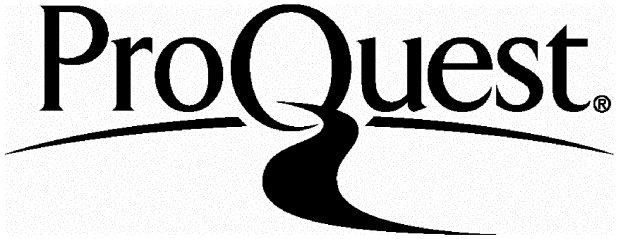
Screening for Oral Cancer and Precancer in General Dental Practice: Evaluation of Machine Learning Software in the Identification of High Risk Groups

Kenneth Lim

2003

Eastman Dental Institute for Oral Health Care Sciences,
University College London, University of London

This thesis is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in the Faculty of Medicine



Abstract

Background: There is much to recommend the early detection of oral cancer and precancer as this could lead to a reduction in mortality and morbidity, an increase in the quality of life, and cost savings for the Health Service. Oral cancer fulfils most of the criteria for a screenable disease; a simple visual examination is well accepted by patients and is a valid means for detecting oral cancer and precancer. However it is widely agreed that population screening for oral cancer and precancer cannot be recommended because of its relatively low prevalence in the United Kingdom. This study investigates the role of general dental practitioners in the detection of lesions, smoking and drinking habits within the population who visit their dentist. It also looks at the usefulness of machine learning software (artificial intelligence) for the identification of groups that could be at high risk of getting oral cancer or precancer. A suite of different machine learning models was trained and used to investigate the possible role of machine learning software in the identification of high-risk groups. If such high-risk groups can be identified and screened at closer intervals, then a bigger harvest of cancer and precancer cases could be expected from these smaller cohorts of individuals. Machine learning models would thus act as a " pre-screen filter" identifying this smaller sub-set of individuals for closer scrutiny thus making screening for oral cancer and precancer more efficient and cost effective.

Design:

i) A prospective demonstration study recruiting patients opportunistically in general dental practices.

ii) Epidemiological comparison of datasets from this current study with one from an earlier study.

iii) The use of these datasets to evaluate the use of machine learning software for the identification of sub-groups at high-risk of oral cancer and precancer.

Methods: This study was carried out by 18 general dental practitioners from practices in north London, south London, Nottingham and Aldershot. Screening dentists attended training sessions and were advised on the criteria of a positive or negative screen. Each then recruited patients over the age of 35 attending at their practices for dental treatment and noted the presence or absence of lesions and conditions. The patients were each invited to complete a health questionnaire. The clinical data and health questionnaires were collected and collated and this information analysed for prevalence of mucosal lesions and correlations with age, gender, smoking and drinking habits. Epidemiological features of this dataset were compared with an earlier dataset obtained from a number of screening studies done in the United Kingdom in 1995.

Results: 2265 cases were available for analysis. Oral lesions were detected in 14.1% (319) of patients with 4.2% (94) of these considered to be positive. This included 2 cases of squamous cell carcinoma one developing from a case of oral submucous fibrosis. The prevalence of white patches was 2.0%, red patches, 0.5% and lichen planus was 1.5%. In this study, the prevalence rate for positive lesions and conditions was 4.2%. 27% of the screened population were smokers and 16.5% consumed more than 5 units of alcohol a week. There were significant correlations between positive lesions and male gender (IRR 1.86, 95%CI 1.22-2.82), heavy smoking males (IRR 3.68, 95%CI 2.10-6.43), heavy smoking female (IRR 3.58, 95% CI1.35-9.50) and heavy alcohol use in males (IRR 2.98, 95% CI 1.06-3.47)

Machine learning models performed well and were capable of achieving high sensitivities of 85% and 80% however always accompanied by low specificities of 27% and 31%. Different types of machine learning methods were used and none performed better than another.

Data from the 1995 studies showed remarkable epidemiological similarities with data from this current study despite the interval between the two projects being some 5 years and studies done in different environments, though all mainly in the London area.

Conclusions: The results suggested that general dental practitioners were able to identify oral mucosal lesions and conditions following screening according to criteria that they had been

taught. The prevalence of risk factors such as smoking and drinking habits within the screened group were similar to those reported in national statistics; the prevalence of lesions and conditions regarded as positive were also comparable to published studies in the UK and other countries. This suggests that the population visiting general dental practices is representative of the general population and that screening in general dental practice is feasible; also that the general dental practice environment appears well suited to opportunistic screening for oral cancer and precancer as about 60% of the population regularly attend at their dentist. Thus the opportunity may exist for screening a large proportion of the general population as part of routine dental patient care, in a cost-effective manner.

Machine learning models performed well in identifying positive cases but with unacceptably high numbers of false positives. At this stage of development machine learning models cannot be recommended for use as a pre-screen filter although it could have a future role as part of an interactive health awareness package to patients. Building on knowledge and experiences gathered from this study, further screening studies are envisaged involving dental practices in other regions to explore possible variations in prevalence of oral cancer and precancer within the UK.

List of tables

List of Figures

Acknowledgements

I am most grateful to Professor Paul Speight for making this study possible. His vision, leadership, inspiration and guidance have been pivotal to this project over these five years. Also to Professor Peter Hammond for his advice and resources in Informatics and Computer technology and to Professor Martin Downer for his wisdom and wit which I greatly enjoyed. Special thanks are due to Drs Aviva Petrie and David Moles for their indispensable help and guidance in the area of statistics. Thanks also to my good friend Irene Lakhani who cheerfully and tirelessly proof-read every line of this work.

I wish to thank the following seventeen general dental practitioners Drs Peter Fine, Nick Unsworth, Richard Pinns, Malcolm Alexander, Malcolm Gold, Rohit Mistry, Colin Gaunt, Douglas Miller, Nigel Monkman, Tushar Patel, Sandip Patel, Sangeeta Patel, Raj Parmar, Peter Jenkinson, Rebecca Roberts, Darren Kowlessar, and Harris Yee-Chong all of whom gave of their busy working schedules to participate in the screening study over 2000/2001.

I am grateful to the Oral and Dental Research Trust for the privilege of winning the Dunhill Medical Trust Fellowship Award over two consecutive years. The generous funding from these awards made possible the purchase of computers and software so necessary for this project.

Finally to my family, my wife, Michie for being there, my two sons, brother Fred, sister Angie for their constant support and encouragement and my late mother who always thought the best of me.

Part A

General Introduction

Chapter 1

Oral Cancer

Chapter 1~Oral Cancer

1.1 Definition and scope

"Oral cancer" is generally defined as squamous cell carcinoma of the lip, tongue, gum, floor of the mouth, oropharynx and other unspecified parts and ill-defined sites of the mouth. These cancers are registered as ICD-9 (International Classification of Diseases) 140, 141, 143, 144, 145, 146 and 149 and constitute over 90% of all malignant diseases in the mouth. More recently, ICD-10 has been introduced (WHO, 1995). Categories relevant to this project would now be listed as C00, C01, C02, C03, C04, C06, C10, and C014. Examples of other cancers that are found in the mouth are salivary gland tumours, sarcomas, liposarcomas (Nunes *et al,* 2002) and lymphomas (Flaitz *et al,* 2002). These are relatively uncommon and tend to have different aetiological factors to squamous cell carcinomas. Since over 90% of cancers presenting in the mouth are squamous cell carcinomas, this overview is presented with an emphasis on squamous cell carcinoma and the more commonly found and significant precancerous conditions and lesions. Table 1.1 is a table of the relevant diseases in their ICD-9 listing along with their updated ICD-10 equivalents.

1.2 Epidemiology

1.2.1 Global incidence

Oral cancer is the sixth most common cancer in the world with an estimated total of 212,000 cases per year (Parkin *et al,* 1999). There is an approximately 20-fold geographical variation in global incidences. Databases collected from some regions tend to be incomplete or unreliable, especially in third world countries where it is most needed; information on the follow-up of cases treated or otherwise is usually not available. The highest rates in the world are found in the Indian sub-continent, Brazil, France and Melanesia. There are even differences between countries in the

same region. Table 1.2 shows a representative sample of age-standardised rates/100,000 for oral and pharyngeal cancer in males and females from a list of 24 countries (Parkin *et al,* 1993). High rates in Melanesia may relate to chewing and smoking habits. In India, Bangladesh, Pakistan and Sri Lanka oral cancer is the most commonly found, and accounts for a third of all cancers. The high incidence in the Bas Rhin region of France (up to 50:100,000 for males) is thought to be associated with the consumption of calvados. Gender difference also influences incidence with males generally 2-3 times more likely to develop intra-oral cancer and lip cancer than females (Table 1.2). Incidence rates of 11-14:100,000 have been reported in white populations, and these are particularly linked to outdoor occupations such as fishing and farming in countries such as Canada and Australia. The incidence for tongue cancer is similar or slightly higher among women. Age is seen to influence incidence as 98% of cases are over the age of 40 years, but in regions of high prevalence many are less than 35 years. There is an observed increase of oral cancer among younger men (Hindle *et al*, 1996). Ethnicity appears to influence incidence rates. Rates of 20.4:100,000 have been observed in African American males in the United States, making it the fourth most common cancer in this racial group (after prostate, lung and colorectal carcinoma). Migrant studies for oral cancer are few. Sverdlow *et al* (1995) showed that mortality rates in Indian migrants in England and Wales were higher than those born in England (odds ratio 2.2: P=95% [CI 1.5-3.1]). However this was based only on 30 cases. An analysis of oral cancer mortality in 24 countries revealed an upward trend in birth cohort related mortality in 19 countries, with the largest increases occurring in Central and Eastern Europe (CRC, 2000; MacFarlane *et al,* 1994). There appears to be an overall increase in the incidence and mortality of oral cancer in the last 25 years (CRC, 2000; Warnakulasuriya, 2001).

1.2.2 United Kingdom incidence

Oral cancer is the 16th most common cancer in the United Kingdom. This is about 1% of all new cancers, on a par with multiple myeloma and cervical cancer. The recent CancerStats-UK (2003)

reported 5099 new cases. The incidence is higher in Scotland and Northern Ireland, and also among such ethnic minorities as Asian and Chinese migrants, and in deprived groups (Warnakulasuriya *et al,* 1999). Table 1.3 compares the age-standardised incidence rates within the UK for oral cancer.

Recent studies examining the epidemiology of oral cancer in England and Wales have shown that the number of cases of oral cancer is increasing particularly among young males in the 35-64 age group (Hindle *et al,* 1996). This suggests that there could be an increase in the numbers of oral cancer in the coming decades as these birth cohorts enter age groups associated with higher frequencies.

1.2.3 Summary

Although not a major health problem in most Western countries, with the exception of some regions of France such as the German border and Brittany, oral cancer is globally an important cancer, especially in developing countries where it is the third most common cancer. Survival is much influenced by the site and tumour stage at presentation and the availability of treatment facilities but generally overall 5-year survival is less than 50%. This decreases to 21% in cases with metastases (Peters, 2002) and trends in mortality do not differ greatly in different parts of the world (Tominaga *et al,* 1998). There is evidence that whilst surgical and rehabilitation techniques have improved in Western countries, mortality from oral cancer has not (Stell and McCormick, 1985). An analysis of oral cancer mortality in 24 countries revealed an upward trend in birth cohort related mortality in 19 countries, with the largest increases occurring in Central and Eastern Europe (Macfarlane *et al,* 1994a). There is also evidence of a trend in the United Kingdom which suggests that incidence may increase amongst males (Hindle et *al,* 1996). Macfarlane *et al* (1994b) also noted increases in incidences for females, especially in central and Eastern Europe, and predicted that, with increasing popularity of smoking among females, this incidence could rise. Indeed, there is already an observed increase of prevalence in females as the

male: female ratio for oral cancer in the United Kingdom has fallen in the past fifty years from 5:1 to 2:1 (CRC, 2000). The increased use of tobacco, whether smoked or chewed, and other high-risk habits, such as the use of paan (betel-quid), is also likely to increase the prevalence of precancerous lesions such as leukoplakia and erythroplakia and conditions such as submucous fibrosis, and this in turn could contribute to future increases in oral cancer cases (Gupta and Nandakumar, 1999). The incidence rate amongst younger age groups also appears to be increasing, shown for example by the incidence of oral cancer in America for males and females under the age of 40 years rising from 0.3 per 100,000 cases in 1975-79 to 1.1 per 100,000 cases in 1987-91, an increase of 267% (Peters, 2002). Llewellyn *et al* (2003) analysed 116 cases of oral cancer in patients under the age of 45 years between 1990 and 1997, providing further evidence of this phenomenon and the role of tobacco use and alcohol consumption as risk factors. These habits were observed in 75% of cases. However they also reported disturbingly that 26% of this population did not have any major risk habits. All indications from global and United Kingdom figures on oral cancer incidence suggest not only that it is currently an important health problem but it is also one likely to increase in prominence over the coming years. Oral cancer is very amenable to primary and secondary prevention. The risk habits are well known and the feasibility of prevention has been demonstrated in population-based studies (Warnakulasuriya *et al,* 1984; Gupta *et al,* 1992).

Table 1.1 A list of relevant diseases and their ICD-9 identification numbers with the equivalent ICD-10 numbers

ICD-9	ICD-10	Disease
140	C00	Malignant neoplasm of the lip
141	C01/02	Malignant neoplasm of the tongue
143	C03	Malignant neoplasm of the gums
144	C04	Malignant neoplasm of the floor of the mouth
	C05	Malignant neoplasm of the palate
145	C06	Malignant neoplasm of other unspecified parts of the mouth
146	C10	Malignant neoplasm of the oropharynx
149	C14	Malignant neoplasm of other and ill-defined sites within the lip, oral cavity and pharynx

Table 1.2 Age-standardised incidence rates for oral and pharyngeal cancer for all areas, plus a representative sample of countries (after Parkin *et al,* 1993)

Country	Male	Female
All areas	14.0	6.5
Melanesia	47.0	25.6
Micronesia/Polynesia	10.0	4.6
Northern Europe	6.5	2.8
Eastern Europe	10.7	2.4
North America	14.2	5.1
South America (tropical)	18.7	4.9
South America (temperate)	8.1	2.3
South East Asia	14.5	9.8
Southern Asia	25.1	14.9

Table 1.3 Age-standardised incidence rates of oral cancer among men and women in the United Kingdom. Northern Ireland and Scotland have higher incidences than England and Wales (CancerStats-UK, 2003)

		Numbers	Rate
United Kingdom	**Male**	3268	10.7
	Female	1831	4.8
England	**Male**	2546	9.9
	Female	1444	4.5
Wales	**Male**	185	11.7
	Female	95	4.9
Scotland	**Male**	446	17.8
	Female	238	7.2
Northern Ireland	**Male**	91	12.1
	Female	54	5.4

1.3 Brief description of oral cancer and its variants

1.3.1 Oral squamous cell carcinoma

Oral squamous cell carcinoma is described as a tumour consisting of irregular nests, columns or strands of malignant cells, infiltrating the subepithelial connective tissues. The tumour cells may resemble any or all of the layers of stratified squamous epithelium. The majority of oral cancers are well or moderately differentiated. Keratin pearl formation is prominent especially in the more superficial parts of the tumour. Prickle cell differentiation is usually conspicuous and there is often a strong host response of lymphocytes, plasma cells and sometimes eosinophils in the underlying connective tissue. The frequency of mitotic figures and the extent of cellular atypia vary, but bizarre forms may be less in evidence in frank carcinomas than in the dysplastic non-invasive epithelium. In large oral carcinomas, the differentiation pattern can vary from one area to another. The pattern of invasion tends to vary with the site; cancers in the floor of the mouth tend to invade in breadth and cancers in the tongue tend to invade in depth. Lymph node spread is a relatively late feature of oral cancer. The site of invasion varies but the tongue appears to have the highest predilection. Table 1.4 compares affected sites for oral cancer from two unrelated studies, Chen *et al* (1990) and Shah *et al* (1990), which show broad agreement in epidemiological patterns of oral cancer.

