78	1.03	0.96	0.78
HGD	1.23	1.13	1.09	1.11	1.00	1.28
OAC	1.00	-0.31	0.75	0.90	0.87	1.29

Table 12: Expression analysis results for patients with matched blood and saliva

The numbers of patients in each diagnostic group are too small to be able to analyse them for use as a diagnostic tool. Thus, one cannot interpret these results and make an assessment of an individual primer's ability to detect those with or at risk of OAC. However, importantly one can see from these results that there are changes in the expression of the primers seen in both blood and saliva samples which would suggest that they are an adequate tissue to attempt to obtain informative data to identify those with or at risk of OAC. One can see from these results that there is aberrant over-expression of two of these primers in the saliva of those with NDBO and 4 of these primers in those with OAC. There is under-expression of 3 primers in HGD and 1 in OAC. For blood aberrant over expression is seen in 1 primer and under expression in 2 primers in those with NDBO, all 6 primers show over expression in those with HGD and 3 show over expression with OAC. Whilst the cohort is too small to determine whether this is truly significant these results do provide promise.

Furthermore, one can use this data to study the trends of the expression of these primers to determine whether blood and saliva have a similar expression pattern, in keeping with the theory behind liquid biopsies and salivary diagnostics. This is achieved by looking at the fold change, in which a difference of 0.1 signifies a 10-fold greater (or lesser) amount of expression. Thus, if a primer has a fold change of 0.9, when compared to the normal, then it is expressed approximately 10-fold higher in that diagnostic group than in the normal group. Looking at the SMAD7 primer one can see that there is a similar trend in terms of expression profile throughout the diagnostic groups (Figure 20), although in the saliva SMAD7 group there is no difference between SMAD7 expression in those with HGD whereas in the blood group there continues to be increased SMAD7 expression in the HGD group.

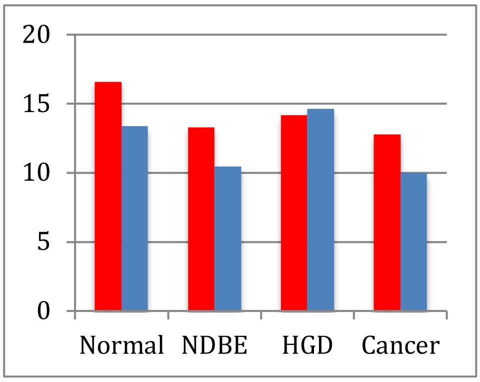


Figure 20
Comparison of blood and saliva across the 4 diagnostic groups using SMAD7 as the primer.

FOLD CHANGE

DIAG.	Blood	Saliva
Norm	1.00	1.00
NDBO	0.80	0.78
HGD	0.85	1.09
OAC	0.77	0.75

Another similar trend, from these results is that in the CDKN2a-r58, CDKN2a-80 and AMY2 primers is that the significant aberrant expression of these primers occurs in the blood of patients with HGD (but not OAC) and in the saliva of patients with OAC (but not in HGD). These results are discussed further in Chapter 6.

	CDKN2a-r58		AMY2		CDKN2a-80	
DIAGNOSIS	Blood	Saliva	Blood	Saliva	Blood	Saliva
HGD	0.56	1.13	0.90	1.11	0.86	1.00
OAC	0.95	-0.31	0.99	0.90	0.88	0.87

Table 13: Table demonstrating the fold change of 3 primers when comparing blood and saliva in patients with HGD and OAC.

Interestingly also, aberrant expression is occurring in primers at points in which published literature support this data. For example, in both blood and saliva one can note that there is aberrantly higher expression of SMAD7 in patients with OAC when compared to those that are normal. SMAD7 is discussed further in Chapter 6, however, it has been published that higher levels of SMAD7 expression are seen in colorectal cancer. Similarly, also, TP53 is significantly aberrantly expressed in the blood of those with HGD which again is in keeping with published work by *Weaver et al* looking at mutations in BO using biopsy samples [1].

5.5 Discussion

When reviewing the literature behind means by which RNA extraction has been approached, in different tissue types, one notes variation in practice. Whilst these variations may appear to be minor, for example 5:1 phenol to chloroform ratio versus 4:1, they do seem to impact on the outcome of the RNA extracted in this work. As such, overall, this work highlights the issue of variation in technique, even on an individual basis, which can significantly impact on the downstream data. In the circumstances of research on small patient numbers this issue matters less as it is likely that only one or two technicians will be performing the work and therefore ensure their methods are identical. However, thinking further ahead should these techniques be used for population screening then one would need to ensure that all extraction techniques are identical which becomes challenging if there are many people, perhaps in different laboratories, performing the work.

One of the early striking aspects of this work were the poor results obtained using spin column-based extraction kits to obtain RNA. The column-based approach is used widely in molecular biology as it allows for a quick and simple means to extract DNA and RNA from varying tissues. Therefore, the use of a column-based approach is ideal as it allows for more consistency in technique which would eliminate the aforementioned issues of individual variation in technique. There is little in the way of published literature on comparing the methods of RNA extraction from saliva. *Pandit et al* published work comparing a similar isopropanol and ethanol precipitation technique to a commercially available spin column-base kit and found that the former out-performed the spin columns providing higher yields and better quality RNA [241]. However, in contrast to this although published literature on utilising salivary RNA is sparse other published work did use spin columns to obtain their RNA [224, 242]. Clearly, therefore, these groups were able to achieve more consistency using spin columns than I was able to replicate. A possible reason behind the poor results achieved with the spin columns is that the kits I used are designed to obtain RNA from tissue types such as cells and tissue in which the RNA is of higher concentration and less degraded. As such it is possible that in highly degraded tissues, such as saliva, the washing steps in the spin columns are too aggressive and cause the RNA not to remain bound to the silica. Whilst the pellet approach with the isopropanol and ethanol precipitation is more labour intensive it does allow for a gentler washing step.

Centrifugation at room temperature yielded significantly worse results than when centrifuged at 4°C. Given how degraded the salivary RNA sample is, an issue that is further discussed in Chapter 6, there is little room for further degradation of the RNA. It is perhaps no surprise then that using chilled reagents and keeping the temperature of the centrifuge at 4°C allows for improved yields by ensuring that any lingering nucleases remain inactive. Centrifugation at 4°C is used in the majority of protocols that are published and also form part of the recommended RNA extraction methods utilised by *Pandit et al.* [241].

It is notable from the Bioanalyzer (Agilent Technologies, Santa Clara, USA) images how poor the quality of the salivary sample is. The average RIN value of the salivary samples extracted was 3.85. This is due to the salivary enzymes and other proteins that degrade RNA and in keeping with published literature in regard to using saliva. However, this poor quality does not prevent its use for both targeted q-RT PCR and sequencing work using molecular biology methods such as shorter fragments and different bioinformatics strategies. There is an abundance of literature using FFPE samples for molecular analysis which often has a RIN value of less than 2. A value of greater than 1.4 is the accepted cut-off for the ability to use the sample [243]. Whilst the use of snap-frozen samples remains the gold-standard there is a plethora of literature in which FFPE samples with similar degradation patterns as that seen in saliva that have been used for whole genome and transcriptome sequencing as well as targeted work [244, 245]. Whilst the field of salivary diagnostics remains new there is even still work using saliva for whole transcriptome sequencing [246].Thus whilst the degradation of the salivary sample does present some problems it does not devalue its use as a diagnostic tissue.

The necessity of performing a DNase step is due to the issues of PCR products becoming contaminated with genomic DNA [247], however, the use of Turbo DNase (Thermo Fisher Scientific, Waltham, USA) to ensure removal of all DNA prior to targeted expression analysis is not performed by all. One reason behind this is that many of the washing reagents found within spin column kits contain DNase and as such it is felt by many that performing a separate DNase step, if a spin column has been used, is unnecessary. However, there is evidence from transcriptome sequencing that the DNase within kits is insufficient to remove all traces of genomic DNA and as such can cause downstream analytical issues especially in highly sensitive methods such as RNA sequencing.

However, even when not using spin columns many people still do not perform a DNase step prior to targeted expression analysis. The justification behind this is that primers can be designed to anneal to sequences in exons on both sides of an intron, thus making any product amplified by genomic DNA too large to cause contamination. Alternatively, primers can be designed to span exon/exon boundaries and thus span more than one intron resulting in them being unable to amplify genomic DNA. However, there are two issues with this approach. Firstly, approximately 5% of genomic DNA is intron-less which thus can cause issues with contamination of the PCR product regardless of primer design [247]. Secondly, when using saliva, one is dealing with highly degraded samples, providing low yields, and two of our target genes used (AMY2 and SMAD7) have low expression levels that are that are mono-exonic. Thus intron-spanning PCR primers cannot be used. This, therefore, makes the Turbo DNase (Thermo Fisher Scientific, Waltham, USA) process vital.

When deciding to compare blood and saliva the decision was made to use whole blood rather than including fractionation into the components prior to RNA extraction. There are issues with this approach. It is noted that there are a high proportion of globin mRNA present in whole blood which can reduce the efficacy of expression profiling by inhibiting the detection of less abundant transcripts. Similarly, it has been noted that using whole blood results in increased noise and reduced sensitivity in gene expression assays [248]. However, there are reasons to study whole blood rather than its subpopulations. Firstly, from a practicality point of view the use of whole blood is quicker and less steps, in regard to fractionating the blood, leads to less opportunity for variation in individual methods and thus minimise sample to sample variability. Secondly, the handling of the sample during fractionating leads to artefacts and cell activation that impacts on downstream analysis. Thirdly, even following fractionating blood globin mRNA often remains the most abundant transcript and thus can affect the analysis. Finally, there are an abundance of subpopulations found within blood that each may contain valuable information when analysed. Analysing each individual population is costly and time-consuming. Thus, given that this work was performed to see whether there is the potential for blood to be used as a screening tissue it was felt most appropriate to use whole blood initially.

Whilst this work did demonstrate similar trends in the expression profile of a few genes between blood and saliva, it was not shown across all the genes. This is to be expected as studies have shown partial concordance of mutations between different tissues [249]. The origins of RNA within saliva are complex with oral epithelial cells, the microbial species and exosomes all contributing, with the majority thought to be from exosomes [250, 251]. Thus, the expression profile found within saliva will likely reflect the exosomal expression profile, although there may be variation depending on other factors in the local environment. Similarly, the expression profile of blood will be determined by a multitude of processes occurring throughout the body. Thus, one cannot expect the expression profile of these two tissues to match exactly on every gene. The RNA found within exosomes is thought to be key for intracellular communication and even play a role in tumour progression and metastasis. It is likely that the salivary expression profile will contain transcripts from a lesion should there be a driver mutations present. Thus, once a process becomes significant one would expect to observe the changes in expression profile within the saliva. This may go some way to explain why the results from this work observe aberrant expression within the blood in patients with HGD that then become significant in the saliva of those with OAC. It is possible that this earlier stage does not result in the release of diagnostically useful exosomes containing RNA from the genes I tested, which then find their way into the saliva. However, within this work we have only tested a limited number of primers, most of which are linked to cancer. One observes this phenomenon predominantly in the CDKN2a gene which is strongly linked to neoplastic change. However, should one want to detect exosomal RNA linked to NDBO or HGD then it is possible that different targets would identify this change. For example, in this work one notes the significant aberrant expression of TLR6 in those with NDBO. This finding is supported by recent work published by *Huhta et al.l* in which it was noted that TLR6 has increasing expression from normal to NDBO, to HGD, and OAC [252]. The reason that this change in salivary expression is no longer significant in those with HGD and OAC may be because more dominant processes are occurring, thus overshadowing TLR6, and are no longer reflected in the salivary exosomal RNA.

5.6 Chapter conclusions

In reference to the aims and objectives of this chapter, this work demonstrated that as proof of concept it is possible to extract sufficient quantity and quality

RNA from blood and saliva for downstream expression profiling. In addition to this, the data produced in this work suggests that blood and saliva do potentially contain useful RNA expression biomarkers that can identify those with or at risk of OAC with CDKN2a and AMY2 giving the most significant fold change from this set of biomarkers. Although the number of patients analysed is too small to determine whether these differences are significant, this early data is encouraging. Whilst we have analysed a small number of patients and only used 6 primers one can also note trends in the expression data that are similar between blood and saliva. This is in keeping with the theory behind salivary diagnostics.

5.7 Complete results for section 5.4.3

Table 14: Mean ΔCt value and fold change for each primer at each diagnosis using blood and saliva

Table 14a: ΔCt

	TP53-6		CDKN2a_r58		SMAD7_118		AMY2		CDKN2a-80		TLR6	
Diagnosis	Blood	Saliva	Blood	Saliva	Blood	Saliva	Blood	Saliva	Blood	Saliva	Blood	Saliva
Normal	13.67	11.39	3.68	2.63	16.6	13.36	19.86	19.44	12.9	10	17.88	10.36
NDBO	16.23	11.81	10.18	6.32	13.3	10.48	21.78	19.95	12.83	9.61	21.02	8.08
HGD	11.44	14.04	2.07	2.96	14.16	14.62	17.79	21.52	11.05	10.01	13.35	13.24
OAC	13.3	11.39	3.5	-0.81	12.8	10	19.66	17.42	11.37	8.72	15.82	13.39

Table 14b: Fold Change

	TP53-6		CDKN2a_r58		SMAD7_118		AMY2		CDKN2a-80		TLR6	
Diagnosis	Blood	Saliva	Blood	Saliva	Blood	Saliva	Blood	Saliva	Blood	Saliva	Blood	Saliva
Normal	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NDBO	1.19	1.04	2.77	2.40	0.80	0.78	1.10	1.03	0.99	0.96	1.18	0.78
HGD	0.84	1.23	0.56	1.13	0.85	1.09	0.90	1.11	0.86	1.00	0.75	1.28
OAC	0.97	1.00	0.95	-0.31	0.77	0.75	0.99	0.90	0.88	0.87	0.88	1.29

Chapter 6

Targeting saliva for population screening of OAC genetic risk

Chapter 6 – Targeting saliva for population screening of OAC risk

6.1 Introduction

The work outlined in chapter 5 demonstrated that saliva has the potential to provide biomarkers linked to the detection of those with or at risk of OAC. Clearly, as discussed earlier, a saliva sample is both safe and acceptable to use as a screening tissue type. It is non-invasive, simple to collect, and low-cost. Saliva has been described as 'the mirror of the body' with published literature demonstrating its potential for providing insight into oral and systemic disease [253, 254]. The field of salivary diagnostics has been around for some time. For example, *Stephen et al* postulated the use of saliva for drug level monitoring in 1976 [255]. However, its use in the field of molecular biology was always limited by the poor quantity and quality of salivary DNA and RNA obtained that analytical techniques at that time could not process. However, the use of saliva as a clinically useful tool, providing insights into prognosis, diagnosis, monitoring and the management of patients is increasing. Bodies of work are utilising saliva and are detecting useful biomarkers in fields such as cardiovascular disease, renal disease, diabetes, infections and cancer [223]. Emerging work is finding both DNA and RNA signatures associated with cancer risk in saliva [224-226, 256]. Therefore, given the promising results outlined in chapter 5, I decided to focus on utilising saliva to determine whether it was appropriate and effective for identifying patients with or at risk of OAC using transcriptomic and epigenetics.

6.2 Aims of chapter

The aim of this chapter is to determine whether saliva is a practical and potentially useful tissue to identify those with or at risk of OAC through transcriptomic and epigenetic analysis. The objectives are:

1. What are the optimum collection and storage processes to ensure the best quality and quantity of salivary RNA is obtained?

2. Does salivary RNA provide significant biomarkers either alone or in combination with clinical data for the detection of those with or at risk of OAC when targeted qRT-PCR is performed on a large cohort of patients?

3. Can additional salivary RNA targets be identified through performing whole mRNA sequencing on selected saliva samples?

4. Can DNA be extracted from the saliva samples and be utilised to provide further useful data through epigenetic analysis?

6.3 Objective 1 - Optimising the collection and storage of saliva

Although the use of saliva in the field of molecular biology is new, the optimum means to collect saliva has been discussed for some time due to its use in other fields. Saliva originates from three paired major glands (parotid, submandibular, and sublingual) and numerous minor glands. It is possible to obtain saliva from the individual major glands although the use of whole saliva, from all glands, is more frequently used and thought to be more clinically relevant [257]. Saliva can be collected under un-stimulated or stimulated conditions. Stimulation of saliva production can be achieved through the use of gustatory agents such as paraffin wax, rubber bands, gum base, and citric acid. However, it is thought that stimulants interfere with the contents of saliva [257, 258]. The flow rate of saliva displays diurnal, seasonal and postural variation and the contents of saliva are affected by eating, drinking, smoking and oral hygiene processes. Thus, ideally saliva collection should be standardised so that individuals are providing saliva at approximately the same time of day having not eaten, drunk, smoked or undertaken oral hygiene processes for 2 hours prior to donation. By standardising these processes, consistent results are achieved [257, 259].

There are four methods discussed for saliva collection:

1. Draining method: Saliva is allowed to drip off the lower lip until the desired quantity is obtained. A collection device with a funnel is used for

this method.

2. Spitting method: Saliva is allowed to accumulate in the floor of the mouth and the subject spits it out until the desired quantity is obtained.

3. Suction method: Saliva is continuously aspirated from the floor of the mouth into a test tube by a saliva ejector or an aspirator such as the RNAPro•SAL system (Oasis Diagnostics, Vancouver, USA)

4. Swab (absorbent) method: Saliva is collected through absorption into a pre-weighed swab, cotton roll, or gauze sponge placed in the mouth. The Salivette (Sarstedt, Nümbrecht, Germany) is a commercially available product for this.

A comparative study of the four methods described found that the suction and swab methods provided the most variability and were thus the least reliable [257]. However, this is an old study and with the production of newer collection devices such as the RNAPro•SAL system (Oasis Diagnostics, Vancouver, USA) and the Salivette (Sarstedt, Nümbrecht, Germany) there is published data demonstrating that these collection methods provide consistent, reliable results [259, 260]. For all the saliva work within this project, the spitting method was adopted.

A significant stumbling block in the use of saliva for molecular biology applications is the poor quality of RNA and DNA in the saliva sample. It was long thought that the presence of useful RNA in saliva was unlikely due to the hostile environment of the mouth complete with endonucleases, such as ribonucleases, that would destroy the fragile salivary RNA. The now established presence of RNA in saliva is thought to be due to a combination of: protective macromolecules; RNA being complexed with lipids, proteins, lipoproteins, or phospholipids; or protection from apoptotic bodies or other vesicular structures. Regardless of this protection, the degradation is still significant, preferentially targeting longer transcripts and randomly occurring at either end. Published literature reports the RIN to be approximately 2.5 and the average fragment length to be approximately 100bp [242, 261]. However, despite the nature of the sample, thanks to the advancing molecular analytical technologies, salivary RNA can be used for targeted analysis and whole mRNA sequencing [242, 246].

6.3.1 Objective 1 – Optimising the collection of saliva samples. Quantity of sample provided. Methods

The means in which saliva is collected can impact on its quality and quantity. Work was performed to test key variables that can affect the quantity and quality of salivary RNA and therefore its downstream application. The initial work focussed on the amount of sample a patient could provide that still yielded consistent results. However, there are two issues with collecting larger quantities of saliva. Firstly, individuals can find it difficult to produce large quantities especially if they are unwell and / or have starved prior to donation. The acceptability of saliva as a population screening tool does also depend on an individual being willing to provide the sample. For example, some individuals do not wish to provide a stool sample for the colorectal screening programme and cite this as their reason not to participate. Secondly, it is logical to assume that at a certain point the aqueous solution content of saliva will increase whilst the other content, that contains the desired DNA and RNA, will decrease. Thus, providing large amounts of saliva may dilute the sample and make RNA extraction more challenging.

For the work to determine how much saliva an individual could provide in one sitting 2 healthy volunteers provided 12 x 1ml of saliva over a 36-minute period (1ml provided every 3 minutes). Neither of these individuals had any known medical issues, took any regular medication or had recently had any short-term illnesses. The individuals had refrained from eating, drinking, smoking and oral hygiene procedures for two hours prior to donation. The subjects were seated comfortably with eyes open and head tilted slightly forward. They were asked to allow saliva to accumulate in the floor of their mouth prior to spitting it into the collection tube until the liquid (rather than the associated foam) reached the 1mL line. Saliva was collected in 15ml Corning® centrifuge tubes (Sigma-Aldrich, St Louis, USA). Unstimulated, whole saliva samples were obtained.

Each 1ml of saliva was immediately stored in a -80°C freezer and underwent RNA extraction 5 days later using the same method optimised and outlined in chapter 5.

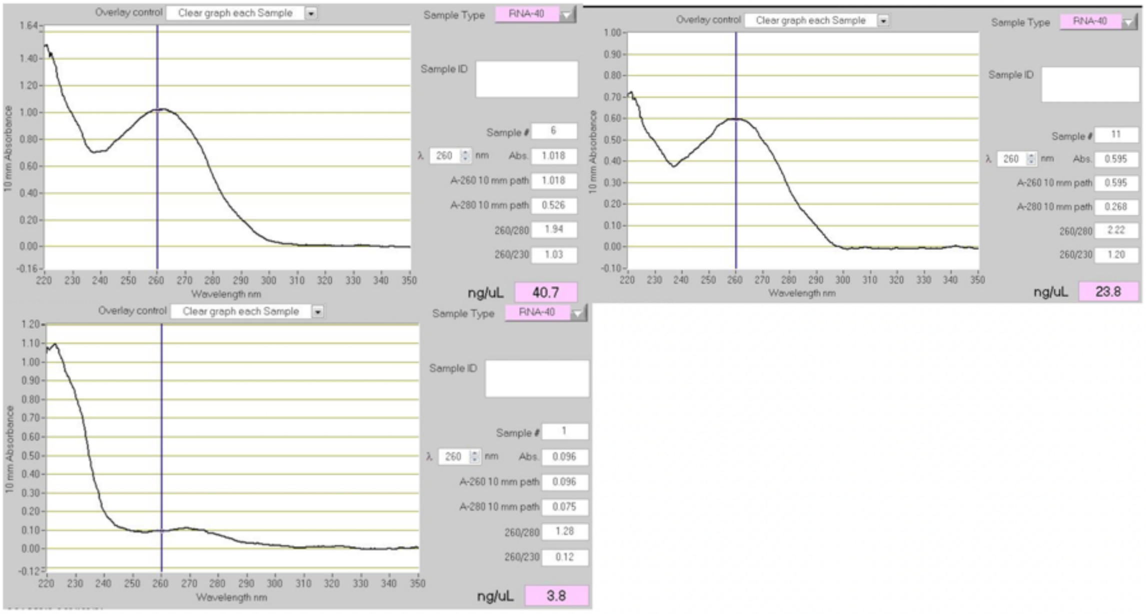
6.3.2 Objective 1 – Optimising the collection of saliva samples. Quantity of sample provided. Results

The below table demonstrates the fall in concentration of RNA extracted as the individual provides more saliva. As one can see from this table the yield of RNA extracted from individual 1 becomes significantly lower after 6ml provided. In individual 2 a similar fate occurs after 8ml.

	Individual 1	Individual 2
Sample no.	**Yield (ng/µL)**	**Yield (ng/µL)**
1	40.7	41.20
2	23.76	27.32
3	38.65	20.99
4	18.57	31.21
5	21.34	24.68
6	23.8	17.64
7	9.01	19.78
8	10.63	19.01
9	7.3	11.11
10	5.11	13.55
11	6.98	4.23
12	3.8	6.69

Table 15: Table demonstrating the concentration of RNA extracted from 12 saliva samples provided consecutively.

When one looks at the NanoDrop (Thermo Fisher Scientific, Waltham, USA) images seen at 1ml and 6mls in individual 1 one can observe that although the yield has diminished significantly, the RNA extracted is likely still of sufficient quality to be usable. However, at 7ml, in individual 1 the NanoDrop (Thermo Fisher Scientific, Waltham, USA) images suggest that the saliva is now of insufficient quality and quantity to be useable. In individual 2 both the quality and quantity of salivary RNA extracted became unusable at 8ml. These observations are demonstrated in the NanoDrop (Thermo Fisher Scientific, Waltham, USA) images below.



Overlay control: Clear graph each Sample
Sample Type: RNA-40
Sample ID:
Sample #: 6
λ 260 nm Abs. 1.018
A-260 10 mm path: 1.018
A-280 10 mm path: 0.526
260/280: 1.94
260/230: 1.03
ng/uL: 40.7

Overlay control: Clear graph each Sample
Sample Type: RNA-40
Sample ID:
Sample #: 11
λ 260 nm Abs. 0.595
A-260 10 mm path: 0.595
A-280 10 mm path: 0.268
260/280: 2.22
260/230: 1.20
ng/uL: 23.8

Overlay control: Clear graph each Sample
Sample Type: RNA-40
Sample ID:
Sample #: 1
λ 260 nm Abs. 0.096
A-260 10 mm path: 0.096
A-280 10 mm path: 0.075
260/280: 1.28
260/230: 0.12
ng/uL: 3.8

Figure 21: NanoDrop™ (Thermo Fisher Scientific, Waltham, USA) results for individual 1 at sample 1, sample 6 and sample 12 respectively

The results of this work demonstrate that individuals could provide 2ml of saliva without affecting the quantity and quality of salivary RNA extracted. Should I wish to collect more saliva then this variable requires further testing on more individuals but, judging on this data alone, 4-5ml is probably a safe amount to donate prior to quantity and quality diminishing. Having reviewed the literature in other studies collecting saliva for RNA or DNA extraction, there does not seem to be any in which an individual has been asked to provide more than 5ml.

There has been no published work exploring the amount of RNA or DNA found in saliva at differing volumes provided. Studies using saliva tend not to ask their volunteers to provide any more than 5ml with saliva collection devices such as the Salivette (Sarstedt, Nümbrecht, Germany) and the RNAPro•SAL system (Oasis Diagnostics, Vancouver, USA) collecting only 1.5ml and 1ml respectively [261, 262]. My results demonstrate that in the two volunteers tested there was a drop in yield the more saliva that was provided with samples after 6ml in one individual and 8ml in another being unusable. The reason for this is unclear. *Tanabe et al* looked at the pH of saliva the more dehydrated an individual is, work that was performed in response to the observation that there is increased decay, missing, and filled teeth and the risk of tooth erosion in athletes. This study found that there was a significant drop in salivary pH following exercise [263]. One could postulate therefore that salivary pH also falls, and thus causes more damage to RNA, the more saliva an individual is asked to provide. However, this seems an unlikely explanation as an individual is not going to become dehydrated following providing over 6mls of saliva. An alternate explanation may be that the content of saliva alters at a certain threshold of volume provided in which the water content overwhelms that of the other constituents that provide the DNA and RNA. Saliva is already 99.5% aqueous solution and thus it would take only a small shift for this to occur. Again, there has been no work exploring this but it seems a reasonable hypothesis that the more demand one places on the salivary glands in short space of time the less cells, electrolytes, and enzymes the saliva will contain.

6.3.3 Objective 1 – Optimising the collection and storage of saliva samples.

Methods

Avoiding the consumption of food and drink for 2 hours prior to donation may not sound like a long time but individuals can find this challenging. Providing saliva following this period of starvation can also become a challenge for some. As such I wanted to determine whether the avoidance of food and drink prior to saliva collection was necessary. The less pre-donation instructions one has to give the more likely that a sample will be received.

The aim of this work is to establish a standardised and reproducible salivary collection method to produce high-quality RNA to improve the performance of downstream translational applications. Here we investigate factors within the collection process which could affect the quality of salivary RNA such as storage at ambient temperatures, fasting windows prior to saliva collection, and collecting saliva with and without the addition of a preservative.

Cohort

Saliva was collected as part of the ethically approved study ‘Saliva To Predict Risk of Disease Using Transcriptomics and Epigenetics’ (SPIT) study (ISRCTN Registration: 11921553). Participants were healthy volunteers (n=9) with no significant past medical history or recent illness. All volunteers gave informed consent.

Preservative comparison sample collection

Each of the nine healthy volunteers who had fasted overnight were given three pre-prepared 30 ml tubes containing either 1 ml of DNA/RNA Shield, 1 ml of RNA*later* or an empty tube (no preservative). Volunteers were allocated randomised orders of donation of saliva to prevent bias. Each volunteer donated 1 ml of unstimulated whole saliva via passive drool per tube, agitating tubes containing preservative throughout collection. The tubes containing no preservative were immediately placed on ice. Samples containing preservatives were left at ambient temperatures for 1 hour. Three pools for each preservative condition were created, each containing saliva from three volunteers, to reduce variability between samples. The pools containing no preservative and DNA/RNA shield were immediately stored at -80°C. The pool containing RNA*later* was left at room temperature overnight before moving to -80°C storage the following day in accordance with the manufacturer's recommendations. This method of saliva collection was used for all following experiments with notable alterations described below.

Storage temperature study and bacterial load study sample collection

For these analyses, saliva samples were collected as described above. Three pools of saliva were then created and initially stored at room temperature. 800 μl of saliva was taken in the morning from each pool on each day for seven days and aliquots stored at -80°C immediately.

Saliva stability after storage at -80°C

1ml saliva samples were collected using the RNA*later* method and stored at room temperature for 24 hours before pooling to create the three sample pools. Four aliquots of 800 μl of saliva was taken from each pool and immediately stored at -80 ^{o}C, then extracted after 1, 2, 3 and 4 weeks.

Fasting study saliva collection

Volunteers (n=9) fasted overnight and then provided 1 ml of saliva upon waking which was collected into RNA*later*. All volunteers then ate and drank and rinsed their mouths with water before donating the first post-prandial saliva sample. They then fasted until the end of the study. All collections were made in the morning at 1, 2, 3 and 4 hours after eating. Volunteers were allowed a few sips of water after each donation. All samples were stored at room temperature for 24 hours before three sample pools were created for each collection time-point, each containing the same three volunteers for each pool. 800 µl of saliva was taken from each pool and stored at -80^{o}C prior to extraction.

RNA extraction and Quality Control

RNA extractions on samples not in preservative were performed using the methods outlined in section 5.3.4. For the samples collected in preservative a different method was required. Given the optimising of the methods outlined in Chapter 5, I decided to continue using isopropanol / ethanol precipitation method and to maintain the methodology processes established in Chapter 5. For samples collected in preservative there was a 1:1 ratio of sample to preservative. Therefore, I made the assumption that in 1ml of saliva / preservative mix there is 500µL of sample which, from the work in Chapter 5, is sufficient to obtain RNA for analysis. Similarly, in the preferred method outlined in Chapter 5 there was three times as much Qiazol (Qiagen, Hilden, Germany) as there was sample (400µL sample to 1.2mL Qiazol). Thus, in the preservative cases, continuing this ratio meant that 1ml of sample required 3ml of Qiazol (Qiagen, Hilden, Germany). Consequently, continuing the 4:1 phenol / chloroform ratio established in Chapter 5 resulted in 750µL of chloroform being required. The initial steps used were as follows:

1. 1mL of the saliva / preservative mix was pipetted into a separate 15ml Corning centrifuge tubes (Sigma-Aldrich, St Louis, USA) following mixing with a pipette.

2. 3mL of Qiazol (Qiagen, Hilden, Germany) was pipetted into the sample / preservative mix. This was then vortexed for 1 minute and left for 5 minutes at room temperature.

3. 750µL of chloroform was then added and again vortexed for 1 minute and left at room temperature for 5 minutes.

4. The sample was then centrifuged at 4° for 10 minutes at 14,000 rpm.

5. The first 400µL of the upper aqueous phase was then removed and placed in a 1.5ml Eppendorf.

6. Following this step, the methods were followed as for the other samples outlined in Chapter 5.

RNA Quality Control methods are outlined in section 5.4.2 and were carried out following extraction. The total RNA yield (ng) and 260/230nm and 260/280nm ratios were analysed using the DeNovix DS-11 FX + Spectrophometer/Fluorometer. RNA integrity (RIN) was analysed using the Agilent Bioanalyzer and Agilent RNA 6000 Nano Chip Kit, using the Eukaryote Total RNA v2.6 assay software. After quality control, all samples were stored at -80^{o}C.

qRT-PCR

cDNA was synthesised and qRT-PCR were performed as per the methods in 5.4.2.

Data Analysis

One-way ANOVA with Tukey's post-test correction statistics were generated using GraphPad Prism™ software, with $p < 0.05$ deemed statistically significant. All averages are shown as the mean ± standard deviation.

6.3.4 *Objective 1 – Optimising the collection and storage of saliva samples.* ***Results***

RNA*later* preservative stabilises salivary nucleotides

To investigate if quality and/or total yield of RNA isolated from saliva could be improved with the addition of commonly available preservatives, saliva was collected either without preservative or in the presence of the preservatives RNA*later* or DNA/RNA Shield (Figure 22).