Table 1.4 Common sites for oral cancer from two studies, Chen *et al* (1990) and Shah *et al* (1990)

Site	Prevalence (%) Chen *et al* (1990)	Prevalence (%) Shah *et al* (1990
Tongue	41.2	36.0
Floor of mouth	24.2	33.0
Palate	11.0	1.0
Gingival	9.3	21.0
Buccal mucosal	6.1	5.0
Retromolar area	2.7	5.0
Miscellaneous sites	5.6	--

Variants

Squamous cell carcinoma may present as a number of variants. Some of these are briefly described under the following headings adapted from Batsakis (2003):

- *Verrucous*: A well-differentiated squamous lesion with a "pushing" invasive pattern and exophytic warty surface. It is found in mucous membranes of the oral cavity and larynx. It is also found on cutaneous surfaces in the upper digestive tract, genitalia and extremities.
- *Basaloid*: This is an aggressive variant of squamous cell carcinoma composed primarily of basaloid epithelial cells. It is found in the upper digestive tract and also the oesophagus and uterine cervix.
- *Sarcomatoid*: this is identified by the presence of malignant epithelial cells admixed with malignant spindle cells. There have been several other names for this variant and these include carcinosarcoma, spindle cell carcinoma, pleomorphic carcinoma, metaplastic carcinoma, polypoid carcinoma and pseudosarcomatous carcinoma.
- *Adenoid squamous cell carcinoma*: this is mainly a neoplasm of the skin especially in head and neck areas exposed to the sun, however it does also occur in mucosal sites which include the vermilion border of the lip, tongue, oral mucosa, nasopharynx, and larynx.
- *Adenosquamous carcinoma*: an unusual lesion defined by the WHO as a malignant tumour with histological features of both adenocarcinoma and squamous cell carcinoma occurring in the head and neck, gastrointestinal tract, upper and lower respiratory tract, urogenital tract and also the oral cavity, larynx and paranasal sinuses. In the mouth, 85% of cases arise in the tongue.
- *Nasopharyngeal type carcinoma*: this is an undifferentiated non-keratinising carcinoma with lymphoid stroma that is histologically like the undifferentiated form of carcinoma found in nasopharyngeal carcinoma except that it is not associated with the Epstein Barr virus. This lesion has also been described as lymphoepithelioma and lymphoepithelial-like carcinoma.

- *Papillary squamous cell carcinoma*: this is an exophytic squamous cell proliferation with the overall architecture of a papilloma except that it displays malignant changes. Lesions are often solitary and found in the oral cavity (usually posteriorly), oropharynx, hypopharynx and larynx. They tend to arise from hyperplastic mucosa rather than white patches.

1.3.2 Second primaries

The predisposition of a cancer patient to develop further primary tumours was recognised over 100 years ago, when multiple malignancies in a patient were first described by Billroth in 1889. It was observed that many of these multiple primary neoplasms tended to develop in the same organs or organ systems, although not exclusively (Hun *et al,* 1933). Distinctive features of head and neck tumours are their multifocality and the high incidence of second primary tumours (Ogden, 1991; 1996).

These multiple cancers are classified into:

- Simultaneous tumours: these are diagnosed at the time of the diagnosis of the primary tumour.
- Synchronous tumours: these are tumours that are not detected at the time of diagnosis of the initial cancer but are diagnosed within six months. These presumably were present initially but were not clinically apparent.
- Metachronous tumours: these are diagnosed more than six months after diagnosis of the initial cancer and are known as second primaries.

The occurrence of multiple primary neoplasms in the head and neck is well documented, having been originally described by Warren and Gates in 1932. While the incidence was thought initially to be approximately 10% of patients, it is now regarded as being as high as 30% of patients with upper aerodigestive tract cancer. Day *et al* (1993) with US data found a 10% prevalence in two

years, and Schwartz *et al* (1994) with French data found a prevalence of 22% over a 5-year period. In populations who continued to smoke the occurrence was as high as 36% over a 5-year period (US data: Gorksky and Silverman, 1994). Multicentric cancers tend to develop predominantly in those patients who have a history of both tobacco and alcohol abuse, with the incidence being much lower in those who do not indulge in these habits (Schantz *et al,* 1988). Of interest is the fact that the second malignancy may develop not only in the mucosa of the upper aerodigestive tract but also in the skin, kidneys, bladder, pancreas, colon and rectum (Berg *et al,* 1976). In other words, the mere existence of one tumour implies an increased susceptibility to the development of further malignant neoplasms in the same or related systems.

The 5-year survival rate tends to be lower in those with second primary tumours. In a recent study of 200 individuals, Cianfriglia *et al* (1999) found a prevalence rate of 14% over 3.2 years; 39% arising in the oral cavity, 18% in the oropharynx, 10% in the lung and 7% in both the lip and larynx. 40% of the second cancers were synchronous, 60% metachronous. This study emphasised the high prevalence of multifocal squamous cell carcinoma and advocated the importance of preventive measures and screening procedures.

1.4 Oral precancer/premalignant lesions

A precancerous lesion is defined as a morphologically altered tissue in which oral cancer is more likely to occur than in its apparently normal counterpart (Axell *et al,* 1996). The term "potentially malignant" has been proposed for lesions that have not actually turned malignant. This term is so similar to the WHO definition of precancer in concept that the terms "precancer" and "premalignant" should be considered synonymous. Whilst there is a wide range of mucosal disorders in which squamous cell carcinoma has been shown to develop, it is difficult to assign histopathological degrees of malignant potential. Pindborg (1980) suggested a division between precancerous lesions, which show clinically and histopathologically detectable mucosal changes,

and precancerous conditions, which are more widespread, or systemic disorders affecting the oral mucosa where oral cancer is statistically more likely to develop. A precancerous condition is defined as a generalised state associated with a significantly increased risk of cancer (WHO, 1978). Examples are syphilis, sideropenic dysphagia and oral submucous fibrosis. In the developed world most oral cancers appear to arise in apparently normal mucosa, whereas in South East Asia there is often a premalignant lesion present prior to the onset of cancer (Scully, 1993). The premalignant lesion is often a leukoplakia, erythroplakia or a combination of both of these: a speckled leukoplakia.

1.4.1 Leukoplakia

Leukoplakia is defined as a " predominantly white lesion of the oral mucosa that cannot be characterised as any other definable lesion; some oral leukoplakias will transform into cancer" (Axell *et al,* 1996). Leukoplakia is the most common premalignant or potentially malignant lesion of the oral mucosa. A provisional diagnosis is made when a lesion at clinical examination cannot be clearly diagnosed as any other disease of the oral mucosa with a white appearance; a definitive diagnosis of oral leukoplakia is usually made after histopathological examination. This would serve to exclude any other definable lesion and also to establish the degree of epithelial dysplasia (Axell *et al*, 1996). According to a study of over 23,000 individuals (Bouquot and Gorlin, 1986), the histological features are those of keratosis and varying degrees of cytological atypia. Although epithelial dysplasia is an important predictive factor of malignant transformation, not all dysplastic lesions will become malignant and non-dysplastic lesions may become malignant (van der Waal *et al,* 1997). Although the prevalence of leukoplakia in the United Kingdom is unknown it has been demonstrated in other studies in the Western world as ranging between 2-4% (Bouquot and Gorlin, 1986; Hogewind and van der Waal, 1988). In an adult Swedish population a 3.6% prevalence rate was recorded (Axell, 1976). Oral cancer has been noted as arising in areas of leukoplakia during follow-up and the histological features of some

leukoplakias are those of oral cancer (Speight and Morgan, 1993). Banoczy and Csiba (1976) found epithelial dysplasia in 24% of leukoplakias and of 68 patients with dysplasia, nine developed oral cancer. They also found that leukoplakia on the tongue had the highest incidence of malignant change. In other studies the malignant transformation rate varies from 3-28% (Pindborg, 1980; Bouquot, 1987). In the United Kingdom, Kramer *et al* (1978) showed a malignant transformation rate of 24% for sublingual keratosis. In two studies from India, rather low annual malignant transformation rates of oral leukoplakia have been reported, 0.3% (Gupta *et al,* 1980) and 0.06% (Silverman *et al,* 1976). On the basis of the lowest reported annual malignant transformation rate of leukoplakia, it can be calculated that patients with leukoplakia carry a 5-fold higher risk of developing oral cancer than the controls (Silverman *et al,* 1976). Certain attributes are associated with an increased risk of malignant transformation. These are:

- Gender: women seem to be at greater risk
- Long duration of the condition
- Leukoplakia in non smokers (idiopathic leukoplakia)
- Location in the floor of the mouth or on the tongue
- Non-homogenous type, the speckled variant
- Presence of Candida albicans
- Presence of epithelial dysplasia

1.4.2 Erythroplakia

Erythroplakia has a similar definition to leukoplakia. It is a red patch that cannot be characterised clinically or histologically as due to any other condition (WHO, 1978). Its prevalence is unknown but it is much less prevalent than leukoplakia, although areas of erythroplakia can arise in leukoplakia (Axell, 1996). The site distribution of erythroplakia matches more faithfully the sites for oral cancer and indeed the majority of lesions diagnosed clinically as erythroplakia show invasive carcinomas on histology or contain areas of severe dysplasia or subsequently become

malignant (Shafer and Waldron, 1975; Mashberg and Meyers, 1976). Erythroplakia has the highest risk of developing cancer (Vedtofte *et al* (1987). At least 90% of these lesions have features of carcinoma in situ or a degree of dysplasia (Shafer and Waldron, 1975). The malignant transformation rate has been estimated as 80% (Speight and Morgan, 1993). Treatment is by surgical excision but the recurrence rate is high, possibly up to 40% according to Vedtofte *et al*, (1987). One reason may be the difficulty in defining the margins at excisions. Like leukoplakia, erythroplakia is closely associated with tobacco smoking and alcohol abuse.

1.5 Potentially malignant conditions

There are several conditions of the oral mucosa which appear to be associated with an increased incidence of oral cancer. These include:

1.5.1 Lichen planus

Lichen planus is a relatively common mucocutaneous disease, which involves the skin and/ or mucous membranes. It was first described in 1869 by Wilson, and an association with malignant transformation was suggested at the beginning of the 20th century (Hallopeau, 1910). Lichen planus has a classic appearance of bilateral white lace-like lesions (papular or reticular) on the buccal mucosa and tongue. Other types of lichen planus include atrophic, ulcerative or bullous lesions. The prevalence is approximately 1% (Scully and El-Kom, 1985; Kaplan, 1991) and it predominantly affects those aged 30-70 years, especially women. Evidence from retrospective studies suggests there is a small but clinically important malignant potential for lichen planus (Barnard *et al*, 1993). The incidence of neoplastic transformation in lichen planus has been documented as ranging from 0.4% to 5.6% (Murti *et al*, 1986; Silverman *et al*, 1985; Holmstrup *et al,* 1988). One of the reasons for the controversy existing over the premalignant status of lichen planus is that there are lesions not having the clinical appearance of lichen planus but having lichenoid features histologically as well as lesions resembling clinical lichen planus but with

dysplasia or atypia histologically (Eisenberg and Krutchkoff, 1992). Diagnosis by biopsy is not universal practice (Bamard *et al*, 1993), and this may explain the malignant change which occurs in some cases of "lichen planus". Eisenberg and Krutchkoff (1992) described a lesion with features of lichen planus, and with cytological atypia, as "lichenoid dysplasia". Eversole (1992) concluded that precancerous lesions with lichenoid features may exist separately to oral lichen planus, although lichen planus may also have a slight tendency to undergo malignant transformation.

1.5.2 Syphilis

In the past syphilitic leukoplakia of the dorsum of the tongue was considered an important precancerous condition, however it is now decreasing in incidence. Banoczy (1977) found that 3 % of leukoplakia patients had syphilis and of those patients who subsequently developed oral cancer 10% had syphilis. Hobaek (1946) proposed that the atrophy of the tongue epithelium caused by syphilis may allow an increased effect by aetiological factors such as smoking and alcohol.

1.5.3 Sideropenic dysphagia (Patterson-Kelly syndrome, Plummer-Vinson syndrome)

Alhbom (1936) first pointed out the relationship between sideropenic dysphagia and oral cancer. This association was demonstrated by Wynder *et al* (1957). Sideropenic dysphagia is a collection of symptoms including low serum iron, diminished iron stores, chronic dysphagia and atrophy of the mucosa of the upper gastrointestinal tract in middle-aged women.

1.5.4 Tylosis

Tylosis is an uncommon genetic disorder characterised by hyperkeratosis of the palms and soles. It is associated with carcinoma of the oesophagus and oral lesions and leukoplakia, and is therefore a condition carrying a high risk of malignant change in the mucosae (Field *et al,* 1997; Risk *et al*, 1999).

1.5.5 Oral submucous fibrosis

Oral submucous fibrosis involves a local disorder of collagen turnover and cross-linkage causing progressive immobility of oral mucosa in response to a complex set of factors (Canniff *et al,* 1986). The epithelium overlying dense connective tissue of reduced vascularity is atrophic and may be dysplastic. This chronic disease of the oral mucosa occurs almost exclusively in Indians and is due to the local action of the areca catechu nut. The inflammatory reaction produced is thought to destroy the underlying connective tissue, which then heals by deposition of thick collagen bands. This thicker underlying connective tissue leads to a reduced blood supply resulting in atrophy of the overlying epithelium. There is a 13-14% frequency of dysplasia in this condition (Pindborg, 1980) and it was found to be present in 40% of Indian patients with oral cancer (Pindborg *et al,* 1967). The clinical features of submucous fibrosis include palpable fibrous bands, tough leathery texture and blanching of the mucosa (Zain *et al,* 1997). The histopathological features include atrophic epithelium with loss of rete-pegs and juxta-epithelial hyalinasation of the lamina propria. Haque *et al* (1997) in an immunohistochemical study observed a high ratio of CD4 to CD8 T-cells in submucous fibrosis suggesting that there is an ongoing immune cellular response and this could lead to an imbalance of immunoregulation and an alteration in the local tissue architecture. Murti *et al* (1985) reported a 7.6% malignant transformation over 10 years in follow-up cases of submucous fibrosis.

1.5.6 Lupus erythematosus

This is a systemic auto-immune disease. Its presentation in the mouth is similar to that of lichen planus. Andreasen (1964) reported a 0.5% malignant transformation rate in patients with skin discoid lupus erythematosus. All cases have been associated with the lower lip and tend to be more common in men.

1.5.7 Actinic keratosis

This is a condition which effects the lower lip and tends to be more common in men. It is due to excessive sun exposure and ultra-violet light. About 10% of all actinic keratoses undergo malignant transformation (Lynch *et al,* 1984).

1.6 Natural history and pathology of oral cancer and precancer

Cancer is a form of uncontrolled cell growth due to inherent defects in the DNA. There is normally a balance between the production of new cells and the loss of cells through the process of apoptosis, or via desquamation. There is therefore a balance of cell proliferation and loss. If there is an accumulation of defects in the balance, malignant changes can occur (Sugerman *et al,* 1995). It is likely that the route to invasive carcinoma is not a consistent one, even within a community exposed to apparently similar aetiological factors. In man the most thoroughly investigated progression to malignancy is colorectal carcinoma in which transformation is accompanied by a relatively small repertoire of chromosomal deletions and oncogenic mutations (Fearon and Vogelstein, 1990). In oral carcinogenesis aetiological factors such as tobacco and alcohol consumption may damage cells at the level of DNA in ways that involve oncogenes and tumour suppressor genes (Field, 1995; Cianfriglia *et al,* 1999). Cancer is understood to arise in stages, a process referred to as multistep carcinogenesis. There is a sequence of changes in the tissues and there is usually a long premalignant phase before malignancy develops. There is a

rapid and usually irreversible alteration to the DNA of somatic target cells. Most carcinogens are mutagens and irritation may be potentiated by genetic disorders, for example disorders affecting DNA repair like Xeroderma pigmentosum, or cell proliferation caused by extrinsic agents and factors may consolidate the irritated state. The next step is promotion whereby groups of cells become morphologically and histologically distinguishable from adjacent cells and may progress or regress depending on previous and subsequent stimuli. Continued exposure to promoters leads to progression to a malignant neoplasm. The development of a malignant neoplasm is therefore multifactorial and proceeds by a multistep process. This can be broadly divided into initiation, promotion and progression. Each step represents an alteration or activation of genetic material culminating in the activation of a gene, which results in uncontrolled and independent cell growth. The most important genes are cellular oncogenes and tumour suppressor genes. Fig 1.1 illustrates the sequence of events leading to the expression of clinical cancer and Fig 1.2 shows a time sequence of tumour growth and detection. The incidence, aetiology and natural history of oral cancer and precancer vary considerably in different population groups. Tobacco and alcohol are the major environmental causes although viral infections may also be a factor. The protective effect of adequate nutrition is well demonstrated. As treatment of the local cancer improves however, opportunities arise for further and recurrent tumours to develop according to the concept of field cancerisation (Slaughter *et al*, 1953). This is based on finding satellites of dysplastic epithelium adjacent to the main lesion. These suggest that changes at subcellular or biochemical levels exert such field changes. It seems possible that the disease process can exert a regional effect or that there might be further unidentified aetiological factors (Ogden, 1996).

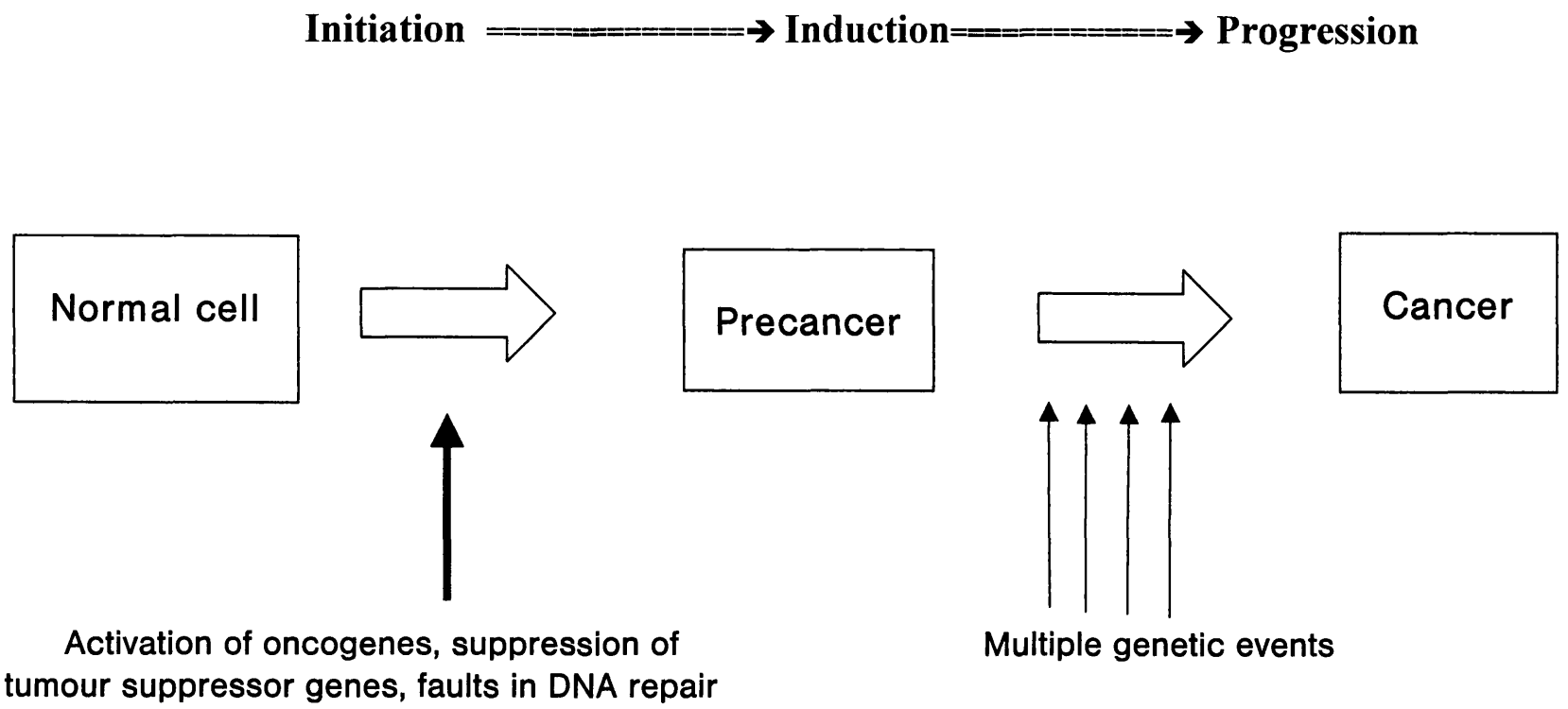


Figure 1.1 Multistage carcinogenesis, the sequence of events leading to cancer

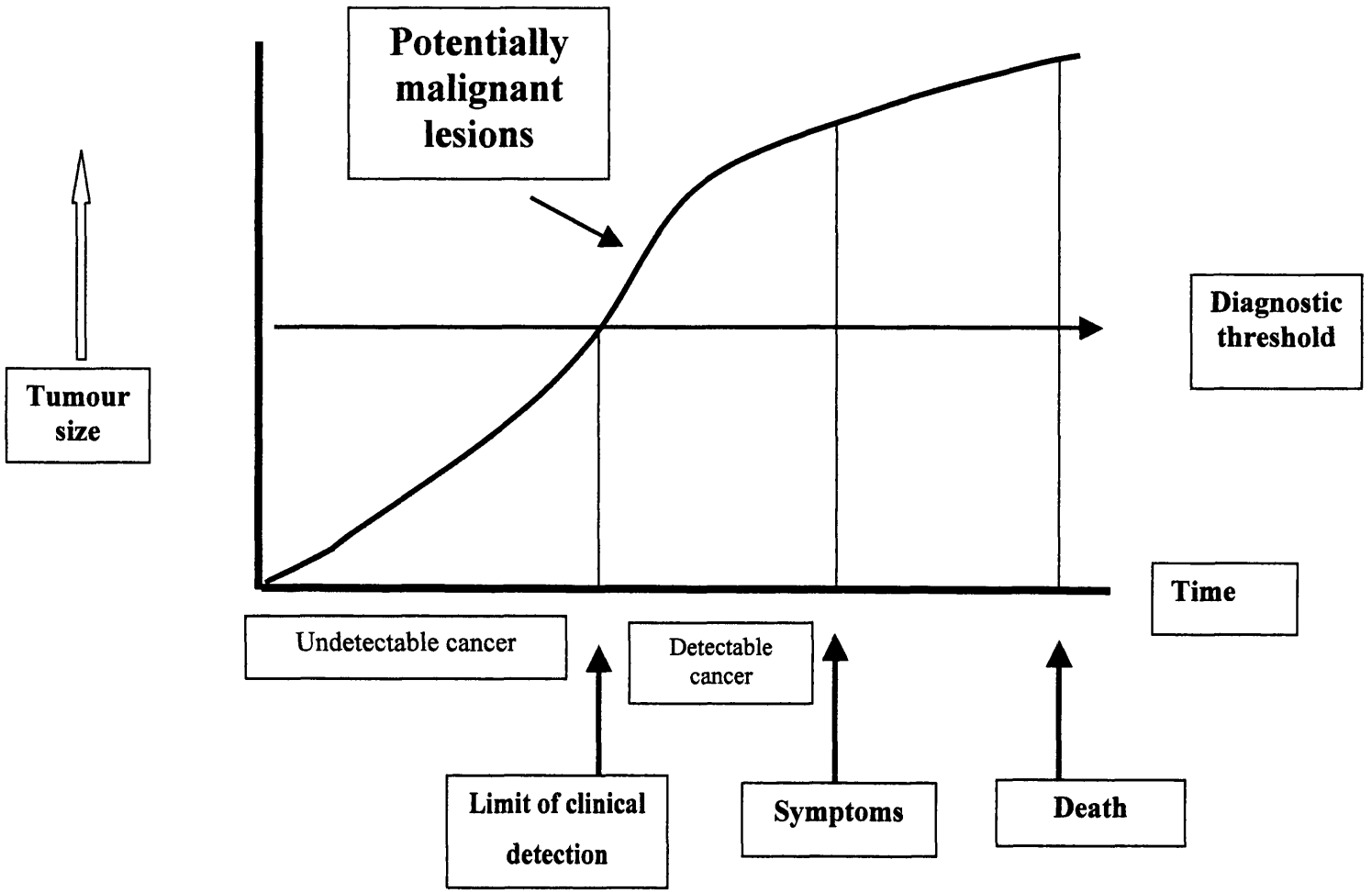


Fig 1.2 Tumour growth and detection

1.7 Aetiology and risk factors

Although it is often not possible to separate the factors and some may exert their effects in a synergistic manner, possible aetiological agents can be grouped as follows:

1.7.1 Tobacco

It is established that tobacco is the most important risk factor for oral cancer. Polycyclic hydrocarbons produced by smoking are broken down by salivary enzymes to produce benzopyrenes, which are proven carcinogens. Smoking in all forms is strongly associated with leukoplakia. Pindborg *et al* (1972) reported floor of mouth lesions in Danish females who smoked cigars. Pipe smokers are likely to develop intraoral keratoses whilst reverse smoking as commonly practised by certain women in the Indian state of Andhra Pradesh is a predisposing factor for carcinoma in the palate. The combined synergistic effect of tobacco and alcohol use is multiplicative in the mouth (Johnson *et al,* 1993). Likewise, tobacco when added to betel quid results in a three-fold increase in the relative risk of developing oral cancer (Johnson *et al,* 1996). Smokeless tobacco is seeing a resurgence of popularity especially among the adolescents in America despite its association with oral cancer (Lewis *et al*, 1999). Snuff dipping is also associated with carcinomas, second primaries, leukoplakia and candidal infection (Sundstrum *et al,* 1982).

1.7.2 Alcohol and related impurities

Pure ethanol has not been shown to be carcinogenic. The role of alcohol is not clearly understood. Most studies and reviews seem to confirm that alcohol does have an independent effect in increasing the risk of oral cancer (Wright and Ogden, 1998). Its role in carcinogenesis may be to simply act as a solvent, facilitating the passage of carcinogens through the cellular membranes. Howie *et al* (2001) suggested that short-term exposure to ethanol may act as a permeability enhancer by molecular rearrangement of the permeability barrier rather than as a result of lipid

extraction. Ethanol also causes liver damage and could activate carcinogens, or may alter intracellular metabolism of the epithelial cells and impair cellular function. There is also the likelihood of direct mucosal damage, which results in chronically increased cell division and thus an increased cancer risk. Bagnardi *et al* (2001) in a meta-analysis found relative risks of developing oral or pharyngeal cancer ranging from 1.8 for the consumption of 20 grams of alcohol a day, to 2.9 for 50 grams a day, to 6.0 for 100 grams a day, not adjusting for smoking. Other carcinogenic sources are some congeners in alcoholic beverages. As alcohol has a high calorific value, its regular and excessive consumption can result in a reduced intake of possible cancer protective substances such as fruit and vegetables.

1.7.3 Alcohol and smoking

The combined effect of smoking and alcohol consumption is synergistic. It is described by Johnson (2001) as super-multiplicative in the mouth, additive in the larynx, and between additive and multiplicative in the oesophagus. Franceschi *et al* (1999), in a study comparing separate and combined effects of smoking and drinking, found extremely elevated risks for oral cancer for the highest joint level of smoking and drinking (>77 drinks/week, >25 cigarettes/day) with an odds ratio of 228 for oral cancer and 100 for pharyngeal cancer. They further concluded that if smoking levels in a population do not change but alcohol consumption increases, the increase in oral cancer would be greater than at any other site in the upper aerodigestive tract, including cancer of the pharynx.

1.7.4 Areca nut/paan chewing

Areca nut is the fourth commonest addictive substance after tobacco, alcohol, and caffeine and is used by 600 million people worldwide (Warnakulasuriya, 2002). Areca nut alone or in combination with other ingredients like tobacco, lime catechu (extract of Acacia catechu), sweeteners, and spices, wrapped in betel leaf (Piper betel) is referred to as paan or betel quid. Areca nut alone or as a component of paan/betel quid is now known to be carcinogenic even

without added tobacco (Van Wyk *et al*, 1993; Thomas and Kearsley, 1993; Warnakulasuriya, 1995). The risk of oral cancers is vastly increased if tobacco is added to the quid and/or smoked; its effect is potentiated by alcohol. Areca nut is therefore seen as an independent risk factor for oral cancer and is also implicated in the development of leukoplakia and oral submucous fibrosis, both of which are potentially malignant (Warnakulasuriya, 2002).

1.7.5 Marijuana use

There have been case reports of oral cancers in marijuana smokers (Donald, 1986; Almadori *et al,* 1990). These have yet to be supported by epidemiological studies. It is also uncertain how much of the risk is due to the smoking per se, although a recent study noted that an increasing number of young patients are being diagnosed with lung and aerodigestive cancers even with no history of tobacco smoking (Fung *et al,* 1999).

1.7.6 Mouthwash use

There is controversy over the possibility that regular use of mouthwashes containing high levels (25% or higher) of alcohol could be dangerous. It would appear that the risk is small and likely only to be expressed in smokers (Johnson *et al*, 1996). However, a large US study (Winn *et al,* 1991) found that, after adjustment for tobacco and alcohol use, the risk from alcohol in mouthwashes is similar, at least qualitatively, to that of alcohol used for drinking, although in terms of attributable risk, the contribution of mouthwash use to oral cancer remains small (Scully, 1995; Winn, 2001).

1.7.7 Other liquids

Some types of tea (mate) consumed in countries like Brazil can be associated with oral cancer (Scully, 1995).

1.7.8 Fungal infection

Dysplastic lesions showing candidal infection have a greater risk of malignant transformation. A mechanism for their aetiological role does exist as enzymes produced by Candida albicans are capable of producing known chemical carcinogens by nitrosination (Krogh *et al,* 1987). O'Grady and Reade (1992) were able to demonstrate neoplastic changes in an experimental group of rats by exposing them to a promoter (a phobol ester) and inducing a candidal infection. Neoplastic changes then occurred in week 34. There appears to be an association between oral candidosis and iron deficiency. Iron deficiency could result in an inadequate host response to the fungus and could lead to a persistence of the infection and its deleterious effects (Binnie *et al,* 1983). Barrett *et al* (1998) reviewed fungal infections in biopsies and found a statistically significant association between fungal infection and epithelial dysplasia. Where subsequent biopsies were available, such dysplasias where fungal infection was present worsened in histological severity.

1.7.9 Viral infection

There is strong evidence to implicate viruses in human cancers (Gross *et al,* 2001). Some well known associations are:

- Epstein Barr virus (EBV) with nasopharyngeal carcinoma and Burkitt's lymphoma (Shimakage *et al,* 2002)
- Hepatitis B and hepatocellular carcinoma (Smukler and Ratner, 2002)
- Human T cell lymphotropic virus (HTLV) with types of lymphoma and leukaemia (Siegel *et al,* 2001)
- Human Papillomavirus (HPV) with cervical carcinoma, and oral and tonsillar cancer (Smith *et al,* 1998). Miller *et al* (2001) in a meta-analysis concluded that HPV is detected with increased frequency in oral dysplastic and carcinomatous epithelium compared with normal mucosa and that oral infection with HPV is an independent risk factor for oral squamous cell carcinoma.

1.7.10 Ultraviolet radiation

Ultraviolet radiation is a carcinogen for the various forms of skin cancer, including basal cell carcinoma, squamous cell carcinoma, and malignant melanoma. This is specially the case in fair skinned individuals living in countries where there is a lot of sunshine, for example Australia (Johnson, 1991). Ultraviolet radiation is also an important factor in the development of squamous cell carcinoma of the vermilion border of the lip. Individuals who have conditions like Xeroderma pigmentosum are especially vulnerable to ultraviolet radiation.

1.7.11 Nutritional deficiencies

Iron is essential for the overall integrity and health of the upper gastro-intestinal tract and its importance may lie in its contribution to normal enzyme systems. There is a clear association between iron deficiency and cancer of the upper aerodigestive tract (Wynder *et al,* 1957). The effect of iron deficiency may in fact be secondary to pyridoxine, folic acid and vitamin B_{12} deficiency. Plasma vitamin A and beta-carotene were found to be significantly lower in patients with oral cancer and remissions from leukoplakia were reported with the administration of beta-carotene and beta-carotene with Vitamin A (Stitch *et al,* 1988). Studies suggest that a diet rich in fruit and vegetables may protect against oral cancer. However, in a recent review by Marshall and Boyle (1996), the case-control evidence was not highly consistent, and individual foods that appeared protective in some studies did not in others. The effects of nutrition were not as strong as the effect of tobacco and alcohol use. In a wider context, a diet rich in fruit and vegetables has consistently been shown to be protective against most epithelial cancers including cancers of the gastro-intestinal system and breast. Reductions of 21% of breast cancers, 60% of stomach cancers and 43% of colorectal cancers have been reported (Mayor, 2002).

1.7.12 Immune deficiency/suppression

There is evidence that HIV-associated immunosuppression increases the risk of specific cancers including lip cancers (Gallagher *et al,* 2001). It is possible that immunosuppression, in the form of HIV infection or drug-induced, predisposes the patient to oncogenic viruses. In some cases patients who have malignancies demonstrate a degree of immunosuppression (Gangal and Tatake, 1991). It is uncertain if the malignancies are in fact the result or the cause of the immunosuppression. Thomas *et al* (1993) reported a case of squamous cell carcinoma of the lower lip associated with immunosuppression. In a study of immunosuppressed patients after having receiving renal transplants, King *et al* (1996) also noted a significantly higher risk of developing oral mucosal disease including lip cancer.

1.7.13 Socio-economic status

Oral cancer is strongly related to social and economic deprivation with the highest rates occurring in the most disadvantaged sections of the population. This association is particularly strong in males, where the risk of men developing a cancer was over four times greater in the most deprived compared to the least deprived categories (CRC, 2000). Pukkala *et al* (1994) found lip cancer five times more common in the lowest social class compared to the highest.

1.7.14. Dental health

Poor oral hygiene is believed to play a role as a risk factor although this has not been conclusively proven (Maier *et al,* 1993; Marshall and Boyle, 1996; Blaram *et al,* 2002).

Part A

General Introduction

Chapter 2

Screening

Chapter 2~Screening

2.1 Introduction

Screening is a part of preventive medicine and arises from the concept that treatment of diseases early in their development offers the best chance of cure. The purpose of prevention is to reduce the risk of a person contracting a disease or to reduce the risk of subsequent disability once a disease has happened (Butler, 1993). There are three types of prevention:

- Primary prevention: which is aimed at preventing the disease from starting, such as health awareness and health promotion campaigns, for example for the cessation of smoking.
- Secondary prevention: which involves the early detection and prevention of further progression of a disease, for example screening.
- Tertiary prevention: which is aimed at reducing any disabilities arising from a disease and instituting appropriate rehabilitation.