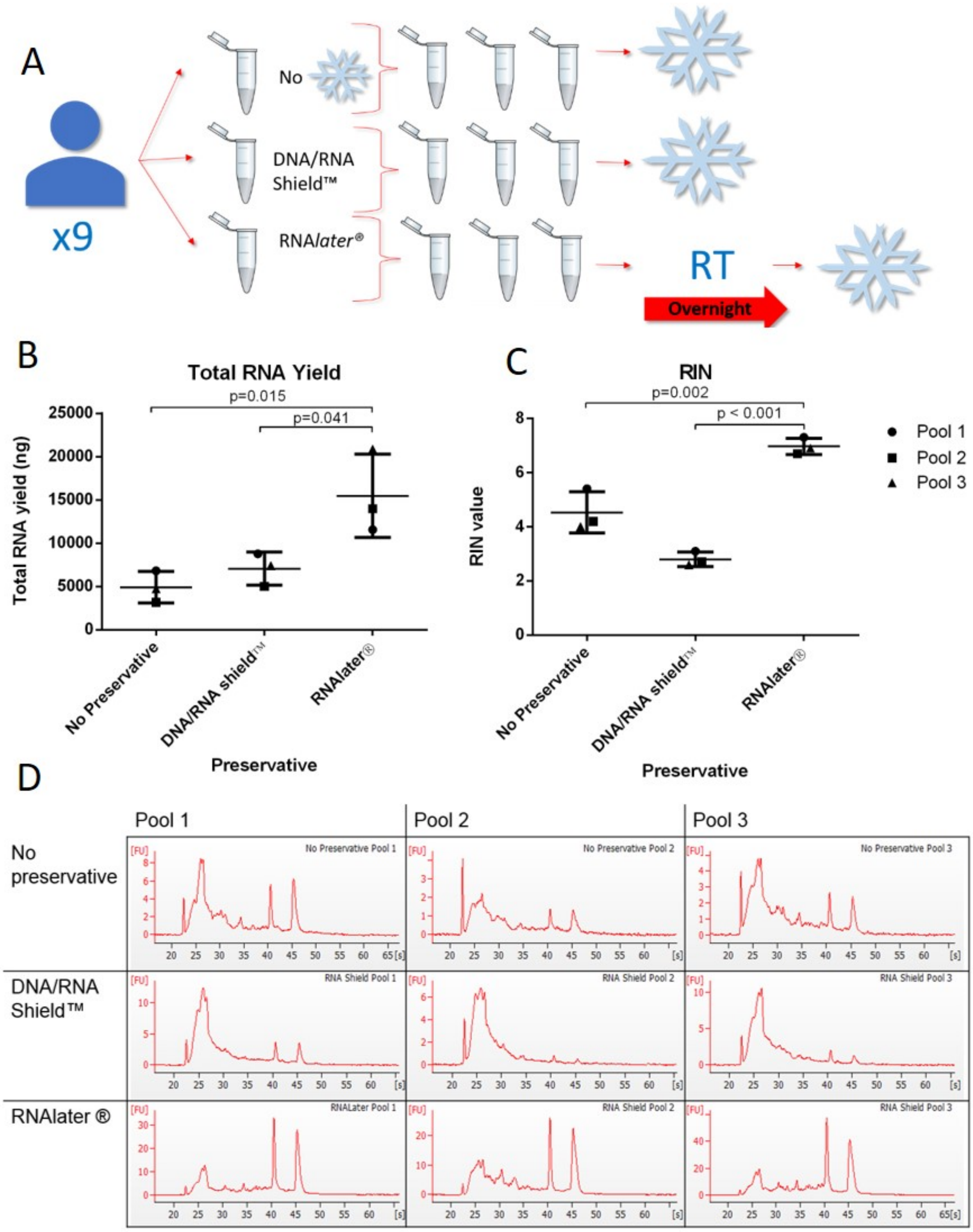


Figure 22: Comparison of RNA yield and quality between when using and not using preservatives

RNA*later* had a total RNA yield of 15473.2 ng (± 4824.7), in comparison to 7073.2 ng (±1916.3, p=0.041) for DNA/RNA Shield and 4911.7 ng (±1824.1, p=0.015) for saliva collected without preservative. To assess RNA quality, samples were assessed using the bioanalyzer which showed saliva collected with RNA*later* gave the highest mean RIN of 7.0 (±0.3), significantly higher than both saliva collected with DNA/RNA Shield (2.8, ±0.26, p=0.0001) and saliva collected without preservative (4.5, ±0.75, p=0.002) (Figure xB). As RNA*later* gave the best results all additional experiments were completed with RNA*later* only.

Room temperature storage does not adversely affect nucleotides

According to the manufacturer, saliva preserved in RNA*later* should be stored at room temperature for one day before freezing. To assess the effects of this, samples were incubated at room temperature for up to seven days, with 800 µl of sample stored at -80 ºC on days 0, 1, 2, 3, 4, 5, 6, and 7. There was no significant change in either RIN or total RNA yield in samples incubated at room temperature for longer than 24 hours although there was an increase in total RNA yield from day 0 to day 1 in two out of the three pools (Figures 23A and 23B).

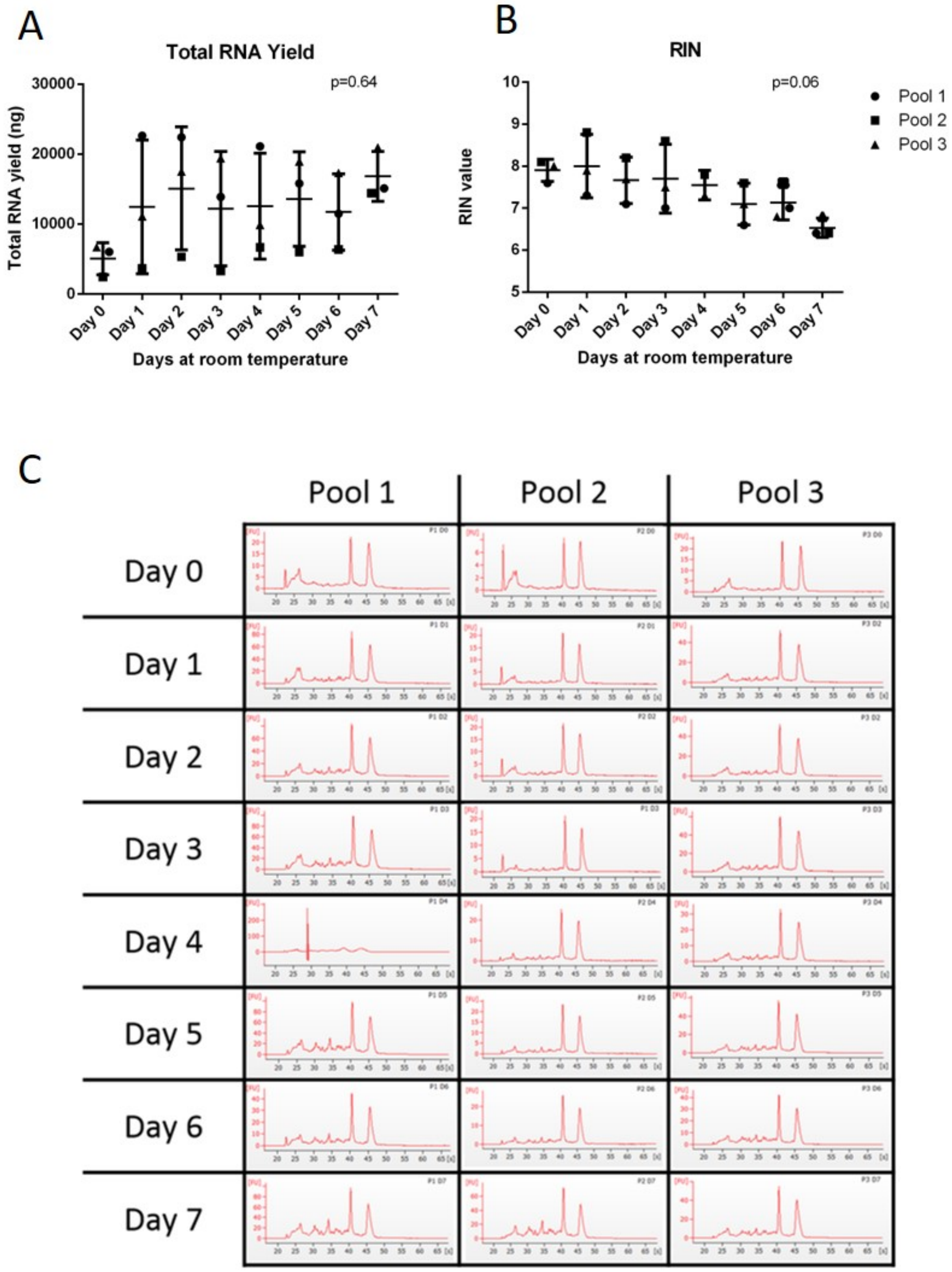


Figure 23: Assessing the impact of storing samples at room temperature on yield and RIN value

Overnight fasting increases salivary nucleotide yield.

Incrementally increasing fasting windows were used to investigate whether fasting has any effect on total RNA yield or quality. RNA yield increased from 1970 ug ± 1458.2 immediately after eating to 8279.3 ng ± 3318.2 (*p* = 0.063) following an overnight fast. This trend towards higher yields did not quite reach statistical significance when a post-ANOVA, TUKEY test, was applied (*p* = 0.06, Figure 24A). There was no change in RIN with increasing fasting length ($p <$ 0.27), (Figure 24B).

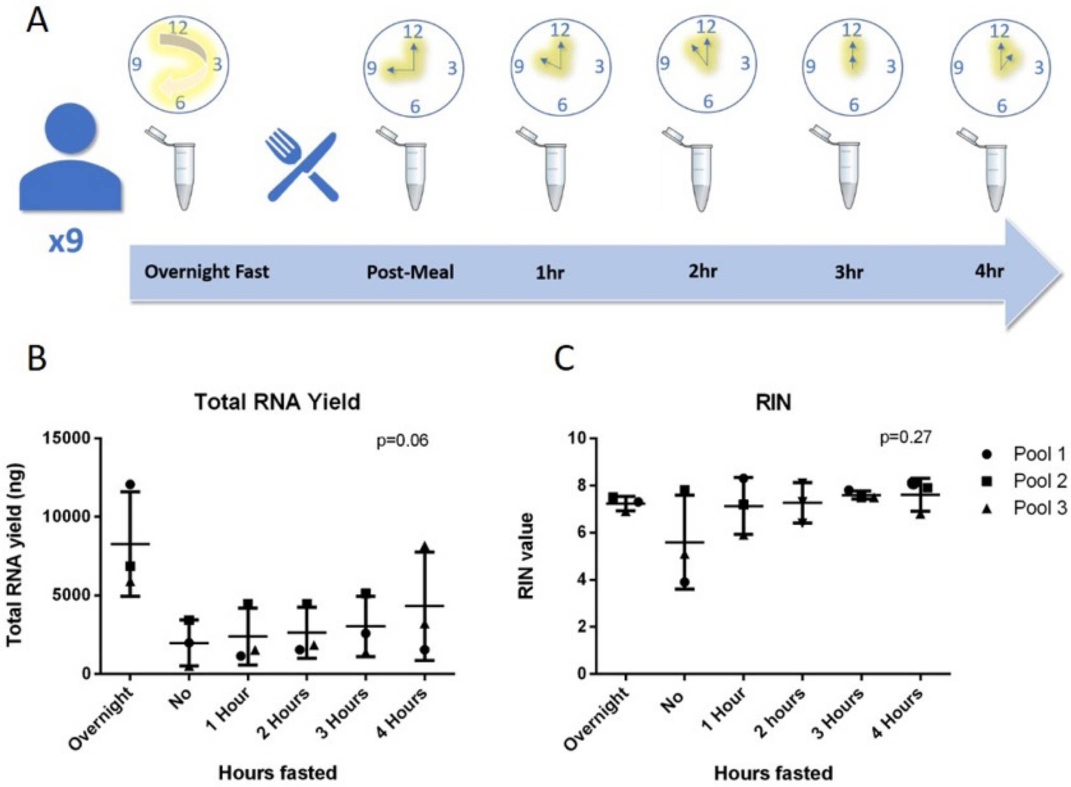


Figure 24: Assessing the impact of overnight fasting on RNA yield and RIN

Duration of freezing does not impact on nucleotide quality or quantity

To investigate whether storing saliva for longer periods of time at -80ºC prior to processing had any effect on total RNA yield or quality, the samples were frozen and extracted each week for four weeks. There are no statistically significant changes in either total RNA yield or RIN between samples that were frozen for up to one month (Figure 25A and 25B).

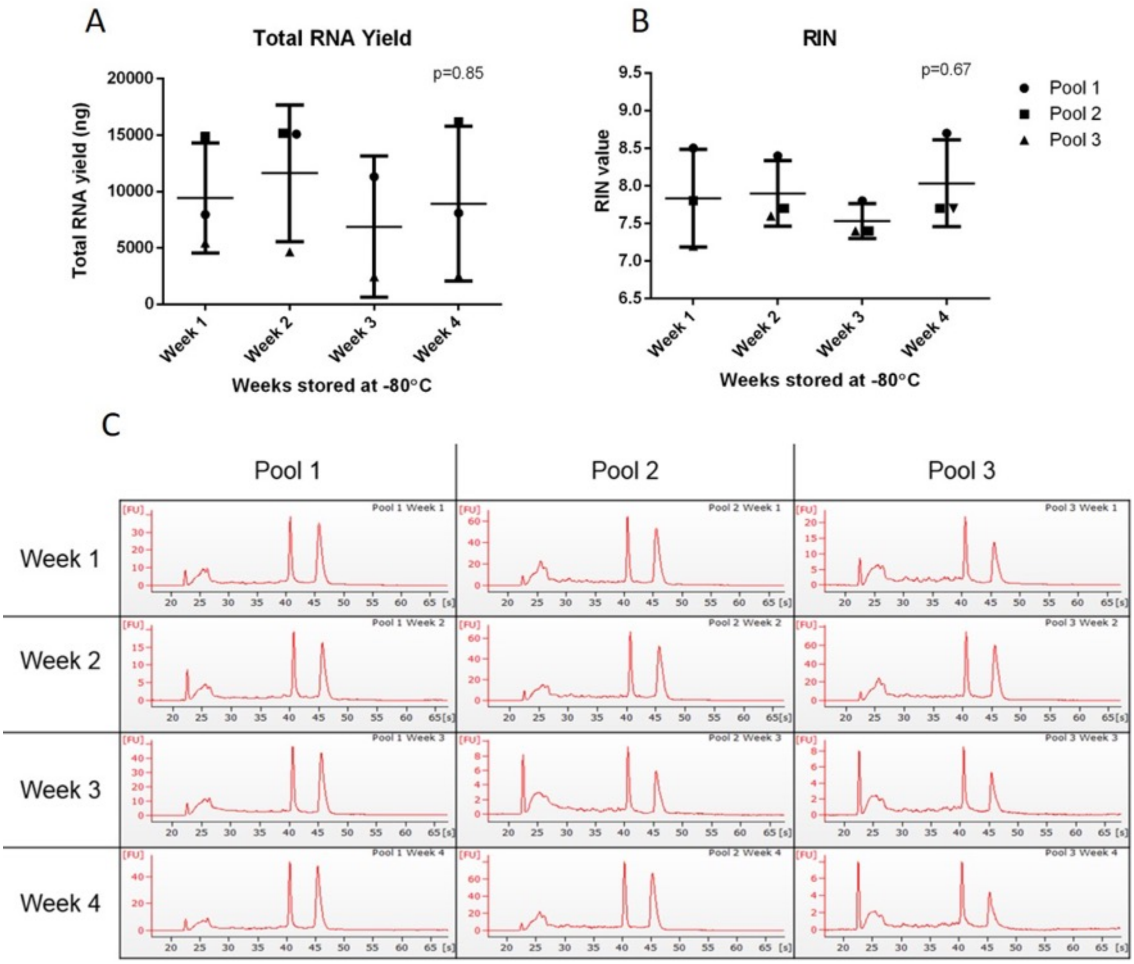


Figure 25: Assessing the impact of freezing sample on RNA yield and RIN

6.3.7 Objective 1 – Optimising salivary sample collection. Discussion

One of the major challenges of working with saliva is the dilution of molecules compared to serum and the rapid degradation of analytes such as RNA. Currently there is no standardisation in saliva collection methods, and a previous study recommended a method of RNA extraction that avoids the use of RNA stabilizers [241].

Studies do not commonly use RNA stabilisers in their saliva collection methods, instead employing collection over ice and ultra-low temperature storage, with RINs of approximately 2.5 from salivary RNA previously reported [241, 242, 253, 264-266]. We have demonstrated that a major advantage of using preservative is a statistically significant increase in total RNA yield and quality with the addition of the RNA stabilizer RNA*later*, in comparison to saliva collected without preservative. Our collection method put the RNA*later* directly into the pre-prepared collection tube, thereby ensuring saliva was mixed with preservative almost immediately on collection. We have reported marked improvement in the median RINs of 7.0 (±0.3) for healthy volunteers.

Saliva has advantages as a matrix due to its ease of collection and accessibility, facilitating the possibility of samples being collected in the primary care setting, or even at home. Our data support the RNA*later* manufacturer's instructions suggesting that samples can be left up to a week without compromising RNA quality, enabling sample postage to a clinical laboratory (Figures 25A and 25B).

As saliva production and molecular content alters after eating food, previous studies have employed a fasting window of 1-1.5 hours prior to saliva collection [264, 265]. Shorter fasting windows minimise patient discomfort. Our results suggest that the quality and yield of RNA is reduced with shorter fasting windows (Figures 24A and 24B). Although no statistical significance was found, our results suggest an overnight fasting window is optimal for RNA yield and quality, possibly because salivary flow rates fall to near zero at night [257]. The notable decrease in RNA total yield ($p = 0.063$) in saliva collected immediately after eating suggests that the stimulation of saliva through the mechanisms associated with eating and drinking do reduce the abundance of RNA. We show that the RIN increases to a consistent level after 1 hour fasting and the total RNA yield steadily increases between 1 – 4 hours fasted. This raises the possibility that only a 1 hour fasting window is required to produce intact RNA, with subsequent hours fasting increasing the yield. Our method can be tailored to the intended downstream application with two standardised fasting windows; a shorter one hour fast for high RINs or an overnight fast for higher yields. For the clinical work, given that for this study patients were being recruited upon their attendance to endoscopy, where they would be fasted for at least 6 hours pre-procedure, these findings suggest we have optimal conditions for collection from our patient cohort.

All studies utilising saliva ask the recruited individuals to avoid eating and drinking 1-2 hours prior to saliva collection. One of the most often cited reasons for this is the fear that the food, in particular animal produce, will contaminate the extracted DNA and RNA. I did not perform any downstream molecular analyses of the RNA extracted from these healthy volunteers who had consumed breakfast and thus I am unable to comment on this. Saliva plays a vital role in digestion. Two of its roles are to lubricate food with mucus and water in order for it not to damage the mucosa of the oesophagus and to begin the digestion process using salivary enzymes that breakdown food [267]. As such, following the consumption of food and drink, it is reasonable to postulate that the contents of saliva alter to respond to these roles. It is likely that the water and mucus content increases alongside an increased presence of salivary enzymes, which will include endonucleases. This creates a more hostile environment for RNA to survive in and thus it is likely that the inconsistent results seen in this aspect of the work are a consequence of the impact in the changes in salivary contents following food and drink consumption has on RNA transcript degradation.

It should be noted that the RIN values of the most degraded group (DNA/RNA Shield) was 2.8, ±0.26. Whilst this represents a degraded sample it is on par with, if not better, than the RIN expected with the use of FFPE samples where this can be as low as 1.4 [268]. Despite this poor quality, both FFPE and saliva tissue samples have been used for targeted analysis using qRT-PCR as well as whole transcriptome sequencing [242, 244, 246, 268]. Thus, whilst we accept that saliva is a degraded sample, its use in molecular biology and its application as a screening tissue type has great potential.

Finally, there are reports that cancer patients have increased RNase activity and thus it may be that the acceptable RIN values found in this work will be lower when applied to clinical practice. However, the promising findings outlined in Chapter 5 suggest saliva is a viable medium for assessing biomarkers in systemic disease, including cancer [269].

6.4 Objective 2 - Detecting transcriptomic biomarkers using salivary RNA on larger cohort of patients

Whilst the early work, outlined in Chapter 5, had demonstrated that there was potential for salivary RNA to provide key biomarkers in the detection of individuals with or at risk of OAC further exploration was required from a larger cohort of patients. Similarly, the initial proof of concept work, comparing tissue types, used only 6 primers which required expansion. Initially I wanted to explore whether biomarkers in saliva were able to differentiate individuals with or at risk of OAC to a significant level and thus work was performed looking at selected primers in a larger cohort of patients.

6.4.1 Objective 2 – Detecting transcriptomic biomarkers using targeted expression analysis on a larger cohort. Methods

Power calculation

Again, this power calculation was performed by Prof. Rifat Hamoudi. Given that this was proof of concept work the standard deviation used was 1 rather than the higher 1.5 used for the proof of concept work described in Chapter 5. R statistical software version 2.14.2 was used to carry out the power calculation with *p* = 0.05 (5% significance testing) and power of 90%, n = 20.0211. As such we aimed to obtain 20 samples in each of the four categories in this targeted study to yield differentially significantly expressed genes across the groups[227, 228].

Sample collection

The ethical approval and inclusion and exclusion criteria for patients recruited for this work is outlined in Chapter 5. In short, patients were recruited upon attendance to UCLH for their endoscopic procedure. As such they had not consumed food for 6 hours and drink for at least two hours prior to collection as part of their preparation for the procedure. Patients were recruited into one of four groups; Normal, NDBO, HGD and cancer. The criteria for these groups are outlined in Chapter 5. Diagnoses were confirmed by myself following review of the endoscopy report, the patient's notes available to me on the hospital computer records, and any relevant histology. All diagnoses were confirmed by another physician.

Recruited patients were asked to complete the specifically designed enhanced questionnaire providing demographic, symptom and risk factor data, outlined in Chapter 4, and provide a 1ml saliva sample. Unstimulated, whole saliva samples were obtained and collected in 15ml Corning centrifuge tubes (Sigma-Aldrich, St Louis, USA). By nature of the fact that patients were recruited prior to endoscopy the saliva was collected at different times of day and the recruitment for this work took place between January 2015 and July 2015. Again, the recruited patients were seated comfortably with eyes open and head tilted slightly forward. They were asked to allow saliva to accumulate in the floor of their mouth prior to spitting it into the collection tube until the liquid (rather than the associated foam) reached the 1mL line.

The samples for this clinical work were collected whilst the optimisation work, outlined in section 5.3, was being performed. At this point as per published work samples were placed on ice after collection and then immediately stored in a -80°^{C} freezer. Following the results of the optimisation work, I acknowledge that ideally for RNA quantity and quality these samples would have been collected in RNA*later*, however, the results from this also demonstrated that collection without preservative and immediate freezing still provided RNA of sufficient quantity and quality for downstream analysis.

RNA extraction and quality control

RNA was extracted and underwent quality control following the same protocol outlined in Chapter 5. The longest a sample was left in the -80°^{C} freezer prior to RNA extraction was 63 days and the shortest time was 6 days. Once the RNA was extracted and had gone through the quality control checks it was stored in a -80°^{C} freezer until all samples for this work were collected, extracted and ready for downstream molecular analysis. From the initial thawing to the final results the samples were freeze / thawed a maximum of 3 times.

Primer target selection and design

Following on from the work outlined in Chapter 5 it was necessary to test further targets on the salivary samples to determine whether these would provide biomarkers for the detection of those with or at risk of OAC. The additional targets were identified using the process outlined in 5.4.1. In total 13 genes were selected with different exons within these genes targeted creating 22 targets altogether. The genes selected were; AMY1, AMY2, CDKN2a, SMAD7, TLR6, TP53, BRAF, EGFR, KIT9, KRAS, NRAS, PIK3CA and PTEN. The design of the primers was in keeping with the methods described in section 5.4.2.

cDNA and qRT PCR

These steps were performed in an identical manner to those carried out and outlined in Chapter 5.

Statistical analysis

Again, once the qRT-PCR was performed the relative expression was calculated, ΔCt, by comparing the Ct of each amplicon against 18S rRNA (housekeeping gene). A ΔΔCt was also calculated in order to compare the four categories of recruited patients (Normal v NDBO v HGD v Cancer). Finally, systematic Mann-Whitney U test using SPSS version 22 was then carried out between all the groups involved and $p < 0.05$ was taken to be significant. This final step was performed by Prof. Hamoudi by systemically comparing the expression of each gene between each of the four groups.

6.4.2 Objective 2 – Detecting transcriptomic biomarkers using targeted expression analysis on a larger cohort. Results

The numbers of patients recruited into each category, with sufficient quality and quantity RNA for downstream molecular analysis, were as follows:

Diagnosis	Numbers recruited
Normal	20
NDBO	20
HGD	19
OAC	21

Table 16: Patient numbers recruited to each diagnostic category

The above table demonstrates that sufficient numbers of patients were recruited into each diagnostic category for conclusions to be drawn as to whether a significant biomarker can be detected using salivary RNA.

The basic demographics of those recruited in each of the four diagnostic categories is outlined in the following table:

	Normal	NDBO	HGD	Cancer
Age	62	70	69	70
Sex	9F, 11M	7F, 13M	3F, 16M	8F, 13M
Av. length of BO	N/A	C3M5	C6M7	C4M6
Stage of cancer	N/A	N/A	N/A	Range T1N0M0 – T3N2M1
Smoking status	3 current, 5 ex-smokers, 12 never	2 current, 6 ex-smokers, 12 never	2 current, 3 ex-smokers, 14 never	3 current, 9 ex-smokers, 9 never
Mean BMI	26.8	29.6	27.3	25.1

Table 17: Demographic data for patients recruited for salivary analysis

This demonstrates that the patient characteristics in all four groups were broadly similar. There were similar ages and numbers of current and ex-smokers seen in all groups and in all four groups the patients BMI would be described as overweight. The slightly lower BMI seen in those with OAC may reflect their disease. The mean length of BO was long in the 3 groups. It should be noted however, that in the normal, NDBO, and OAC there was a relatively equal distribution of males to females whereas in the HGD there are considerably more males.

The following tables demonstrate the average ΔCt values and fold change in each category, for each of the 22 targets.

	NORMAL		NDBO		HGD		OAC	
	ΔCt	Fold change	ΔCt	Fold change	ΔCt	Fold change	ΔCt	Fold change
AMY1	17.78	1.00	18.02	1.01	14.41	0.81	16.74	0.94
AMY2	18.24	1.00	20.01	1.10	15.79	0.87	14.22	0.78
CDKN2a_r58	3.79	1.00	0.31	0.08	-3.88	-1.02	-2.29	-0.60
CDKN2a-80	6.60	1.00	7.63	1.16	3.15	0.48	6.18	0.94
SMAD7-118	9.83	1.00	11.41	1.16	6.25	0.64	8.93	0.91
SMAD7-129	19.68	1.00	21.14	1.07	16.58	0.84	16.38	0.83
TLR6	16.74	1.00	15.44	0.92	10.94	0.65	13.66	0.82
TP53-c63	15.79	1.00	15.73	1.00	11.72	0.74	14.80	0.94
TP53-3	15.53	1.00	15.66	1.01	16.56	1.07	13.28	0.85
TP53-5	16.62	1.00	19.86	1.20	19.24	1.16	14.16	0.86
TP53-6	13.28	1.00	15.58	1.17	16.79	1.26	9.64	0.73
TP53-7	8.15	1.00	10.26	1.26	7.74	0.95	7.06	0.87
TP53-9	16.47	1.00	20.36	1.24	19.29	1.17	14.55	0.88
TP53-10	13.63	1.00	14.43	1.06	18.48	1.36	12.22	0.9
BRAF	14.14	1.00	16.59	1.17	16.38	1.16	13.46	0.95
EGFR-d19	14.41	1.00	17.38	1.21	17.26	1.20	11.93	0.83
KIT9	14.15	1.00	17.35	1.23	17.83	1.26	12.80	0.90
KRAS	13.39	1.00	17.35	1.30	17.84	1.33	12.43	0.93
KRAS-q61	15.93	1.00	17.96	1.13	16.08	1.01	14.28	0.90
NRAS-1	16.63	1.00	20.35	1.22	19.25	1.16	16.15	0.97
PIK3CA-7	16.42	1.00	20.07	1.22	19.21	1.17	16.17	0.98
PTEN-3	6.35	1.00	7.78	1.23	5.89	0.93	6.17	0.97

Table 18: Demonstrating the ΔCt and the fold change seen in each primer

As discussed previously, a fold change of greater than 0.1 (i.e. 1.00 to 0.9) is around a ten-fold difference in expression. The lower value indicating increased expression. A difference of 0.1 in the fold change is considered to be significant. From this data one can note that there is significant aberrant expression of some of these primers at different stages of progression. Most notably in those individuals with HGD or OAC.

As one can appreciate from the above data not all the primers that were tested had significant differences in their expression throughout the progression to OAC. This is in keeping with our knowledge of the progression towards cancer in which different pathways are activated at different times. A nice example of this is seen in the *TP53* primers that have a significant difference in their expression late on, observed in the patients with OAC. There has been published work demonstrating the late stage of *TP53* mutation and consequent high expression at the latter stages of cancer development [1].

A striking observation from the presented data is the significant expression change seen at all stages of progression in CDKN2a-r58. In this presented data CDKN2a-r58, which is linked to the *P16* tumour suppressor gene, appears to be a key primer in identifying those with or at risk of OAC. Below is a diagram demonstrating the striking observation of the increasing expression of CDKN2a in patients with NDBO, HGD and OAC compared to the normal controls. In this diagram you will note that as the expression levels increase one can observe a significant proportion of the truly high-risk cases (those with HGD or OAC):

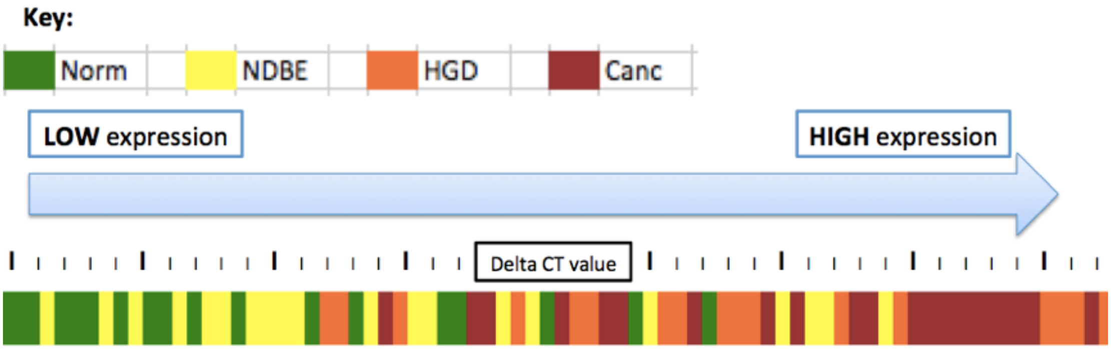


Figure 26: Demonstrating increasing expression of CDKN2a-r58 in patients from normal to OAC. Each square represents one patient's diagnosis

A further observation from this data is the significant fold changes seen in SMAD7 and AMY2. SMAD7 is known to be linked to cancer whereas the role of AMY2 in this process has been far less studied. This is further addressed in section 6.4.3. Figure 27 is a graphical representation of the significant changes in expression observed in these two primers.

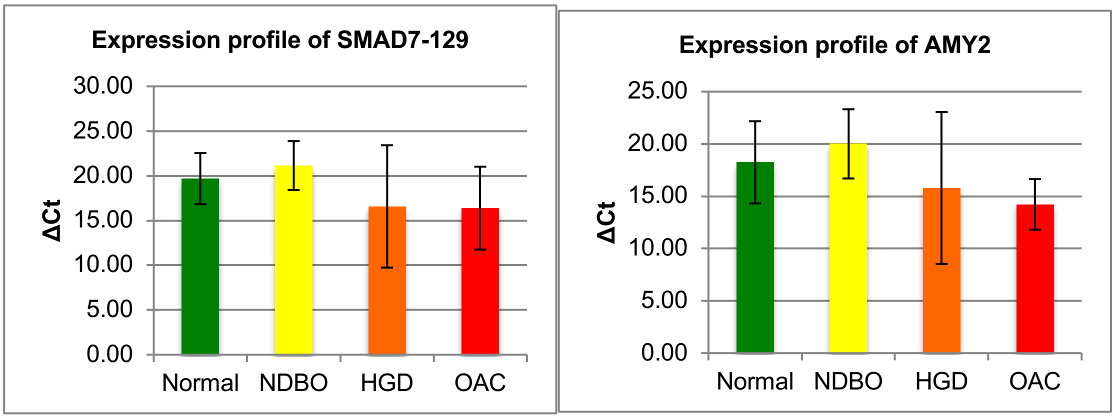


Figure 27: Graphs demonstrating the changes in expression between patients in each of the 4 diagnostic categories for SMAD7-129 and AMY2 primers.

The following tables demonstrate the primers with statistically significant changes in expression when the diagnostic categories are compared. A value of less than *p* 0.05 was taken as significant:

Table 19: List of primers with statistically significant changes in expression when diagnostic categories are compared

Table 19a: **NORMAL v NDBO**

Gene	P value
SMAD7-118	0.0468
EGFR-d19	0.0253
KIT9	0.0253
KRAS	0.0179
TP53-7	0.0217
NRAS-1	0.0139
PIK3CA-7	0.0139
PTEN-3	0.0228
TP53-5	0.0282

Table 19b: **NORMAL v HGD**

Gene	P value
CDKN2a-r58	0.0139
SMAD7-118	0.05
TLR6	0.0127
KIT9	0.0333
KRAS	0.0248
TP53-10	0.01

Table 19c: **NORMAL v OAC**

Gene	P value
AMY2	0.0011
CDKN2a-r58	0.003
SMAD7-129	0.044
TP53-6	0.0384

Table 19d: NDBO v HGD

Gene	P value
AMY2	**0.00001**
CDKN2a-r58	**0.0118**
CDKN2a-80	**0.0051**
SMAD7-118	**0.0004**
BRAF	**0.0432**
EGFR-d19	**0.0158**
KIT9	**0.0314**
KRAS	**0.0257**
TP53-6	**0.00136**
TP53-7	**0.0043**

Table 19e: NDBO v OAC

Gene	P value
AMY2	0.05
CDKN2a-r58	0.0355
CDKN2a-80	0.0036
SMAD7-118	0.0011
SMAD7-129	0.00261
PTEN-3	0.0072

In total 12 of the 22 genes that were tested had significant differences in expression observed at one or more stage of progression. Of these, again, the most striking observations are the significant differences in CDKN2a, TLR6, SMAD7 and AMY2. Whilst one acknowledges that this is a small cohort of patients and will require validation on a larger group, it appears that RNA biomarkers can be found in saliva that are linked to OAC and in particular these four genes seem to be of key significance.