Screening is part of secondary prevention and can be defined as "the application of a test or tests to people who are apparently free of the disease in question in order to sort out those who probably have the disease from those who probably do not. Screening is not intended to be diagnostic and individuals with the disease or suspicious findings are to be referred to their health professionals for diagnosis and necessary treatment" (Junger and Wilson, 1968). This should therefore interrupt the natural history of the disease and prevent progression to advanced disease and death (Chamberlain, 1993).

There are two main strategies or approaches for any preventative programme (Downer, 1994): the population-based approach and the high-risk strategy.

- The population-based approach involves altering the general factors in a population which contribute to the overall causes of the disease. The main advantage is that it has a greater potential of being effective as it targets the underlying cause as well as being behaviourally appropriate for the population, but suffers the disadvantage of the

prevention paradox, which is that although this approach may provide large benefits to the community, only small benefits may be obtained on an individual basis.

- The high-risk strategy involves the identification of individuals at high risk of contracting the disease. Targeted screening provides a more efficient means of case finding. The advantage is that individuals identified can be given very specific counselling which could be done on a one to one basis between the patient and the health professional. The disadvantages are that it may be difficult to identify or reach the vulnerable groups for targeting, the cost may be high and the compliance low. Targeted screening may be an efficient means of identifying cases in diseases of low prevalence, for example oral cancer in the United Kingdom.

2.2 History of screening

In Britain, screening has been practised for more than 80 years in the routine examination of school-age children and in routine postnatal care. Screening for communicable diseases such as pulmonary tuberculosis was introduced in 1943, enabling treatment at an earlier stage and allowing affected persons to be removed from the community to prevent further spread. Screening for endemic industrial disease, for example silicosis by chest X-rays, was also implemented around that time. With increased medical knowledge and widespread use of antibiotics, more emphasis was placed on non-communicable diseases such as cancer and heart disease. Screening for cervical cancer in the United Kingdom commenced in the 1960s and may have resulted in a 20-30% reduction of mortality. Hakama (1990) reported a 60% reduction in cervical cancer incidence. Screening for breast cancer in the United Kingdom began in the mid-1980s; a Swedish study reported that population screening in the 40-69 age groups was associated with a 21% reduction in mortality from breast cancer. There are however, some recent dissenting opinions expressed about the efficacy of breast screening (Baum, 2002).

2.3 Principles and criteria for screening

A series of general principles for screening were proposed by Wilson and Junger in 1968 and these have been widely adopted. These are:

- The condition for which screening is undertaken should be an important health problem.
- There should be an accepted treatment for the cases identified.
- Facilities for diagnosis and treatment should be available.
- There should be a latent or early asymptomatic stage.
- There should be a suitable test or examination.
- This test should be acceptable to the population.
- The natural history of the condition should be understood.
- There should be an agreed policy on whom to treat as patients.
- The cost of case finding should be non-wastefully balanced in relationship to expenditure on medical care as a whole.
- Case finding should be a continuing process and not a once and for all event.

2.4 Methods of screening

There are many situations where individuals could be screened, for example risk assessment of an applicant for life insurance, or for employment in health care professions where applicants are screened for hepatitis and tuberculosis, or to diagnose conditions like hypertension and phenylketonuria. As this study deals with cancer, it would be appropriate to consider methods of screening for cancer.

2.4.1 Observation

This is the most widely available method for the detection of cancer. It is useful in identifying suspicious lesions in the skin, retina, lip, mouth, larynx, and external genitalia. Internal cancers

require an extension of observation using clinical instruments like endoscopes, radiographs, magnetic resonance imaging and ultrasound. The use of metachromatic dyes further aids observation and was described by Rickard (1963). This can be used to detect neoplastic areas at mucosal sites and has been used on the cervix and in the oral mucosa (Warnakulasuriya and Johnson, 1996).

2.4.2 Palpation

The second most available detection procedure is palpation. This is particularly valuable in detecting lumps, nodules, or tumours in the breast, lip, mouth salivary glands, thyroid, subcutaneous tissues, anus, prostate, testes and enlarged lymph nodes in the head and neck, axilla and groin.

2.4.3 Laboratory-based tests

The use of laboratory-based tests such as exfoliative cytology, needle and brush biopsy techniques and occult blood testing of faeces have proved useful.

2.5 Methods of recruitment

There are two basic ways of recruiting individuals to a screening programme (Fowler and Mant, 1990):

2.5.1 Population screening

This is one of the most popular and possibly the most efficient method. Individuals are specifically invited to attend for a screen. The population selected is usually based on a general medical practitioner list. Under the present NHS contract, general medical practitioners are required to invite patients to attend for health screens at specific mandatory intervals. Population screening usually involves some form of targeting whereby the appropriate populations, for

example women between the ages of 50 and 64 in the case of breast cancer, are invited to be screened.

2.5.2 Opportunistic screening

This method involves offering a screen when patients attend a clinic for some other, unrelated reason. This is ideally augmented by patient information leaflets to raise public awareness of the disease to be screened for. The public likes to receive information about their health and such information tends to result in a more positive attitude to health and disease prevention (Marteau, 1990). Other opportunities for recruitment include offering screens in the workplace, funfairs, residential homes, and the armed forces. A variant of opportunistic screening is targeted opportunistic screening as in a study by Dombi (2001), in which oral cancer and precancer screening was done at a lung clinic and screened attendees who were mainly heavy smokers with lung diseases.

2.6 Potential advantages and disadvantages of screening

2.6.1 Potential advantages

2.6.1.1 Decrease in mortality

The major objective of screening is to "reduce the risk of death from the disease being screened for" (Hakama, 1990). Thus the most significant potential advantage of screening is a reduction in mortality. Adami *et al* (1994) compared the survival rates of invasive cervical cancer between 1960 and 1964 (pre-screening) and 1980 and 1984 in Sweden. They found that screening reduced mortality by more than half in women less than 40 years of age. In older women (where screening was less extensive) there was only a slight improvement between the two time periods. They concluded that although there may be other explanations for the improved 5-year survival rates over the 20 years, the most obvious explanation was that screening reduced

mortality by earlier diagnosis of invasive disease. For breast cancer, Blamey *et al* (1994) stated that there was an overall reduction in breast cancer mortality in Sweden over a 12-year follow up period of 29% in women older than 50 who were invited to screen and 13% in younger women. Mandel *et al* (1993, 1999), in a study in Minnesota, demonstrated that annual faecal occult blood testing (FOBT) decreased mortality from colorectal cancer by 33% and biennial testing resulted in a relative mortality reduction of 21%.

2.6.1.2 Decrease in morbidity

Beyond the potential benefits of avoided deaths, screening may reduce cancer morbidity in the population. Generally, treatment for early stage cancers tends to be less aggressive than for more advanced cancers. However, like mortality, as an end point morbidity may be difficult to measure.

2.6.1.3 Cost savings

Cancers detected at their early stages tend to require less invasive treatment. Oral cancer and malignant melanomas are examples where, if a lesion is identified early, the treatment is a simple excision, whereas local invasion or metastatic spread could significantly increase the overall costs of treatment.

2.6.1.4 Improving quality of life

The benefits of screening leading to earlier diagnosis and successful treatment can be assessed not only in terms of financial cost but also in terms of the benefits of restoring a person to a normal quality of working and family life.

2.6.1.5 Reassurance for those screened negative

Individuals screening negative would be reassured and have positive habits reinforced especially if the screening programme included health counselling.

2.6.2. Potential disadvantages

2.6.2.1 Selection bias

The outcome of a screening programme can be influenced by selection bias. The more health aware an individual is, the more likely they are to attend a screening programme. This phenomenon is favourable for screening for breast cancer since this is more commonly found in patients of Social class I and II who are more likely to seek preventative care. In screening for cervical cancer, the uptake is low since it is more prevalent in the lower social classes (Williams, 1992). Oral cancer is in this category as it occurs up to three times more frequently in social class V as compared to social class I (Townsend *et al,* 1988).

2.6.2.2 Length time bias

Length time bias is related to the natural history of a disease and may also be unique to each individual. A screening programme is more likely to detect a disease with a long clinical phase than a short preclinical phase. It is thought that diseases with a long preclinical phase tend to have a long clinical phase and therefore have a better prognosis; screening therefore will have little benefit. Conversely there are certain types which are aggressive with short clinical phases which are less likely to be detected by screening. This is a problem encountered in screening for a disease such as neuroblastoma where there are variant types, some which have a good prognosis and regress spontaneously (in which case screening is not useful) and others which are aggressive with short clinical phases and therefore will not easily be detectable by screening (Gordis, 1994).

2.6.2.3 Lead time bias

Is the disease merely detected earlier? This argument relates to the benefit that a patient may derive by being diagnosed at an earlier stage of their disease so that the clinical survival time may appear to be longer than if the lesion had actually surfaced by self presentation (Eddy, 1980). One may argue that screening could have deprived the patient of time spent in ignorance of the disease

and free of medication. For this bias to be minimised, it is important that there be a noticeable increase in survival and quality of life by early detection.

2.6.2.4 Over-diagnosis bias

The phenomena of over-diagnosis relates to the diagnosis of a disease detectable by screening which may never have been clinically diagnosed later. A mass screening programme for childhood neuroblastoma by assessment of urinary catecholamine metabolites was conducted in Saitama Prefecture, Japan from 1981-1992 for 6-month-old infants. Over that 12-year period, the annual incidence of neuroblastoma in children under the age of one increased from about 28 to 260 per million (Yamamoto *et al*, 1995). There was also no reduction of mortality from the disease. Similar experiences have been reported elsewhere in Japan (Bessho, 1996). It is thought that some premalignant states may regress spontaneously, meaning that treatment was not required. It can then be argued that these states or lesions must be treated since a "wait and see" policy would defeat the objective of early detection if a lesion were allowed to progress. A major problem of screening for any type of cancer is the lack of absolute knowledge of progression and regression rates for precancerous lesions. Further studies of disease progression are therefore essential in order to truly assess the value of detecting and treating precancerous lesions (Speight and Morgan, 1993).

2.6.2.5 Psychological trauma

Individuals with a false positive screen would be subject to anxiety and the possible further trauma of further diagnostic tests. Individuals screened negative but who indulge in high risk habits like areca nut chewing, heavy drinking and heavy smoking are at risk of having these habits reinforced.

2.6.2.6 Costs

There is no accepted gold standard for estimating the costs of an intervention like screening and many assumptions need to be made in deriving the costs of such an exercise. The cost for avoiding a death from cervical cancer was estimated to be £300,000 and £80,000 for breast cancer. These far exceed the benefit: premium ratio in the National Health Service that was set in 1985 at £14,000 per person (Downer and Speight, 1993). Screening therefore can be expensive and costs are difficult to quantify, as they often have to be made on numerous assumptions (Kernick, 2000).

2.7 Ethical and psychological problems of screening

It is generally accepted that if a screening programme is to be implemented, there is an ethical obligation to ensure that it does more good than harm to participants and that any adverse psychological effects are negligible (Flynn, 1991; Austoker, 1994). The lung cancer screening programme in the United Kingdom, for example, was abandoned since it failed to show any benefits between the screened and unscreened populations in terms of mortality rates (Fontana, 1985). Marteau (1990) demonstrated high levels of anxiety in patients participating in screening programmes. She suggested that this screen-related anxiety could be reduced by explaining the reason for selecting a particular patient and the benefits to that individual. Edwards and Hall (1992) stated the necessity of informed consent. Lerman and Rimer (1993) concluded that no screening programme should be initiated without planning how the communication of an abnormal result and its follow-up were to be dealt with. This was to avoid the distress associated with receiving abnormal screening results. This in turn was seen to adversely influence participation in subsequent screening and diagnostic follow-up. Brandberg *et al* (1993), however, in a study in which psychological reactions to melanoma screening were measured, showed no increase in psychosomatic problems, anxiety, or false sense of security, but an increased subjective susceptibility to melanoma was demonstrated. The potential harm of a screening programme can be reduced by careful planning and by careful consideration of the effects on

individuals and the population. The psychological costs include increased levels of anxiety, and possible further unnecessary investigations. The exact reason for the tests should be explained to the patient, in writing and, if possible, verbally. The nature of the disease, the level of risk, and the disorder being screened for must be clearly explained. The patient must be told the positive benefits of the screen, the chances of a positive result, and the nature and availability of effective treatment. All patients screened positive or negative must be told of their result and patients screened positive should be referred for specialist attention within days of receiving the result. Negative patients should be told when the next screen is due and can be given advice on how to maintain their low risk state.

2.8 Screening for cancer

Screening is regarded as a cost effective and clinically useful approach for the early diagnosis of several malignancies, especially breast and cervical cancer (Miller *et al,* 1991). A screening programme should have:

- Set protocols
- Adequately trained staff and calibrated screeners
- A clear definition of what constitutes a positive lesion
- Good patient management which would include health education, reinforcement of good habits, discouragement of bad ones and counselling for those found to have a positive condition or lesion
- The capacity to repeat the screening procedures at set intervals, depending on the length of the pre-clinical detectable phase and the known behaviour of the disease. This is often a compromise which balances the number of patients who develop the disease in between screens (the interval cases) and the costs to the health service.
- A specialist facility to manage and treat those found to have positive lesions or conditions

2.9 Examples of screening programmes

2.9.1 Breast cancer screening

The NHS breast screening programme was set up in 1988 by the Department of Health and became the first population screening programme of its kind in the world. The programme began in 1990 by inviting women aged 50 and over for mammography and achieved national coverage in 1993. Women aged between 50 and 64 are routinely invited for breast screening every three years. Recent meta-analyses have demonstrated a mortality benefit approaching 30% in women over the age of 50 seven to nine years from the start of the trials (Kerlikowske *et al*, 1995; Kerlikowske, 1997).

2.9.2 Cervical cancer screening

Cervical cancer is the 12th most common cause of death in women in the UK. Screening for cancer of the cervix by cytology has been shown to be effective in reducing mortality from the disease (Hakama, 1990; Parkin *et al*, 1985; Adami *et al*, 1994). In Britain, the cervical screening programme began in the mid 1960s. Women between the ages of 20 and 64 are eligible for a free cervical smear at least once every 5 years, the actual intervals depending on the individual Health Authorities. There were 2,740 new cases of invasive cervical cancer in England and Wales in 1997; this was a fall of 26% over the previous 5 years with 9.3 cases per 100,000 (Office of National Statistics, 1998). The risk of developing cervical cancer is increased four-fold in women who are on the contraceptive pill who also test positive for the DNA of HPV 9 (WHO, 2002). 99% of all women in Britain in whom cervical cancer is diagnosed are positive for HPV, an infection that affects women in their twenties (Dyer, 2002). Since 1991 mortality has fallen at an increasing rate, which is believed to reflect the improvements made to the screening programmes. Deaths from cervical cancer are falling by 7% a year (National Office of NHS cancer screening programme 2002). Screening is estimated to save 1,300 lives a year (Patnick, 1999).

2.9.3 Colorectal cancer

In the UK, 33,000 people are diagnosed with colorectal cancer each year. This number is increasing by 1% every year for men while staying the same for women. In Scotland the number of men diagnosed is staying the same with slight increases for women. Colorectal cancer is the second most common cause of death in the UK. The National Screening Committee reviewed the evidence in colorectal cancer and found evidence that population screening of people over 50 for occult blood in the faeces may reduce the mortality rate for this cancer (Hardcastle *et al*, 1996). Two pilot schemes were set up in 1999, one in Coventry, Warwickshire in England and one in Fife, Tayside and Grampian in Scotland; these were seen to be representative of the UK population and included a mix of urban and rural populations.

2.9.4 Gastric cancer

Gastric cancer is a major cause of death worldwide, especially in developing countries. In Japan there has been a significant decrease in the gastric cancer death rate which has been attributed to the national mass screening programme that was initiated in the 1960s, although clinical evidence is lacking. Most studies show a twofold decrease in screened versus unscreened populations (Murakami *et al*, 1990; Kampschoer *et al,* 1989; Oshima *et al*, 1986).

2.9.5 Other cancers considered for screening

There are a number of cancers that have been advocated as good candidates for screening including malignant melanoma, ovarian cancer and prostate cancer. There is no screening programme in place for malignant melanoma in the UK although the Cancer Research UK website advocates self examination for those who are considered to be at high risk - the fair skinned, freckled, and those who burn easily in the sun.

There are presently two main screening strategies for ovarian cancer, one using CA 125, a serum tumour marker as a primary test and the other using transvaginal scanning as a primary test (Lewis and Menon, 2003). A pilot randomized trial for ovarian cancer in the United Kingdom

allocated 10,977 women to a control group and 10,958 women to a screened group in 1989 (Jacobs, 1999). The primary screen was CA 125, followed by ultrasonography if CA 125 was elevated. Women were offered 3 annual screening rounds and both groups were followed for 7 years. Compliance was 70.7% for all 3 screenings and 85.5% for at least 1 screening. There were 20 ovarian cancers in the control group and 16 in the screened group, only 6 of which were detected by screening. However, the outcome for women with ovarian cancer in the control group was unexpectedly poor. A larger trial is required to fully assess screening benefits.

Screening for prostate cancer in the UK is not available as there is no evidence that it is justified, although it is under consideration. This causes concern both with the public and within the medical profession because the incidence of the disease is rising, and about 25% of patients present with metastatic disease. Underlying the debate about prostate cancer screening is the fact that many early tumours may not be life-threatening, with patients dying with the disease rather than of it. It is difficult to predict which ones will become life-threatening. In addition, there appears to be no consensus about the best strategy for treatment of early tumours. Conservative therapy with no immediate treatment is a viable option for some cases, but each form of active treatment has significant side effects. There is no evidence that screening and treatment will reduce mortality from prostate cancer. Screening trials are in progress in Europe and the USA and results are expected in about ten years time.

2.10 Oral cancer screening

Oral cancer is easy to detect at all stages of development and diagnosis of small lesions results in simple treatment, less morbidity, maintenance of function and better survival rates (Warnakulasuriya and Johnson, 1996). Oral cancer may be preceded by a clinically detectable precancerous lesion like leukoplakia (Pindborg *et al,* 1991) and screening may offer the opportunity to reduce the incidence of invasive lesions. The majority (95%) of oral cancers are squamous cell carcinomas and should be amenable to primary prevention by health education on minimising lifestyle habits such as excessive alcohol consumption and smoking.

Much has been reported on the year-on-year increase in new registrations of oral cancer (Boyle *et al,* 1993; Renson, 1990; Hindle and Nally, 1991; Hindle *et al*, 1996); and there are concerns that a large percentage of patients present late for treatment with lesions larger than 4cms when the prognosis is significantly worse (Speight and Morgan, 1993). The need for screening has certainly been established, although it is widely agreed that due to the relatively low incidence of oral cancer population screening cannot be advocated as it appears not to be cost-effective (Speight *et al*, 1993;Warnakulasuryia *et al* 1996; Rodrigues *et al*, 1998).

2.10.1 Is oral cancer a screenable disease?

The suitability of oral cancer as a screenable disease is measured against the criteria laid down by Wilson and Junger (1968) and arranged in three tables: first, those criteria which are fully met (Table 2.1a); second, those criteria which are partially met (Table 2.1b); and finally those criteria which are not met or are not applicable (Table 2.1c).

Oral cancer therefore satisfies many, though not all, of the criteria of a disease suitable for screening according to those set by Junger and Wilson (1968). Due to its relatively low incidence and prevalence, however, it is difficult to envisage acceptance of population screening without further studies into natural history, cost benefits, prevalence and available human resources.

2.11 Previous oral cancer screening studies

It is relevant at this point to differentiate between observational prevalence studies which record the presence and absence of oral cancer and precancerous lesions and conditions in a population group, and screening studies where individuals are screened and the performance is evaluated for validity or accuracy of diagnosis. The actual number of such screening studies for oral cancer with measures of performance have been few. Moles *et al* (2002) in a meta-analysis of six studies involving seven such screening programmes, possibly the only significant work done in this respect, reported good overall sensitivity and specificity despite differences in population groups, locations, and types of screener. The pooled sensitivity was 0.796 with specificity of 0.977.

This suggests that screening for oral cancer and precancerous lesions and conditions can be a viable proposition in a general practice environment. Table 2.2 shows the performance of these screening studies according to author(s), year, country, and types of screener involved, numbers screened, prevalence, sensitivity and specificity. "Primary Health Care workers" refers to district nurses, health visitors and other hospital staff. Dental surgeons referred to in the Julien *et al* (1995) studies were dental surgeons working in the hospital and university service. "GDPs" mentioned for the Ikeda *et al* 1995 work refers to general dental practitioners. These, and further prevalence studies are discussed.

Table 2.1a The suitability of oral cancer as a screenable disease: the criteria which are *fully met.*

The condition for which screening is undertaken should be an important health problem	Oral cancer is a serious and increasing problem in developing countries. Although oral cancer is relatively uncommon in the United Kingdom, there are around 3,500 new cases a year, which is about 1% of all neoplasms and is comparable to multiple myeloma or cervical cancer (CRC, 2000). The incidence is rising in many countries in Western Europe and increases in incidence and mortality have occurred in younger males during the last 30 years (Hindle *et al,* 1996). These (although for females the trend is not as pronounced) are associated with clear birth cohorts. The implication of this study is that there may be a future increase in new cases as these birth cohorts enter the age groups with higher disease frequencies. It would appear that oral cancer is an important health problem now and will become even more so in the future.
Facilities for diagnosis and treatment should be available	Cases of oral cancer are usually referred to maxillofacial surgeons, ear nose and throat specialists, plastic or general surgeons depending on the referring practitioner. There are many centres in the United Kingdom equipped to deal with oral cancer and precancer. Patients undergo a number of investigations to determine a definitive diagnosis. The use of tolonium chloride and a brush biopsy technique are recent additions to the armamentarium, but a surgical biopsy is regarded as the gold standard.
There should be a suitable test or examination	The most appropriate test for oral cancer is a simple visual examination. This is generally well accepted by patients and can be carried out effectively by health professionals and health workers. In a recent meta-analysis, Moles *et al,* (2002) described seven substantive or pilot screening programmes and found sensitivities ranging from 0.60-0.95, with specificities ranging from 0.81-0.99. The method used in all of these programmes was a visual examination. The brush biopsy technique has not been evaluated in screening tests although there is currently a screening study in general dental practice involving the use of tolonium chloride.

Table 2.1a (cont'd) The suitability of oral cancer as a screenable disease: the criteria which are *fully met.*

The test or examination should be acceptable to the population	The accepted test for oral cancer and precancer is a visual examination. Examining the mouth is generally acceptable since almost everyone in the western world would probably have had their teeth examined at some point in their lives. An oral examination is not invasive and is generally regarded as being neither unpleasant nor unsafe.
There should be an accepted treatment for the cases identified	Small oral cancer lesions are simply excised with a healthy margin and have a better prognosis than those larger than 4 cms (Speight and Morgan, 1993; Platz *et al,* 1986; Silverman, 1988). Larger and more aggressive tumours are extensively resected, followed up by chemotherapy, radiotherapy and rehabilitation, generally involving a team approach including surgeons, oncologists, physiotherapists and speech therapists. In recent years surgical techniques and rehabilitation have improved resulting in reducing morbidity, although there has been no significant effect on mortality (Stell and McCormick, 1985).

Table 2.1b The suitability of oral cancer as a screenable disease: the criteria which are *partially met.*

There should be a latent or asymptomatic stage	It is recognised that some oral cancers are preceded by a precancerous or potentially malignant lesion and this may make oral cancer amenable for early detection (Pindborg, 1991). The ideal time to detect oral cancer would be the latest point at which further progression could be prevented by treatment (Chamberlain, 1993) or by removal of the risk factors such as tobacco (Gupta *et al,* 1986). The point of transformation for oral cancer is not yet known. Some oral cancers indeed can arise *de novo* (Guggenheimer *et al,* 1994). What is also unclear is the percentage of precancerous lesions which undergo malignant transformation and what the malignant potential of precancerous conditions or lesions is (Boyle *et al,* 1995).
There should be an agreed policy on whom to treat as patients	Histologically confirmed oral cancer (squamous cell carcinoma) is uncontroversial, however differences of opinion exist in the management of precancerous lesions. Kleinmann *et al* (1993) called for an internationally agreed consensus on what should be considered as a cancerous or precancerous lesion, as this would allow for comparisons between studies and lead to more uniform treatment modalities.

Table 2.1c The suitability of oral cancer as a screenable disease: the criteria which are *not applicable or not met.*

The natural history of the condition should be understood	The natural history of oral cancer is not well understood and it may not always be possible to detect oral cancer in its initial stages. There appears to exist a phase, described as the detectable preclinical phase, at which point a sensitive test will detect the disease before the patient would normally have developed symptoms (Eddy, 1980). There are several states associated with a significantly increased risk of oral cancer including leukoplakia and erythroplakia. Another problem is the unknown rate of progression from the precancerous state to the cancerous state. Speight and Morgan (1993) estimated that it may take up to 15 years for a white patch to develop into oral cancer; therefore detecting a precancerous state in a 70-year-old person may be of interest but not of consequence since the treatment cannot be justified in terms of prevention or cost saving. Franceschi *et al* (1997) suggested that the sojourn time for preclinical cancer could be in the order of one year.
Case finding should be a continuing process and not a once-and-for-all event	Once a screening programme is in place, it is important to continue follow-up. As there are presently no national screening initiatives for oral cancer, this criterion is not yet met. However there is a case for a programme of opportunistic screening in general dental practice. This, if in place, would allow longitudinal follow-up of individuals already screened.

Table 2.1c (cont'd) The suitability of oral cancer as a screenable disease: the criteria which are *not applicable or not met.*

The cost of case finding should be balanced in relationship to expenditure on medical care as a whole	There are no documented studies in the United Kingdom for the cost of oral cancer screening versus the cost of treatment of lesions. The relatively low incidence of oral cancer and the high cost of human resources and accurate monitoring make it difficult to justify. Downer and Speight (1993) described a method based on that formulated by Eddy (1986) which compares the effects of primary and secondary intervention in oral cancer in Sri Lanka. Based on hospital statistics, Sankaranarayanan *et al* (2000) estimated the population sizes required to be screened in a high incidence country. Assuming a mortality reduction of 50% in the intervention arm compared with the control arm, they recommend a sample size of 30,000 in each year period. Assuming a compliance of 80%, they predict early detection should result in 50% fewer deaths in the screened population. In a recent UK study using simulation modelling, it was demonstrated that selective screening could avoid 15 deaths per 100,000 individuals examined. Assuming a compliance rate of 50% the net benefit of screening was the equivalent of 2.8 lives saved (Downer *et al* 1997 a,b, 1998).

Table 2.2 Screening studies in oral cancer which have shown measures of performance (after Moles *et al,* 2000).

Authors	Year	Country	Screeners	Numbers Screened	Prevalence	Sensitivity	Specificity
Metha *et al*	1986	Sri Lanka	Primary Health Care workers	39331	1.41	0.59	0.98
Warnakulasuriya and Pindborg	1990	India, Kerala	Primary Health Care workers	29265	21.63	0.95	0.81
Mathew *et al*	1997	India, Kerala	Primary Health Care workers	2069	0.25	0.94	0.98
Ikeda *et al*	1995	Japan, Tokoname	GDPs	802	9.74	0.6	0.94
Downer *et al*	1995	UK, London	Dental Surgeons	309	5.5	0.71	0.99
Julien *et al*	1995	UK, London	Dental Surgeons	985	2.23	0.64	0.99
Julien *et al*	1995	UK, London	Dental Surgeons	1042	3.07	0.81	0.99

2.12 Discussion

Oral cancer fulfils most but not all of the criteria set for a screenable disease and there is a good case for screening and early diagnosis. In this context, the main cancer referred to would be squamous cell carcinoma which constitutes the large majority (>90%) of cancers in the mouth and is amenable to primary prevention. Despite the finding that symptoms can present at any stage of the disease and appear to be independent of the tumour stage (Guggenheimer *et al,* 1994; Kowalski *et al,* 1994) Kaufman *et al* (1980) showed that, although late stage tumours had a shorter interval between onset of symptoms and diagnosis, most lesions may have a long detectable clinical phase. There seems a good argument, therefore, for screening or case finding as a means of early detection. It has already been shown that levels of sensitivities and specificities achieved by means of a simple visual examination by experienced and calibrated health professionals can make such programmes potentially worthwhile.

Health education raising awareness of oral cancer is an important component in achieving the objectives of screening. Despite the fact that there have been numerous publicity campaigns about oral cancer and precancer, there appears however to be a "gap" in public awareness of this set of diseases in the United Kingdom.

In a recent study, a random sample of 1,894 members of the public over the age of 16 was interviewed and oral cancer was the least heard of type of cancer. Only 56% of the participants had heard of oral cancer as opposed to 96% for skin cancer, 97% for lung cancer, and 86% for cervical cancer. 76% were aware of the link between smoking and oral cancer but only 19% were aware of the link to alcohol consumption (Warnakulasuriya *et al,* 1999).

There is a suggestion that screening should be linked with oral health promotion instead of disease. Such programmes would then encourage screening to be linked to general health and lifestyle messages rather than just to disease prevention (Blinkhorn and Jones, 1993). Apart from ignorance and unawareness of health issues which tend to lead to poor compliance, cultural factors could also be relevant. Pearson *et al* (1999), in a multi centre cross-sectional study

involving four medical practices, noted that 25% of a sample of Bangladeshi subjects had never visited a dentist. These were more likely to be women who thought that dental check-ups were of little value. Language difficulties were also cited as a problem.

The criteria for a positive lesion must depend on subjective decision and the ability of the screeners to recognise aberrations from the normal. The screeners therefore need to be agreed on what constitutes a positive test. Warnakulasuriya and Pindborg (1990) defined a positive test as the presence of a red or white lesion or an area of ulceration. These criteria can further be modified by defining a number of specific lesions to be included or excluded as positive (Speight *et al,* 1992). There also appears to be a need for agreement and standardisation as to which areas in the mouth are to be included in the screen. Some oral cancer screening studies, for example, exclude the tongue and lip whereas others include the oropharynx and larynx. This standardisation would allow for a better comparison of national and international trends. A controlled randomised trial (in which the population at risk was randomly divided into a study group and a control group) would give the best evidence of the efficacy of a screening programme. Members of the study group would receive the screening protocol, while members of the control group would follow the currently accepted program of follow-up and symptom-directed diagnostic intervention. A study would then evaluate the two groups during the same period (thereby eliminating any bias caused by different time periods). Such a study, especially for a disease of low incidence like oral cancer and precancer, would, of course, be subject to serious logistical financial and ethical problems.

The way ahead appears to consist of being able to target sub-populations at high risk. In the United Kingdom there needs to be a means of identifying sub-populations at higher risk of oral cancer and precancer and a means whereby these individuals can be regularly monitored, as each screen only gives a "snap-shot" in time of such an individual. The general dental practice appears to be the ideal environment for longitudinal studies to be done on these cohorts.

Part A

General Introduction

Chapter 3

Machine Learning

Chapter 3~Machine Learning

3.1 Definitions and overview

Machine learning is an aspect of computer programming; a computer program is said to learn when it changes its behaviour in a way that makes it perform better in the future (Witten and Frank, 2000). Mitchell (1997) defines an experience (E) in respect of a task (T) with a performance measure (P) in which a computer program is said to learn if its performance at tasks T, as measured by P, improves with experience E.

This is often done by going through a number of programmed steps collectively described as an algorithm. Machine learning is a multidisciplinary field and draws from many fields such as biology, statistics, philosophy, and the sciences. For example, an artificial neural network is inspired by the physiological function of a neuron. Machine learning is especially used in areas where humans can easily and intuitively perform well, but where traditional disciplines such as mathematics or statistics cannot, for example, in language ability or pattern recognition. It can be used in procedures that are repetitive and laborious and where humans can easily tire or make mistakes, such as in the identification of stained dysplastic cells on histolopathological specimens in PAP smears. Here machine learning is able to repeatedly perform such learned tasks with high accuracy.

3.1.1 Machine learning in the context of artificial intelligence

The term *artificial intelligence* is sometimes used synonymously with *machine learning* and can be defined as "a science that is concerned with building machines that can act and react appropriately, adapting their response to the demands of the situation. Such machines should display behaviour comparable with that considered to require intelligence in humans" (Finlay and Dix, 1996). Strictly speaking, machine learning is a branch of artificial intelligence. Other branches of computer science listed under artificial intelligence include heuristics, planning, common sense, reasoning, inference and mathematical logic. Applications of artificial intelligence

include game playing, particularly chess, speech recognition, understanding natural language, computer vision, and application of knowledge in expert systems such as the histological identification of cell dysplasias or the interpretation of ECG tracings.

3.2 Main machine learning methods

There are several and some overlapping learning paradigms. The following is a brief overview, with a focus on the learning methods relevant to this study. These are decision trees with their variants, rule induction and classification and regression trees, also neural networks and logistic regression. Other learning methods not discussed here include Bayesian modelling, instance based learning, genetic algorithms, and reinforcement learning. One of the aims of this project was to examine a smaller range of machine learning tools which are well documented and readily supported.

3.2.1 Decision trees

Decision trees, as described by Quinlan (1988), are a specific decision analysis technique and have been widely used over a wide range of tasks such as diagnosis of medical cases and assessing credit risks for loan applications. This is a method of reaching solutions (known technically as "target functions") in which the learned function is represented by a decision tree. Following on from the analogy of a tree, decision trees classify instances by sorting them "down" the tree to "leaves and nodes". Each leaf is a classification, for example gender, and each node is a decision on this classification or attribute, in this instance, male or female. One of such algorithms is the C5 algorithm. C5 is a commercial version of a decision tree algorithm where classifiers are expressed in the form of a decision tree or "if-then" rules.