6.4.4 Objective 2 – Detecting transcriptomic biomarkers using targeted expression analysis on a larger cohort. Discussion

As established in section 6.3 the optimum means by which to collect the saliva is with the patient having refrained from eating, drinking, smoking and oral hygiene processes for 1-2 hours prior to collection. In this patient group this was adhered to as they were all recruited prior to undergoing their endoscopy. However, as established in section 6.3 the optimum means of storing the saliva was by using preservative whereas in this group preservative was not used and instead the samples were placed immediately into a -80°C freezer. This occurred because the recruitment of patients coincided with the work on optimising collection and storage procedures. As such the results were produced after these patients had already been recruited and their samples collected. The work in section 6.3 demonstrates that there is a definite improvement in the quality of RNA obtained in samples stored in preservative. However, although the average quality of the RNA sample was likely not to have been optimal the targeted nature of this qRT-PCR analysis means the results likely remain valid. The primers have been designed to work with degraded samples and the methods included using reverse primer specific rather than oligo DT. Oligo DT relies on the presence of less degraded samples with longer chains of RNA as it binds to the poly-A tail. Thus this requisite is avoided by using gene-specific reverse primers.

Of the 13 genes tested in this work, eight provided a statistically significant result at least at one stage of the progression towards cancer. This will need validation with a larger cohort of patients, however, 5 of those genes demonstrated significance across multiple stages. The significant genes and their known relevance to BO and OAC are discussed here:

CDKN2a had statistically significant expression differences when normal patients were compared with those with HGD and OAC and when NDBO patients were compared with those with HGD and OAC. In general, we observed increased expression of CDKN2a when the low risk patient (normal or NDBO) were compared with the higher risk patients (HGD or OAC).

CDKN2a is a gene that encodes for the tumour suppressor protein P16. Its role in the cell cycle is to regulate the process by decelerating cells progression from G1 phase to S phase. Its role in cancer development has been well established with the inactivation of CDKN2A/p16 tumour-suppressor gene being one of the most common genetic abnormalities in human neoplasia [270, 271]. *Igaki et al* noted a high frequency of point mutations combined with loss of heterozygosity of CDKN2a in SCC of the oesophagus, however this was not observed in OAC [272, 273]. *Bian et al* demonstrated that the means by which p16 is inactivated in OAC is through hypermethylation of CDKN2a promoter. Importantly, this was observed in high frequency in the latter stages of progression, but can also occur earlier on i.e. during metaplasia or dysplasia [272]. Logically speaking should there be evidence that the CDKN2A/p16 axis becomes inactivated in patients that have OAC or are progressing towards it then there should be less expression of the gene seen as patients progress towards cancer. This is not observed in our work. Reasons for this are unclear although we note some have suggested that p16 may be linked to clonal expansion *per se*, but not progression risk [MARTINEZ et al, Nature Comms 2016].

Two different exons along the **SMAD7** gene were tested in this work and when combined the differences in expression of this gene were statistically significant when all stages of progression were compared. When normal patients were compared to NDBO there was a lower expression of SMAD7 in those with NDBO. However, when normal patients and those with NDBO were compared with those with HGD or OAC, there was an increased expression of SMAD7 seen in those with HGD and OAC.

The SMAD proteins transduce extracellular signals from beta ligands (TGFβ) to the nucleus where they activate downstream gene transcription[274]. The SMAD7 protein, encoded by the SMAD7 gene, is an inhibitory protein that blocks formation of SMAD2 and SMAD4. SMAD7 has been demonstrated to have both pro and anti-tumour actions depending on the type of cancer and the stage of progression towards that cancer. It can exert tumour suppressive action through TGFβ by restricting growth of epithelial cells and maintaining their differentiation state. Alternatively, SMAD7 can encourage cancer progression and metastasis by increasing angiogenesis and inducing epithelial-mesenchymal transition [275]. There has been little studied on the role of SMAD7 in OAC and BO although there is evidence in oesophageal SCC that increased expression of SMAD7 correlates with a worse prognosis [276]. In colorectal cancer *Stolfi et al* demonstrated that there was increased expression of SMAD7 in colorectal cancer tumours and their work on cell lines showed that silencing SMAD7 prevented colorectal cancer cell growth [277]. It is likely that these pro-tumour actions are why we are seeing increased SMAD7 expression in those with HGD and OAC. However, this does not explain why there is a decrease in SMAD7 expression noted between those who are normal and those with NDBE. *Onwugebusi et al* noted that there was a significantly reduced expression of SMAD4 and SMAD2 in those with NDBE when compared to those who were normal. They noted that this was likely to be the result of several different mechanisms, including methylation, deletion, and protein modification[278]. The role of SMAD7 is to inhibit the formation of SMAD2 and SMAD4. Thus, if there is a decrease in their expression during the pathogenesis of BO then it is logical to assume that there would also be a decrease in SMAD7 expression as it would no longer be required to provide the negative feedback loop and inhibit their formation.

AMY2 was noted to have statistically significant differences in expression despite it, unlike the other genes tested, being a low copy gene. This was observed when normal patients were compared to those with OAC and when NDBO patients were compared to those with HGD and OAC. It was noted that the expression of AMY2 was increased in those with the higher risk lesions (HGD and OAC). There is little in the way of published data on AMY2 and its link to cancer. However, work by *Kang et al* suggested it could play a role as a tumour suppressor gene in gastric carcinoma [279].

Five TP53 exons were tested within this work. Generally speaking the statistically significant expression differences were demonstrated when normal or NDBO patients were compared to those with HGD and OAC. In these groups there was noted to be an increased expression of TP53 in those with higher risk (HGD and OAC). Two of the TP53 primers also demonstrated a statistically significant expression difference when normal patients were compared to those with NDBO. In these cases, there was a reduced expression of TP53 in those with NDBO than compared with the normal patients.

TP53 is a tumour suppressor gene that has the ability to induce cell cycle arrest at G1/S regulation point, repair DNA, is involved in senescence, and can initiate apoptosis. Despite the huge diversity in tumourigenesis amongst various cancers, TP53 mutation is reported to occur in almost every type of cancer ranging from 10% of haematopoietic malignancies to almost 100% of ovarian carcinomas [280]. This cements its position as a key regulator of various signalling pathways involved in tumourigenesis. In OAC the mutation and over-expression of TP53 has been demonstrated to occur late in its pathogenesis. *Weaver et al* noted that TP53 mutation was found in 72% of cases with HGD and 69% of cases of OAC, but only 1 of their 40 cases (2.5%) of NDBO [1]. *Bian et al* also demonstrated that over-expression of TP53 occurs predominantly in the dysplastic phase with 85% of their cases of HGD and 71% of their cases of LGD having TP53 over-expression. 67% of their cases of OAC had TP53 over-expression [115]. Other studies have the rates of TP53 over-expression lower with LGD being 10-20% and HGD and OAC being 60% [281].The importance of TP53 over-expression has been replicated by others in which it has been demonstrated that individuals with over-expression of TP53 had an increased risk of malignant progression [282]. Thus, our finding of increased TP53 expression in the latter stages of progression (HGD and OAC) has been replicated in other work. It should be noted that *Weaver et al* found one case of TP53 mutation in NDBO, *Bian et al* found no TP53 mutation or over-expression in any of their NDBO cases and other studies have shown a maximum of 5% of cases of NDBO with TP53 over-expression [1, 115, 281]. However, *Bian et al* and others outlined by Flejou demonstrated this by using immunohistochemical techniques and thus would not have been able to observe a reduction in expression seen if using qRT-PCR techniques. Thus, our observation that there is reduced expression of TP53 in those with NDBO compared to normal individuals may demonstrate a stage in the process of the development of BO in which normal individuals do not benefit from the protective actions of TP53 and subsequently develop NDBO.

TLR6 was noted to have statistically significant differences in expression when normal patients were compared to those with HGD and OAC. Again, it was noted that there was increased expression of TLR6 in the higher risk groups (HGD and OAC). TLR6 is linked to NF*k*B which has been shown to be involved in the progression to OAC [283].

13 TLR receptors have been identified which act as receptors for the innate immune system to recognise and respond to pathogen or damage associated molecular patterns [284]. TLR6 responds to pathogen associated molecular patterns. TLRs play a key role in mediating the inflammatory response and have

been demonstrated to also be involved in the establishment and maintenance of disease [285]. In cancer TLRs have been noted to demonstrate pro-carcinogenic and pro-inflammatory actions that contribute to the tumour microenvironment and are associated with poor prognosis [284]. It has been noted that the microbiome of the distal oesophagus is altered in BO and OAC and along with the abnormal TLR expression this supports the potential for the role of TLRs and their response to these changes in bacteria in the pathogenesis of OAC [284, 286]. Thus, our finding of increased TLR6 expression in those with HGD and OAC may reflect these changes in the microbiome, the innate immune system's response and the consequent inflammation.

When the expression of **EGFR** of patients with NDBO was compared to those with OAC there was a statistically significant difference in expression. Those with OAC demonstrated a higher level of expression of EGFR. There was also noted to be a statistically significant reduction in expression of EGFR when the normal individual was compared to those with NDBO.

EGFR acts as a receptor tyrosine kinase involved in cell signalling and controlling proliferation and differentiation. It has been demonstrated to be involved in the tumourigenesis of many cancers and is linked to poor outcomes [287]. The expression of EGFR in OAC was noted by *Cronin et al* to be increased 13-fold in those with OAC when compared to those with NDBO with 80% of OAC specimens examined having increased EGFR expression [287]. This increased expression of EGFR in OAC has been demonstrated in other studies as well including work by *Pretto et al* [288]. Thus, our finding of increased EGFR expression in those with OAC when compared to those with NDBO is certainly in keeping with published literature. However, *Pretto et al* also demonstrated that there was increased expression of EGFR in those with NDBO, although to a lesser extent than in those with OAC, when compared to individuals with GORD. This is not in keeping with our findings in which there was reduced expression of EGFR when normal patients were compared to those with NDBO. This finding may reflect that the exon on the EGFR gene that was tested in this work was not over expressed during the pathogenesis of BO and as such was not over expressed.

6.4.5 Objective 2 – Creating a risk prediction tool utilising transcriptomic and questionnaire data

Utilising the data obtained through this targeted expression analysis the question then became as to whether one could accurately predict who had or was at risk of developing OAC from salivary RNA samples when combined with the patient demographic, risk factor and/or symptom data collected in the enhanced questionnaires, discussed in Chapter 5. Therefore, this data was provided to Professor Rosenfeld alongside the salivary RNA expression data and was analysed using the AI techniques outlined in Chapter 4. This work yielded encouraging results. By utilising 6 RNA amplicons (including AMY2, CDKN2A, TP53 and SMAD7) and 5 questionnaire data (including waist / hip ratio, PPI use and previous cancer) we were able to accurately differentiate between the four groups as well as identify those with or at risk of OAC with 93% sensitivity, 73% specificity and AUC 0.88. The full details of the RNA amplicons and questionnaire data used to create this tool have been patented by University College London (patent number: WO2017137427, https://patents.google.com/patent/WO2017137427A1/en).

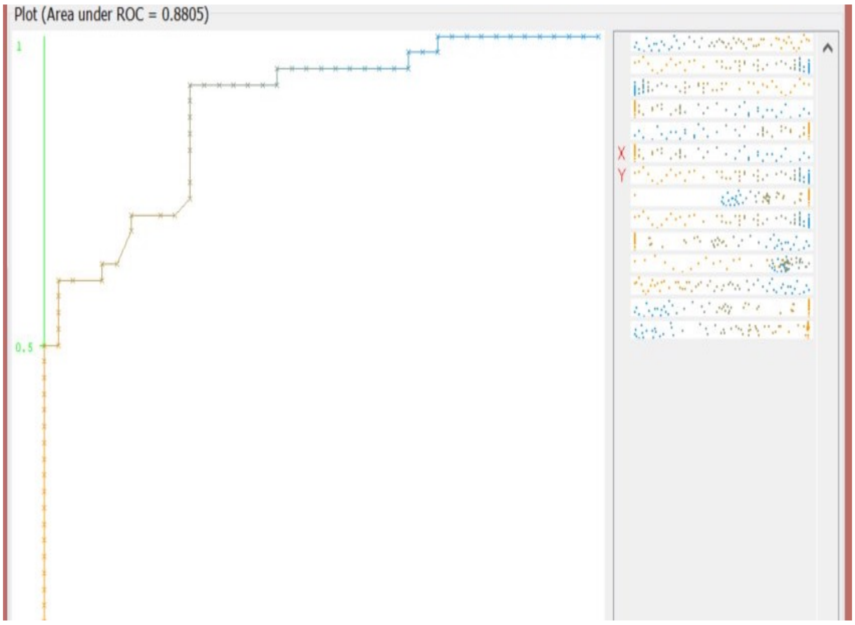


Figure 28: Utilising both salivary RNA expression and questionnaire data the developed tool's AUC is 0.88.

6.5 Objective 3 - Identification of further targets through whole mRNA sequencing

Whilst the above work does appear to show promise, it is obviously only preliminary data requiring further validation with a larger cohort of patients. Similarly, this preliminary work has been hypothesis-led with only 13 genes explored. As such, following on from the success of this preliminary work, it was felt important to perform a discovery set to ascertain if there are other genes that require investigation and expression analysis using qRT-PCR. Therefore, work was undertaken to perform whole mRNA sequencing (RNA-seq) on patient saliva samples.

6.5.1 Objective 3 – Whole mRNA sequencing methods

RNA-seq was performed using Ion AmpliSeq RNA Library Kit (4482335, ThermoFisher Scientific, Waltham, MA, USA) in accordance to manufacturer's standards. This kit was designed for targeted amplification of more than 20,000 distinct human RNA transcripts concurrently within the same primer pool, with 150bp amplicon amplified for each targeted gene (ThermoFisher Scientific, Waltham, MA, USA).

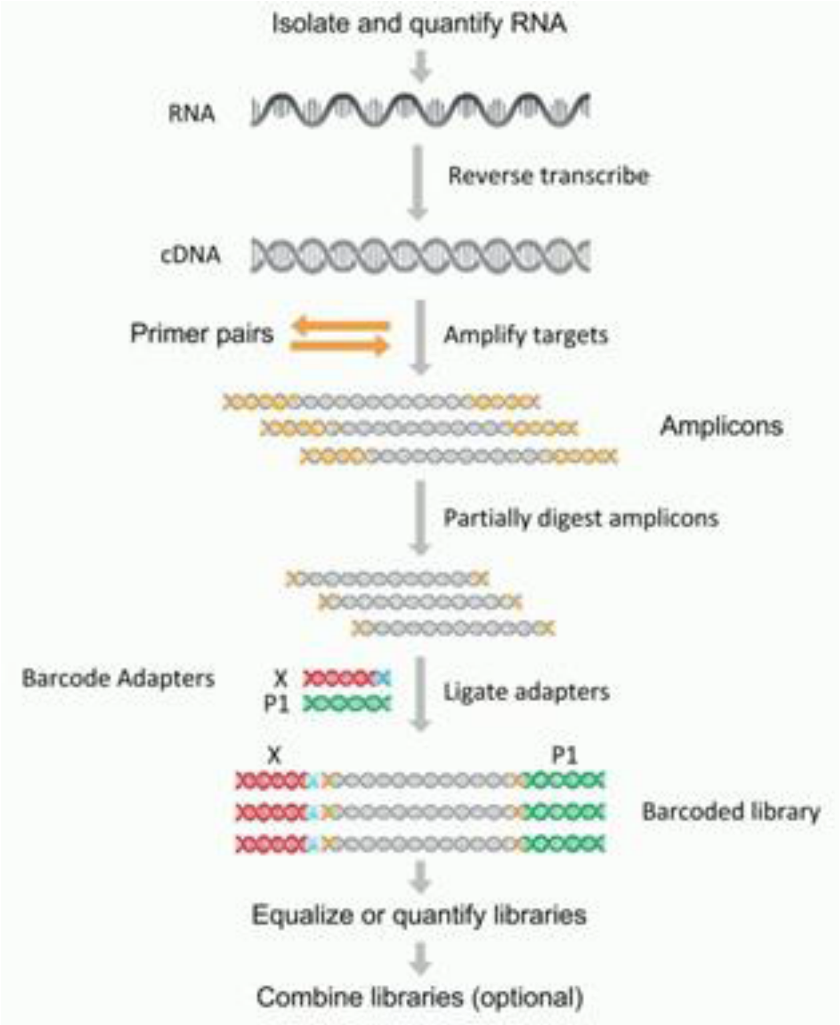


Figure 29: Workflow diagram of whole transcriptome amplification. *cDNA synthesised from RNA via reverse transcription is amplified using random primers. The amplicons are partially digested to remove part of the primer sequences, allowing for the binding of P1 adapters (allow amplicons to bind to ion spheres for sequencing) and barcodes (needed for library recognition by sequencing software).*

The power calculation for this work was performed by Prof Rifat Hamoudi who based this on whole transcriptomic data, where it has been shown that the standard deviation for detection of differentially expressed genes is 0.35 ($\sigma = 0.35$) as more genes are being interrogated and the delta (the measure of effect) is 1 [227, 228]. The power calculation was carried out using R statistical software version 2.14.2 with p = 0.01 (1% significance testing) and power of 90%. This found that n = 5.55289 suggesting we required 6 samples in each group in order for a significant transcriptomic signature to be discovered.

Patient samples selected to undergo this process had to be of sufficient quality. Quality control was performed on RNA samples as per Chapter 5. It was determined that samples needed to have a fragment mean distribution of RNA transcripts of at least 110bp, demonstrated using the Bioanalyser (Agilent Technologies, Santa Clara, California, USA), to be of sufficient quality. The selected samples then underwent the following processes.

First-strand cDNA synthesis

Eluted RNA samples were incubated at 80°^{C} for 10 minutes and then placed on ice for 2 minutes. A master mix was added to the RNA samples (Table 20). The RNA mixture was incubated at 42°^{C} for 30 minutes, 85°^{C} for 5 minutes and 4°^{C} for 5 minutes (Figure 30).

Components of reaction master mixture	Volumes added for each reaction (µl)
5x VILO™ RT Reaction Mix	1
10x SuperScript® Enzyme Mix *(includes SuperScript® III RT, RNase OUT™ Recombinant Ribonuclease Inhibitor and proprietary helper protein)*	0.5
Eluted sample RNA	2
Nuclease free water	1.5
Total	5

Table 20: Volumes of various components in master mixture for first-strand cDNA synthesis for whole transcriptome extraction

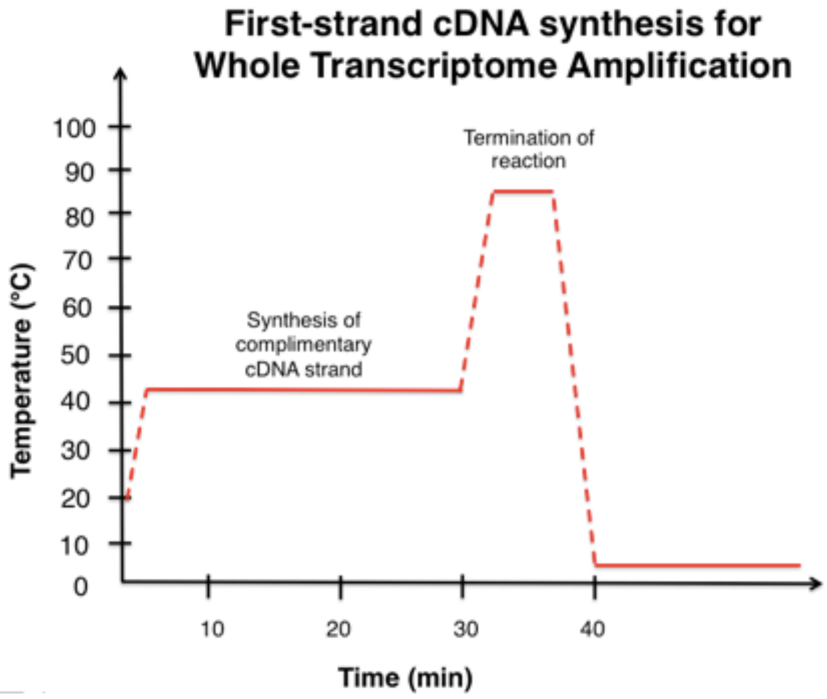


Figure 30: Protocol of first-strand cDNA synthesis for whole transcriptome amplification. Samples were heated at 42°^{C} for 30minutes, 85°^{C} for 5minutes and at 4°^{C} for 5minutes

Amplification of targets and primer digestion

Following the creation of cDNA, 15µl of a master mix (Table 21) was added to the cDNA and the samples were incubated at 99°C for 2 minutes, cycled 40 times at 99°C for 15 seconds followed by 60°C for 16 minutes and held at 15°C (Figure 22). 2µl FuPa reagent was added to the samples and then this mixture was incubated at 50°C for 10 minutes, 55°C for 10 minutes, 60°C for 10 minutes and held at 10°C (Figure 31).

Components of reaction master mixture	Volumes added for each reaction (µl)
5x Ion Ampliseq HiFi™ Mix	4
Ampliseq Expression Core Panel	8
Nuclease free water	3
Total	15

Table 21: Volumes of various components in reaction master mixture for target amplification

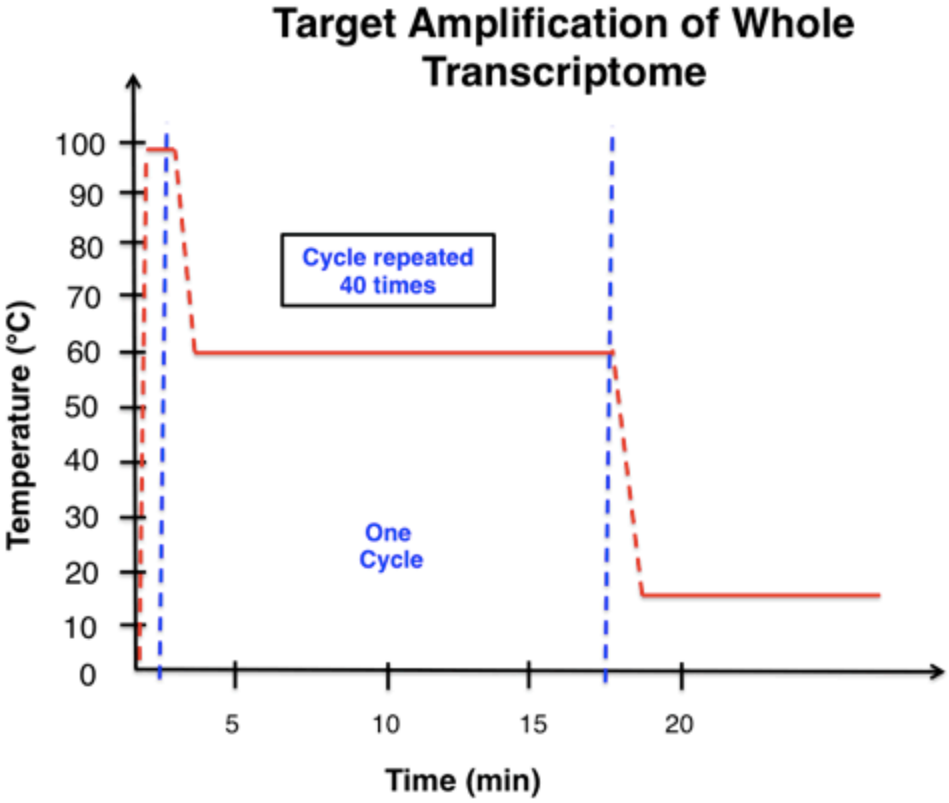


Figure 31: Protocol of target amplification for whole transcriptome. Samples were incubated in the thermal cycler at 99°C for 2minutes, cycled 40 times at 99°C for 15seconds followed by 60°C for 16minutes, and held at 15°C

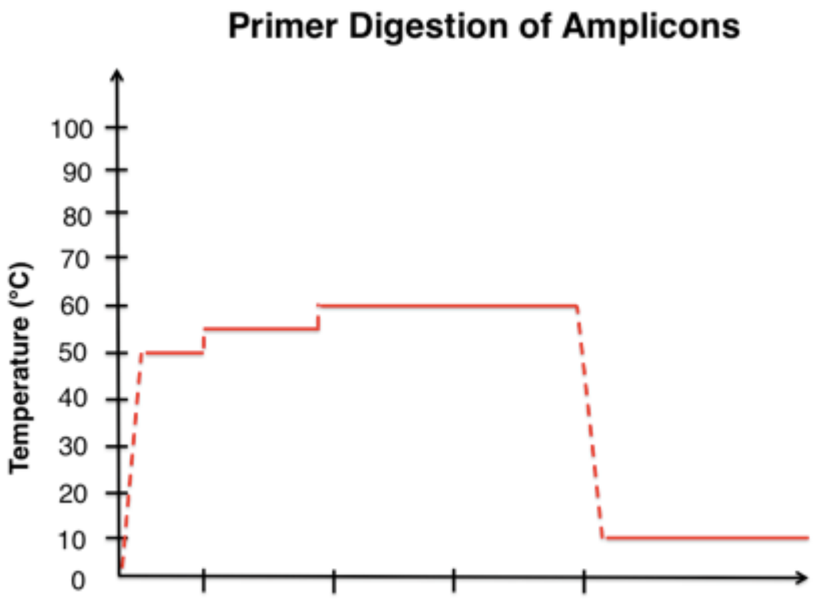


Figure 32: Protocol of primer digestion of amplified transcriptome. Samples were incubated at 50°C for 10minutes, 55°C for 10minutes, 60°C for 10minutes and held at 10°C

Ligation of adapters to amplicons

Ion P1 Adapter and Ion Xpress Barcodes were prepared in the same 1:4 diluted mixture. 2µl barcode-adapter mixture and 4µl switch solution were added to the samples, vortexed vigorously and pulsed to collect condensate. 2µl DNA ligase was added and the samples were incubated at 22°C for 70 minutes, 72°C for 5 minutes and held at 10°C.

Amplification and purification of libraries

45µl Agencourt AMPure XP reagent (A63880) from Beckman Coulter, Inc. was added to each library. The libraries were incubated at room temperature for 5minutes before being placed on the DynaMag-96 Side Magnet rack (12331D) for 3 minutes to separate the bead pellet from the supernatant. The supernatant was discarded and the beads were washed with 150µl of fresh 70% ethanol and returned to the magnetic rack for 3 minutes. This step was repeated and then the beads were air-dried for 3 minutes in order for the residual ethanol to evaporate. 50µl of 1X Library Amp Mix and 2µl 25X Library Amp primers were mixed with each bead pellet and the 50µl supernatant were transferred to fresh tubes and incubated at 99°C for 2minutes, in 5 cycles of 99°C for 15 seconds followed by 64°C for 1 minute, and held at 10°C. The beads were purified with several washes with ethanol and air-dried before 50µl of low TE solution was added to each pellet to disperse the beads. The tubes were returned to the magnetic rack for 3 minutes before 45µl of the supernatant was transferred to new tubes.

Quality control of whole mRNA library using High Sensitivity DNA Assay

Quality control of the whole mRNA library was carried out using High Sensitivity DNA Kit (5067-4626, Agilent, Inc, Santa Carla, CA, USA) in accordance with the manufacturer's protocol. Briefly, 15µl blue dye concentrate was added to the High Sensitivity DNA gel matrix vial. The gel-dye mixture was then centrifuged for 10 minutes at room temperature at 14,000rpm.

The High sensitivity DNA chip was primed on a base plate of the chip priming station with 9µl of gel-dye remix pipetted into the well labelled "G" in the third row. 9µl gel-dye mix was also pipetted into the other 2 "G" wells. 5µl of High Sensitivity DNA marker was pipetted into the ladder well and the sample wells. 1µl DNA ladder was pipetted into the ladder well and 1µl of the library samples were pipetted into each well, with 1µl of the marker in the unused wells. The DNA chip was then vortexed for 60 seconds and then read on the Bioanalyzer (Agilent Technologies, Santa Clara, USA). Reasonable quality samples were determined by a visible smear seen around the 100bp mark, indicating that the average transcripts are 100bp in length and thus the library amplification was successful.

Library Quantitation using qPCR

Ion Library TaqMan Quantitation Kit (4468802) was obtained from Thermo Fisher Scientific. Five 10-fold serial dilutions of DH10B Ion Control library (68pM) were performed at 1:10 (6.8pM) , 1:100 (0.68pM), 1:1000 (0.068pM), 1:10,000 (0.0068pM), and 1:100,000 (0.00068pM) to form a standard curve. 1:100 dilutions of libraries were made and 11µl of the master mixture (Table 22) was pipetted into each sample well of the PCR plate with 9µl of diluted library and standard samples. The plate was incubated in the CFX Connect thermal cycler (BioRad) at 50$^{\circ C}$ for 2 minutes, in 5 cycles of 95$^{\circ C}$ for 20 seconds, followed by 40 cycles of 95$^{\circ C}$ for 1 second with 60$^{\circ C}$ for 20 seconds.

Components of master mixture	Volume added for each reaction (µl)
Ion Library TaqMan® qPCR Mix (2x)	10
Ion Library TaqMan® Quantitation Assay (20x)	1
Total	11

Table 22: Volumes of various components in reaction master mixture for qPCR to quantify libraries

Next Generation Sequencing

Preparation of Template-Positive Ion Sphere Particles (ISPs) by Emulsion PCR

Ion PI Template OT2 200 Kit v3 (4488318, Thermo Fisher Scientific, Waltham, MA, USA) was used to produce template-positive ISPs from transcriptome libraries for semiconductor sequencing.150µL Ion OneTouch Breaking Solution was added to each of 2 Recovery Tubes placed within the centrifuge component of the Ion OneTouch 2 Instrument (Figure 20). An amplification plate was installed into the device and Reagent Tubes were filled with Ion OneTouch Oil and Recovery Solution.

Libraries were grouped in 3s for emulsion PCR (ePCR) and enrichment to allow 3 samples to be sequenced per chip. For ePCR, transcriptome libraries were diluted to either 100pM (if all 3 library concentrations were above 100pM) or to the lowest molarity of the 3 samples being run together (if one or more samples had molarity below 100pM). For those groups of libraries diluted to 100pM, 4µL of each diluted library was combined in one tube. 8µL of the combined libraries was then placed in a fresh tube and made up to 100µL with nuclease free water (NFW). For those groups diluted to concentrations below100pM, the total volume of combined libraries required was calculated using the equation: volume = 800 ÷ molarity. Library groups were then also made up to 100µL with NFW.

An amplification solution was prepared by adding the components listed in Table 23 to a tube containing 2µL of Ion PI™ Master Mix. The amplification solution was loaded onto the IonOneTouch Reaction Filter alongside 200µL of Ion OneTouch Reaction Oil. The filter was the installed into the Ion OneTouch 2 Instrument for ePCR.

The supernatant from both Recovery Tubes was removed, leaving behind approximately 100µL of solution containing a pellet of template-positive ISPs. The ISPs were resuspended in solution and combined into one fresh tube. 100µL NFW was added to each Recovery Tube to recover any residual beads and the solution transferred to the tube containing ISPs. The template-positive ISP solution was made up to 1µL with NFW.

Order	Reagent	Volume added (µl)
1	**NFW**	80
2	**Ion PI™ Enzyme Mix**	120
3	**Ion PI™ ISPs**	100
4	**Diluted combined libraries**	100
	Total Volume incl. Ion PI™ Master Mix	2400

Table 23: Reagents added to Ion PI™ Master Mix to make amplification solution for ePCR.

Quality Control using Qubit 2.0 Fluorometer

The percentage of template-positive ISPs in solution was calculated using Qubit 2.0 Fluorometer (Q32866, Thermo Fisher Scientific, Waltham, MA, USA) as per manufacturer guidelines. Samples were aimed to have 16-30% template-positive ISPs in order to proceed with enrichment and sequencing.

Enrichment of Template-Positive ISPs

Enrichment of ISPs was carried out using the Ion OneTouch ES (Thermo Fisher Scientific, Waltham, MA, USA). Melt-Off Solution was prepared by combining 280µL Tween Solution with 40µL 1MNaOH. 100µL of resuspended Dynabeads MyOne Streptavidin C1 Beads were transferred to anew tube and placed on the DynaMag-2 magnet (12321D, Thermo Fisher Scientific, Waltham, MA,USA) for 2 minutes. Supernatant was discarded and 1mL Ion OneTouch Wash Solution added to beads and vortexed for 30 seconds. The tube was returned to the magnet for 2 minutes and supernatant discarded. 130µL of MyOne Beads Capture Solution was added to the beads and the mixture vortexed for 30 seconds.

An 8-well strip was prepared using the components listed in Table 24. A new pipette tip and 0.2mLPCR tube were loaded onto the OneTouch ES alongside the filled wells and machine turned on for enrichment of ISPs.