Decision trees provide an effective method of decision making because they clearly lay out the problem so that all choices can be viewed, discussed and challenged. They also provide a

framework to quantify the values of outcomes and the probabilities of achieving them so that the best decisions or best guesses can be made. Some disadvantages and limitations of decision trees are:

- They need to break numeric fields into fixed ranges, hence missing patterns, and providing less information. A small change in a numeric value therefore can have a large impact on the outcome.
- They work best on small samples of data and cannot easily approach large data sets. This often results in significant loss of information because of the fact that the information once considered is then discarded as the algorithm moves down the "flow chart" and is not considered again.
- They look at very simple combinations of attributes within a table, and hence may miss patterns or the "bigger picture".

Since by the nature of the process, some attributes get ignored or discarded, decision trees may make less accurate predictions, and worse yet, if some values are missing from the new data item, they may make unpredictable decisions or no decisions at all. Thus their predictive power is questionable in datasets that contain missing data. A simple decision tree is illustrated in Figure 3.1.

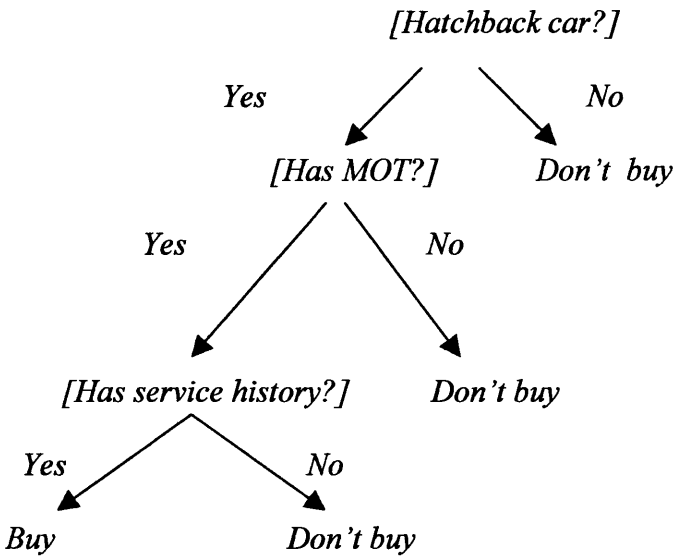


Figure 3.1 A decision process for the purchase of a second hatchback car as an example of a simple decision tree.

3.2.2 Rule induction

Rule induction is a similar technique and is a variant of the C5 decision tree algorithm.

It is possible to establish a decision list with classification rules within a dataset. The following is an example of such a series regarding playing a game of tennis (adapted from Mitchell, 1997):

If outlook= sunny and temperature = hot and humidity =high, then play tennis= no

If outlook= sunny and temperature = mild and humidity =high, then play tennis= no

If outlook= sunny and temperature = mild and humidity =normal, then play tennis= yes

If outlook= sunny and temperature = cool and humidity =normal, then play tennis= yes

Such is an example of a decision list that contains classification rules, or rulesets which assign output classes or verdicts like *yes*, or *no* to each instance. An "instance" here is synonymous with the term "attribute". The decision tree presentation is useful when attributes in the data can be split or populations partitioned into subsets relevant to the problem. The ruleset presentation is useful for viewing how particular groups of items relate to a specific conclusion.

3.2.3 Classification and regression trees

Classification and regression trees, sometimes described as CARTs and described by Breiman-Friedman-Olshen-Stone (1984), are a variant of the decision tree model. Similar to decision trees, they split the training records into segments or partitions with similar output fields and repeatedly apply this splitting procedure on each new derived partition until some stop criteria is reached, for example a minimum number of samples in the partition. This is known technically as recursive partitioning. The basic algorithm starts by finding the best split measured by the reduction of an "impurity index" given by the split. "Impurity" is a technical measure which itself has a number of ways of being measured. The aim is to arrive at groups of samples recognised as being similar, so that the smaller the samples in a group, the less impure that sample set is. The use of this impurity index differentiates this approach from the other decision tree algorithms. CARTs are often considered as an improvement of the decision tree algorithms because they can deal with incomplete data, they can handle large datasets and the output fields can be numeric as well as symbolic.

3.2.4 Artificial neural networks

A neural network can be defined as "an interconnected assembly of simple processing elements known as nodes or units whose functionality is loosely based on the animal neuron. The processing ability of the network is stored in the interunit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns" (Gurney, 1997).

The nodes are very simple computing elements and are based on the function of a biological neuron. The neuron behaves like an on-off switch: when sufficient neurotransmitter has accumulated in a synapse, an action potential is generated and the node "fires". This has been modelled mathematically as a weighted sum of all incoming signals to a node, which compares to a threshold value. If the threshold is exceeded, the node fires, otherwise it remains quiescent. Computational power in an artificial neural network derives not from the complexity of each

processing unit (as in a conventional computer) but from the density and complexity of the interconnections. The memory in an artificial neural network is distributed throughout its structure in weights, and is modified as the neural network learns. There are many kinds of neural network and this appears to be a developing science as more and more new ones, or variations of existing ones, are being invented. However, the main categorisation of these methods is the distinction between *Supervised* and *Unsupervised* learning. *Supervised* and *Unsupervised* learning are generic descriptives not necessarily exclusive to neural networks.

In *Supervised learning* there is a "teacher" who, in the learning phase, "tells" the net how well it performs (this is actually done by feeding back the errors made by the learning process and having a means of readjusting the various "weights" within the nodes to reduce this error value). This variant is referred to as reinforcement learning or fully supervised learning. In *Unsupervised learning* the net is autonomous; it examines the given data and looks for properties within the dataset and learns to cluster groups which bear similarities together. This is also described as self-organisation. Within these two main categories, there is usually a further subdivision of the network topologies into *feedforward only* and *feedback* nets.

3.2.4.1 Some popular neural network designs

The following are among the most popular and widely used neural network variants:

- The Perceptron:

The Perceptron was invented in the fifties by Frank Rosenblatt and was intended to be a computational model for the retina. It was based on the events that take place in the human eye. The retina absorbs the stimuli from the environment and passes them onto a layer which forms an internal "pattern" called the association layer. The pattern is passed on to the next layer, which reacts to the input stimuli. There are therefore input layers, hidden layers and the output layer.

- Backpropagation:

Backpropagation is a very researched and popular learning mechanism. This method was used in our earlier studies (Speight *et al,* 1995). Backpropagation allows weights on nodes within the

hidden layers to be changed during learning. These changes of weights influence the global input and therefore influence the activation and finally the output of a neuron. The actual network output is measured against the target output, that is, the desired result. The difference is then treated as an error to be minimised by readjusting the weights within the hidden nodes. These errors are propagated back from the output nodes into the network of hidden units in order that the weights within the various nodes can be readjusted with the intention of minimising this error value, hence the name backpropagation. The backpropagation algorithm is also known as error backpropagation or back error propagation.

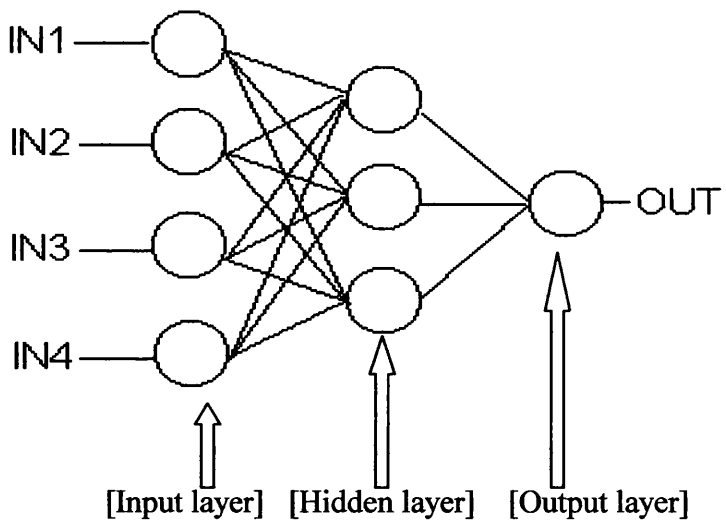


Figure 3.2 The layout of a simple, typical neural network with one hidden layer.

Figure 3.2 shows the basic layout of a typical neural network with interconnections. One hidden layer is shown although there are usually several hidden layers. There are a number of input nodes in this figure labelled as IN1, IN2, IN3, and IN4. These represent the various input factors. All the nodes are interconnected in both directions and lead to one output node which is the desired result. For example, IN1 could represent age, IN2 gender, IN3 smoking habit, and IN4 the diagnostic outcome positive or negative for oral dysplasia. The input "weights" are fed back and forth through the hidden layer(s), and changing weights are assigned to each node which

influence the resultant output. These weights change according to the value of an error value which is propagated backwards to the various nodes. A "final picture" is arrived at, where a resultant output value is given, which represents the least error arrived at by the adjustment of the weights at the various nodes; in this instance, the diagnosis which could then be a positive or a negative depending upon what the network has learned from the characteristics of positive and negative cases within this set.

- Hopfield nets

A Hopfield net is an example of a class of dynamic physical systems used for associative memory (Hopfield 1982). Hopfield nets are used to learn patterns. If the patterns to be learned are known, then the weights on each of the nodes in the neural net can be calculated. This is in contrast to the previous system of learning where the weights are continually adapted.

- Kohonen networks

Teuvo Kohonen is one of the most famous and prolific researchers in neurocomputing. These networks owe their name to him. There are three main types of Kohonen network, the most well known is the Self Organising Map, or the SOM. This, with its variants, is the most popular artificial neural network algorithm in the unsupervised learning category.
About 4000 research articles on SOMs have appeared in the open literature, and many industrial projects use the SOM as a tool for solving problems. Many fields of science as diverse as statistics, signal processing, control theory, financial analyses, experimental physics, chemistry and medicine have adopted the SOM as a standard analytical tool.

3.2.4.2 Features and advantages of the artificial neural network

Artificial neural networks work by combining signals and by generating new ones, which is in contrast to the execution of instructions stored in a memory as in a conventional computer. Data input is stored as sets of (electronic) weights rather than as a program. In general, the weights adapt by continually presenting examples from a set of training vectors to the net. Processing in a

network therefore occurs in a parallel rather than a serial fashion. There is no single computer-processing unit (microchip) and each node can operate independently or simultaneously with other nodes in the net. Nets are robust in the presence of noise; small changes in an input signal will not drastically affect a node's output. Nets are also robust in the presence of hardware failure: a change in a weight may only affect the output for a few of the possible input patterns. Artificial neural networks are especially useful for classification, patterns, and perceptual and mapping problems, and are tolerant of some imprecision as found in some datasets to which hard and fast rules cannot be easily applied. A characteristic feature of their operation is that neural nets work by extracting statistical regularities or features from the given set for training. This allows the net to respond to input not seen during training by classifying it appropriately with one of the previously seen patterns, or by assigning it to new classes. This ability to generalise is one of the key reasons for using neural nets. It is necessary here to mention a common criticism of the neural network, which is that the internal workings remain as a "black box", and investigators are unable to understand or observe how conclusions and results are arrived at.

3.2.5 Logistic regression

Linear regression models attempt to find a straight line or surface through the range of input fields that minimises the discrepancies between predicted and observed output values. Logistic regression models use a similar strategy to generate equations for predicting probabilities associated with each possible value of a symbolic output field. Logistic regression is used to predict the probability of an outcome variable from a set of independent variables. It is widely used throughout epidemiology, mainly in case control studies, to calculate odds ratios that are used to relate disease risk to exposure. Brugere *et al* (1986), using this method, demonstrated a relationship between the site of oral cancer and the level of daily alcohol consumption whilst Mashberg *et al* (1993) were able to demonstrate an increasing odds ratio of risk of oral cancer with tobacco (35 cigarettes a day) and alcohol (21 whisky equivalents a day). By evaluating the

relative risk of different lifestyle habits it is possible to use this information for developing more sophisticated means of identifying individuals who could be targeted for screening or health education. Wilkinson *et al* (1994) evaluated a risk scoring system for cervical cancer to be used in primary care. The system enabled identification of 75% of those women with cervical neoplasia. They concluded that further research was required to assess the effectiveness of risk targeting. Logistic regression is a useful tool for assessing risks and is one of the machine learning methods provided in some commercially available datamining software packages.

3.3 The use of machine learning in medicine

Many biological, pathological and physiological processes behave in ways that are not best analysed by classical linear statistical processing. Multidimensional relationships often seen in clinical data are not always apparent to other forms of analyses. Cross *et al* (1995) explained how non-linear processing as afforded by machine learning models, particularly neural networks, might improve upon the predictive power of other approaches. Non-linear statistical methods have been tried but they tended to be too complex and computationally intensive and never became widely accepted. The recognition of complex non-linear interactions within datasets in medicine has led to the development of additional methods of analysis, ranging from regression models to different forms of machine learning such as decision trees and neural networks. These are typically used when multiple items of data for each case produce clusters of cases who exhibit the diagnosis within a larger number of cases who do not exhibit the diagnosis; it is very difficult to separate the two populations by multivariate discriminant analysis (Cross, 1995). Machine learning is suitable in the field of medicine because almost all clinical situations and decisions are usually based on more than one item of data. Machine learning systems are used in a wide range of applications. A search on PubMed in May 2003 using the words "Machine Learning" and "Artificial Intelligence" resulted in 11,935 and 12,076 entries respectively.

3.3.1 Some illustrative examples of machine learning in medicine

Machine learning, especially neural networks, is used for decision support and as a diagnostic aid in a wide variety of medical applications. It includes aids to clinical diagnosis, for example in cases of myocardial infarction. Baxt *et al* (2002) tested the ability of a neural network to recognise the presence of cardiac ischemia in patients presenting with chest pain in a hospital emergency department and found that with 40 variables, a neural network achieved sensitivities of 88% and specificities of 86.2%, suggesting that it was useful as an aid to diagnosis. Djavan *et al* (2002), using neural networks for the early detection of prostate cancer, found that the predictive accuracy of the neural network was better than that achieved by using conventional prostate specific antigen parameters. Machine learning was used to predict which patients seen in a psychiatric emergency room needed hospitalisation and which did not (Somoza and Somoza, 1993). A neural network achieved a sensitivity of 70% and a specificity of 94% and was in broad agreement with the clinicians. Machine learning was also used to process complex patterns and trends in data to identify abnormality in vision fields that would lead to glaucomatous change much earlier than by traditional methods (Sample *et al*, 2002). Adam *et al* (2002) used a decision tree model to distinguish between prostate cancer and benign prostate hyperplasia by using serum protein fingerprinting coupled with pattern matching algorithms. Neural networks are often used to recognise images and patterns. Balaji *et al* (2002) obtained three-dimensional images of prostate tumours to depict areas and volumes of cancer foci for treatment planning, staging and radiotherapy. Machine learning was also used in waveform analysis, for example in wave patterns in electroencephalograms (EEGs). An example of this is the work of Jando *et al* (1993) who used a backpropagation network to recognise high voltage spike and wave spindle patterns (HVS) in a rodent model of human petit mal epilepsy. The HVS patterns detected correctly by the neural network reached 93-99% of their manually marked HVS patterns. Castellaro *et al* (2002) used neural networks to classify and analyse EEG tracings and concluded that such systems appear useful and can improve the quality and reliability of the EEG exam but also concluded that they cannot totally substitute the involvement of a physician. In the field of dentistry, computer

intuition (CI) was used to guide scientific research. In this study, CI indicated that computer tomography was the diagnostic imaging modality of choice in placements of oral implants (Almog and Heisler, 1997).

3.3.2 A discussion of the role of machine learning in medicine

Machine learning is a useful means of analysing complex data and has been applied in fields such as decision support, early detection, wave form analysis, outcome prediction, prediction of patient responses to drugs, education, and identification of pathological specimens. There are instances where a neural network can yield still more information about their data. Lucht (2002), using neural networks to study magnetic resonance mammographic images, found that the neural network not only performed well but yielded a unique classification of just three tissue classes. Whilst the use of machine learning appears to be gaining popularity, caution has been expressed and care advocated in the use of these new tools. Comparisons should be made with newer analytical methods, for example alternative statistical techniques for unravelling non-linear relationships commonly found in clinical data. It is also wise to recognise that these new "black box" technologies (where one inputs data and does not know how the data is processed within the machine learning process) could obviate the necessity for the usual protocols in the design of the experiments and in scientific rigor (Drew and Monson, 1999). Lillehaug and Lajoie (1998) suggested that whilst machine learning systems can perform well, health professionals should maintain the role of independent and self-sufficient problem solvers and decision-makers.