A summary of the enrichment process is shown in Figure 33.

Well	Reagent	Volume added (µl)
1	**Template-Positive ISP Solution (in Resuspension Solution)**	100
2	**Dynabeads® MyOne™ Streptavidin C1 Beads (resuspended in Capture Solution)**	130
3	**Ion OneTouch™ ES Wash Solution**	300
4	**Ion OneTouch™ ES Wash Solution**	300
5	**Ion OneTouch™ ES Wash Solution**	300
6	**Empty**	-
7	**Melt-Off Solution**	300
8	**Empty**	-

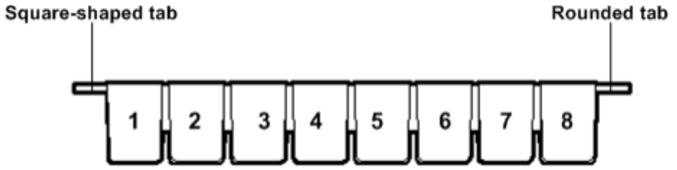


Table 24: Reagents added to 8-well strip for enrichment of ISPs. Image from Thermo Fisher Scientific 2011.

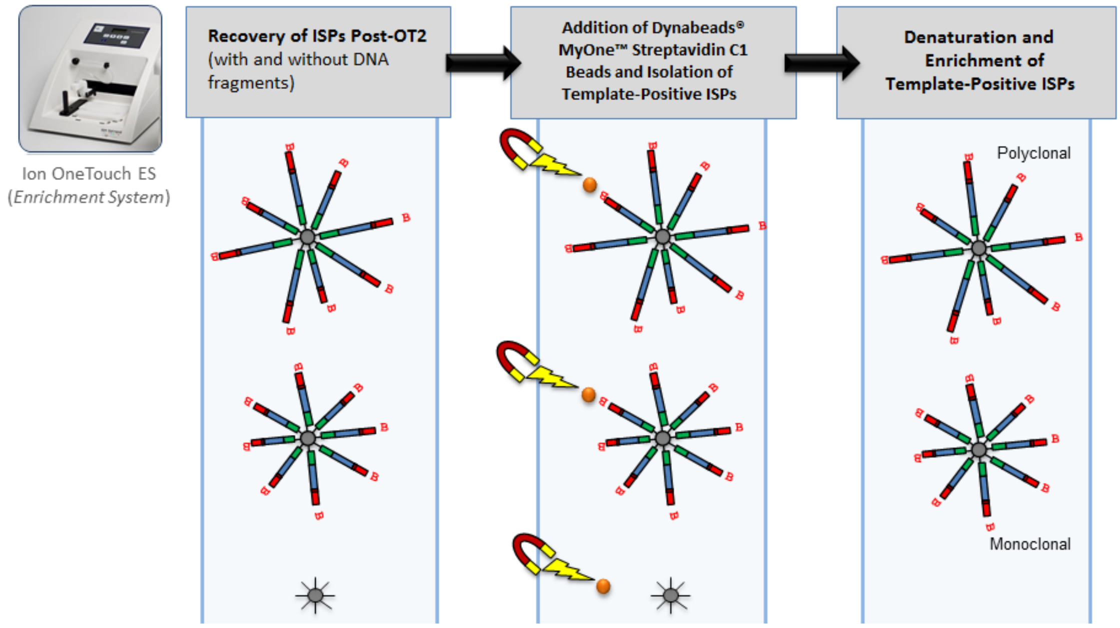

Ion OneTouch ES
(Enrichment System)
Recovery of ISPs Post-OT2
(with and without DNA fragments)
Addition of Dynabeads® MyOne™ Streptavidin C1 Beads and Isolation of Template-Positive ISPs
Denaturation and Enrichment of Template-Positive ISPs
Polyclonal
Monoclonal

Figure 33: Summary diagram of enrichment of Template-Positive IPSs using Ion OneTouch™ ES. *Dynabeads® MyOne™ Streptavidin C1 Beads bind Biotin component of amplicons, isolating these from unbound ISPs when placed in the magnet. Adapted from Blervaque 2013.*

Ion Semiconductor Sequencing

Ion PI Hi-Q Sequencing 200 Kit (A26433) and Ion PI™ Chip Kit v3 (A26771) from Thermo FisherScientific were used for semiconductor sequencing. The Ion Proton System (4476610) (Figure 34) was initialised with Wash and Clean Solutions as well as 4 deoxyribonucleotide (dNTP) solutions (Figure 34).

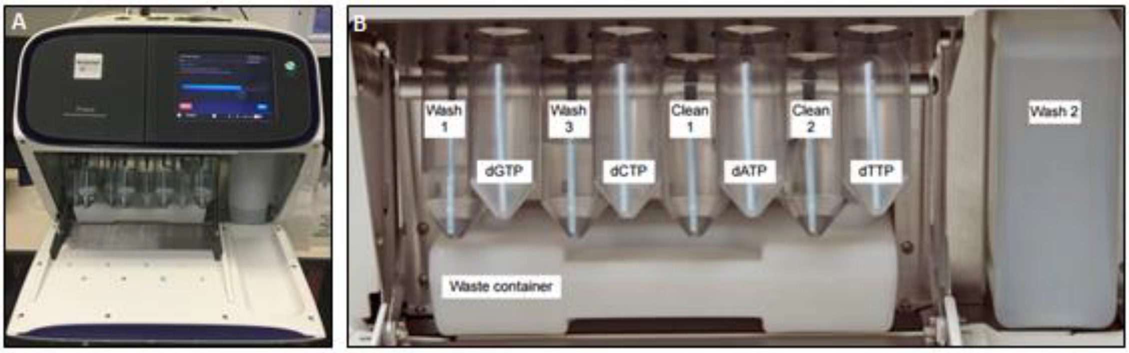


Figure 34: (A) Ion Proton™ System for semiconductor sequencing. (B) List of reagents placed in containers and reagent tubes in Ion Proton™ System. dGTP, dCTP, dATP and dTTP represent the 4 dNTPs. Taken from Thermo Fisher Scientific 2013.

Flushing Solution and 50% Annealing Buffer were prepared as per manufacturer guidelines. 5μL IonPI Control ISPs were added to the enriched ISP samples previously made. The mixture was centrifuged for 5 minutes at 15,500 x g and supernatant removed leaving 10μL solution behind with the pellet. Ion PI Annealing Buffer (15μL) was added alongside 20μL of Ion PI Sequencing Primer.

After vortexing the mixture, samples were placed in the thermal cycler for 2 minutes at 95°C and 2minutes at 37°C. Ion PI Loading Buffer (10μL) was then added to the mixture creating a total volume of 55μL.

Samples were loaded onto the Ion PI Chip v3 (Figure 35) and centrifuged for 10 minutes in the IonChip Minifuge (4479672, Thermo Fisher Scientific, Waltham, MA, USA). In a 1.5mL tube, 49μL of 50% Annealing Buffer and 1μL Foaming Solution (10% Triton X-100) were injected with air by pipetting to create a foam. 100μL of foam was then injected into the chip loading port, with expelled liquid removed from the opposite port. 55μL of 50% Annealing Buffer was placed in the chip loading well and centrifuged for 30 seconds. The foam and Annealing Buffer steps were repeated once again. The chip was injected with 100μL Flushing Solution twice followed by 100μL Annealing Buffer three times. Finally, 6μL Ion PI Hi-Q Sequencing Polymerase and 60μL 50% Annealing Buffer were combined and injected into the chip loading port with any excess solution removed from the opposite well. The chip was incubated at RT for 5 minutes and then loaded onto the IonProton System for sequencing.

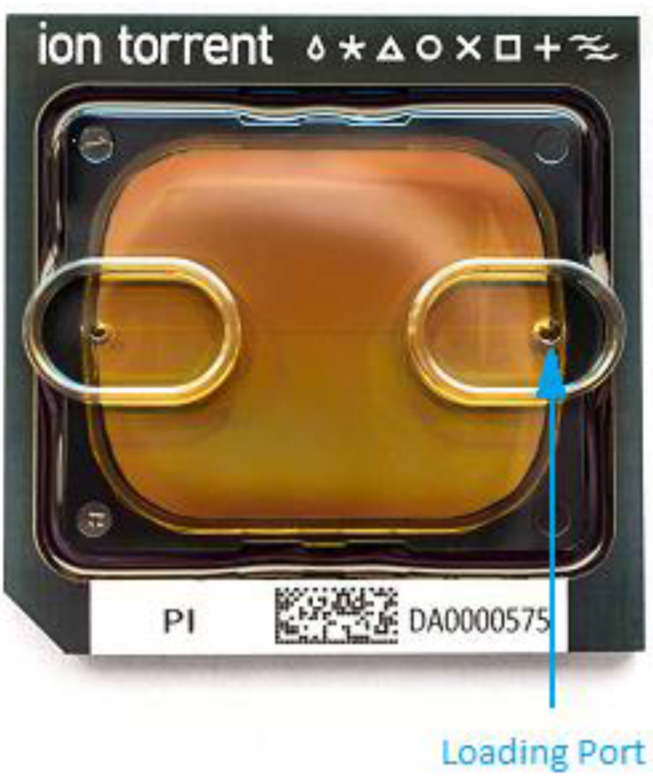


Figure 35: Ion PI Chip v3. The loading port located within the loading well is demonstrated with the blue arrow. Image adapted from Thermo Fisher Scientific 2015.

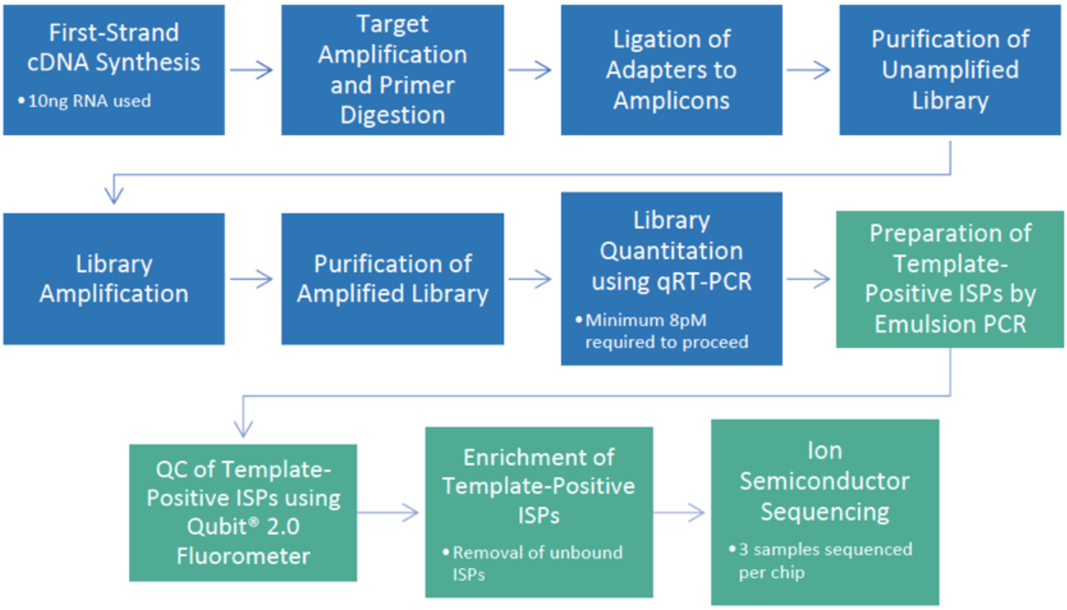

First-Strand cDNA Synthesis
•10ng RNA used
Target Amplification and Primer Digestion
Ligation of Adapters to Amplicons
Purification of Unamplified Library
Library Amplification
Purification of Amplified Library
Library Quantitation using qRT-PCR
•Minimum 8pM required to proceed
Preparation of Template-Positive ISPs by Emulsion PCR
QC of Template-Positive ISPs using Qubit® 2.0 Fluorometer
Enrichment of Template-Positive ISPs
•Removal of unbound ISPs
Ion Semiconductor Sequencing
•3 samples sequenced per chip

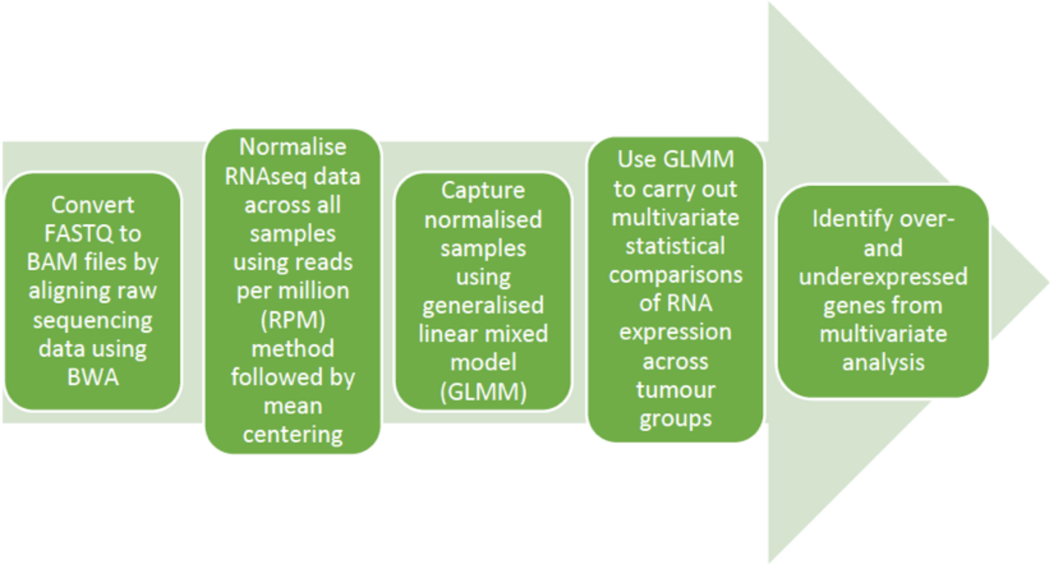
Figure 36: Flow diagram of steps involved in whole transcriptome amplification (blue boxes) and NGS (green boxes). QC= quality control.

Statistical Analysis

Statistical analysis was performed by Prof. Rifat Hamoudi using Microsoft Excel 2013 and IBM SPSS software 22.0. CFXManager (Bio-Rad, UK) was used for standard curve fitting and regression equation in qRT-PCR. Expert 2100 version B.02.08 (Agilent, CA, USA) was used for bioanalysis.

Bioinformatics Analysis

Bioinformatics analysis was completed by Prof. Rifat Hamoudi using the methods described in Figure 37. Whole RNA sequencing analysis was performed using Torrent Suite version 5.0.1 (ThermoFisher Scientific, MA, USA) and R/Bioconductor version 3.2. DESeq2 (https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8) was used to normalise the whole transcriptome data. The Database for Annotation, Visualization and Integrated Discovery (DAVID) (https://david.ncifcrf.gov/home.jsp) and metascape (https://metascape.org/gp/index.html) was used for annotation and visualisation of gene expression data. The Benjamini-Hochberg correction method was used to assess the False Discovery Rate (FDR), where p<0.05 and FDR<0.25 were deemed to be statistically significant.



Convert FASTQ to BAM files by aligning raw sequencing data using BWA

Normalise RNAseq data across all samples using reads per million (RPM) method followed by mean centering

Capture normalised samples using generalised linear mixed model (GLMM)

Use GLMM to carry out multivariate statistical comparisons of RNA expression across tumour groups

Identify over- and underexpressed genes from multivariate analysis

Figure 37: Sequential method used for bioinformatics analysis. FASTQ: file format containing raw nucleotide data, BAM (Binary Alignment Map): file containing sequence alignment data, BWA (Burrows-Wheeler Aligner): software used to map raw sequencing data to reference genome.

Reference transcriptome

In order to interpret the results of this work it was necessary to have a reference transcriptome to provide comparison. Following discussion in regards of using saliva or oesophageal biopsies from normal individuals as the reference it was decided that given that our aims are to identify patients with or at risk of OAC we should use normal oesophagus as reference. As such snap frozen biopsies were obtained from three patients found to have a normal upper gastrointestinal tract following endoscopic examination. These were processed using the previously outlined methods in order to create the reference transcriptome. The transcriptomes obtained from the OAC data were compared with the reference to ensure that they are drivers for the risk to developing OAC.

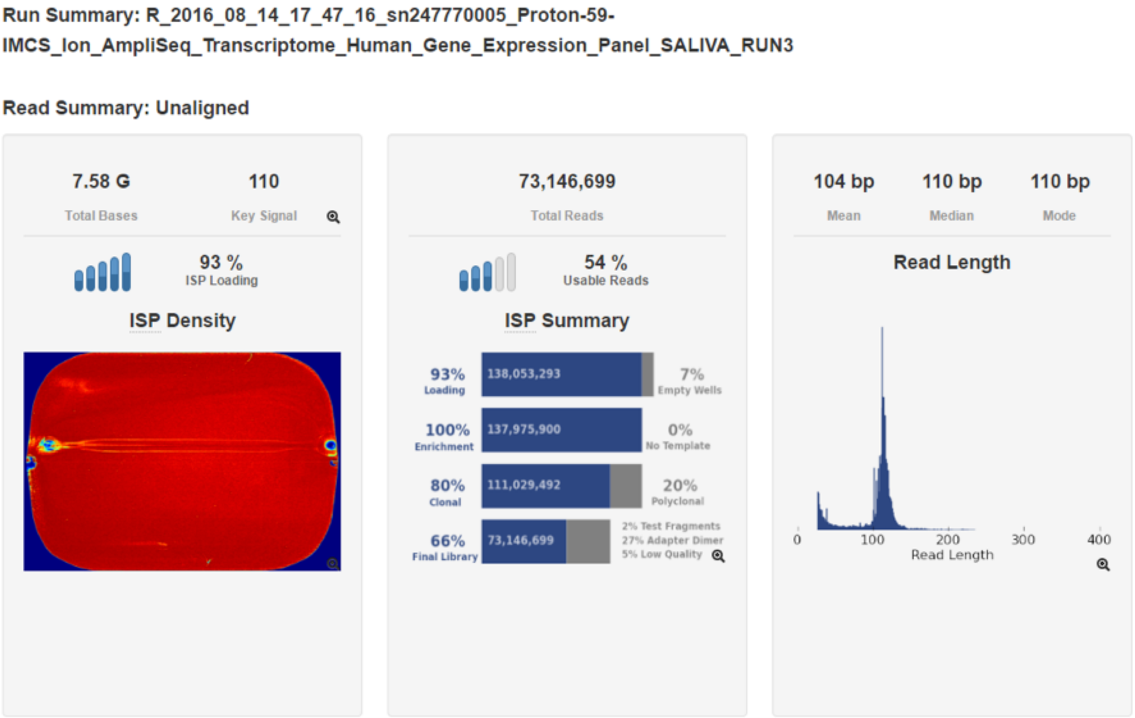
6.5.2 Objective 3 – Whole mRNA sequencing results

Whilst the power calculation determined that 4 samples from each diagnostic group were required for statistical significance due to time constraints only 2 per group were performed as part of this work. The further samples will be analysed at a later date.

Quality control of Ion PI™ NGS Chips

Qubit 2.0 Fluorometer (Q32866, Thermo Fisher Scientific, Waltham, MA, USA) analysis was performed for quality control of ISPs. Percentage templated ISPs are shown in the below table, ranging from 20-30%. Given these values, it was possible to proceed with enrichment and NGS for all samples.

An example of NGS chip quality control parameters is shown in Figure 38 below.
Table 25 shows the total usable transcript reads for each sample as well as base call quality and mean read length. For each chip, it was aimed to have around 1,000,000 good quality sequencing reads per sample in order for bioinformatics analysis to be performed adequately.



Run Summary: R_2016_08_14_17_47_16_sn247770005_Proton-59-IMCS_Ion_AmpliSeq_Transcriptome_Human_Gene_Expression_Panel_SALIVA_RUN3

Read Summary: Unaligned

7.58 G Total Bases

110 Key Signal

93 % ISP Loading

ISP Density

73,146,699 Total Reads

54 % Usable Reads

ISP Summary

93% Loading	138,053,293	7% Empty Wells
100% Enrichment	137,975,900	0% No Template
80% Clonal	111,029,492	20% Polyclonal
66% Final Library	73,146,699	2% Test Fragments 27% Adapter Dimer 5% Low Quality

104 bp Mean

110 bp Median

110 bp Mode

Read Length

Figure 38: Example of NGS chip quality as displayed following sequencing

Sample number	Diagnosis	Reads	Mean read length (bp)
400	Normal	21,042,188	103
402	Normal	10,321,584	96
1186	Normal	5,398,939	83
1259	Normal	4,583,510	50
1229	Normal	2,361,540	33
1260	Normal	6,678,010	72
1205	NDBO	3,801,012	75
1212	NDBO	3,587,223	72
1231	NDBO	9,225,973	89
1275	NDBO	9,343,506	45
1256	NDBO	11,896,935	78
1164	NDBO	10,076,580	95
286	NDBO	7,773,873	92
1098	NDBO	40,921,874	108
410	HGD	17,774,573	89
1085	HGD	14,042,854	78
1234	HGD	4,367,633	67
1235	HGD	7,288,375	42
1253	HGD	11,285,009	75
244	OAC	9,833,575	87
263	OAC	12,460,236	84
1207	OAC	28,055,265	100
1195	OAC	47,390,240	107
1261	OAC	13,401,333	70
1240	OAC	60,370,169	109

Table 25: Usable transcript reads and mean read length for each sample

The initial work using targeted biomarkers, presented in section 6.4, demonstrated that salivary RNA could be used as a means to predict OAC risk. Taking this forward it was important to explore whether there were other biomarkers that could improve on this work and consequently improve the accuracy of the risk prediction. As a demonstration of the potential for this approach, two well characterised patients from each diagnostic group were chosen and their RNA extracted as previously outlined in Chapter 5. Whole transcriptome RNAseq was performed using the Ion Proton.

Global Quality Control Measures of NGS

An important aspect of the quality control is to determine the quality of alignment to the transcriptome and the genes that are significantly differentially expressed.

The patient samples that were analysed for this work were aligned to the reference transcriptome. The following histogram demonstrates that the samples processed had good alignment to the transcriptome. A *p* value < 0.05 is taken as significant and this histogram demonstrates that the highest frequency peaks (seen in the first two columns) occurs at this level. Thus, for this work around 2,400 genes were differentially expressed between the different groups of the Barrett metaplasia-dysplasia-adenocarcinoma sequence.

The initial quality control test to ensure the methodology worked correctly is to identify the number of differentially expressed genes from the whole transcriptome analysis across all the groups. This is carried out using *p*<0.05 as an indicator of statistical significance, a *p*-value histogram was generated to determine the proportion of transcripts which were significantly differentially expressed across all samples (Figure 39). Around 2,400 genes were found to be significantly differentially expressed between the groups.

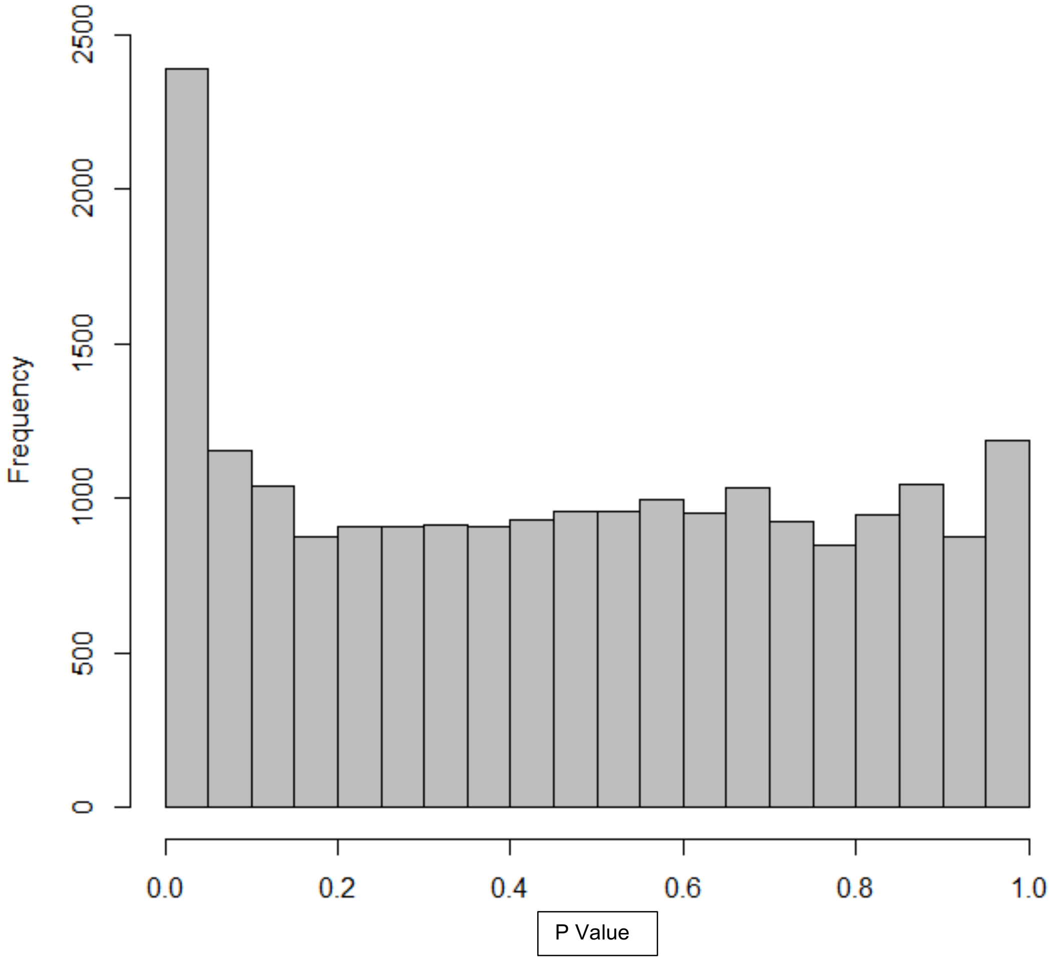


Figure 39: Histogram demonstrating number of genes with their respective P-values aligned to transcriptome

This correlation plot provides an overview of how closely related the samples are to each other. The darker the colour on the plot, the more closely they are related. Interestingly, although perhaps not surprisingly given the incurred risk, one can observe from this plot that the samples that are Normal and NDBE are most closely correlated to each other but there is a mixture between the samples, indicating the high degree of heterogeneity that exist in the Barrett's metaplasia-dysplasia-adenocarcinoma sequence.

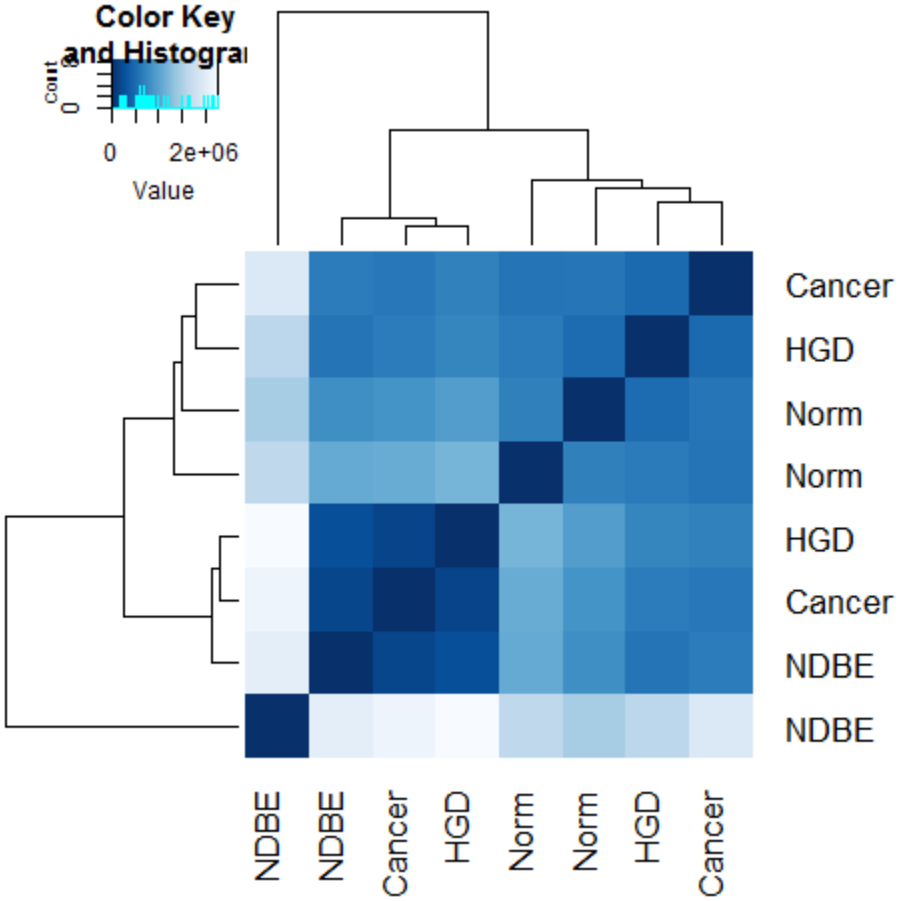


Figure 40: Correlation plot between samples following RNA-seq analysis

Finally, as part of the quality control the following histogram represents the hierarchical unsupervised clustering of the 8 samples. Each column represents the genes that were aligned to the transcriptome and significantly differentially expressed ($p < 0.05$, Figure 41) with red indicating high expression and blue being low expression. Again, one can see that one of the NDBE does not relate to any other sample. One also notes that there is a large amount of heterogeneity between each sample which is in keeping with the literature [133]. One would also expect this heterogeneity to be more significant given that we are using a saliva sample, rather than a biopsy of the lesion, that is affected by a variety of local and systemic factors. In order to ascertain useful information from these samples, therefore, one must use multivariate statistical analysis to detect genes of significance. This approach has been published [229].

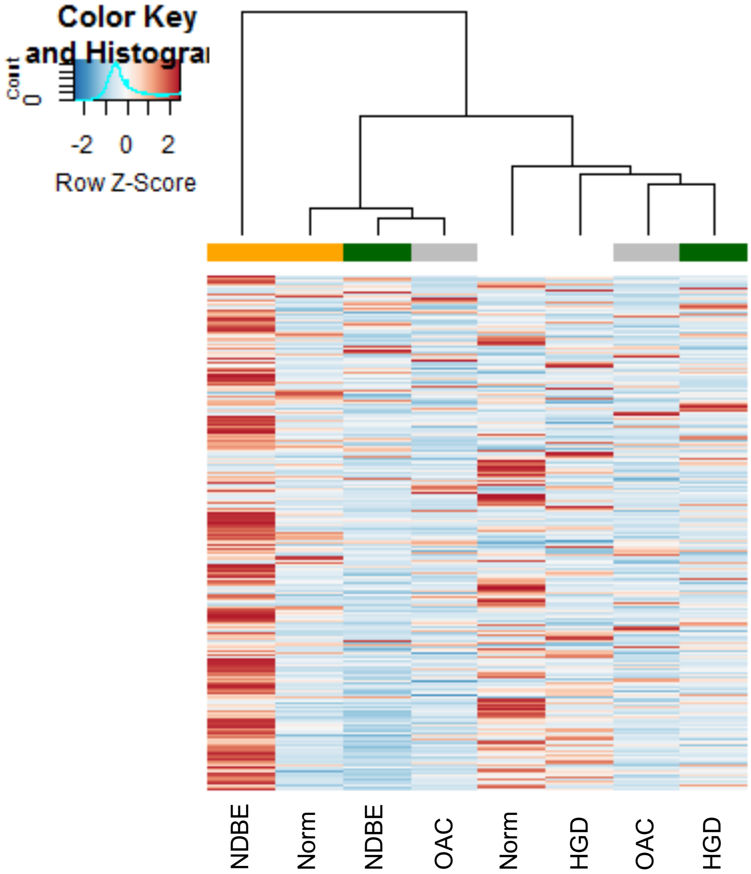


Figure 41: Histogram of hierarchical unsupervised clustering between samples

Differential Gene Expression Analysis

In order to assess differential gene expression across different groups, transcript copies for each gene were combined across all members of individual groups (e.g. NDBO). Comparisons were then made between varying sets of 2 groups using the unpaired *t*-test. Statistically significant differential expression was defined by a *p*-value <0.05. An adjusted *p*-value was then derived using Benjamini Hochberg correction method. GeneCards® (www.genecards.org) was used to determine gene annotation and any previously known association with cancer development.

The bioinformatics analysis resulted in around 2,400 differentially expressed genes across all 8 samples. Multivariate statistics and variation filtering reduced it to 134 genes common between NDBE, HGD and OAC. Applying AI (MIAT) and pathway analysis to the 134 genes identified narrowed them down to 40 genes.

The below table outlines the number of genes found to be significantly aberrantly expressed between the diagnostic groups and the below list highlights some of the genes of interest.

Diagnostic groups	Number of genes aberrantly expressed (adj. p <0.05)
Normal v NDBE	1024
Normal v HGD	1147
Normal v OAC	1347

Gene	Gene Name
SERPINB1	Serpin family B member 1(SERPINB1)
TLR6	Toll like receptor
TM2D2	TM2 domain containing 2(TM2D2)
SUSD4	Sushi domain containing 4(SUSD4)
FOXRED2	FAD dependent oxidoreductase domain containing 2(FOXRED2)
HEXIM2	Hexamethylene bisacetamide inducible 2(HEXIM2)
LILRB2	Leukocyte immunoglobulin like receptor B2(LILRB2)
SIGLEC14	Sialic acid binding Ig like lectin 14(SIGLEC14)
BCL2L14	BCL2 like 14(BCL2L14)
BPIFB1	BPI fold containing family B member 1(BPIFB1)
CD44	CD44 molecule (Indian blood group)(CD44)
COCH	Cochlin(COCH)
ECSIT	ECSIT signalling integrator(ECSIT)
FAM187B	Family with sequence similarity 187 member B(FAM187B)
H1FNT	H1 histone family member N, testis specific(H1FNT)
HOXA13	Homeobox A13(HOXA13)
HS1BP3	HCLS1 binding protein 3(HS1BP3)
NDRG3	NDRG family member 3(NDRG3)
SERPINC1	Serpin family C member 1(SERPINC1)
TCEAL4	Transcription elongation factor A like 4(TCEAL4)
CDK2	Cyclin dependent kinase 2(CDK2)
CLEC12A	C-type lectin domain family 12 member A(CLEC12A)
CD80	CD80 molecule(CD80)
CD82	CD82 molecule(CD82)
H3F3A	H3 histone family member 3A(H3F3A)
MAP4K3	mitogen-activated protein kinase kinase kinase kinase 3(MAP4K3)
RUNX3	Runt related transcription factor 3(RUNX3)
TLR1	Toll like receptor 1(TLR1)
ZER1	Zyg-11 related cell cycle regulator(ZER1)
CCND2	Cyclin D2(CCND2)
SNORA9	Small Nucleolar RNA, H/ACA Box 9
SNORA34	Small Nucleolar RNA
SMAD7	SMAD family member 7(SMAD7)
CDKN2A	Cyclin dependent kinase inhibitor 2A(CDKN2A)
TP53	Tumor protein p53(TP53)

Table 26: Genes significantly aberrantly expressed between the diagnostic groups during RNAseq and a list of some of the genes of interest

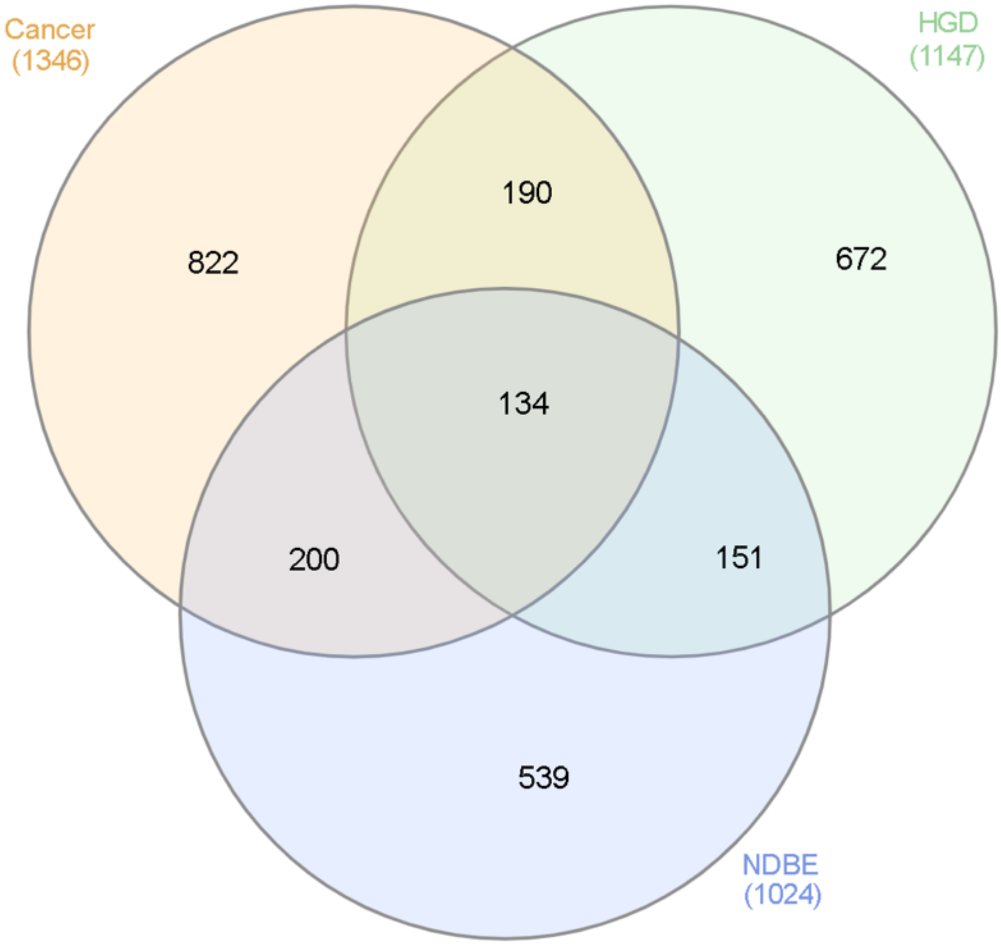


Figure 42: Venn diagram showing the comparison of the significant differentially expressed transcriptomes across the Barrett's sequence groups

The following figures focus on some key differentially expressed pathways that were found common to all the diagnostic groups or at each diagnostic stage.

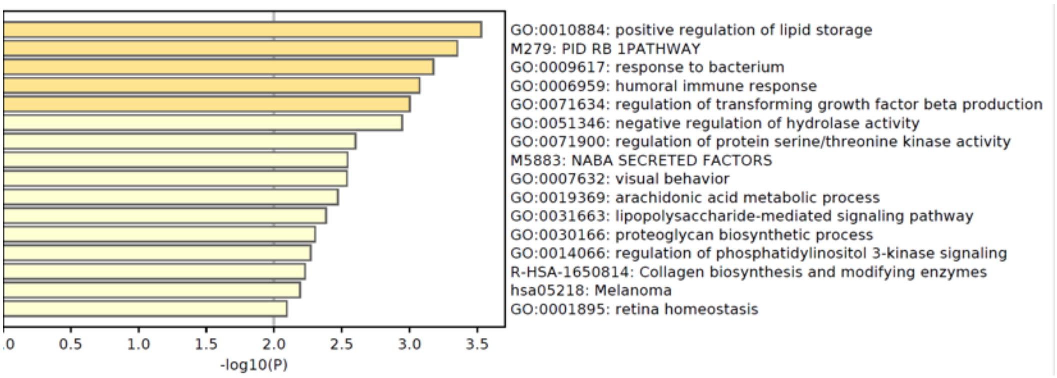


Figure 43: Differentially expressed pathways common between NDBO, HGD and OAC

This shows the retinoblastoma pathway to be over expressed (M279: PID RB 1PATHWAY). The retinoblastoma pathway plays a key role in regulating the cell cycle by interacting with dimerization partner 1 (DP1) leading to transcriptional activation [289]. Other interesting pathways are GO:0006959: humoral immune response suggesting that acquired immunity is important in progression from NDBO to OAC and GO:0009617: response to bacterium suggesting that at the initial stages innate immunity probably play a role through activating inflammatory pathways either via bacterial infection or more likely via acid reflux or similar irritant.

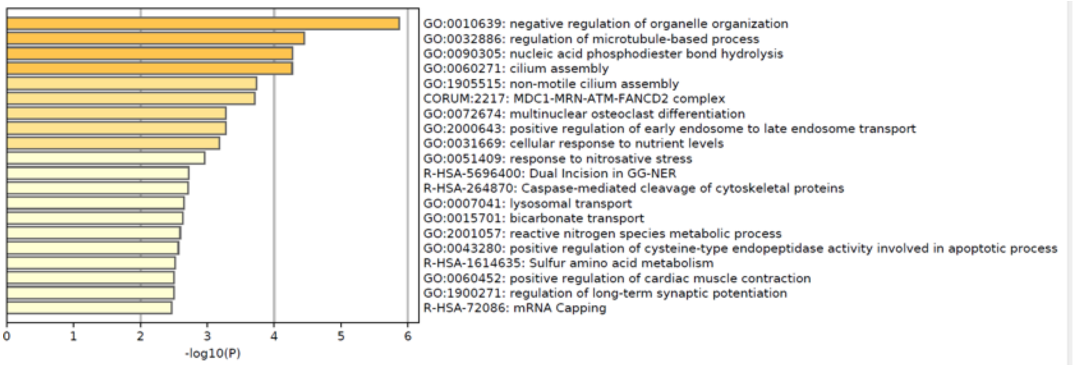


Figure 44: Differentially expressed pathways unique to NDBE

An interesting pathway in this analysis is CORUM:2217: MDC1-MRN-ATM-FANCD2 complex that relates to DNA repair. This possibly shows that the tissue is exposed to constant damage resulting in DNA damage and consequently mutations that eventually lead to OAC.

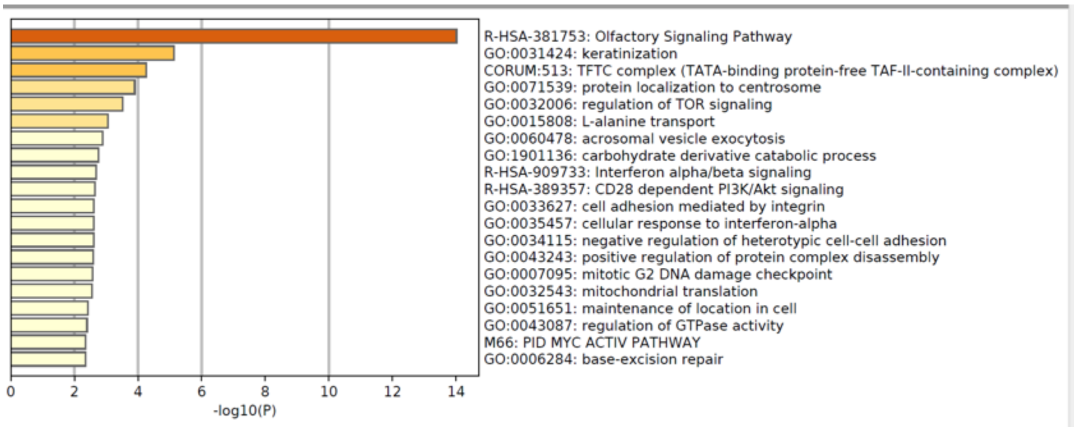


Figure 45: Differentially expressed pathways unique to HGD

At this stage the pathways expressed suggest that there is a lot of attempted DNA repair occurring. However, failure to effectively repair the damage will then result in OAC. The aberrant expression of CORUM:513: TFTC complex (TATA-binding protein-free TAF-II-containing complex), GO:0006284: base-excision repair pathway demonstrates this as it is involved with DNA repair. GO:0007095: mitotic G2 DNA damage checkpoint is aberrantly expressed showing that DNA damage checkpoint at G2 is activated to repair the DNA before progressing to mitosis. R-HSA-909733: Interferon alpha/beta signalling and R-HSA-389357: CD28 dependent PI3K/Akt signalling are immune pathways mostly at the checkpoint between B and T cell interaction and if those are not properly regulated they can lead to OAC as shown by the presence of activated GO:0032006: regulation of TOR signalling and M66: PID MYC ACTIV PATHWAY which are common cancer related pathways. This set of aberrantly expressed genes show a mixture of NDBO and OAC-related processes of DNA repair and cancer pathways as well as the interface of immune surveillance and checkpoint immune markers.

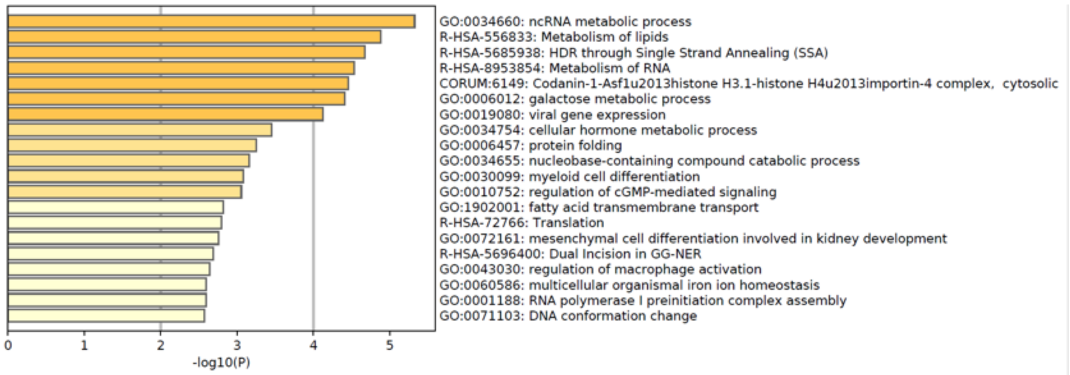


Figure 46: Differentially expressed pathways unique to OAC

This figure demonstrates the pathways that are related to tissue remodelling in cancer such as GO:0030099: myeloid cell differentiation and GO:0072161: mesenchymal cell differentiation involved in kidney development pathways are aberrantly expressed. Interestingly, the CORUM:6149: Codanin-1-Asf1u2013histone H3.1-histone H4u2013importin-4 complex, cytosolic pathway is aberrantly expressed indicating that histones are involved. Examining the OAC differentially expressed genes shows a family of histones to be differentially expressed including HIST1H3E, HIST1H2BF, HIST1H2BC, HIST1H3E, HIST1H2BF and HIST1H2BC further indicating the involvement of histones at the later stage of the Barrett's metaplasia-dysplasia-carcinoma sequence. Together, this demonstrates the end point shift in cellular pathways if the DNA repair and homeostatic regulation mechanisms fail at the HGD stage.

The figures below demonstrate the expression profile of some chosen genes within the 4 diagnosis groups:

Figure 47: Expression profile of selected genes within each diagnostic group

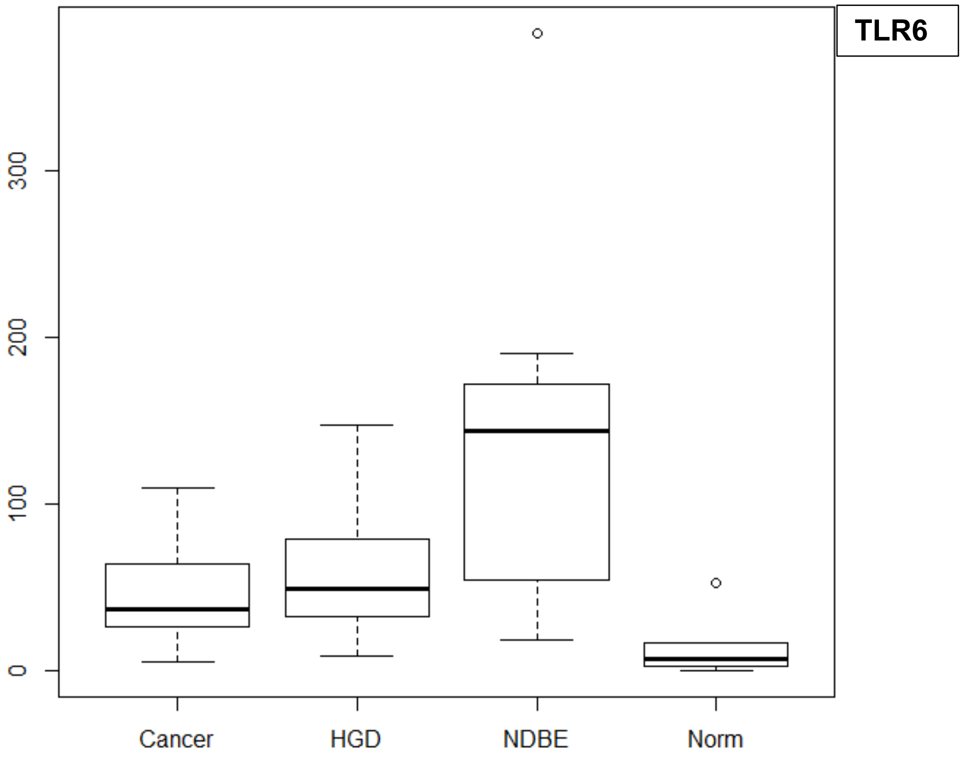


Figure 47a: TLR6

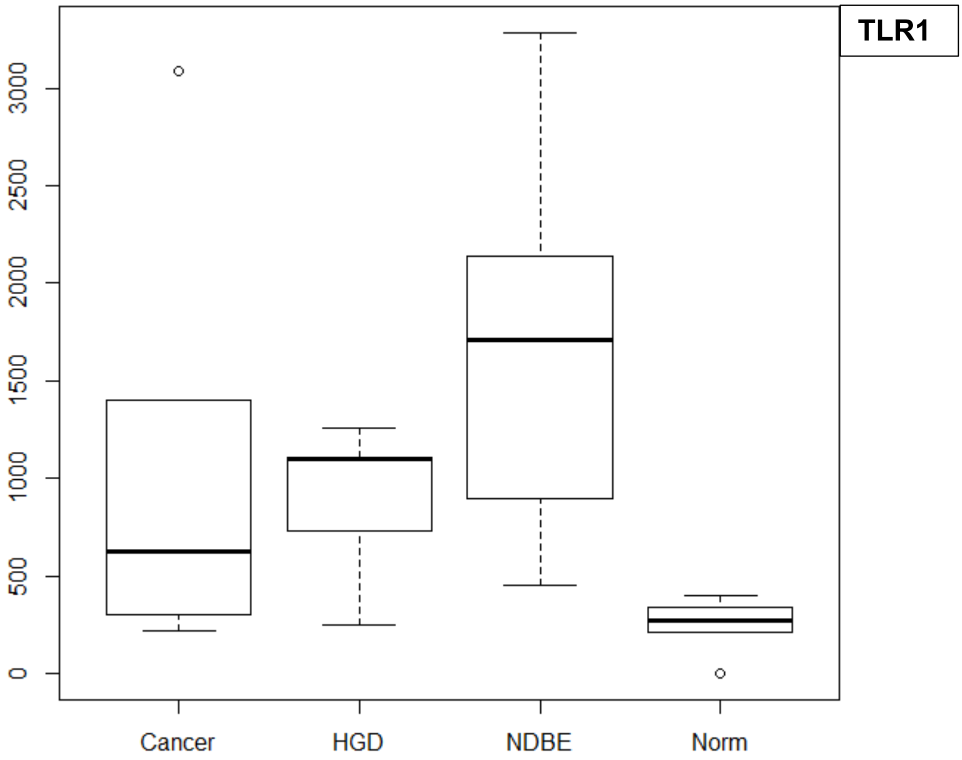


Figure 47b: TLR1

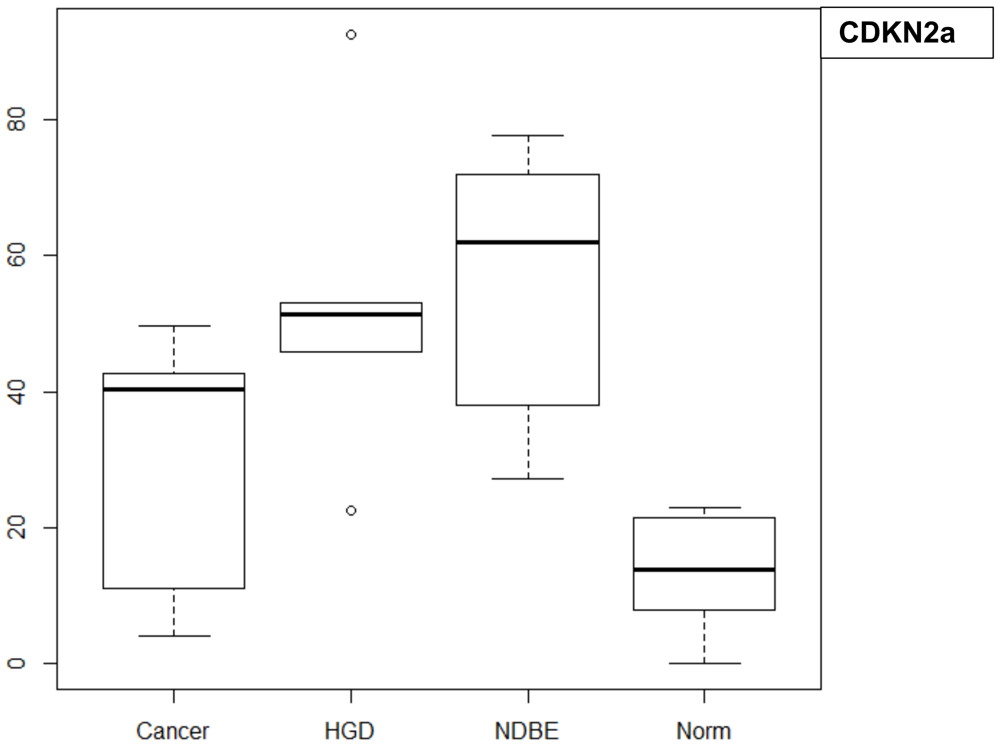


Figure 47c: CDKN2a

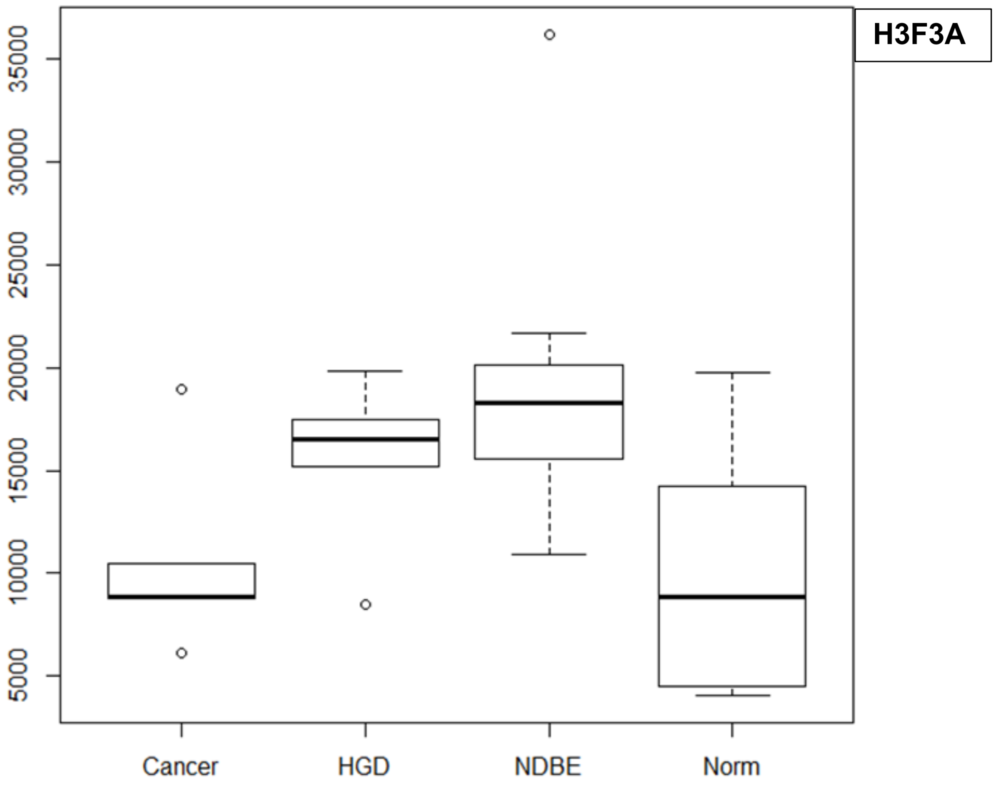


Figure 47d: H3F3A

Of the 13 genes originally tested in the 80 patient group it was notable, as highlighted in Figure 26, how effective CDKN2a was in differentiating patients with or at risk of OAC. Following the sequencing of eight patients (two samples in each diagnostic category) it is again notable that when comparing the normal patients to any of the other diagnostic groups (NDBO, HGD, OAC) CDKN2a is significantly aberrantly expressed on each occasion. Whilst this sample size if not significantly powered it does continue to suggest that this marker may be of benefit in the early identification of those with or at risk of OAC.

Of the other genes found to be significantly aberrantly expressed in the 80 patient group, when compared to this sequencing group, there is little overlap, although again this sequencing group is not sufficiently powered. When the normal patients are compared to those with OAC TP53 is significantly aberrantly expressed which is in keeping with the literature.

From this sequencing group alone there are some other additional interesting findings. Of particular note is Cyclin D2. The expression of this becomes progressively higher as samples progress from normal to NDBO and HGD before falling in patients with OAC although still remaining higher expressed than in normal patients. This is discussed below.

6.5.3 Objective 3 – Whole mRNA sequencing. Discussion

The sample size used for this work is too small to allow definitive conclusions. However, this work does demonstrate that, despite the poor quality of salivary samples in regard to its RIN value, one can still perform discovery RNA-Seq and generate potentially meaningful results. Further work will be performed to complete this discovery set with the results used to interrogate whether they can be utilised to accurately identify individuals with or at risk of OAC.

Despite this small sample size, it is interesting to note that of the samples processed two key genes found in the targeted expression analysis work, CDKN2A and TP53, were again found to be of importance. The role of these genes in the progression towards OAC has been discussed previously within this thesis. The identification of these genes in the saliva samples of patients with or at risk of OAC again suggests that these salivary markers may be of key importance in future predictive tool creation.

Cyclin D2 was noted to be of particular interest within this work. This became increasingly expressed as patients progressed from normal to NDBO and HGD. Although the expression became lower in the OAC group it remained up-regulated in comparison to normal. The Cyclin D family is involved in the cell cycle regulation process as well as growth factor dependent intracellular pathways. Thus, defects in its regulation could lead to an absence of growth regulation of cancer cells. Cyclin D2 has been linked to cancer in other work including its overexpression in gastric cancer [290, 291]. There has been little work exploring Cyclin D2 in BO and OAC although published data demonstrates Cyclin D1 to be of importance in BO and SCC of the oesophagus [292, 293].

Within the patients with HGD there are a further two genes of interest. PRKCA is noted to be significantly overexpressed in those with HGD. PRKCA is known to be a pan activator of the NF*k*β pathway and as such is involved in the chronic inflammation linked to the progression to cancer. The role of this pathway is well-established to be linked to cancer, particularly within the gastrointestinal tract, and as such further exploration of this genes role in OAC is indicated [294]. Furthermore, recent work in animal models suggests that this gene is linked to BO and OAC [295]. The second gene of note within the HGD patients was AMZ2P1 which again was noted to be significantly overexpressed in those with HGD. This gene is known to be linked to the regulation of SMAD which has already been discussed to be important in the pathogenesis of BO and OAC. Additionally, AMZ2P1 has been linked to colorectal cancer [296].

Within the patients with OAC another gene of note is RRAGB which was noted to be significantly overexpressed. This gene is known as a signal transducer that essentially acts as a switch for the regulation of certain pathways including the mTOR, AKT and PI3K pathways. These pathways have been linked to cancer as well as BO and as such further work exploring whether RRAGB can act as a salivary biomarker should be performed should it remain significant following completing the discovery set.

The identification of histone family of genes (HIST1H3E, HIST1H2BF, HIST1H2BC, HIST1H3E, HIST1H2BF and HIST1H2BC) in OAC is also of interest. *Graber et al* demonstrated this to be expressed in 4/6 patients with OAC and 11/12 cancer cell lines (compared to 5/14 normal cell lines) [297]. This gene links to histone production which suggests an epigenetic role in the pathogenesis of OAC. This area is further explored in section 6.6.

Finally, BPIFA1 and BPIFB1 were noted to be down-regulated in the progression towards OAC from normal. These genes are linked to the innate immune system and are known to be involved in the inflammatory response as well as linked to some cancers in particular lung cancer. It should be noted, however, that these genes are particularly linked with oral and nasopharyngeal disease and as such salivary samples will be particularly vulnerable to local disease.

6.6 Objective 4 - Detecting epigenetic biomarkers using salivary DNA

Following on from the discussion in Chapter 2 there have been limited genomic biomarkers discovered that have identified the high-risk individual and it is clear that the epigenomic regulation of the genome plays a key role in the pathogenesis of BO and OAC. As such I was keen then to explore whether the salivary epigenome could provide invaluable data that could identify the at-risk individual. Epigenetics is of particular interest in the development of BO and OAC as it provides the potential link of environmental variables in the pathogenesis of disease. We know that factors such as gastro-oesophageal reflux disease are heavily linked to BO and OAC and as such studies utilising epigenetic data may provide valuable insights. Additionally, it is well established that DNA is more robust than RNA because of its double helix structure whereas RNA is single stranded. As such should it prove that DNA analysis provides adequate data for the identification of those with or at risk of OAC then the population screening tool may utilise DNA only as it may allow for easier collection and extraction conditions when using this for large scale whole population work.

6.6.1 Objective 4 - Performing epigenetic analysis on salivary samples. Methods

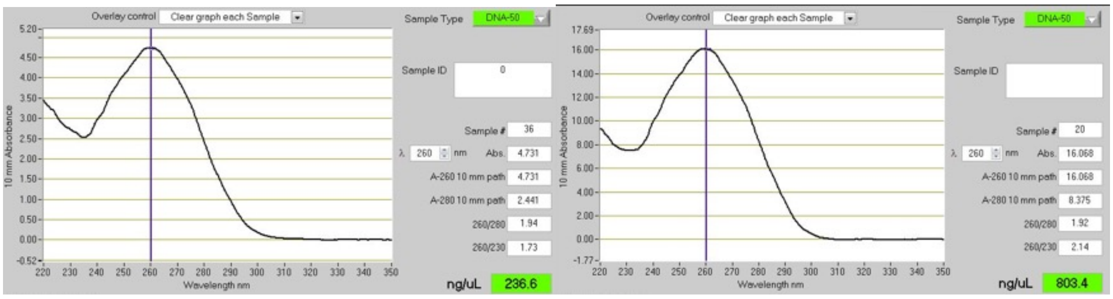
The samples used for the extraction of DNA were those also used for the extraction of RNA. Thus, the relevant processes are outlined in 6.4

DNA extraction and quality control

In order to extract the RNA only 400µl of the 1ml saliva sample was used. It would have been possible, therefore, to use the remainder of the sample for the DNA extraction process. However, ideally, we wanted to leave some patient sample remaining in case of the need to do future work on these samples. As such we decided to use the same sample that had undergone RNA extraction. Therefore, once the upper 400µl of the aqueous phase had been removed to undergo further RNA extraction processes the rest of the sample, containing Qiazol (Qiagen, Hilden, Germany), chloroform and patient saliva was placed on ice. Once the RNA extraction processes had been completed, the sample was retrieved from the ice and the DNA extraction processes began. The first step of this was to remove the upper 100µl of the aqueous phase to minimise the chances of RNA contamination.

Back extraction buffer (BEB) was made by combining 4M Guanidine Thiocyanate, 50 mM Sodium Citrate and 1M Tris (free base) dissolved in millipore water. 500µl of the BEB was then placed in an Eppendorf containing the Qiazol (Qiagen, Hilden, Germany), chloroform and patient saliva mix. This was then inverted for 10 minutes prior to being spun in the centrifuge for 30 minutes at $4^{\circ C}$. Similar to the RNA extraction process, the upper 400µl of the aqueous phase was removed following centrifugation and placed in a 1.5ml Eppendorf. An equal volume of isopropanol was added to the sample, vortexed for 1 minute and then left in ice for 2 hours. Following this it was spun in the centrifuge again for 30 minutes at $4^{\circ C}$ and then the supernatant was removed whilst carefully avoiding the pellet. The pellet was then washed using 500µl of

70% ETOH, vortexed for 1 minute and spun in the centrifuge for 7 minutes at 4°C. Finally, again, the supernatant was removed and the sample air-dried for 10 minutes prior to 32µl of DPEC being added and the sample mixed using a pipette. The samples underwent initial quality control by using NanoDrop (Thermo Fisher Scientific, Waltham, USA). However, it was noted at this point that a significant proportion of the DNA samples (34 out of 80) had nanodrop curves with significant contamination from the extraction process that had potential to affect the downstream processes. Therefore, all the samples were further cleaned and concentrated using the Zymo DNA Clean & Concentrator™-5 (Zymo Research, Irvine, CA, USA). This was performed following the protocol supplied by the manufacturer. Following completion, the DNA was of better quality as determined by the NanoDrop (Thermo Fisher Scientific, Waltham, USA) images. Once passed through the quality control processes the samples were stored in a -80°C freezer.



Overlay control: Clear graph each Sample
Sample Type: DNA-50
Sample ID: 0
Sample #: 36
λ 260 nm Abs. 4.731
A-260 10 mm path: 4.731
A-280 10 mm path: 2.441
260/280: 1.94
260/230: 1.73
ng/uL: 236.6

Overlay control: Clear graph each Sample
Sample Type: DNA-50
Sample ID:
Sample #: 20
λ 260 nm Abs. 16.068
A-260 10 mm path: 16.068
A-280 10 mm path: 8.375
260/280: 1.92
260/230: 2.14
ng/uL: 803.4

Figure 48: NanoDrop™ (Thermo Fisher Scientific, Waltham, USA) images of DNA extracted from patient samples demonstrating good quantity and quality obtained.

Bisulfite conversion and epigenetic analysis

Dr Amy Webster performed this aspect of the work. The DNA samples were thawed and underwent further quality control using Qubit Fluorometric Quantitation (ThermoFisher Scientific, Waltham, MA, USA). Suitable samples underwent bisulfite conversion using the Zymo EZ-96 DNA Methylation MagPrep (Zymo Research, Irvine, CA, USA) as per the manufacturer's protocol prior to being sent to the UCL Genomics lab for epigenetic analysis using the Infinium MethylationEPIC BeadChip (Illumina, San Diego, CA, USA). This chip provides over 850,000 of methylation sites per sample at single nucleotide resolution.