3.4 Use of machine learning in screening

Machine learning has been used in many aspects of screening. Using data from screening studies as a starting point, machine learning systems are usually used to identify features in the data that would lead to the prediction of a positive or negative diagnosis, the most suitable applications being those which involve many clinical and pathological features, of which none is

pathognomonic. The following are some examples of this. A neural network was used to predict breast cancer malignancy using mammograms and biopsies as the gold standard. Floyd *et al* (1994) studied 260 cases and found that a neural network achieved a sensitivity of 1.0 and a specificity of 0.59. Ercal *et al* (1994) found that a neural network designed to detect malignant melanomas from colour images was able to classify correctly over 80% of the malignant and benign tumours on real skin images. Snow *et al* (1994) used a neural network to assess the diagnosis and prognosis of prostate cancer. The network was able to predict a biopsy result with 87% overall accuracy from serum prostate specific antigen levels. It was also able to predict tumour recurrence with a 90% overall accuracy. Perhaps the most well-known application of machine learning in screening is the Papnet ® testing system developed by Neuromedical Systems. This is an augmentation of the Pap smear test developed by Dr George Papanicolaou for cervical cancer screening where machine learning is used to detect signs of malignancy. Papnet® uses two separate neural networks, one trained to recognise suspicious single cells and the other trained to recognise abnormal cell clusters. Farnsworth *et al* (1996) evaluated the Papnet® screening system and found it to detect higher numbers of abnormal PAP smears than manual screening. Koss *et al* (1997) reported a significant reduction of false negative cases with the PAP smears in conjunction with neural network-based technology. A recent study by Cenci *et al* (2002) reported on 20,154 smears and confirmed the usefulness of neural networks in the identification of positive cases in primary screening as well as in rescreening of cervical cancer.

3.5 The use of machine learning in screening for oral cancer and precancer

As there is no single appropriate test for the detection of malignant and premalignant disease in the mouth (Zakrzewska *et al,* 1993), it is difficult to define suitable criteria over this wide range of diseases and conditions for use by machine learning software in order that positive and negative cases can be predicted with significant accuracy. Machine learning has been used successfully for aiding the diagnosis of oral cytological smears (Brickley *et al*, 1996) and brush biopsy tissue samples (Scuibba *et al*, 1999), where a neural network program is used to identify

dysplastic tissue from histologically stained tissue samples. These, as with almost all applications of machine learning in medicine, are to support the determination of a clinical decision or a definitive diagnosis.

3.5.1 The case for the use of machine learning in oral cancer screening

It is generally advised and accepted that population screening for oral cancer and precancer in the United Kingdom is not justifiable and this is due to the relatively low prevalence of these diseases and high cost both in finance and available resources (Warnakulasuriya and Johnson, 1995; Rodrigues *et al,* 1998). An alternative strategy is to target individuals that are considered to be at high risk by using factors known to influence risk such as age, gender, smoking and alcohol consumption. In a simulation model of an opportunistic screening programme based on best evidence from data derived from previous screening studies, Downer *et al* (1997b) demonstrated a health gain equivalent to 5.2 lives saved out of a hypothetical population of 100,000. Adjusting for the fact that only 50% of the population attend the dentist, this would give the potential gain of 2.8 or about 3 lives. If a machine learning model were to be used to preselect a population at high risk, the gain was 8 lives saved. Therefore if machine learning software can be trained to identify this sub-population, this form of selected or targeted screening might produce a greater "harvest" of positive cases and therefore make a screening study more efficient and cost effective. Speight *et al* (1995) showed that a neural network specifically written for the purpose, trained on age, gender and lifestyle habits, was able to predict the presence of positive and negative lesions with a sensitivity of 0.8 and a specificity of 0.77 over 365 cases. This current project develops this concept further and examines the ability of machine learning models to identify such high-risk individuals. It also seeks to examine the proposition usually regarded as being self-evident that heavy smokers and heavy alcohol consumers are automatically at higher risk. Individuals at high risk, once identified, could then be given closer and more regular screening. The uniqueness of such an approach lies in the fact that this risk assessment would operate at the start of the screening or case finding and may be operated by reception or nursing staff, or even by the

patients themselves The questions answered form part of initial clinical data and provide a risk assessment with no extra effort from the clinical or medical expert. Such a system could also be used in the public domain as part of a health information or awareness package.

3.5.2 The use of a purpose-written neural network pilot program for oral cancer screening

Speight *et al* (1995) evaluated the ability of a neural network program written for the purpose to identify a subset of the population at risk of oral cancer and precancer. This was achieved by using lifestyle habits including smoking, drinking and dental attendance as indicators of risk. A neural network program was written by one of the authors in the Turbo Pascal programming language. The model was a standard three layer feed forward neural network using a backpropagation algorithm.

A dataset of 2027 subjects was derived from previous screening programmes in which 54 cases (2.7%) were identified to have positive lesions (Julien *et al,* 1995; Downer *et al* 1995; Kulasegaram *et al,* 1995). The computer was instructed to randomly generate two groups of subjects, each with the same prevalence of positive lesions. These comprised a training set of 1662 individuals and a test set of 365 individuals. The test set consisted of 365 cases of which 10 were positive. This gave a positive prevalence (2.74%) similar to the overall positive prevalence of the dataset. The neural network was then trained on items of personal information relating to age, gender and habits selected from the questionnaire data. The additional input fields were the smoking and drinking habits as well as dental attendance. The lesions and conditions deemed as positive are tabulated in Table 3.1 and the fields used are tabulated in Table 3.2

Table 3.1 The lesions and conditions to be considered as positive

Precancerous conditions	Precancerous lesions	Cancer
Lichen Planus Lupus Erythematosus Submucous fibrosis Actinic keratosis	Leukoplakia Erythroplakia Red/white speckled lesions	Squamous cell carcinoma Basal cell carcinoma

Table 3.2 The fields used for the training of the neural network set out as used in the earlier study by Speight *et al* (1995).

Question	Definition
Male	
Female	
Non-smoker	Never smoked or not for over 10 years
Moderate smoker	<20 cigarettes /day
Heavy smoker	> 20 cigarettes/day
Non-drinker	Never drinks or <5 units /week
Moderate drinker	Female < 14, Male < 21 units a/week
Heavy drinker	Female >14, Male >21
Age	Continuous variable from 40 years
Irregular Dental attendee	Has not visited the dentist in the last year.

3.5.2.1 Results and discussion of this pilot study

The neural network correctly identified 8 out of the10 positives within the test set of 365 individuals giving a sensitivity of 0.8 and identified 273 negatives giving a specificity of 0.77. These results were very encouraging and formed the basis for the project described in this thesis which used commercial software incorporating machine learning facilities. This was because of an awareness of the limitations of a neural network program specially written for the purpose. This could suffer from being over-trained and the data could be over-fitted, meaning that such a program could be capable of "memorising" the features of the entire dataset and would therefore perform very well within this set but may be less accurate when presented with new data. As there were only 10 positives in the test set, getting one right or wrong could result in a sensitivity difference of 10%. The neural network for the previous study was written in a form and code that was difficult for a clinician without specialised training in the science of machine learning to reproduce.

Neural networks and other machine learning models may be built into existing computer hardware, but such purpose-built microcircuits so far available are expensive and of low capacity (Cross *et al,* 1995). Commercially available datamining software packages almost always have machine learning models integrated into the software package. They also have the overall benefit of better user support and on-going improvements, developments and enhancements than a software package written for the purpose. Examples of these such as IBM Intelligent miner, KXEN, Microsoft SQL Server/Excel, SPSS/Answertree and CART/MARS are shown in Figure 3.3 and Table 3.3, which displays a recent online poll of users of various datamining systems by KD Nuggets in 2002. KD Nuggets is a website and news resource dedicated to the many aspects of datamining. It can be seen from this recent poll that Clementine® was the most popular system, albeit out of a small total of 880 users who responded to this poll. Clementine® was chosen for this study because of its relatively low cost, user-friendliness and efficient technical support facilities; also because it incorporated a good selection of machine learning types unlike some other packages that depended on just one type of machine learning model. Clementine®, initially

a product of Integrated Solutions Limited, was taken over by SPSS Inc. who developed and incorporated improvements to the system. It continues to be refined and adapted according to customer needs. This ongoing development of the product by a large parent company was a very strong factor in its favour. Although desirable, it was not practicable for economic reasons to evaluate a large range of datamining packages available, and Clementine® was taken as a good representative example of these packages. The use of Clementine® was therefore dictated by availability and there was no proof that it would have performed better or worse than the other packages.

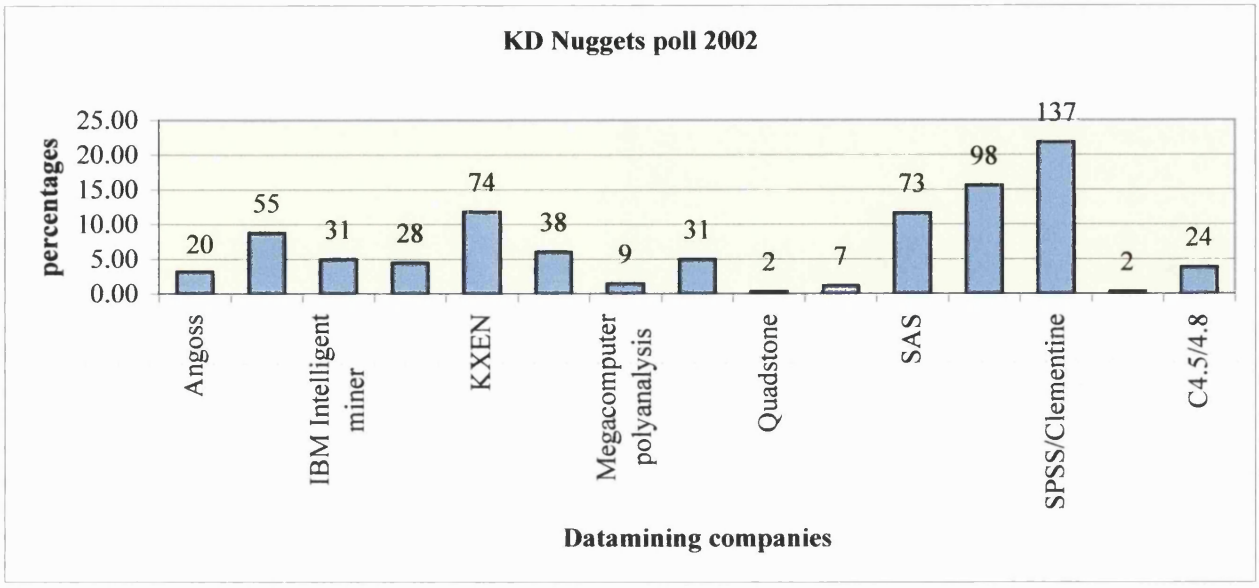


Figure 3.3 The popular datamining packages showing a selection of numbers (above bars) and percentages (Y-axis) of users that responded to the online poll by KD Nuggets in 2002.

Table 3.3 The popular datamining companies showing the complete list of the numbers and percentage of users represented as bars in Figure 3.3

Datamining companies	Percentages	Numbers
Angoss	3.18	20
CART/MARS	8.74	55
IBM Intelligent miner	4.93	31
Insightful miner	4.45	28
KXEN	11.76	74
MATLAB	6.04	38
Megacomputer polyanalysis	1.43	9
Microsoft SQL	4.93	31
Quadstone	0.32	2
Oracle Suite	1.11	7
SAS	11.61	73
SPSS/Answertree	15.58	98
SPSS/Clementine	21.78	137
Thinkanalytics	0.32	2
C4.5/4.8	3.82	24

Part A

General Introduction

Chapter 4

Aims and Objectives

Chapter 4~Aims and Objectives

4.1 Introduction

4.1.1 The case for targeted screening

There appears to be a cogent case for screening for oral cancer and precancer in the general population yet due to the relatively low numbers of such cases in the United Kingdom, it is broadly agreed that population screening is not justifiable in terms of manpower and cost (Warnakulasuriya and Johnson, 1996; Rodrigues, 1998). An alternative strategy seems to be necessary. This could be based on the identification of groups of individuals seen to be at higher risk of developing oral cancer. It has been widely accepted that heavy smoking and high alcohol consumption can be risk factors while age and gender also appear to influence the prevalence of positive lesions and conditions within a population group. This study explores the utilisation of these factors in the identification of populations at higher risk of oral cancer and precancer. Such a group, if identified and targeted for screening and closer scrutiny, could yield a better return of positive cases and make oral cancer screening more efficient and more feasible.

4.1.2 The feasibility of screening in general dental practice

General dental practice would seem to be the ideal environment to screen individuals for oral cancer and precancer. There is a question that arises regarding the suitability of such a population of dental patients. They are usually seen to be a self-selecting cohort, more highly motivated about their teeth and oral hygiene and therefore less likely to be at risk of oral cancer and precancer than other sub-populations such as the homeless and alcoholics, or indeed the population at large. Previous screening studies have recorded prevalence figures and noted risk habits amongst those with oral cancer and precancer. Such data was used as a comparison for this study. In this study, the prevalence of relevant oral lesions and high-risk habits within the population attending their dentist for routine treatment was recorded.

4.1.3 The use of machine learning software programs in screening for oral cancer and precancer

Many strategies are being proposed in an attempt to circumvent the problems of large-scale screening for a disease of low prevalence. One such is the use of machine learning computer programs trained in identifying those seen to be at high risk of developing oral cancer or precancer. This extends the concept of targeted screening by using the computer as a means of identifying sub-populations considered to be at high risk and therefore worthy of closer attention and follow up. An earlier dataset was available for use as a set for training machine learning software and the performance was tested on hitherto unseen test data obtained from our opportunistic screening study. Further methods of training involved combining the two datasets and separating a smaller randomly selected set of about 1000 cases and training on the other 3000 cases.

4.2 Aims

The broad aims of this study were:

- To determine the feasibility of opportunistic screening in general dental practice and, in particular, to observe the prevalence of relevant oral lesions as well as age, gender and smoking and alcohol consumption habits in this population group.
- To determine if the risk factors are related to the likelihood of having a relevant lesion
- To evaluate the role of machine learning software in the prediction of high-risk groups or patients with relevant lesions or conditions.

4.3 Objectives

The objectives of this study were:

- To determine the prevalence of relevant oral lesions within the population attending their general dental practice for routine treatment.

- To determine the prevalence of smoking and alcohol consumption, seen as high-risk habits, within this population.
- To observe age, gender and racial distribution in this population.
- To determine the performance of a neural network program written for the purpose of identifying such high-risk groups in an earlier demonstration study.
- To evaluate the performance of commercially available machine learning software with data from that earlier study.
- To compare the data between two datasets, one from the earlier study, and the other obtained from the opportunistic screening study.
- To report on the capability of the commercially available machine learning software package to identify high-risk groups, and to do this in two ways; one by using the earlier dataset as a training set and testing its performance on the unseen dataset obtained from the opportunistic screening study in general dental practice; the other by pooling the two sets of data and drawing training and test sets from them.
- To extend the use of such derived machine learning models for applications such as interactive health education packages.
- To note conclusions and make suggestions based on findings for the case for targeted oral cancer screening in the United Kingdom.

Part B

Materials and Methods

Chapter 5~Materials and Methods

5.1 Prevalence of lesions in general dental practice

5.1.1 Introduction

This study investigated the prevalence of oral cancer and precancerous lesions and conditions in the sub-population that attend their dentist for routine treatment and was done in the form of a multicentre demonstration study. To date, prevalence of dysplastic and benign lesions and conditions in such populations in the United Kingdom have not been studied. The value of screening for oral cancer and precancer in such a population has been in debate as it has been assumed that this population group, being more highly motivated about oral and dental hygiene, were less likely to have relevant lesions and conditions and, as such, will not be representative of the population at large. This study set out to determine the risk habits and prevalences of oral lesions and conditions in such a population.

5.1.2 Aims

- To carry out a multicentre opportunistic screening study in a general dental practice setting in order to observe the prevalence of malignant and potentially malignant lesions and conditions, and benign mucosal lesions in an adult population over 35 years of age.
- To observe age, gender, smoking and alcohol consumption habits in this population, also to record their dental attendance and ethnicities.
- To obtain a dataset of patients attending general dental practices for routine dental attention and treatment, in order to compare the descriptive epidemiology of their lifestyle habits of smoking and drinking and dental attendance with a dataset from earlier studies (Julien *et al*, 1995; Downer *et al,* 1995; Kulasegaram *et al,* 1995).

5.1.3 Materials and methods

This study was carried out in general dental practices in the UK. The patients attending at these dental practices were opportunistically selected on the basis of being aged 35 years and over. No other targeting criteria were used. A broad selection of general dental practices was selected. Practices ranged from all National Health Service practices to Mixed (NHS and Private), to All-Private practices. The majority of the practices, however, tended to offer both a mixture of NHS and private styles of treatments. Eighteen general dental practitioners took part, sixteen located in Greater London, in practices around the Barnet area of north London and the Streatham and Brixton area of south London. There were two other participants, one in Nottingham and one in Aldershot. Each dentist attended a training session to standardise criteria for the identification of lesions and to receive instruction on the protocol of the study. The dentists were advised of the criteria for a positive and negative screen. A positive screen was defined as the presence of a white patch, a red patch or an ulcer of longer than two weeks duration (Warnakulasuriya 1990; Jullien *et al,* 1995). These basic criteria were modified by defining a number of well-known clinical entities that might have these appearances but could be included as positive or negative lesions (Speight *et al,* 1992) (Table 5.1). For example, ulcers and white lesions with an obvious traumatic aetiology and recurrent ulcers (aphthae) were regarded as negative.