Data analysis

In order to analyse the data Dr Rifat Hamoudi developed a specific pipeline using R 3.3 for the EPIC array methylome analysis. The pipeline has the following steps:

1. Calculate the detection p-values
2. Examine mean detection p-values across all samples to identify any failed samples
3. QC Report
4. Remove poor quality samples
5. Normalisation
6. Examine higher dimensions to look at other sources of variation
7. Filtering
8. Probe-wise differential methylation analysis using t-statistics (ANOVA)
9. Clustering
10. Differential methylation analysis of regions (DMRs)
11. Pathway analysis using Gene Ontology

6.6.2 Objective 4 - Performing epigenetic analysis on salivary samples.

Results

The initial analysis of this work looked to ensure that sample quality was sufficient to yield meaningful results. As such as a methylation map was created to compare patients with OAC to those that are normal. This is demonstrated below and confirms that there was good sample quality.

For normalisation the data generated was used to test different normalisation functions.

Figure 49: Testing different normalisation functions.

Epigenetic methylation map showing density of CpG sites. The peak at zero shows unmethylated sites and at 1 shows methylated sites. Low values between confirm high sample quality

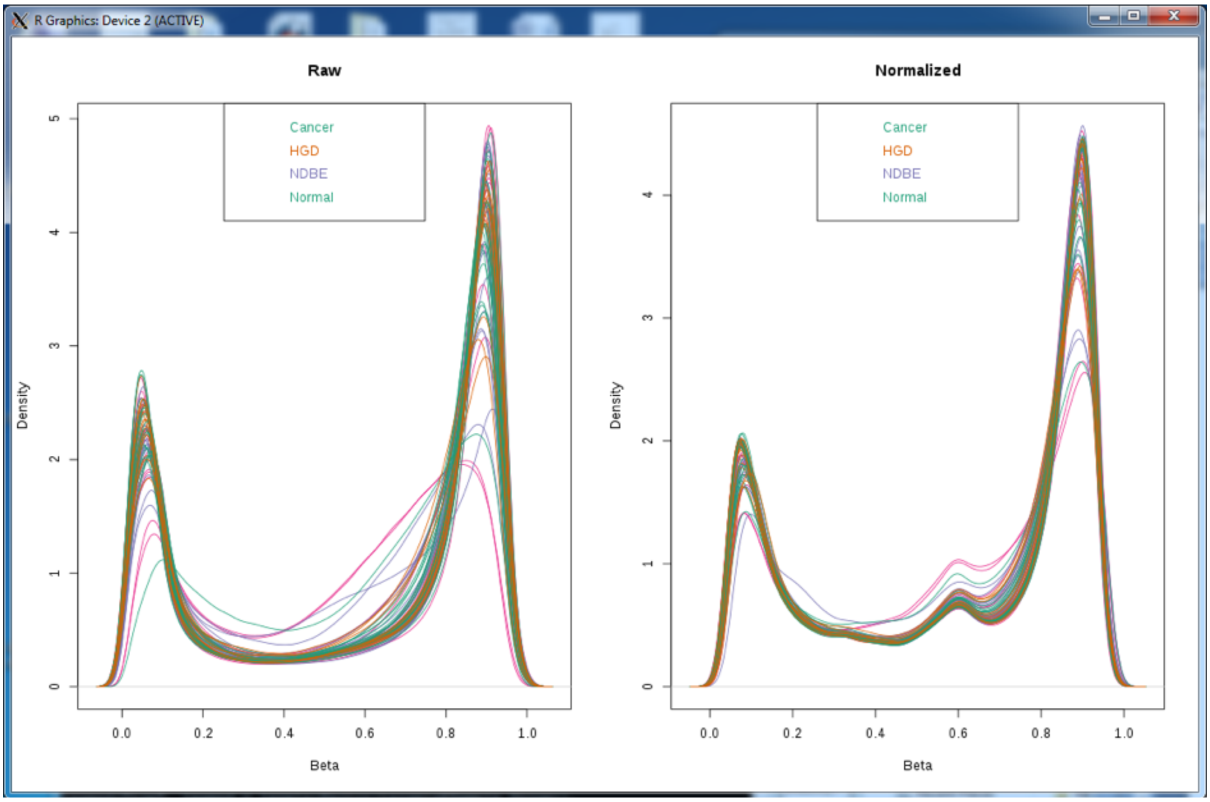


Figure 49a: BMIQ quantile normalisation

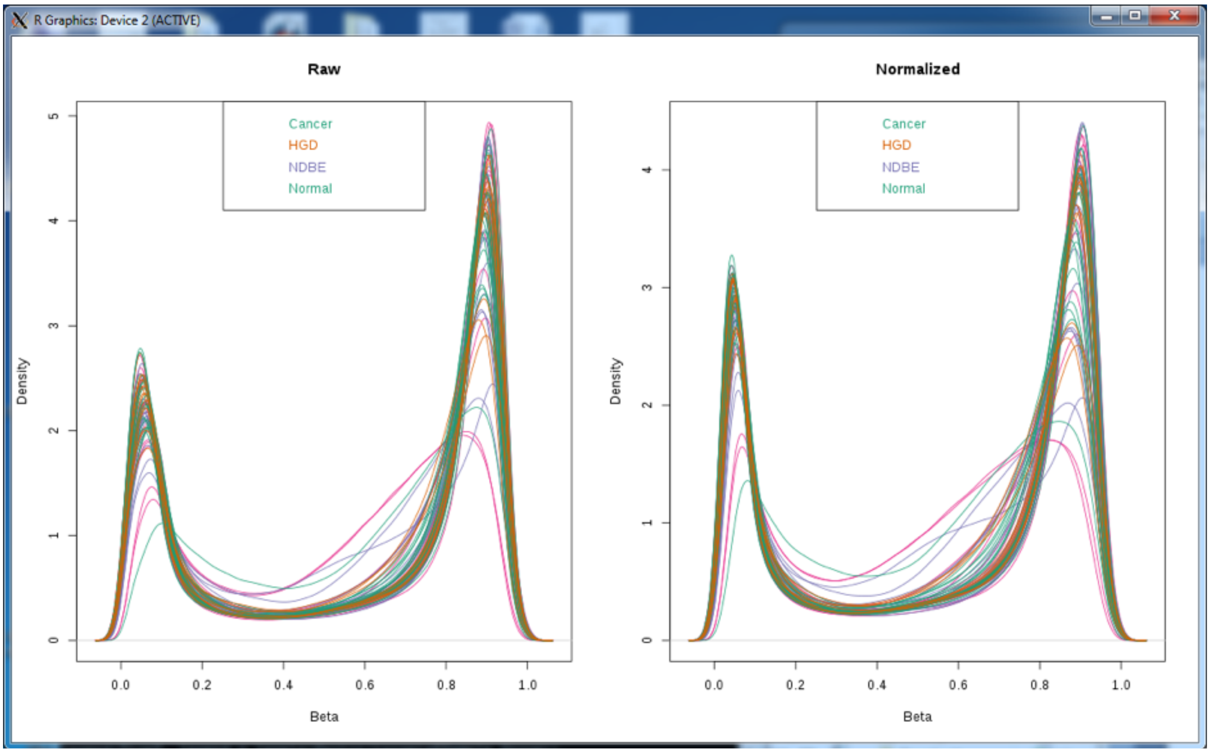


Figure 49b: SWAN normalisation

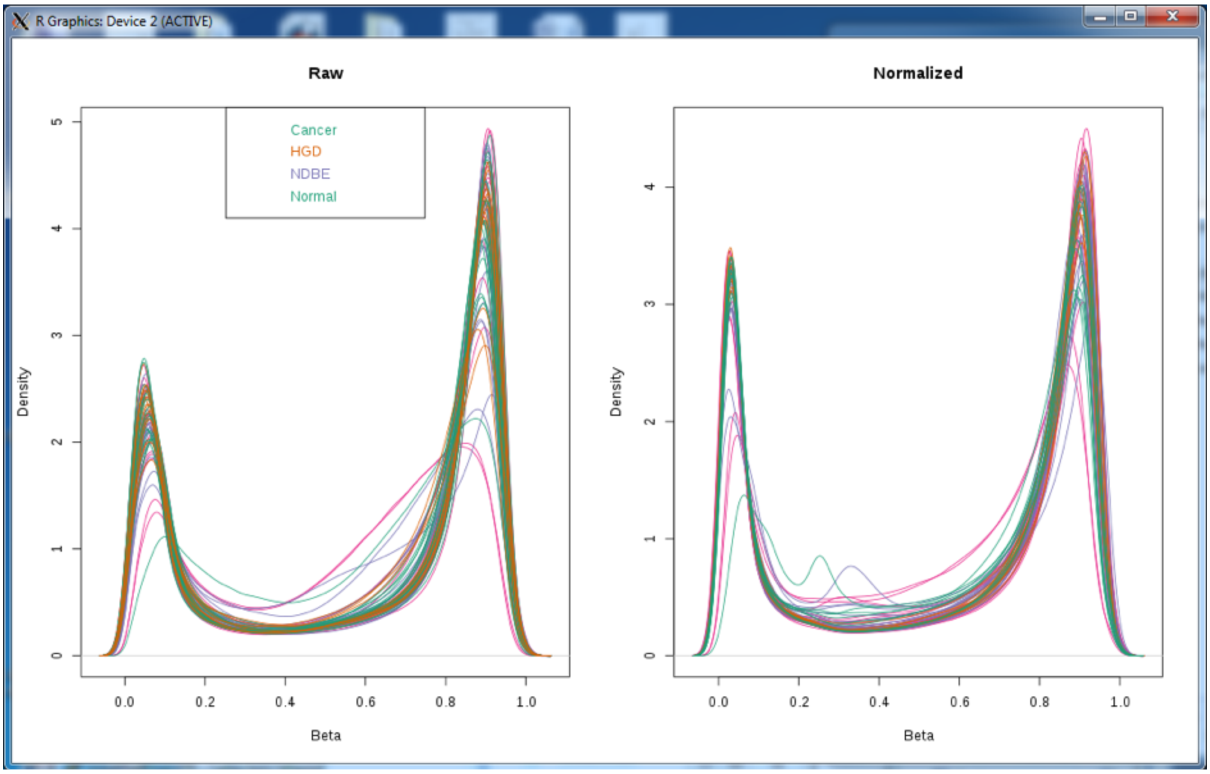


Figure 49c: FUNNorm normalisation

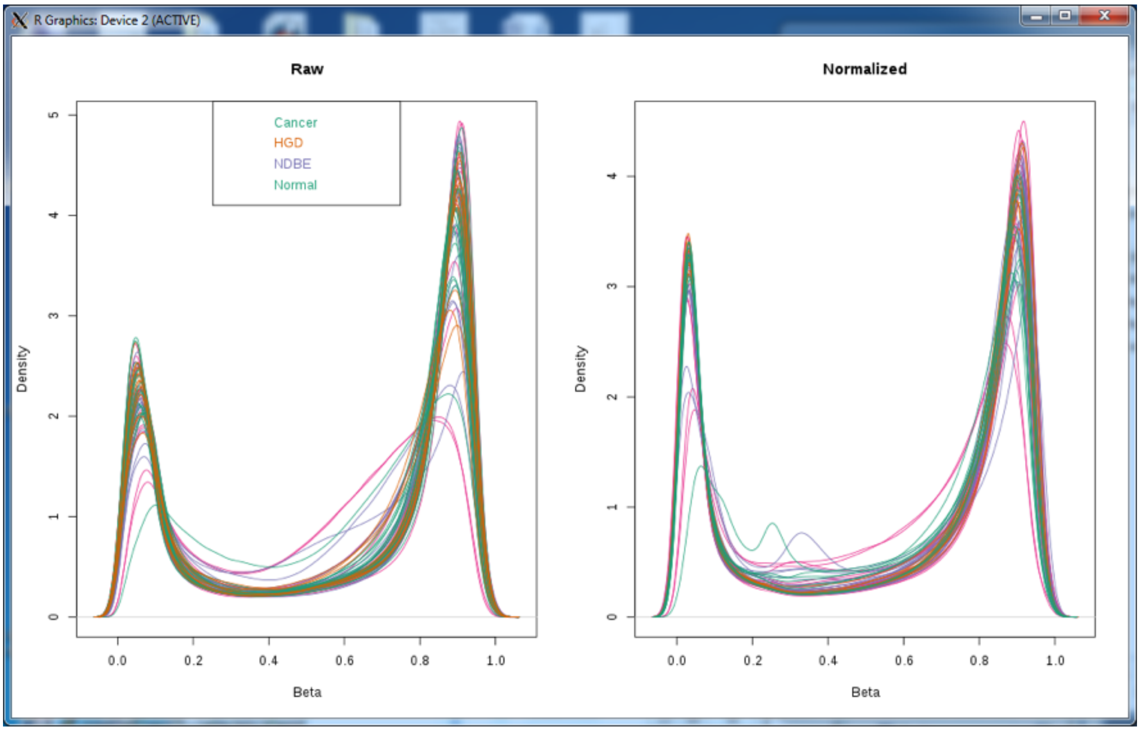


Figure 49d: Pre-process Illumina method normalisation

The above figures demonstrate that the best normalisation function is FUNNorm, therefore this was used to normalise the EPIC array data. Following filtering differential methylation was carried out on DMPs using ANOVA and compared using IEVORA (performed by Professor Teschendorff) and MIAT (performed by Professor Rosenfeld) across the different sets.

A heat map was then produced of the top 50 differentially methylated sites when comparing normal patients to those with OAC. This demonstrated some clear differences between the two groups.

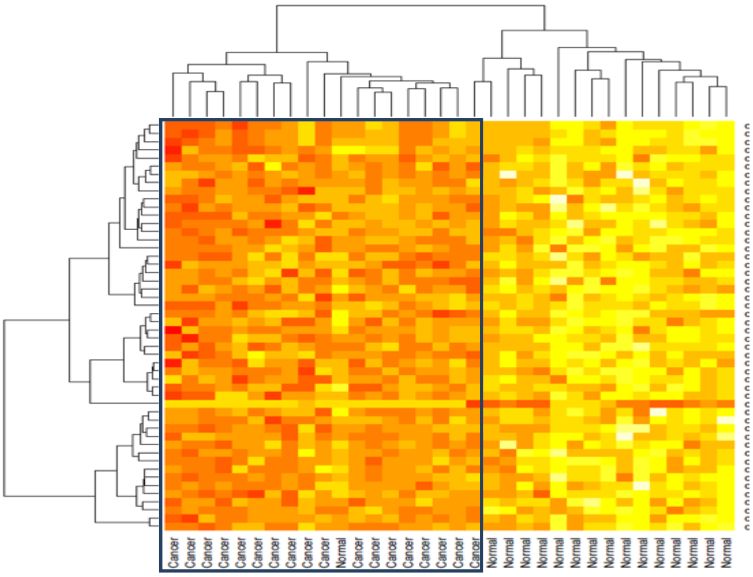


Figure 50: Heat map of top 50 methylation sites for oesophageal cancer patient samples (box), compared to normal patients. Clear differences are seen between the two groups

However, upon further analysis of these results it was noted that there were some confounding data. This was caused by cell-type composition, sex, ethnic group and even bead-chip. This is highlighted by the singular value decomposition plot below in particular in PC-2.

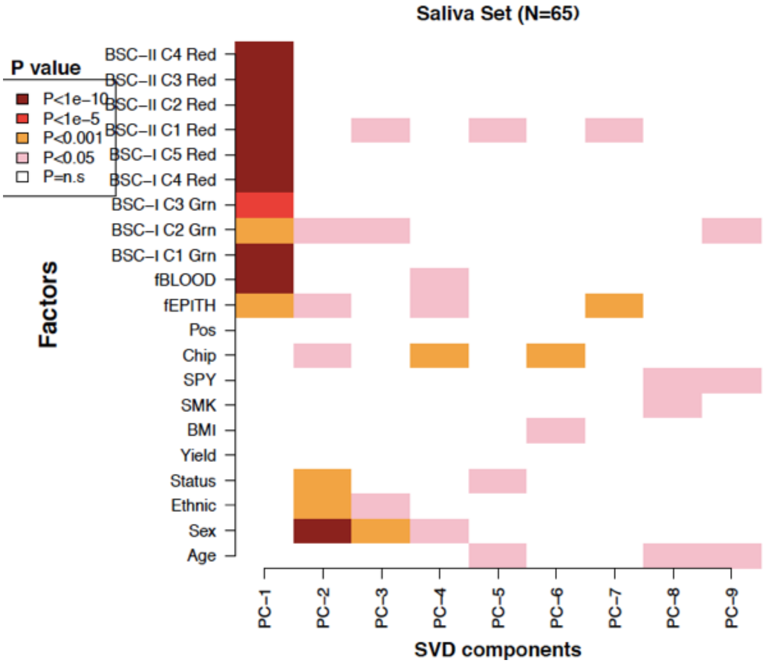


Figure 51: Depicting singular value decomposition plot for epigenetic data

The sex confounding factor reflects the significantly higher prevalence of BO and OAC in males. It was dealt with by excluding XY chromosomes which is an established approach. It was also noted that that the majority of patients in this study were white-Caucasian which reflects the known racial disparity in the prevalence of BO and OAC. However, any patients recruited who were either from black or Asian ethnic backgrounds were invariably normal. As such this became a confounding factor. It was therefore decided to exclude any non-white patients from this analysis. Following exclusion of all non-white racial groups the samples sizes were as follows; 9 normal patients, 17 NDBO, 15 dysplasia and 15 OAC. The singular value decomposition plot below demonstrates the significance of associations between the principal component analysis and singular value decomposition after restriction to white-Caucasians and removing the XY chromosomes.

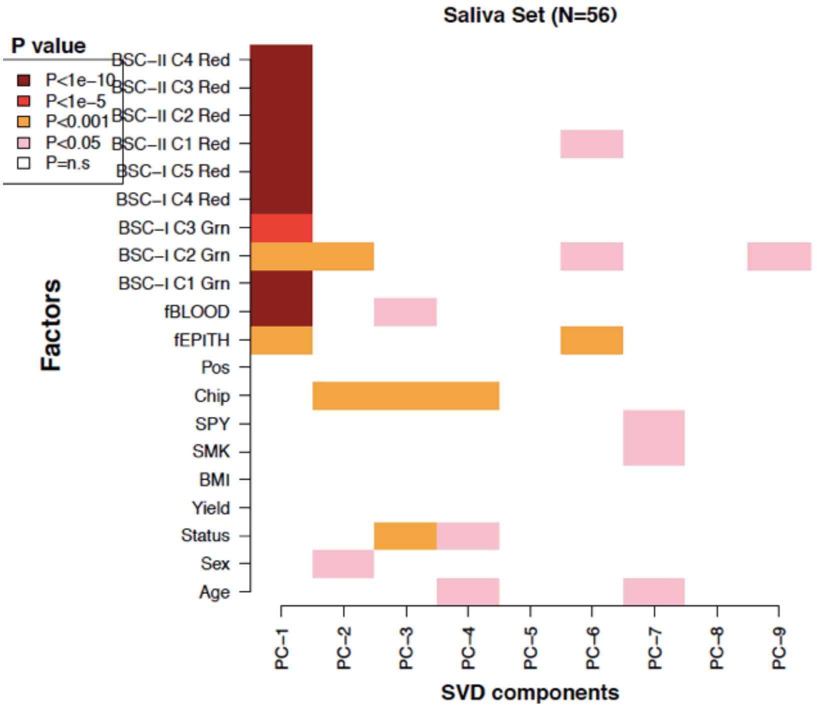


Figure 52: Depicting singular value decomposition plot for epigenetic data following removal of non-Caucasians and XY chromosomes

This plot demonstrates that the remaining top component correlates to “fBLOOD” which is the estimated fraction of immune cells in the samples. It was again decided to adjust the data for this as it removes the issues of our estimation of immune cells being inaccurate. Linear models (t-tests) were then run for comparison of each pair of diagnostic groups in order to create a histogram of *p* values. These are depicted below and demonstrate that, particularly when comparing OAC to normal and dysplasia to normal, we obtain encouraging results. If one uses a false discovery rate (FDR) of 0.2, meaning we have 80% confidence that the CpG’s with FDR’s of less than 0.2 are true positives then we note that there are approximately 1,600 differentially methylated regions (DMR) when comparing normal to OAC. This FDR threshold is an acceptable cut-off and has been used in published work including work creating a gene-expression signature predicting survival in breast cancer [298].

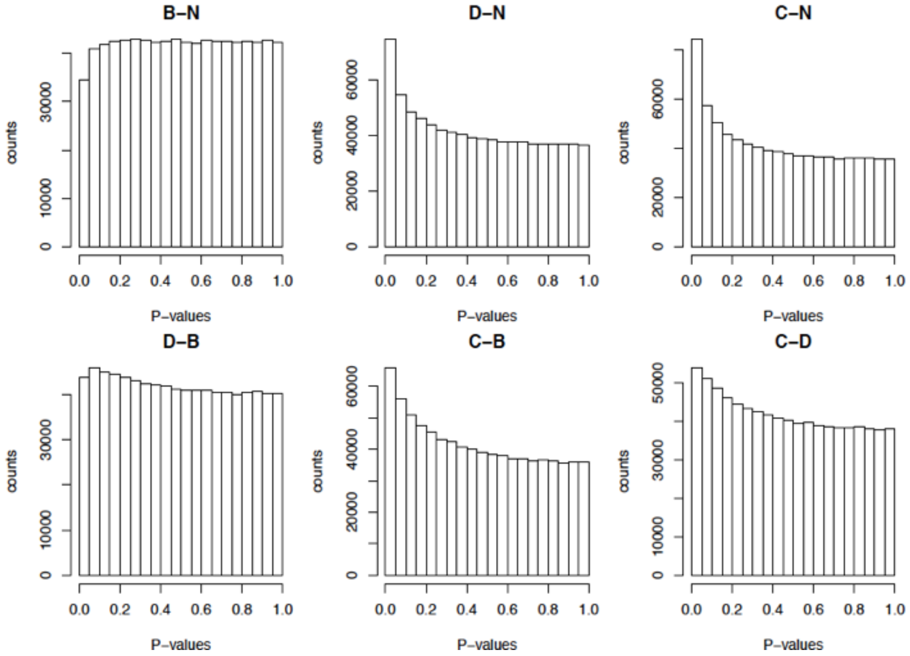


Figure 53: Depicting the histograms of P-values when comparing the various diagnostic groups.

The decision was then made to explore whether the approximately 1600 DMRs found when comparing normal patients to those with OAC demonstrated a similar pattern of differential methylation when comparing normal patients with those with dysplasia. This was performed using a scatterplot of the t-statistics of differential methylation between OAC and normal (x-axis) again as the corresponding statistics between normal and dysplasia. This highlights that the CpGs that are hypermethylated in OAC when compared to normal also tend to be hypermethylated in dysplasia when compared to normal. Likewise, this is also seen when comparing areas of hypomethylation.

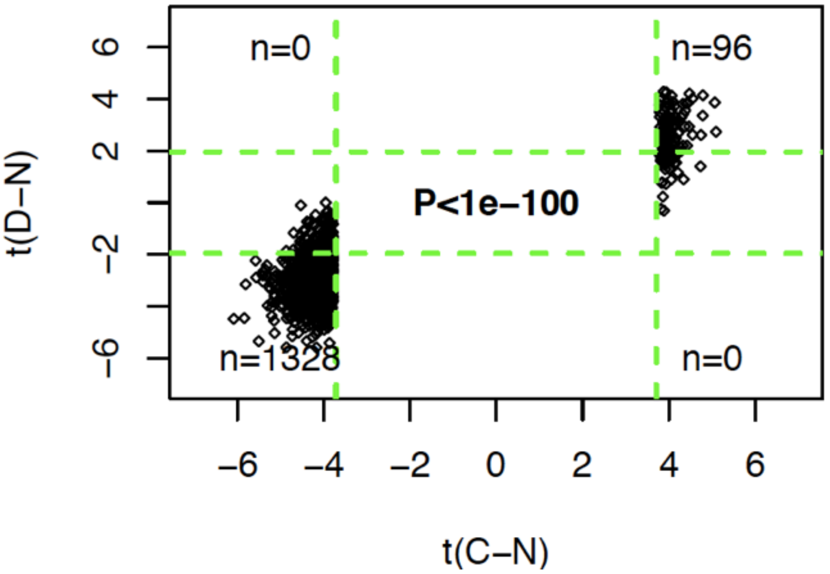


Figure 54: Scatterplot comparing the CpGs that are hyper and hypomethylated when comparing OAC to normal and dysplasia to normal.

The DNA methylation profiles for the top 25 DMRs when comparing OAC to normal patients were then created. This yielded exciting results, depicted below, in which a clear progression is seen when comparing all four diagnostic groups.

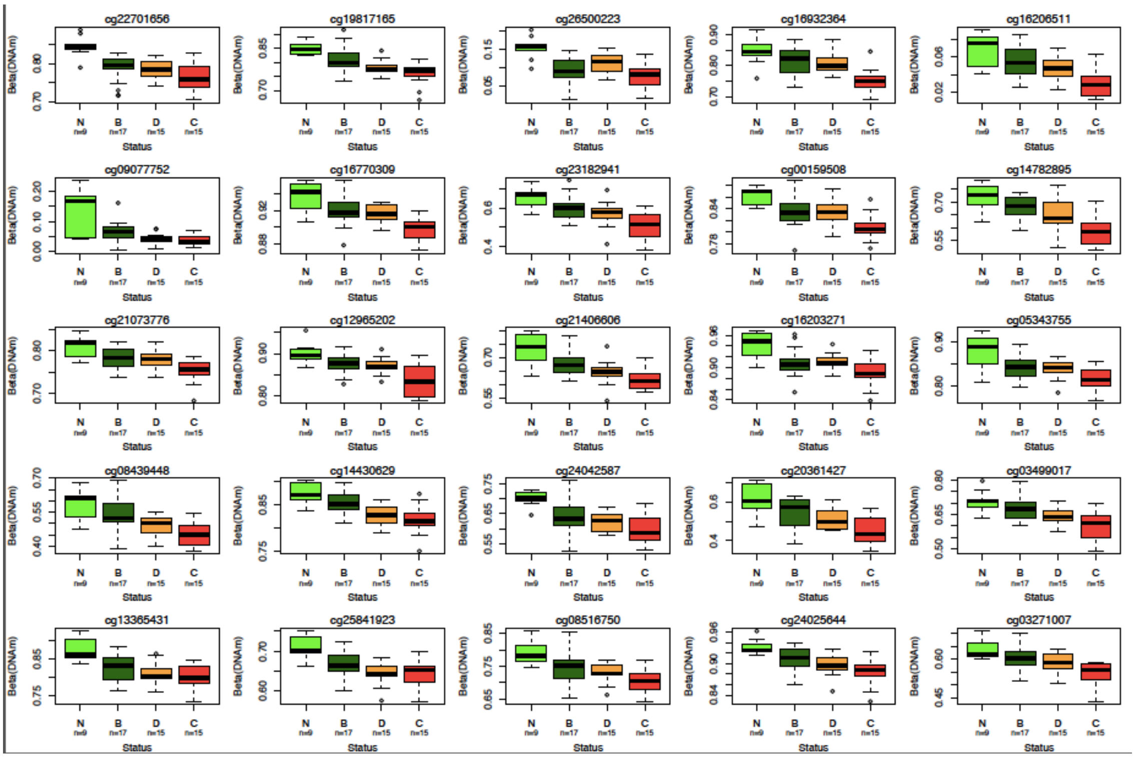


Figure 55: The top 25 DMRs for OAC v normal patients when utilised across all diagnostic groups. A clear progression is seen as patients go from normal to NDBO, dysplasia and OAC.

When utilising these top 25 DMRs as a gene signature one is able to accurately differentiate between the four diagnostic groups with 99.9% accuracy. However, clearly this sample size is too small to be able to draw significant conclusions on their potential accuracy as a diagnostic tool, but it does show great potential.

The genes in figure 55 above were mapped to DMR using the pipeline discussed above. The DMR's mapped to the following genes (description from https://www.genecards.org/):

DMR	Location	Gene	Gene description
cg19817165	N_Shore	DYRK2	Family of protein kinases whose members are presumed to be involved in cellular growth and/or development. Among its related pathways are Regulation of TP53 Activity and Gene Expression. Also involved in tyrosine kinase activity. Phosphorylates p53/TP53 at 'Ser-46', and thereby contributes to the induction of apoptosis in response to DNA damage.
cg16932364		NOS1	Belongs to the family of nitric oxide synthases, which synthesize nitric oxide from L-arginine. Nitric oxide is a reactive free radical, which acts as a biologic mediator in several processes, including neurotransmission, and antimicrobial and antitumoral activities. Diseases associated with NOS1 include Achalasia and Familial Oesophageal and Pyloric Stenosis
cg20361427	S_Shelf	MIOS	Among its related pathways are mTOR signalling pathway (KEGG) and PI3K / Akt Signalling. Diseases associated with MIOS include Hereditary Non-Polyposis Colorectal Cancer.
cg13365431		SLC4A2	The encoded protein regulates intracellular pH, biliary bicarbonate secretion, and chloride uptake. Reduced expression of this gene may be associated with primary biliary cirrhosis (PBC) in human patients, while differential expression of this gene may be associated with malignant hepatocellular carcinoma, colon and gastric cancers.
cg16206511	Island	MON2	Little known about this gene. Involved in binding and ARF guanyl-nucleotide exchange factor activity
cg12965202	S_Shore	GTF2A1	Plays an important role in transcriptional activation.
cg21406606	N_Shore	ADAT2	Related to Gene Expression and tRNA processing pathways.
cg16203271	N_Shore	EMILIN1	The encoded protein associates with elastic fibers and may play a role in the development of elastic tissues
cg14430629		RNF112	Plays an important role in neuronal differentiation, including neurogenesis and gliogenesis, during brain development. Involved in the maintenance of neural functions and protection of the nervous tissue cells from oxidative stress-induced damage.
cg24025644	N_Shore	CRYAA	Involved in autokinase activity and participation in the intracellular architecture.
g03271007		RNF145	In response to bacterial infection, negatively regulates the phagocyte oxidative burst by controlling the turnover of the NADPH oxidase complex subunits.

Table 27: Overview and description of the genes the DMR's were mapped to.

6.6.3 Objective 4 - Performing epigenetic analysis on salivary samples. Discussion

The epigenetic analysis on salivary DNA samples yielded exciting results. The strong correlation between the DMRs found when comparing those with OAC to normal and those with dysplasia to normal patients is in keeping with the literature on the epigenome in BO and OAC and demonstrates that these DMRs are likely to be true positives. Furthermore, when selecting the top 25 DMRs for the comparison between normal and OAC we see a clear progression from normal along the metaplasia-dysplasia-carcinoma sequence when all diagnostic groups are looked at. This is despite not explicitly searching for DMRs that provided this progression. This is significant and with this small sample of patients (n=56) the epigenetic signature allows for differentiation between diagnostic groups with 99.9% accuracy. However, this is obviously a small sample, with only 9 normal controls, and work on a larger cohort is clearly required.

Interestingly when the DMRs were mapped to genes it is notable that some of these are related to well-known pathways involved in the progression of BO to OAC. DYRK2 is related to TP53 activity and tyrosine kinase activity that have been discussed within this thesis and well published in relation to OAC. Additionally, MIOS is related to PI3K / AKT signalling which has also been discussed within this thesis as known to have links with OAC. Alongside this SLC4A2 and NOS1 have been found to be relevant in other GI cancers / disease. The presence of these within the findings suggest that the DMR's found in the salivary epigenome are relevant and worthy of further investigation.

A further limitation of this work is that ethnicity was clearly a cofounding factor during the analysis. It is well-established that there is a huge variation in incidence of BO and OAC depending on ethnicity with white men and women being of particular risk. As such the vast majority of patients recruited with BO and OAC were white compared to the majority of those who were normal being non-white. It is well-established DNA methylation profiles vary between ethnic groups and that these differences are even present from birth [299]. As such the fact that the majority of normal patients, being used as the controls within this study, were non-white and accordingly would have a different baseline methylation profile became problematic. Accordingly, because of the small sample size we had to exclude the non-white patients in order to analyse this data. These results are therefore only applicable to white-Caucasian patients rather than a whole, diverse population. Clearly further work with larger numbers, with sufficient representation of all ethnic groups is required. It may be that a gene signature for each ethnic group will need to be established, although ideally one for a whole population would be preferred for population screening.

Another cofounding factor that had to be dealt with was the disparity between males and females and the more advanced disease. Within the normal control group there was an almost equal split of males and females (F=9, M=11) whereas in all other groups there are significantly more males. This again reflects the known higher incidence in males but causes issues when analysing epigenetic data due to the presence of sex chromosomes. The exclusion of the sex chromosomes for epigenetic analysis is not perfect and potentially ignores the role these play in the gender disparity within the disease. However, again because of the small sample sizes and the variation of male to female ratio within our diagnostic groups this became important in order to overcome this cofounding factor. Sex differences in DNA methylation have been studied with some data reporting higher methylation found in males over females. Work has demonstrated that the active female chromosome displayed similar DNA methylation patterns to that of the male X chromosome whereas others work found there to be higher methylation levels seen in the inactive female X chromosome compared to the active X. Analysis of sex differences in DNA methylation on the autosomal chromosomes have either revealed no, few or small changes whilst in cells found in saliva, females tend to have higher DNA methylation levels on both the X chromosome as well as the autosomes [300].

The use of a panel of epigenetic markers for the detection or prediction of outcome in cancer is not new with promising data being found in areas such as lung and breast cancer[301, 302]. The majority of work using epigenetic signatures for early detection utilises blood whereas this work using saliva offers great potential from a population screening tool point of view.

Furthermore, as discussed earlier, the use of the more stable and more easily extracted DNA rather than RNA also offers great potential moving forward. It should be noted though that at present, with these results from this small sample of patients, fewer genes and thus a smaller panel of transcriptomic biomarkers have been required to differentiate patients. Alongside this the biomarkers used are those with well-known links to cancer progression and thus may provide more meaningful data.

Further work is clearly needed on a larger cohort of patients to determine the true accuracy of this epigenetic panel. Additionally, one should explore the other DMRs highlighted within this work and perform a deeper pathway analysis through machine learning in order to ascertain how these DMRs fit into the pathogenesis of OAC.

6.7 Chapter conclusions

The work within this chapter has demonstrated that saliva has great potential as a biofluid containing invaluable biomarkers for the early detection of those with or at risk of OAC.

Even though there were a limited number of genes targeted in the expression analysis it produced notable data with some genes like CDKN2A, H3FA and CDK2 providing exciting results. The further work completing the discovery RNA-Seq may highlight further genes of interest and potentially allowing for the accurate identification of individuals with or at risk of OAC using salivary RNA alone. At present, however, the data available exploring these 40 genes look promising but needs validation on the 80 cohort of patients that was used to carry out the proof of concept work. A patent was granted on this (https://patents.google.com/patent/WO2017137427A1/en) so further work may result in commercially approved tool for non-invasive diagnosis of onset of OAC using saliva. Having said that, an accurate tool was produced with the addition of limited patient questionnaire data. It is possible that the outcome of the RNA-seq validation will provide a subset of genes resulting in the ability to create a transcriptomic based diagnostic tool rather than having to use the additional questionnaire data. The advantage of this would be that one could potentially create a single colorimetric assay that could be used as a predictive test for BO or OAC. This is something that is not possible if one utilises an epigenetic diagnostic tool.

However, it should be stated that the epigenetic data obtained within this work demonstrated great promise in the accurate identification of those with or at risk of OAC. Using just the top 25 DMRs when comparing those with OAC and those that were normal we were then able to demonstrate a progression throughout the metaplasia-dysplasia-carcinoma sequence and the ability to identify those with or at risk of OAC with 99.9% accuracy. It should be noted that 25 DMRs equates to a large number of individual data points, most not linked to gene regions, and as such this may affect its feasibility as a predictive test. Having said that, the epigenetic Thiswork demonstrates the clear role that environmental factors play in the pathogenesis of BO and OAC, however, given the small number of patients within this preliminary data one could see that the significance of these DMRs alter with changes in factors such as medication, smoking and GORD. Clearly further work is needed to determine the role of the epigenetic markers identified in the progression through the Barrett's metaplasia-dysplasia-carcinoma sequence.

Chapter 7

Conclusions and future work

Chapter 7 – Conclusions and future work

There is little dispute that the early detection of individuals with or at high-risk of developing cancer is a key goal in improving cancer outcome. This becomes increasingly challenging, however, in cancers such as OAC in which the symptoms and signs of the disease often present in the later stages when the cancer is incurable [9]. As physicians we have long been taught to and have relied upon our clinical acumen to detect at-risk patients, however, in cancers such as OAC the data suggests that this is insufficient. The almost static 5-year survival rates for OAC over the past 40 years despite the advances in areas such as imaging, endoscopy, surgery and oncology demonstrate that the existing strategies for the early detection of OAC are ineffective [5]. It is clear that new strategies have to be developed and that these strategies must be acceptable from a patient perspective as well not placing further strain on limited healthcare resource.

With this in mind the work within this thesis aimed to address those issues and explore the possibility of a new strategy for the detection of BO and OAC utilising AI analysis of patient data. Initial work explored the use of questionnaire data alone which, although would be low-cost, is susceptible to manipulation and inaccurate data entry alongside not proving specific enough to ensure vast numbers of individuals were not subjected to invasive testing unnecessarily. As such additional, more robust data was required. The use of biomarkers within medicine is now widely reported and data is accumulating at a rapid pace. With patient acceptability and cost as important factors in the uptake of any diagnostic test, once our early data demonstrated potential utilising saliva for biomarker detection, our work focussed on this biofluid.

The work within this thesis has presented exciting data on the use of saliva as a diagnostic tool. I have demonstrated that adequate quantity and quality RNA and DNA can be extracted from saliva for further analysis. Furthermore, both the transcriptomic and epigenetic studies have yielded encouraging results in which we have been able to differentiate individuals with BO and OAC from normal individuals. Additionally, and perhaps importantly, this work has also demonstrated promise in identifying individuals who are at high-risk of developing OAC with HGD and can potentially differentiate them and those with OAC from those with NDBO. This may be of importance as it would potentially allow us to survey those with BO through non-invasive salivary sampling rather than subject them to regular endoscopic examination. This provides both patient and healthcare benefits.

Finally, the work outlined in this thesis can theoretically be applied to any cancer or disease process and thus one could hypothesise the future potential for individuals to be screened and surveyed for a large number of diseases simply by providing a sample of their saliva. The consequence of this could alter the manner in which we deliver healthcare through strengthening prevention and monitoring strategies.

Clearly the data presented in this thesis is preliminary and more work is required. This thesis provides opportunity for numerous further studies gathering more data on what has been demonstrated so far and also exploring other potential uses of saliva as a liquid biopsy. In order to take this thesis forward the further studies would focus on four areas; increasing the patient cohort, biomarker discovery, understanding BO and OAC evolution and disease monitoring.

Increasing the patient cohort

This thesis has developed two diagnostic tools that require further validation with an increased patient cohort. The predictive tool using transcriptomic and questionnaire data (patent number: WO2017137427) was developed on 80 patients (20 in each diagnostic category) and is now being tested on a larger number of patients as part of the SPIT study.

The epigenetic data gathered within this work identified 25 DMRs that were able to accurately identify those with or at-risk of OAC with high accuracy. However, this work also needs further validation with a larger cohort of patients including variation in ethnicity. Our work was limited due to the small number of females and non-Caucasians represented and as such these had to be excluded from the analysis in addition to heterogenous environmental factor affecting those patients.

One would be able to follow the existing methods discussed in this thesis on the larger cohort of patients being gathered in the SPIT study in order to explore whether these predictive tools were an effective diagnostic tool.

Biomarker discovery

The transcriptomic work within this thesis was based upon only 13 genes that were targeted due to known links with BO and OAC following a literature search. The results of the RNA-seq validation performed within this thesis highlight more genes that require exploration. However, we need to perform RNA-seq on a sufficient number of patients in order to effectively highlight all genes that require testing as they may be of value in enhancing the salivary diagnostic tool. Once these genes have been identified one would create primers and perform qRT-PCR on patient samples as described within this thesis. The data generated could be analysed alongside the existing data to determine whether these additional genes enhance the tool. At present the transcriptomic work relies upon some questionnaire data in order to provide sufficient accuracy and ultimately one would aim to identify a panel of genes that would be able to identify the at-risk individual through salivary sampling alone. This would require the development of logistic mathematical model to integrate the biomarkers validated into a transcriptomic signature that can show predictive power for the Barrett’s sequence.

Due to the limitations of the epigenetic data gathered, in particular due to the lack of diversity in the patients included, we will need to perform EPIC array methylome analysis on the more diverse population obtained through the SPIT study. This is likely to generate further DMRs of importance that will require testing on the larger cohort of patients.

Understanding BO and OAC evolution

In regard to the transcriptomic findings within this work and the future RNA-seq data one would like to be able to perform functional validation of the biomarkers identified using cell lines. This would hopefully enable us to identify the role these genes play within the progression of BO to OAC and shed light on the disease pathogenesis and mechanistic insights into the development of OAC. This may also provide avenues for future work exploring potential therapeutic targets.

The DMRs found within this study need to be mapped to genes in order to identify their potential role within the progression of BO and again identify potential targets for therapeutic intervention. Further DMRs will likely be identified when EPIC array methylome analysis is performed on the more diverse population gathered in the SPIT study which will again need to be mapped to genes and provides insights into progression. Included within the SPIT study is the collection of questionnaire data which will allow us to explore how the patient symptoms, risk factor and demographic data impacts on these DMRs. This would again potentially provide further insights into pathogenesis and progression of BO and OAC.

Disease monitoring

It would be beneficial to study the concordance of transcriptomic and epigenetic biomarkers between blood, saliva and biopsy samples in OAC and BO in order to provide insights into their future use within clinical practice in areas such as screening, disease monitoring, treatment response, and prognosis. Ideally one would hope to be able to obtain this data in a longitudinal fashion throughout a patient’s lifetime in order to gain these valuable insights into determining how these biomarkers alter as the disease progresses, undergoes treatment and potentially recurs.

Whilst the work within this thesis provides multiple avenues for further research I believe that the data presented demonstrates that the transcriptomic and / or epigenetic information found in salivary samples could be utilised as a population screening tool to identify individuals at risk. It is possible that this data will need to be combined with demographic and / or symptom data, such as smoking history or measurements of obesity, in order to improve its accuracy. I feel it is likely that a predictive tool composed of a combination of biomarkers is likely to yield highest accuracy. Further work in identifying biomarkers, gathering data on larger cohorts of patients and testing the predictive tool developed is vital for success. As a population screening tool, I believe that the provision of saliva will likely be acceptable to individuals to perform and thus key to its applicability within healthcare will be finding a low-cost, high throughput means to perform the test.

This thesis demonstrates the great potential of saliva not only as a population screening tool but it may also have a role in many other aspects of disease management that are vital in altering the grave prognosis of OAC.

References

1. Weaver, J.M., et al., *Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis.* Nat Genet, 2014. **46**(8): p. 837-43.
2. Institute, N.H.G.R. *The Cost of Sequencing a Human Genome*. Available from: https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/.
3. DeGregori, J., *Challenging the axiom: does the occurrence of oncogenic mutations truly limit cancer development with age?* Oncogene, 2013. **32**(15): p. 1869-75.
4. BBC. *"Quiet epidemic" of male cancer in the UK*. 2013; Available from: http://www.bbc.co.uk/news/health-22940088.
5. UK, C.R. *Cancer Statistics*. Available from: http://www.cancerresearchuk.org/cancer-info/cancerstats.
6. Thrift, A.P., *The epidemic of oesophageal carcinoma: Where are we now?* Cancer Epidemiol, 2016. **41**: p. 88-95.
7. Pohl, H., B. Sirovich, and H.G. Welch, *Esophageal adenocarcinoma incidence: are we reaching the peak?* Cancer Epidemiol Biomarkers Prev, 2010. **19**(6): p. 1468-70.
8. Thrift, A.P. and D.C. Whiteman, *The incidence of esophageal adenocarcinoma continues to rise: analysis of period and birth cohort effects on recent trends.* Ann Oncol, 2012. **23**(12): p. 3155-62.
9. Schlansky, B., et al., *A survey of oesophageal cancer: pathology, stage and clinical presentation.* Aliment Pharmacol Ther, 2006. **23**(5): p. 587-93.
10. Health, I., *Saving lives, averting costs*. 2014, Cancer Research UK.
11. Barrett, N.R., *The lower esophagus lined by columnar epithelium.* Surgery, 1957. **41**(6): p. 881-94.
12. Naef, A.P., M. Savary, and L. Ozzello, *Columnar-lined lower esophagus: an acquired lesion with malignant predisposition. Report on 140 cases of Barrett's esophagus with 12 adenocarcinomas.* J Thorac Cardiovasc Surg, 1975. **70**(5): p. 826-35.
13. Borrie, J. and L. Goldwater, *Columnar cell-lined esophagus: assessment of etiology and treatment. A 22 year experience.* J Thorac Cardiovasc Surg, 1976. **71**(6): p. 825-34.
14. Cameron, A.J., et al., *Adenocarcinoma of the esophagogastric junction and Barrett's esophagus.* Gastroenterology, 1995. **109**(5): p. 1541-6.
15. Hamilton, S.R., R.R. Smith, and J.L. Cameron, *Prevalence and characteristics of Barrett esophagus in patients with adenocarcinoma of the esophagus or esophagogastric junction.* Hum Pathol, 1988. **19**(8): p. 942-8.
16. Fitzgerald, R.C., et al., *British Society of Gastroenterology guidelines on the diagnosis and management of Barrett's oesophagus.* Gut, 2014. **63**(1): p. 7-42.
17. Sharma, P. and E.I. Sidorenko, *Are screening and surveillance for Barrett's oesophagus really worthwhile?* Gut, 2005. **54 Suppl 1**: p. i27-32.
18. Reid, B.J., et al., *Barrett's oesophagus and oesophageal adenocarcinoma: time for a new synthesis.* Nat Rev Cancer, 2010. **10**(2): p. 87-101.
19. Reid, B.J., et al., *Flow-cytometric and histological progression to malignancy in Barrett's esophagus: prospective endoscopic surveillance of a cohort.* Gastroenterology, 1992. **102**(4 Pt 1): p. 1212-9.

20. Shen, C.N., Z.D. Burke, and D. Tosh, *Transdifferentiation, metaplasia and tissue regeneration.* Organogenesis, 2004. **1**(2): p. 36-44.

21. Lavery, D.L., et al., *Evolution of oesophageal adenocarcinoma from metaplastic columnar epithelium without goblet cells in Barrett's oesophagus.* Gut, 2016. **65**(6): p. 907-13.

22. Spechler, S.J., et al., *American Gastroenterological Association technical review on the management of Barrett's esophagus.* Gastroenterology, 2011. **140**(3): p. e18-52; quiz e13.

23. Takubo, K., et al., *Cardiac rather than intestinal-type background in endoscopic resection specimens of minute Barrett adenocarcinoma.* Hum Pathol, 2009. **40**(1): p. 65-74.

24. Kelty, C.J., et al., *Barrett's oesophagus: intestinal metaplasia is not essential for cancer risk.* Scand J Gastroenterol, 2007. **42**(11): p. 1271-4.

25. de Jonge, P.J., et al., *Barrett's oesophagus: epidemiology, cancer risk and implications for management.* Gut, 2014. **63**(1): p. 191-202.

26. Schlemper, R.J., et al., *The Vienna classification of gastrointestinal epithelial neoplasia.* Gut, 2000. **47**(2): p. 251-5.

27. Kerkhof, M., et al., *Grading of dysplasia in Barrett's oesophagus: substantial interobserver variation between general and gastrointestinal pathologists.* Histopathology, 2007. **50**(7): p. 920-7.

28. Odze, R.D., *Diagnosis and grading of dysplasia in Barrett's oesophagus.* J Clin Pathol, 2006. **59**(10): p. 1029-38.

29. Downs-Kelly, E., et al., *Poor interobserver agreement in the distinction of high-grade dysplasia and adenocarcinoma in pretreatment Barrett's esophagus biopsies.* Am J Gastroenterol, 2008. **103**(9): p. 2333-40; quiz 2341.

30. Curvers, W.L., et al., *Low-grade dysplasia in Barrett's esophagus: overdiagnosed and underestimated.* Am J Gastroenterol, 2010. **105**(7): p. 1523-30.

31. Sharma, P., et al., *Dysplasia and cancer in a large multicenter cohort of patients with Barrett's esophagus.* Clin Gastroenterol Hepatol, 2006. **4**(5): p. 566-72.

32. Schnell, T.G., et al., *Long-term nonsurgical management of Barrett's esophagus with high-grade dysplasia.* Gastroenterology, 2001. **120**(7): p. 1607-19.

33. Reid, B.J., et al., *Predictors of progression to cancer in Barrett's esophagus: baseline histology and flow cytometry identify low- and high-risk patient subsets.* Am J Gastroenterol, 2000. **95**(7): p. 1669-76.

34. Buttar, N.S., et al., *Extent of high-grade dysplasia in Barrett's esophagus correlates with risk of adenocarcinoma.* Gastroenterology, 2001. **120**(7): p. 1630-9.

35. Heitmiller, R.F., M. Redmond, and S.R. Hamilton, *Barrett's esophagus with high-grade dysplasia. An indication for prophylactic esophagectomy.* Ann Surg, 1996. **224**(1): p. 66-71.

36. Fernando, H.C., et al., *The Society of Thoracic Surgeons practice guideline series: guidelines for the management of Barrett's esophagus with high-grade dysplasia.* Ann Thorac Surg, 2009. **87**(6): p. 1993-2002.

37. in *Barrett's Oesophagus: Ablative Therapy for the Treatment of Barrett's Oesophagus*. 2010: London.

38. Haidry, R.J., et al., *Improvement over time in outcomes for patients undergoing endoscopic therapy for Barrett's oesophagus-related neoplasia: 6-*

year experience from the first 500 patients treated in the UK patient registry. Gut, 2015. **64**(8): p. 1192-9.

39. Bartels, H., H.J. Stein, and J.R. Siewert, *Preoperative risk analysis and postoperative mortality of oesophagectomy for resectable oesophageal cancer.* Br J Surg, 1998. **85**(6): p. 840-4.
40. Manner, H., et al., *Efficacy, safety, and long-term results of endoscopic treatment for early stage adenocarcinoma of the esophagus with low-risk sm1 invasion.* Clin Gastroenterol Hepatol, 2013. **11**(6): p. 630-5; quiz e45.
41. Fotis, D., et al., *Submucosal invasion and risk of lymph node invasion in early Barrett's cancer: potential impact of different classification systems on patient management.* United European Gastroenterol J, 2015. **3**(6): p. 505-13.
42. Manner, H., et al., *The frequency of lymph node metastasis in early-stage adenocarcinoma of the esophagus with incipient submucosal invasion (pT1b sm1) depending on histological risk patterns.* Surg Endosc, 2015. **29**(7): p. 1888-96.
43. Manner, H., et al., *Early-stage adenocarcinoma of the esophagus with mid to deep submucosal invasion (pT1b sm2-3): the frequency of lymph-node metastasis depends on macroscopic and histological risk patterns.* Dis Esophagus, 2017. **30**(3): p. 1-11.
44. Graham, D., et al., *Risk of lymph node metastases in patients with T1b oesophageal adenocarcinoma: A retrospective single centre experience.* World J Gastroenterol, 2018. **24**(41): p. 4698-4707.
45. Dunkin, B.J., et al., *Thin-layer ablation of human esophageal epithelium using a bipolar radiofrequency balloon device.* Surg Endosc, 2006. **20**(1): p. 125-30.
46. Bulsiewicz, W.J., et al., *Safety and efficacy of endoscopic mucosal therapy with radiofrequency ablation for patients with neoplastic Barrett's esophagus.* Clin Gastroenterol Hepatol, 2013. **11**(6): p. 636-42.
47. Phoa, K.N., et al., *Remission of Barrett's esophagus with early neoplasia 5 years after radiofrequency ablation with endoscopic resection: a Netherlands cohort study.* Gastroenterology, 2013. **145**(1): p. 96-104.
48. Haidry, R.J., et al., *Radiofrequency ablation and endoscopic mucosal resection for dysplastic barrett's esophagus and early esophageal adenocarcinoma: outcomes of the UK National Halo RFA Registry.* Gastroenterology, 2013. **145**(1): p. 87-95.
49. Shaheen, N.J., et al., *Durability of radiofrequency ablation in Barrett's esophagus with dysplasia.* Gastroenterology, 2011. **141**(2): p. 460-8.
50. Orman, E.S., et al., *Intestinal metaplasia recurs infrequently in patients successfully treated for Barrett's esophagus with radiofrequency ablation.* Am J Gastroenterol, 2013. **108**(2): p. 187-95; quiz 196.
51. Committee, A.S.o.P., et al., *The role of endoscopy in Barrett's esophagus and other premalignant conditions of the esophagus.* Gastrointest Endosc, 2012. **76**(6): p. 1087-94.
52. Committee, A.T., et al., *Endoscopic mucosal resection.* Gastrointest Endosc, 2015. **82**(2): p. 215-26.
53. Filby, A., et al., *Cost-effectiveness analysis of endoscopic eradication therapy for treatment of high-grade dysplasia in Barrett's esophagus.* J Comp Eff Res, 2017.

54. Pollit, V., et al., *A cost-effectiveness analysis of endoscopic eradication therapy for management of dysplasia arising in patients with Barrett's oesophagus in the United Kingdom.* Curr Med Res Opin, 2019. **35**(5): p. 805-815.
55. Gerson, L.B., K. Shetler, and G. Triadafilopoulos, *Prevalence of Barrett's esophagus in asymptomatic individuals.* Gastroenterology, 2002. **123**(2): p. 461-7.
56. Chen, Q., H. Zhuang, and Y. Liu, *The association between obesity factor and esophageal caner.* J Gastrointest Oncol, 2012. **3**(3): p. 226-31.
57. Dorak, M.T. and E. Karpuzoglu, *Gender differences in cancer susceptibility: an inadequately addressed issue.* Front Genet, 2012. **3**: p. 268.
58. Lindblad, M., et al., *Estrogen and risk of gastric cancer: a protective effect in a nationwide cohort study of patients with prostate cancer in Sweden.* Cancer Epidemiol Biomarkers Prev, 2004. **13**(12): p. 2203-7.
59. Cook, M.B., et al., *Sex disparities in cancer mortality and survival.* Cancer Epidemiol Biomarkers Prev, 2011. **20**(8): p. 1629-37.
60. Xu, Z. and J.A. Taylor, *Genome-wide age-related DNA methylation changes in blood and other tissues relate to histone modification, expression and cancer.* Carcinogenesis, 2014. **35**(2): p. 356-64.
61. Zhang, H.Z., G.F. Jin, and H.B. Shen, *Epidemiologic differences in esophageal cancer between Asian and Western populations.* Chin J Cancer, 2012. **31**(6): p. 281-6.
62. Stein, H.J. and J.R. Siewert, *Barrett's esophagus: pathogenesis, epidemiology, functional abnormalities, malignant degeneration, and surgical management.* Dysphagia, 1993. **8**(3): p. 276-88.
63. *Public Health England*. 2013; Available from: http://www.noo.org.uk/NOO_about_obesity/trends.
64. Freeman, H.J., *Risk of gastrointestinal malignancies and mechanisms of cancer development with obesity and its treatment.* Best Pract Res Clin Gastroenterol, 2004. **18**(6): p. 1167-75.
65. Lumeng, C.N. and A.R. Saltiel, *Inflammatory links between obesity and metabolic disease.* J Clin Invest, 2011. **121**(6): p. 2111-7.
66. Donohoe, C.L., et al., *Obesity and gastrointestinal cancer.* Br J Surg, 2010. **97**(5): p. 628-42.
67. Ryan, A.M., et al., *Barrett esophagus: prevalence of central adiposity, metabolic syndrome, and a proinflammatory state.* Ann Surg, 2008. **247**(6): p. 909-15.
68. Barak, N., et al., *Gastro-oesophageal reflux disease in obesity: pathophysiological and therapeutic considerations.* Obes Rev, 2002. **3**(1): p. 9-15.
69. El-Serag, H., *The association between obesity and GERD: a review of the epidemiological evidence.* Dig Dis Sci, 2008. **53**(9): p. 2307-12.
70. Fantuzzi, G., *Adipose tissue, adipokines, and inflammation.* J Allergy Clin Immunol, 2005. **115**(5): p. 911-9; quiz 920.
71. Kershaw, E.E. and J.S. Flier, *Adipose tissue as an endocrine organ.* J Clin Endocrinol Metab, 2004. **89**(6): p. 2548-56.
72. Khan, T., et al., *Metabolic dysregulation and adipose tissue fibrosis: role of collagen VI.* Mol Cell Biol, 2009. **29**(6): p. 1575-91.
73. Wang, Z. and T. Nakayama, *Inflammation, a link between obesity and cardiovascular disease.* Mediators Inflamm, 2010. **2010**: p. 535918.
74. Mayi, T.H., et al., *Human adipose tissue macrophages display activation of cancer-related pathways.* J Biol Chem, 2012. **287**(26): p. 21904-13.

75. Cook, M.B., et al., *Cigarette smoking increases risk of Barrett's esophagus: an analysis of the Barrett's and Esophageal Adenocarcinoma Consortium.* Gastroenterology, 2012. **142**(4): p. 744-53.
76. Nomura, A.M., et al., *The association of cigarette smoking with gastric cancer: the multiethnic cohort study.* Cancer Causes Control, 2012. **23**(1): p. 51-8.
77. Pfeifer, G.P., et al., *Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers.* Oncogene, 2002. **21**(48): p. 7435-51.
78. Findlay, J.M., M.R. Middleton, and I. Tomlinson, *Genetic susceptibility to Barrett's oesophagus: Lessons from early studies.* United European Gastroenterol J, 2016. **4**(4): p. 485-92.
79. Potter, S., et al., *Referral patterns, cancer diagnoses, and waiting times after introduction of two week wait rule for breast cancer: prospective cohort study.* BMJ, 2007. **335**(7614): p. 288.
80. Hanna, S.J., A. Muneer, and K.H. Khalil, *The 2-week wait for suspected cancer: time for a rethink?* Int J Clin Pract, 2005. **59**(11): p. 1334-9.
81. Fentiman, I.S., *Two week wait for suspected cancer: milestone or millstone?* Int J Clin Pract, 2005. **59**(11): p. 1251-2.
82. Redaniel, M.T., et al., *Rapid diagnostic pathways for suspected colorectal cancer: views of primary and secondary care clinicians on challenges and their potential solutions.* BMJ Open, 2015. **5**(10): p. e008577.
83. GSZ Tun, M.S., S Anwar, A Masri, A Shirazi-Nejad, N Rezwan, D Bullas, R Atkinson, K Kapur, A Soliman, V Sathyanarayana, E Said, *Taking NICE guidelines further: Straight to test for dysphagia.* Gut, 2016. **65**(1).
84. Aslam, M.I., et al., *The "two-week wait" referral pathway is not associated with improved survival for patients with colorectal cancer.* Int J Surg, 2017. **43**: p. 181-185.
85. Sharpe, D., et al., *The "two-week wait" referral pathway allows prompt treatment but does not improve outcome for patients with oesophago-gastric cancer.* Eur J Surg Oncol, 2010. **36**(10): p. 977-81.
86. England, P.H., *Trends in Cancer Waiting Times metrics, England, 2009/10 to 2014/15.* National Cancer Intelligence Network Data Briefing, 2016(PHE publications gateway number: 2015711).
87. Keith Siau, A.C.Y., Samantha Hingley, James Rees, Nigel John Trudgill, Andrew M Veitch, Neil C Fisher, *The 2015 upper gastrointestinal "Be Clear on Cancer" campaign: its impact on gastroenterology services and malignant and premalignant diagnoses.* Frontline Gastroenterology, 2017.
88. England, P.H., *National Cancer Intelligence Network*
Be Clear on Cancer: Oesophago-gastric cancer awareness regional pilot campaign: Interim evaluation report. 2015, Publich Health England.
89. Brown, J. and P. Sharma, *From Prague to Seattle: Improved Endoscopic Technique and Reporting Improves Outcomes in Patients with Barrett's Esophagus.* Dig Dis Sci, 2016. **61**(1): p. 4-5.
90. Corley, D.A., et al., *Surveillance and survival in Barrett's adenocarcinomas: a population-based study.* Gastroenterology, 2002. **122**(3): p. 633-40.
91. Corley, D.A., et al., *Impact of endoscopic surveillance on mortality from Barrett's esophagus-associated esophageal adenocarcinomas.* Gastroenterology, 2013. **145**(2): p. 312-9 e1.
92. Vaughan, T.L. and R.C. Fitzgerald, *Precision prevention of oesophageal adenocarcinoma.* Nat Rev Gastroenterol Hepatol, 2015. **12**(4): p. 243-8.

93. Tomizawa, Y. and K.K. Wang, *Screening, surveillance, and prevention for esophageal cancer.* Gastroenterol Clin North Am, 2009. **38**(1): p. 59-73, viii.
94. Old, O., et al., *Barrett's Oesophagus Surveillance versus endoscopy at need Study (BOSS): protocol and analysis plan for a multicentre randomized controlled trial.* J Med Screen, 2015. **22**(3): p. 158-64.
95. Ross-Innes, C.S., et al., *Evaluation of a minimally invasive cell sampling device coupled with assessment of trefoil factor 3 expression for diagnosing Barrett's esophagus: a multi-center case-control study.* PLoS Med, 2015. **12**(1): p. e1001780.
96. Fitzgerald, R.C., et al., *Cytosponge-trefoil factor 3 versus usual care to identify Barrett's oesophagus in a primary care setting: a multicentre, pragmatic, randomised controlled trial.* Lancet, 2020. **396**(10247): p. 333-344.
97. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
98. Paynter, N.P., et al., *Association between a literature-based genetic risk score and cardiovascular events in women.* JAMA, 2010. **303**(7): p. 631-7.
99. Yu, Y., et al., *Exome and whole-genome sequencing as clinical tests: a transformative practice in molecular diagnostics.* Clin Chem, 2012. **58**(11): p. 1507-9.
100. Turajlic, S., et al., *Author Correction: Resolving genetic heterogeneity in cancer.* Nat Rev Genet, 2020. **21**(1): p. 65.
101. Swanton, C., *Intratumor heterogeneity: evolution through space and time.* Cancer Res, 2012. **72**(19): p. 4875-82.
102. Graham Brock, E.C.-R., Lan Hu, Christine Coticchia, Johan Skog, *Liquid biopsy for cancer screening, patient stratification and monitoring.* Translational Cancer Research, 2015. **4**(3): p. 280-290.
103. Siddiqui, A.A., et al., *Relationship of pancreatic mass size and diagnostic yield of endoscopic ultrasound-guided fine needle aspiration.* Dig Dis Sci, 2011. **56**(11): p. 3370-5.
104. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-92.
105. Aichler, M. and A. Walch, *In brief: the (molecular) pathogenesis of Barrett's oesophagus.* J Pathol, 2014. **232**(4): p. 383-5.
106. Gindea, C., et al., *Barrett esophagus: history, definition and etiopathogeny.* J Med Life, 2014. **7 Spec No. 3**: p. 23-30.
107. Spechler, S.J., et al., *History, molecular mechanisms, and endoscopic treatment of Barrett's esophagus.* Gastroenterology, 2010. **138**(3): p. 854-69.
108. Ji, Z., et al., *Inflammatory regulatory network mediated by the joint action of NF-kB, STAT3, and AP-1 factors is involved in many human cancers.* Proc Natl Acad Sci U S A, 2019. **116**(19): p. 9453-9462.
109. Agrawal, N., et al., *Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma.* Cancer Discov, 2012. **2**(10): p. 899-905.
110. Ross-Innes, C.S., et al., *Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma.* Nat Genet, 2015. **47**(9): p. 1038-46.
111. Gregson, E.M., J. Bornschein, and R.C. Fitzgerald, *Genetic progression of Barrett's oesophagus to oesophageal adenocarcinoma.* Br J Cancer, 2016. **115**(4): p. 403-10.

112. Maley, C.C., *Multistage carcinogenesis in Barrett's esophagus.* Cancer Lett, 2007. **245**(1-2): p. 22-32.

113. Paulson, T.G., et al., *p16 mutation spectrum in the premalignant condition Barrett's esophagus.* PLoS One, 2008. **3**(11): p. e3809.

114. Keswani, R.N., et al., *Clinical use of p53 in Barrett's esophagus.* Cancer Epidemiol Biomarkers Prev, 2006. **15**(7): p. 1243-9.

115. Bian, Y.S., et al., *p53 gene mutation and protein accumulation during neoplastic progression in Barrett's esophagus.* Mod Pathol, 2001. **14**(5): p. 397-403.

116. Cancer Genome Atlas Research, N., et al., *Integrated genomic characterization of oesophageal carcinoma.* Nature, 2017. **541**(7636): p. 169-175.

117. Gharahkhani, P., et al., *Genome-wide association studies in oesophageal adenocarcinoma and Barrett's oesophagus: a large-scale meta-analysis.* Lancet Oncol, 2016. **17**(10): p. 1363-1373.

118. Reid, B.J., T.G. Paulson, and X. Li, *Genetic Insights in Barrett's Esophagus and Esophageal Adenocarcinoma.* Gastroenterology, 2015. **149**(5): p. 1142-1152 e3.

119. Kanwal, R. and S. Gupta, *Epigenetic modifications in cancer.* Clin Genet, 2012. **81**(4): p. 303-11.

120. Agarwal, A., et al., *Role of epigenetic alterations in the pathogenesis of Barrett's esophagus and esophageal adenocarcinoma.* Int J Clin Exp Pathol, 2012. **5**(5): p. 382-96.

121. Kaz, A.M., et al., *Global DNA methylation patterns in Barrett's esophagus, dysplastic Barrett's, and esophageal adenocarcinoma are associated with BMI, gender, and tobacco use.* Clin Epigenetics, 2016. **8**: p. 111.

122. Agarwal, R., et al., *Epigenomic program of Barrett's-associated neoplastic progression reveals possible involvement of insulin signaling pathways.* Endocr Relat Cancer, 2012. **19**(1): p. L5-9.

123. Jin, Z., et al., *Hypermethylation of the AKAP12 promoter is a biomarker of Barrett's-associated esophageal neoplastic progression.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(1): p. 111-7.

124. Jin, Z., et al., *A multicenter, double-blinded validation study of methylation biomarkers for progression prediction in Barrett's esophagus.* Cancer Res, 2009. **69**(10): p. 4112-5.

125. Kaz, A.M., et al., *Genetic and Epigenetic Alterations in Barrett's Esophagus and Esophageal Adenocarcinoma.* Gastroenterol Clin North Am, 2015. **44**(2): p. 473-89.

126. Kailasam, A., S.K. Mittal, and D.K. Agrawal, *Epigenetics in the Pathogenesis of Esophageal Adenocarcinoma.* Clin Transl Sci, 2015. **8**(4): p. 394-402.

127. Slaby, O., et al., *Dynamic changes in microRNA expression profiles reflect progression of Barrett's esophagus to esophageal adenocarcinoma.* Carcinogenesis, 2015. **36**(5): p. 521-7.

128. Drahos, J., et al., *MicroRNA Profiles of Barrett's Esophagus and Esophageal Adenocarcinoma: Differences in Glandular Non-native Epithelium.* Cancer Epidemiol Biomarkers Prev, 2016. **25**(3): p. 429-37.

129. Zeng, J., et al., *Transcriptional regulation by normal epithelium of premalignant to malignant progression in Barrett's esophagus.* Sci Rep, 2016. **6**: p. 35227.

130. Selaru, F.M., et al., *Global gene expression profiling in Barrett's esophagus and esophageal cancer: a comparative analysis using cDNA microarrays.* Oncogene, 2002. **21**(3): p. 475-8.
131. Greenawalt, D.M., et al., *Gene expression profiling of esophageal cancer: comparative analysis of Barrett's esophagus, adenocarcinoma, and squamous cell carcinoma.* Int J Cancer, 2007. **120**(9): p. 1914-21.
132. Hyland, P.L., et al., *Global changes in gene expression of Barrett's esophagus compared to normal squamous esophagus and gastric cardia tissues.* PLoS One, 2014. **9**(4): p. e93219.
133. Wang, J., et al., *Differential gene expression in normal esophagus and Barrett's esophagus.* J Gastroenterol, 2009. **44**(9): p. 897-911.
134. Sabo, E., et al., *Expression analysis of Barrett's esophagus-associated high-grade dysplasia in laser capture microdissected archival tissue.* Clin Cancer Res, 2008. **14**(20): p. 6440-8.
135. Murao, T., et al., *Overexpression of CD55 from Barrett's esophagus is associated with esophageal adenocarcinoma risk.* J Gastroenterol Hepatol, 2016. **31**(1): p. 99-106.
136. Visser, E., et al., *Prognostic gene expression profiling in esophageal cancer: a systematic review.* Oncotarget, 2017. **8**(3): p. 5566-5577.
137. Adil Butt, M., et al., *Upregulation of mucin glycoprotein MUC1 in the progression to esophageal adenocarcinoma and therapeutic potential with a targeted photoactive antibody-drug conjugate.* Oncotarget, 2017. **8**(15): p. 25080-25096.
138. Jia, Y., et al., *miR-25 is upregulated before the occurrence of esophageal squamous cell carcinoma.* Am J Transl Res, 2017. **9**(10): p. 4458-4469.
139. Sun, J., et al., *Targeting of miR-150 on Gli1 gene to inhibit proliferation and cell cycle of esophageal carcinoma EC9706.* Cancer Biomark, 2017. **21**(1): p. 203-210.
140. Zhang, K., et al., *Circulating miRNA profile in esophageal adenocarcinoma.* Am J Cancer Res, 2016. **6**(11): p. 2713-2721.
141. Fearon, E.R. and B. Vogelstein, *A genetic model for colorectal tumorigenesis.* Cell, 1990. **61**(5): p. 759-67.
142. Feinberg, A.P., R. Ohlsson, and S. Henikoff, *The epigenetic progenitor origin of human cancer.* Nat Rev Genet, 2006. **7**(1): p. 21-33.
143. Herceg, Z. and P. Hainaut, *Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis.* Mol Oncol, 2007. **1**(1): p. 26-41.
144. Hao, Y., et al., *Gene expression profiling reveals stromal genes expressed in common between Barrett's esophagus and adenocarcinoma.* Gastroenterology, 2006. **131**(3): p. 925-33.
145. Gorodeski, E.Z., et al., *Use of hundreds of electrocardiographic biomarkers for prediction of mortality in postmenopausal women: the Women's Health Initiative.* Circ Cardiovasc Qual Outcomes, 2011. **4**(5): p. 521-32.
146. Kuo, W.J., et al., *Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images.* Breast Cancer Res Treat, 2001. **66**(1): p. 51-7.
147. Vlahou, A., et al., *Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data.* J Biomed Biotechnol, 2003. **2003**(5): p. 308-314.
148. Sierra, B. and P. Larranaga, *Predicting survival in malignant skin melanoma using Bayesian networks automatically induced by genetic algorithms. An*

empirical comparison between different approaches. Artif Intell Med, 1998. **14**(1-2): p. 215-30.

149. Duen-Yian, Y., C. Ching-Hsue, and C. Yen-Wen, *A predictive model for cerebrovascular disease using data mining.* Expert Systems with Applications, 2011. **38**(7): p. 8970-8977.
150. Kononenko, I., *Inductive and bayesian learning in medical diagnosis.* Applied Artificial Intelligence: An International Journal, 1993. **7**(4): p. 317-337.
151. van der Sommen, F., et al., *Machine learning in GI endoscopy: practical guidance in how to interpret a novel field.* Gut, 2020.
152. Beck, A.H., et al., *Systematic analysis of breast cancer morphology uncovers stromal features associated with survival.* Sci Transl Med, 2011. **3**(108): p. 108ra113.
153. Deo, R.C., *Machine Learning in Medicine.* Circulation, 2015. **132**(20): p. 1920-30.
154. Bejnordi BE, V.M., Johannes van Diest P, CAMELYON16 Consortium, *Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer.* JAMA, 2017. **318**(2219 - 2210): p. 2199.
155. Williams, A.M., et al., *Artificial intelligence, physiological genomics, and precision medicine.* Physiol Genomics, 2018. **50**(4): p. 237-243.
156. Saeys, Y., I. Inza, and P. Larranaga, *A review of feature selection techniques in bioinformatics.* Bioinformatics, 2007. **23**(19): p. 2507-2517.
157. Rosenfeld A, G.D., Hamoudi R, Butawan R, Eneh V, Khan S, Miah H, Niranjan M, Lovat L, *MIAT: A novel attribute selection approach to better predict upper gastrointestinal cancer.* Data Science and Advanced Analytics, 2015.
158. Gerson, L.B., et al., *Use of a simple symptom questionnaire to predict Barrett's esophagus in patients with symptoms of gastroesophageal reflux.* Am J Gastroenterol, 2001. **96**(7): p. 2005-12.
159. Thrift, A.P., et al., *A clinical risk prediction model for Barrett esophagus.* Cancer Prev Res (Phila), 2012. **5**(9): p. 1115-23.
160. Eloubeidi, M.A. and D. Provenzale, *Clinical and demographic predictors of Barrett's esophagus among patients with gastroesophageal reflux disease: a multivariable analysis in veterans.* J Clin Gastroenterol, 2001. **33**(4): p. 306-9.
161. Ford, A.C., et al., *Ethnicity, gender, and socioeconomic status as risk factors for esophagitis and Barrett's esophagus.* Am J Epidemiol, 2005. **162**(5): p. 454-60.
162. Ward, E.M., et al., *Barrett's esophagus is common in older men and women undergoing screening colonoscopy regardless of reflux symptoms.* Am J Gastroenterol, 2006. **101**(1): p. 12-7.
163. Thukkani, N. and A. Sonnenberg, *The influence of environmental risk factors in hospitalization for gastro-oesophageal reflux disease-related diagnoses in the United States.* Aliment Pharmacol Ther, 2010. **31**(8): p. 852-61.
164. Anderson, L.A., et al., *Risk factors for Barrett's oesophagus and oesophageal adenocarcinoma: results from the FINBAR study.* World J Gastroenterol, 2007. **13**(10): p. 1585-94.
165. Johansson, J., et al., *Risk factors for Barrett's oesophagus: a population-based approach.* Scand J Gastroenterol, 2007. **42**(2): p. 148-56.
166. Steevens, J., et al., *A prospective cohort study on overweight, smoking, alcohol consumption, and risk of Barrett's esophagus.* Cancer Epidemiol Biomarkers Prev, 2011. **20**(2): p. 345-58.

167. Sun, X., et al., *Predicting Barrett's Esophagus in Families: An Esophagus Translational Research Network (BETRNet) Model Fitting Clinical Data to a Familial Paradigm.* Cancer Epidemiol Biomarkers Prev, 2016. **25**(5): p. 727-35.

168. Lipman, G., et al., *Systematic assessment with I-SCAN magnification endoscopy and acetic acid improves dysplasia detection in patients with Barrett's esophagus.* Endoscopy, 2017. **49**(12): p. 1219-1228.

169. Collins, G.S., et al., *Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD).* Ann Intern Med, 2015. **162**(10): p. 735-6.

170. Nie F, X.S., Jia Y, Zhang C, Yan S., *Trace ratio criterion for feature selection.* Proc Natl Conf Artif Intell., 2008. **2**: p. 671-676.

171. Khalilia, M., S. Chakraborty, and M. Popescu, *Predicting disease risks from highly imbalanced data using random forest.* BMC Med Inform Decis Mak, 2011. **11**: p. 51.

172. Moturu ST, J.W., Liu H, *Predicting future high-cost patients: A real-world risk modeling application.*, in *Proc - 2007 IEEE Int Conf Bioinforma Biomed BIBM*. 2007. p. 202-208.

173. Maroco, J., et al., *Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests.* BMC Res Notes, 2011. **4**: p. 299.

174. Langley, N.R., B. Dudzik, and A. Cloutier, *A Decision Tree for Nonmetric Sex Assessment from the Skull.* J Forensic Sci, 2018. **63**(1): p. 31-37.

175. Krittanawong, C., et al., *Artificial Intelligence in Precision Cardiovascular Medicine.* J Am Coll Cardiol, 2017. **69**(21): p. 2657-2664.

176. Zhang W, Z.F., Wu X, Zhang X, Jiang R., *comparative study of ensemble learning approaches in the classification of breast cancer metastasis.*, in *Int Jt Conf Bioinformatics, Syst Biol Intell Comput IJCBS*. 2009. p. 242-245.

177. DR, C., *Machine Learning in Medicine.* Circulation, 2015. **132**(20): p. 1920-1930.

178. Eftekhar, B., et al., *Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data.* BMC Med Inform Decis Mak, 2005. **5**: p. 3.

179. Shahid, N., T. Rappon, and W. Berta, *Applications of artificial neural networks in health care organizational decision-making: A scoping review.* PLoS One, 2019. **14**(2): p. e0212356.

180. Avi Rosenfeld, D.G.G., Sarah Jevons, Jose Ariza, Daryl Hagan, Ash Wilson, Samuel J Lovat, Sarmed S Sami, Omer F Ahmad, Marco Novelli, Manuel Rodriguez Justo, Alison Winstanley, Eliyahu M Heifetz, Mordehy Ben-Zecharia, Uria Noiman, Rebecca C Fitzgerald, Peter Sasieni, Laurence B Lovat, *Development and validation of a risk prediction model to diagnose Barrett's oesophagus (MARK-BE): a case-control machine learning approach.* Lancet Digital Health, 2020. **2**(1): p. 37-48.

181. Ross-Innes, C.S., et al., *Risk stratification of Barrett's oesophagus using a non-endoscopic sampling method coupled with a biomarker panel: a cohort study.* Lancet Gastroenterol Hepatol, 2017. **2**(1): p. 23-31.

182. Wang, H. and G. Li, *Extreme learning machine Cox model for high-dimensional survival analysis.* Stat Med, 2019. **38**(12): p. 2139-2156.

183. Rosenfeld A, R.A., *Explainability in Human–Agent Systems.* Springer US, 2019. **33**.
184. Hippisley-Cox, J. and C. Coupland, *Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm.* Br J Gen Pract, 2013. **63**(606): p. e11-21.
185. Haidry, R.J., et al., *Comparing outcome of radiofrequency ablation in Barrett's with high grade dysplasia and intramucosal carcinoma: a prospective multicenter UK registry.* Endoscopy, 2015. **47**(11): p. 980-7.
186. Di Caro, S., et al., *Role of body composition and metabolic profile in Barrett's oesophagus and progression to cancer.* Eur J Gastroenterol Hepatol, 2016. **28**(3): p. 251-60.
187. Lagergren, J., et al., *Symptomatic gastroesophageal reflux as a risk factor for esophageal adenocarcinoma.* N Engl J Med, 1999. **340**(11): p. 825-31.
188. Nason, K.S., et al., *Gastroesophageal reflux disease symptom severity, proton pump inhibitor use, and esophageal carcinogenesis.* Arch Surg, 2011. **146**(7): p. 851-8.
189. Thrift, A.P., J.M. Garcia, and H.B. El-Serag, *A multibiomarker risk score helps predict risk for Barrett's esophagus.* Clin Gastroenterol Hepatol, 2014. **12**(8): p. 1267-71.
190. Rubenstein, J.H., et al., *Prediction of Barrett's esophagus among men.* Am J Gastroenterol, 2013. **108**(3): p. 353-62.
191. Edelstein, Z.R., et al., *Risk factors for Barrett's esophagus among patients with gastroesophageal reflux disease: a community clinic-based case-control study.* Am J Gastroenterol, 2009. **104**(4): p. 834-42.
192. Abrams, J.A., et al., *Racial and ethnic disparities in the prevalence of Barrett's esophagus among patients who undergo upper endoscopy.* Clin Gastroenterol Hepatol, 2008. **6**(1): p. 30-4.
193. Kubo, A., et al., *Sex-specific associations between body mass index, waist circumference and the risk of Barrett's oesophagus: a pooled analysis from the international BEACON consortium.* Gut, 2013. **62**(12): p. 1684-91.
194. Xie, S.H., et al., *Assessing the feasibility of targeted screening for esophageal adenocarcinoma based on individual risk assessment in a population-based cohort study in Norway (The HUNT Study).* Am J Gastroenterol, 2018. **113**(6): p. 829-835.
195. Kunzmann, A.T., et al., *Model for Identifying Individuals at Risk for Esophageal Adenocarcinoma.* Clin Gastroenterol Hepatol, 2018. **16**(8): p. 1229-1236 e4.
196. Ireland, C.J., et al., *Validation of a risk prediction model for Barrett's esophagus in an Australian population.* Clin Exp Gastroenterol, 2018. **11**: p. 135-142.
197. Ireland, C.J., et al., *Development of a risk prediction model for Barrett's esophagus in an Australian population.* Dis Esophagus, 2017. **30**(11): p. 1-8.
198. Herrera Elizondo, J.L., et al., *Prevalence of Barrett's esophagus: An observational study from a gastroenterology clinic.* Rev Gastroenterol Mex, 2017. **82**(4): p. 296-300.
199. Lemaitre G, N.F., Aridas CK, *Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning.* J Mach Learn Res., 2017. **18**(1): p. 559-563.
200. Offman, J., et al., *Barrett's oESophagus trial 3 (BEST3): study protocol for a randomised controlled trial comparing the Cytosponge-TFF3 test with usual*

care to facilitate the diagnosis of oesophageal pre-cancer in primary care patients with chronic acid reflux. BMC Cancer, 2018. **18**(1): p. 784.

201. Kalousis A, P.J., Hilario M, *Stability of feature selection algorithms: a study on high-dimensional spaces.* Knowl Inf Syst, 2007. **12**(1): p. 95-116.

202. Park T, C.G., *The Bayesian Lasso.* J Am Stat Assoc., 2008. **103**(482): p. 681-686.

203. Zhang, X., et al., *Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data.* BMC Bioinformatics, 2006. **7**: p. 197.

204. Offman, J. and R.C. Fitzgerald, *Alternatives to Traditional Per-Oral Endoscopy for Screening.* Gastrointest Endosc Clin N Am, 2017. **27**(3): p. 379-396.

205. Coste, J., et al., *Non response, incomplete and inconsistent responses to self-administered health-related quality of life measures in the general population: patterns, determinants and impact on the validity of estimates - a population-based study in France using the MOS SF-36.* Health Qual Life Outcomes, 2013. **11**: p. 44.

206. Sager, M., et al., *Transcriptomics in cancer diagnostics: developments in technology, clinical research and commercialization.* Expert Rev Mol Diagn, 2015. **15**(12): p. 1589-603.

207. Hatzimichael, E. and T. Crook, *Cancer epigenetics: new therapies and new challenges.* J Drug Deliv, 2013. **2013**: p. 529312.

208. Mandel, P. and P. Metais, *[Not Available].* C R Seances Soc Biol Fil, 1948. **142**(3-4): p. 241-3.

209. Shao, W., et al., *Evaluation of genome-wide genotyping concordance between tumor tissues and peripheral blood.* Genomics, 2017. **109**(2): p. 108-112.

210. Allard, W.J., et al., *Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases.* Clin Cancer Res, 2004. **10**(20): p. 6897-904.

211. Zhe, X., M.L. Cher, and R.D. Bonfil, *Circulating tumor cells: finding the needle in the haystack.* Am J Cancer Res, 2011. **1**(6): p. 740-51.

212. Crowley, E., et al., *Liquid biopsy: monitoring cancer-genetics in the blood.* Nat Rev Clin Oncol, 2013. **10**(8): p. 472-84.

213. Perrone, F., et al., *Circulating free DNA in a screening program for early colorectal cancer detection.* Tumori, 2014. **100**(2): p. 115-21.

214. Yanez-Mo, M., et al., *Biological properties of extracellular vesicles and their physiological functions.* J Extracell Vesicles, 2015. **4**: p. 27066.

215. Huang, X., et al., *Characterization of human plasma-derived exosomal RNAs by deep sequencing.* BMC Genomics, 2013. **14**: p. 319.

216. Karachaliou, N., et al., *Real-time liquid biopsies become a reality in cancer treatment.* Ann Transl Med, 2015. **3**(3): p. 36.

217. Tie, J., et al., *Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer.* Sci Transl Med, 2016. **8**(346): p. 346ra92.

218. Harris, F.R., et al., *Quantification of Somatic Chromosomal Rearrangements in Circulating Cell-Free DNA from Ovarian Cancers.* Sci Rep, 2016. **6**: p. 29831.

219. Hamilton, J.G., *Needle phobia: a neglected diagnosis.* J Fam Pract, 1995. **41**(2): p. 169-75.

220. Tabak, L.A., *A revolution in biomedical assessment: the development of salivary diagnostics.* J Dent Educ, 2001. **65**(12): p. 1335-9.

221. Greabu, M., et al., *Saliva--a diagnostic window to the body, both in health and in disease.* J Med Life, 2009. **2**(2): p. 124-32.

222. Lee, Y.H. and D.T. Wong, *Saliva: an emerging biofluid for early detection of diseases.* Am J Dent, 2009. **22**(4): p. 241-8.

223. Malamud, D., *Saliva as a diagnostic fluid.* Dent Clin North Am, 2011. **55**(1): p. 159-78.

224. Elashoff, D., et al., *Prevalidation of salivary biomarkers for oral cancer detection.* Cancer Epidemiol Biomarkers Prev, 2012. **21**(4): p. 664-72.

225. Xie, Z., et al., *Salivary microRNAs as promising biomarkers for detection of esophageal cancer.* PLoS One, 2013. **8**(4): p. e57502.

226. Zhang, L., et al., *Salivary transcriptomic biomarkers for detection of resectable pancreatic cancer.* Gastroenterology, 2010. **138**(3): p. 949-57 e1-7.

227. Wei, C., J. Li, and R.E. Bumgarner, *Sample size for detecting differentially expressed genes in microarray experiments.* BMC Genomics, 2004. **5**: p. 87.

228. Flaherty P, N.G., Muralidharan O, Winters M, Buenrostro J, Bell J, Brown S, Holodniy M, Zhang N, Ji HP. , *Ultrasensitive detection of rare mutations using next-generation targeted resequencing.* Nucleic Acid Research, 2012. **40**(1): p. 1-12.

229. Hamoudi, R.A., et al., *Differential expression of NF-kappaB target genes in MALT lymphoma with and without chromosome translocation: insights into molecular mechanism.* Leukemia, 2010. **24**(8): p. 1487-97.

230. Rosebeck, S., et al., *Cleavage of NIK by the API2-MALT1 fusion oncoprotein leads to noncanonical NF-kappaB activation.* Science, 2011. **331**(6016): p. 468-72.

231. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Res, 2015. **43**(Database issue): p. D805-11.

232. Wang, Q., C. Ma, and W. Kemmner, *Wdr66 is a novel marker for risk stratification and involved in epithelial-mesenchymal transition of esophageal squamous cell carcinoma.* BMC Cancer, 2013. **13**: p. 137.

233. Kimchi, E.T., et al., *Progression of Barrett's metaplasia to adenocarcinoma is associated with the suppression of the transcriptional programs of epidermal differentiation.* Cancer Res, 2005. **65**(8): p. 3146-54.

234. Stairs, D.B., et al., *Cdx1 and c-Myc foster the initiation of transdifferentiation of the normal esophageal squamous epithelium toward Barrett's esophagus.* PLoS One, 2008. **3**(10): p. e3534.

235. Saadi, A., et al., *Stromal genes discriminate preinvasive from invasive disease, predict outcome, and highlight inflammatory pathways in digestive cancers.* Proc Natl Acad Sci U S A, 2010. **107**(5): p. 2177-82.

236. Ernst, P., *Review article: the role of inflammation in the pathogenesis of gastric cancer.* Aliment Pharmacol Ther, 1999. **13 Suppl 1**: p. 13-8.

237. Howe, L.R., et al., *Molecular pathways: adipose inflammation as a mediator of obesity-associated cancer.* Clin Cancer Res, 2013. **19**(22): p. 6074-83.

238. Gukovsky, I., et al., *Inflammation, autophagy, and obesity: common features in the pathogenesis of pancreatitis and pancreatic cancer.* Gastroenterology, 2013. **144**(6): p. 1199-209 e4.

239. Hamidi, A.E., et al., *Archival cervical smears: a versatile resource for molecular investigations.* Cytopathology, 2002. **13**(5): p. 291-9.

240. Suresh V Kuchipudi, M.T., Rahul K Nelli, Gavin A White, Belinda Baquero Perez, Sujith Sebastian, Marek J Slomka, Sharon M Brookes, Ian H Brown, Stephen P Dunham and Kin-Chow Chang, *18s rRNA is a reliable*

normalisation gene for real time PCR based on influenza virus infected cells. Virology journal, 2012. **9**(1): p. 230.

241. Pandit, P., J. Cooper-White, and C. Punyadeera, *High-yield RNA-extraction method for saliva.* Clin Chem, 2013. **59**(7): p. 1118-22.

242. Palanisamy, V. and D.T. Wong, *Transcriptomic analyses of saliva.* Methods Mol Biol, 2010. **666**: p. 43-51.

243. Ludyga, N., et al., *Nucleic acids from long-term preserved FFPE tissues are suitable for downstream analyses.* Virchows Arch, 2012. **460**(2): p. 131-40.

244. Li, P., et al., *Whole-Transcriptome profiling of formalin-fixed, paraffin-embedded renal cell carcinoma by RNA-seq.* BMC Genomics, 2014. **15**: p. 1087.

245. Munchel, S., et al., *Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics.* Oncotarget, 2015. **6**(28): p. 25943-61.

246. Spielmann, N., et al., *The human salivary RNA transcriptome revealed by massively parallel sequencing.* Clin Chem, 2012. **58**(9): p. 1314-21.

247. Flohr, A.M., et al., *DNase I treatment of cDNA first strands prevents RT-PCR amplification of contaminating DNA sequences.* Biotechniques, 2003. **35**(5): p. 920-2, 924, 926.

248. Vartanian, K., et al., *Gene expression profiling of whole blood: comparison of target preparation methods for accurate and reproducible microarray analysis.* BMC Genomics, 2009. **10**: p. 2.

249. Schwaederle, M., et al., *Use of Liquid Biopsies in Clinical Oncology: Pilot Experience in 168 Patients.* Clin Cancer Res, 2016. **22**(22): p. 5497-5505.

250. Fabryova, H. and P. Celec, *On the origin and diagnostic use of salivary RNA.* Oral Dis, 2014. **20**(2): p. 146-52.

251. Yang, J., et al., *Detection of tumor cell-specific mRNA and protein in exosome-like microvesicles from blood and saliva.* PLoS One, 2014. **9**(11): p. e110641.

252. Huhta, H., et al., *Toll-like receptors 1, 2, 4 and 6 in esophageal epithelium, Barrett's esophagus, dysplasia and adenocarcinoma.* Oncotarget, 2016. **7**(17): p. 23658-67.

253. Henson, B.S. and D.T. Wong, *Collection, storage, and processing of saliva samples for downstream molecular applications.* Methods Mol Biol, 2010. **666**: p. 21-30.

254. Segal, A. and D.T. Wong, *Salivary diagnostics: enhancing disease detection and making medicine better.* Eur J Dent Educ, 2008. **12 Suppl 1**: p. 22-9.

255. Stephen, K.W. and C.F. Speirs, *Methods for collecting individual components of mixed saliva: the relevance to clinical pharmacology.* Br J Clin Pharmacol, 1976. **3**(2): p. 315-9.

256. Bahn, J.H., et al., *The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva.* Clin Chem, 2015. **61**(1): p. 221-30.

257. Navazesh, M., *Methods for collecting saliva.* Ann N Y Acad Sci, 1993. **694**: p. 72-7.

258. Talge, N.M., et al., *It's not that bad: error introduced by oral stimulants in salivary cortisol research.* Dev Psychobiol, 2005. **47**(4): p. 369-76.

259. Golatowski, C., et al., *Comparative evaluation of saliva collection methods for proteome analysis.* Clin Chim Acta, 2013. **419**: p. 42-6.

260. Chiang, S.H., et al., *RNAPro*SAL: a device for rapid and standardized collection of saliva RNA and proteins.* Biotechniques, 2015. **58**(2): p. 69-76.

261. Park, N.J., et al., *Characterization of salivary RNA by cDNA library analysis.* Arch Oral Biol, 2007. **52**(1): p. 30-5.

262. Zubakov, D., et al., *Stable RNA markers for identification of blood and saliva stains revealed from whole genome expression analysis of time-wise degraded samples.* Int J Legal Med, 2008. **122**(2): p. 135-42.

263. Tanabe, M., et al., *Effects of rehydration and food consumption on salivary flow, pH and buffering capacity in young adult volunteers during ergometer exercise.* J Int Soc Sports Nutr, 2013. **10**(1): p. 49.

264. Laidi, F., et al., *Significant correlation between salivary and serum Ca 15-3 in healthy women and breast cancer patients.* Asian Pac J Cancer Prev, 2014. **15**(11): p. 4659-62.

265. Ishikawa, S., et al., *Identification of salivary metabolomic biomarkers for oral cancer screening.* Sci Rep, 2016. **6**: p. 31520.

266. Li, Y., et al., *Salivary transcriptome diagnostics for oral cancer detection.* Clin Cancer Res, 2004. **10**(24): p. 8442-50.

267. Tiwari, M., *Science behind human saliva.* J Nat Sci Biol Med, 2011. **2**(1): p. 53-8.

268. Ribeiro-Silva, A., H. Zhang, and S.S. Jeffrey, *RNA extraction from ten year old formalin-fixed paraffin-embedded breast cancer samples: a comparison of column purification and magnetic bead-based technologies.* BMC Mol Biol, 2007. **8**: p. 118.

269. Kharchenko, S.V. and A.A. Shpakov, *[Regulation of the RNAse activity of the saliva in healthy subjects and in stomach cancer].* Izv Akad Nauk SSSR Biol, 1989(1): p. 58-63.

270. Nobori, T., et al., *Deletions of the cyclin-dependent kinase-4 inhibitor gene in multiple human cancers.* Nature, 1994. **368**(6473): p. 753-6.

271. Sherr, C.J., *Cancer cell cycles.* Science, 1996. **274**(5293): p. 1672-7.

272. Bian, Y.S., et al., *p16 inactivation by methylation of the CDKN2A promoter occurs early during neoplastic progression in Barrett's esophagus.* Gastroenterology, 2002. **122**(4): p. 1113-21.

273. Igaki, H., et al., *Mutation frequency of the p16/CDKN2 gene in primary cancers in the upper digestive tract.* Cancer Res, 1995. **55**(15): p. 3421-3.

274. Heldin, C.H., K. Miyazono, and P. ten Dijke, *TGF-beta signalling from cell membrane to nucleus through SMAD proteins.* Nature, 1997. **390**(6659): p. 465-71.

275. Stolfi, C., et al., *The dual role of Smad7 in the control of cancer growth and metastasis.* Int J Mol Sci, 2013. **14**(12): p. 23774-90.

276. Osawa, H., et al., *Prognostic value of the expression of Smad6 and Smad7, as inhibitory Smads of the TGF-beta superfamily, in esophageal squamous cell carcinoma.* Anticancer Res, 2004. **24**(6): p. 3703-9.

277. Stolfi, C., et al., *A functional role for Smad7 in sustaining colon cancer cell growth and survival.* Cell Death Dis, 2014. **5**: p. e1073.

278. Onwuegbusi, B.A., et al., *Impaired transforming growth factor beta signalling in Barrett's carcinogenesis due to frequent SMAD4 inactivation.* Gut, 2006. **55**(6): p. 764-74.

279. Kang, J.U., et al., *AMY2A: a possible tumor-suppressor gene of 1p21.1 loss in gastric carcinoma.* Int J Oncol, 2010. **36**(6): p. 1429-35.

280. Rivlin, N., et al., *Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis.* Genes Cancer, 2011. **2**(4): p. 466-74.

281. Flejou, J.F., *Barrett's oesophagus: from metaplasia to dysplasia and cancer.* Gut, 2005. **54 Suppl 1**: p. i6-12.

282. Kastelein, F., et al., *Aberrant p53 protein expression is associated with an increased risk of neoplastic progression in patients with Barrett's oesophagus.* Gut, 2013. **62**(12): p. 1676-83.

283. Puccio, I., et al., *Immunohistochemical assessment of Survivin and Bcl3 expression as potential biomarkers for NF-kappaB activation in the Barrett metaplasia-dysplasia-adenocarcinoma sequence.* Int J Exp Pathol, 2018. **99**(1): p. 10-14.

284. Kauppila, J.H. and K.S. Selander, *Toll-like receptors in esophageal cancer.* Front Immunol, 2014. **5**: p. 200.

285. Sabroe, I., et al., *The role of TLR activation in inflammation.* J Pathol, 2008. **214**(2): p. 126-35.

286. Yang, L., et al., *Microbiome in reflux disorders and esophageal adenocarcinoma.* Cancer J, 2014. **20**(3): p. 207-10.

287. Cronin, J., et al., *Epidermal growth factor receptor (EGFR) is overexpressed in high-grade dysplasia and adenocarcinoma of the esophagus and may represent a biomarker of histological progression in Barrett's esophagus (BE).* Am J Gastroenterol, 2011. **106**(1): p. 46-56.

288. Pretto, G., et al., *Increase of epidermal growth factor receptor expression in progression of GERD, Barrett, and adenocarcinoma of esophagus.* Dig Dis Sci, 2013. **58**(1): p. 115-22.

289. Hallmann, A., *Key elements of the retinoblastoma tumor suppressor pathway in Volvox carteri.* Commun Integr Biol, 2009. **2**(5): p. 396-9.

290. Takano, Y., et al., *Cyclin D2, but not cyclin D1, overexpression closely correlates with gastric cancer progression and prognosis.* J Pathol, 1999. **189**(2): p. 194-200.

291. Deshpande, A., P. Sicinski, and P.W. Hinds, *Cyclins and cdks in development and cancer: a perspective.* Oncogene, 2005. **24**(17): p. 2909-15.

292. Arber, N., et al., *Increased expression of the cyclin D1 gene in Barrett's esophagus.* Cancer Epidemiol Biomarkers Prev, 1996. **5**(6): p. 457-9.

293. Dey, B., et al., *Expression of Cyclin D1 and P16 in Esophageal Squamous Cell Carcinoma.* Middle East J Dig Dis, 2015. **7**(4): p. 220-5.

294. Merga, Y.J., et al., *Importance of the alternative NF-kappaB activation pathway in inflammation-associated gastrointestinal carcinogenesis.* Am J Physiol Gastrointest Liver Physiol, 2016. **310**(11): p. G1081-90.

295. Guo, Y., et al., *Clinical significance of the correlation between PLCE 1 and PRKCA in esophageal inflammation and esophageal carcinoma.* Oncotarget, 2017. **8**(20): p. 33285-33299.

296. Bazzocco, S., et al., *Highly Expressed Genes in Rapidly Proliferating Tumor Cells as New Targets for Colorectal Cancer Treatment.* Clin Cancer Res, 2015. **21**(16): p. 3695-704.

297. Graber, M.W., et al., *Isolation of differentially expressed genes in carcinoma of the esophagus.* Ann Surg Oncol, 1996. **3**(2): p. 192-7.

298. Naderi, A., et al., *A gene-expression signature to predict survival in breast cancer across independent data sets.* Oncogene, 2007. **26**(10): p. 1507-16.

299. Adkins, R.M., et al., *Racial differences in gene-specific DNA methylation levels are present at birth.* Birth Defects Res A Clin Mol Teratol, 2011. **91**(8): p. 728-36.

300. Hall, E., et al., *Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets.* Genome Biol, 2014. **15**(12): p. 522.

301. Diaz-Lagares, A., et al., *A Novel Epigenetic Signature for Early Diagnosis in Lung Cancer.* Clin Cancer Res, 2016. **22**(13): p. 3361-71.

302. Nogueira da Costa, A. and Z. Herceg, *Detection of cancer-specific epigenomic changes in biofluids: powerful tools in biomarker discovery and application.* Mol Oncol, 2012. **6**(6): p. 704-15.