It was not possible to prescribe set criteria for the method of recruitment and each dentist selected the most appropriate method to suit their practice routine. Thus, some recruited sequentially while others recruited randomly or on fixed days per week. The aim was for each dentist to attempt to recruit approximately 200 patients into the study. Each patient was first asked to complete a questionnaire, Form A (Appendix 1), concerning his or her age, gender and smoking and drinking habits. This was completed in the waiting room under the guidance of a practice nurse or receptionist. Before commencing routine treatment the dentist then recorded the presence or absence of lesions independently on a second form, Form B (Appendix 1). The dentist was blinded to the results of the questionnaire at the time of examination. The two forms were collected and sent in batches to the Eastman Dental Institute to be collated. Patients with lesions

were referred as appropriate for secondary care, according to the normal protocols of each practice. There was no attempt, in this study, to record the results of secondary care for any patient. The data collected were entered into a spreadsheet (Excel), datasets with inconsistencies were excluded, and the data were set into tables showing lesions observed, profile of people examined, and relative risks, and histograms of age, drinking and smoking habits and ethnicity relative to the prevalence of positive conditions. Data were analysed using the STATA statistical software package (Version 5.0, STATA Corporation, Texas, USA). Incidence rate ratios were calculated using Poisson regression. Age was considered as a continuous independent variable since 98% of all oral cancer cases occur in those aged over 40 (Hindle and Nally, 1991). As there is evidence that the prevalence amongst lower age groups is increasing (Hindle *et al*, 1994; OPCS, 1990, 2002), it was decided to include those aged over 35 years in this study. Alcohol and tobacco consumption are considered to be risk factors for oral cancer and precancer and subjects were divided into three groups according to levels of use (Table 5.2). The levels were set according to the recommendations for smoking and drinking levels for males and females in the United Kingdom. The patients were given an Information sheet (Appendix 1) detailing the reason for this project in order that their consent and cooperation was informed. Ethics Committee approval was given by the Eastman Ethics Committee and the relevant Local Ethics Committees.

Table 5.1 Conditions and lesions considered as positive (Speight *et al*, 1992).

Precancerous conditions	**Precancerous lesions**	**Cancer**
Lichen planus Lupus erythematosus Submucous fibrosis Actinic keratosis	Leukoplakia Erythroplakia Red/white speckled lesions	Squamous cell carcinoma Basal cell carcinoma

Table 5.2 Definitions for smoking and drinking levels. A unit in the United Kingdom is equal to 10ml or 10 grams of alcohol (ethanol).

Smoking	Heavy smoker	>20 cigarettes/day.
	Moderate smoker	<20 cigarettes/day
	Non-smoker	Never smoked or not smoked in the last 10 years
Alcohol Consumption	Heavy drinker	>21 units/week Males, >14units/week, Females
	Moderate drinker	>5units/week
	Light drinker	<5units/week

5.2 The use of Clementine®, a datamining software package, as a tool for the detection of groups at high risk of having positive lesions and conditions

5.2.1 Introduction

This study investigated the ability of a commercially available datamining software package with machine learning capabilities to identify sub-populations at high risk of oral cancer by using age, gender, smoking and drinking habits. It would be unrealistic to expect to be able to identify positive cases just on these factors. The aim would be to identify a sub-group of individuals wherein most or all of the positives lie. It is therefore envisaged that the role of machine learning could be that of a pre-screen filter.

In an earlier study, Speight *et al* (1995) used a neural network program written for the purpose to identify those at high risk of having a positive lesion or condition and this neural network performed with a sensitivity of 80% and a specificity of 77%. There were however shortcomings

encountered by the use of bespoke software programs as discussed in Chapter 3. This experiment sets out to investigate the ability of Clementine® to overcome these shortcomings.

The term "datamining" refers to the use of a variety of techniques used to identify decision-making knowledge in bodies of data. These datasets may be obtained from sources such as applicants for insurance, customers at supermarkets, users of the national rail service, telephone service and taxpayers. Datamining is the study of the trends and patterns and relationships within such datasets so that these patterns and characteristics can be applied in areas such as prediction of trends, fraud detection, customer preferences, market analysis and sales forecasts, thus enabling decision support, sales strategy and financial resources to be applied where they are most needed. This is usually achieved by the use of passive and active data manipulation. Data can simply be displayed, for example, in the form of graphs and histograms for scrutiny, or data can be actively manipulated, usually by the use of machine learning models which would "learn" from the characteristics and patterns within that set and be able to produce predictive results or be able to cluster data in ways that would not necessarily be intuitive. Machine learning models are therefore commonly found within datamining packages. Neural networks and/or decision tree type algorithms are commonly included. It was for this feature that the datamining package Clementine® was chosen; also for its good visual interface, ease of handling of data and inclusion of a range of machine learning models designed to be utilised without the need for detailed technical knowledge of the underlying algorithms. The end-user who does not need to be an experienced programmer or necessarily have any programming skills, can easily develop machine learning models, and modify and re-engineer derived models if necessary.

5.2.2 Aims

- To test the efficacy of Clementine 6 ® in the identification of sub-populations at higher risk of oral cancer and precancer.

- To find the optimum negative to positive ratio of cases that would yield the best results for the machine learning models.

5.2.3 The structure and use of Clementine®

The features of Clementine®'s datamining package are presented via a visual programming interface. The main area of Clementine®'s screen is known as the *stream pane* and occupies about two-thirds of the screen (Fig.5.1). The lower third of the screen, described as a palette, contains a series of available operations which are represented by icons. These icons are described as *nodes* which are variously nested in six sections and described as: Sources, Record operations, Field operations, Graphs, Modelling (node for deriving the machine learning models), and Output operations. These nodes can be brought onto the stream pane (the working area) by pointing and clicking on the relevant icon via the mouse button and setting it in place on the stream pane by clicking the cursor again. These nodes can be connected up (by using the middle mouse button) and operated as a data stream representing the flow of data through each operation.

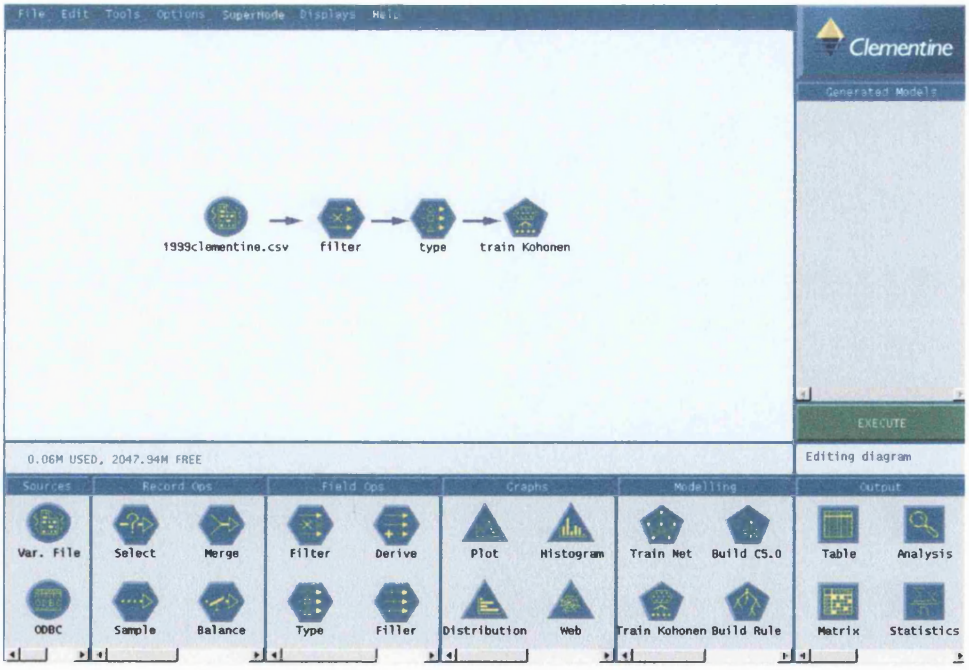


Figure 5.1 A typical Clementine® desktop image showing the main working pane, a representative information stream from the left: a circular Source node, streamed to a Filter node

thence to a Type node and finally to a pentagonal Train Kohonen node. The suite of available functions in the form of nodes is shown on the lower section of the pane

5.2.4 The functional abilities of Clementine ®

5.2.4.1 Overview

The following is an overview of the functional abilities of this software system. As Clementine® is designed to have a broad commercial application and is thus capable of several other functions, this is an overview concentrating mainly on features as applied to the present study.

The main machine learning paradigms found in Clementine® are:

1) The Connectionist approach as exemplified by the neural networks and Kohonen nets.

2) The Inductive approach as exemplified by the C5.0 and C&R Tree algorithms.

3) Logistic regression, a statistical technique useful for non-linear multivariate data.

5.2.4.2 The operations palettes

These are subdivided into different categories and are described according to their function. These are the sources, records, field, graphs, modelling, generated and output palettes.

5.2.4.2.1 Sources palette

This is the first collection of nodes and is represented in the left corner of the lower section of the Clementine® screen. Nodes, shown as circular icons, are processing operations in Clementine®'s visual programming environment. Data flows from, into, or through a node. The sources palette contains nodes that can be used to gain access to a wide variety of data sources. There are 6 variants:

- Variable File - used for free-field ASCII data
- Fixed File - used for fixed-field ASCII data
- ODBC -open database connectivity, a data exchange interface which allows programs of various types to exchange data with each other

- SPSS import - reads data from SPSS files
- SAS import - reads data from SAS files
- User Input- used to replace existing created source nodes. When the data are imported into a source node, they can be examined more closely by displaying them in a table or by running a statistics report. (These nodes would be found in the Output palette.) This phase is called the *data understanding* phase. Further insight into the data can be gained by using further output nodes such as Tables and Graphs.

5.2.4.2.2 Records palette

Record operation nodes are used to make changes to the data set at the record level. These operations are important during the data preparation phase because they allow the data to be tailored to the particular need. The record operations contain the following nodes, which are able to select and manipulate data according to a wide range of criteria chosen by the user. These are:

Select: This node allows an "include" or "discard" selection based on user defined conditions. For example, one could use this node to include or discard all cases identified as "heavy smokers".

Sample: A Sample node will allow the sampling of a maximum number of cases chosen by the user in the form of i) a random percentage, ii) a 1-in –n, where "n" is an integer, or iii) the first batch of a number decided by the user.

Merge: The Merge node can take multiple inputs in order to create a single output record containing some or all of the input fields.

Balance: The Balance node is used to correct imbalances in datasets. It increases or decreases the proportion of records where a specified condition is true, for example "Diagnosis = Positive"

Sort: These nodes sort records into ascending or descending order based on values of one or more fields.

Distinct: A Distinct node checks for records that have the same values in selected fields and either passes the first, or all but the first distinct record found. The mode of operation would then allow

whether to include only the first of such records found, or to discard this and pass all others.

Aggregate: This node replaces the sequence of input nodes with summary aggregated output records. The record structure downstream of an aggregate node will normally be different from that upstream as fields not involved in aggregation are dropped. Key fields allow the specification of a "key" for aggregation. The values of all key fields are combined to produce one key value and one aggregated record is returned for each unique set of values.

Append: The Append node is used to concatenate sets of records. This node is very similar to the merge node except that, unlike the merge node which joins records from different sources together, the append node reads and passes all the records downstream from one source until there are no more, then goes on to read from the next source.

5.2.4.2.3 Field operation palette

Field operation nodes prepare the data for modelling and any further downstream operations. The Field operation palette contains the following six nodes:

- *Filter node:* This node has two functions, filtering out selected unwanted fields from records which pass through them, and renaming fields.
- *Type node*: This node specifies three important properties of fields: the type, direction, and blank definition. A field's type can be an integer range defined by upper and lower bounds, for example 0-100, or a real range, for example 0-1, or a set defined by multiple symbolic values such as, moderate drinker, heavy drinker. A flag as defined by two symbolic values such as, positive, negative, or true, false or typeless, where no information is gathered about the data, such as record numbers.

The direction of a field is relevant to modelling nodes. A field's direction can be:

In: where the field will be used as an input to machine learning.

Out: where the field will be used as the output that the machine learning will try to predict. *Both*: where the field will be used as both an input and output and this is relevant for the GRI and *aprori* nodes. All other modelling nodes will ignore this field.

None: the information in this field is ignored by the machine learning models.

Blank : This allows the user to control what to do with values that may represent missing values. The operations are able to treat certain data values as blanks and still process the other data without system conflicts.

Derive Node: The derive node adds a new field to every record that passes through it. There are six variants to this node:

- *Any*: The new field is the result of an arbitrary expression. For example a Derive node for a test set can be given a formula of "Training set" = datasets assigned numbers >1.
- *Flag*: The new field is a flag reflecting a specified condition. For example, True value = "P", False value = "N". True if: Diagnosis = "Positive"
- *Set*: The new field is a member of a specified set of values.
- *Count*: The new field is based on the number of times a condition has been true.
- *State*: The new field is one of two states. Switching between these states is triggered by specified conditions.
- *Conditional:* This derived field is in the form of an "If...", "Then...", "Else..." format

Filler node: These nodes allow the facility to select fields within the data whose values can be examined and potentially replaced.

History node: History nodes are for use mainly with sequential data, for example, time series or temperature series data. Their function is to place previous values of one or more such fields in current records.

Set-to-flag node: This node provides the means to flag or label values within datasets with more than one attribute so that those flagged values can be cross referred according to differing criteria. A commercial example of this is when there is a product within a large department shop that can be bought from more than one department within the same shop. A set-to-flag key would enable the user to see which department that product was bought from.

5.2.4.2.4 Graphs palette

These are displayed as triangular icons and there are seven types of graphs:

Plot node: Plot nodes are used to show the relationship between values of two numeric fields.

Distribution node: This shows the occurrence of values for non-numeric fields.

Histogram node: This shows the occurrence of values for numeric fields. All records are read and a histogram displayed. The range can be the complete range of values or user defined.

Web node: This node allows the strength of relationships between values of two or more symbolic fields to be displayed. The connections are shown graphically as dotted normal or heavy lines of varying colours in order of strength of association.

Collection nodes: These are similar to the Histogram node except that rather than showing the occurrence of the values for a single field, they show the distribution of values of one numeric field relative to values of another. An example of this in commerce (Clementine® is predominately a package with wide industrial applications) is the instance of records of job lots passing through a factory. If each record has fields describing lot size and the time at which the lot was completed, the varying output of the factory through the day can be plotted by collecting lot sizes over completion time.

Multiplot node: This is a variant of the Plot node, the difference being that multiple "Y" fields, displayed in different colours, can be displayed simultaneously over a single "X" field

Directed web: this is a variant of the Web node which shows only connections from any number of "from" fields to a single "to" field. Similar to the Web node, connections are shown graphically with dotted, normal and heavy lines in order of increasing strength.

5.2.4.2.5 Modelling palette

There are eleven modelling nodes; the following is a brief overview of these nodes.

- *Train Net:* This node is used to create and train a neural network. This is in the form of a Multi-layer perceptron and typically has an input layer representing the input fields, one or

more hidden layers and an output layer. This node is able to handle numeric, symbolic and flag inputs and outputs. Several training methods are allowed:

- *Quick* method, where the program chooses an appropriate topology and working parameters for the network.
- *Dynamic* method, where the program begins with an initial topology and modifies this topology by adding or removing hidden units as the training progresses.
- *Multiple* method where several separate networks of differing topologies are created and these are trained simultaneously. At the end of training the model with the lowest error is presented as the final model.
- *Prune* method which starts with a large network and removes (prunes) the weakest units in the input and hidden layers as training proceeds.
-*Radial Basis Function Network (RBFN)*. This method uses a technique similar to k-means clustering. These clusters form the single "hidden layer" and feed onto the "output" layer.

With the exception of the Dynamic method, all these methods have an expert options mode where low-level parameters for training can be selected. This allows the training process to be fine-tuned. Within this option, the number of layers can be specified, as well as the Persistence, the momentum of the training, that is the updating of the weight changes during the training (Alpha), and the Learning rate (Eta).

- *Build C5.0:* This node uses a C5.0 algorithm. Two kinds of model can be produced, a decision tree and a rule set. An added feature is the boosting method, which increases the accuracy of classification. This works by building a sequence of models that focus on the errors or misclassifications of the earlier model. Finally, cases are classified by applying the whole set of trees to them, using a weighted voting procedure to combine the separate predictions into one overall prediction. There is also an option to favour accuracy or generality. A greater accuracy from the training data could result in overfitting and therefore could lead to poor performance when the model is applied to new data.

Expert options allow the severity of pruning to be increased or decreased. An increased pruning severity would result in a smaller tree and a decrease would tend to obtain a more accurate tree. The minimum records per branch can be limited. This helps to prevent overtraining with noisy data. Misclassification costs can be set such that the relative importance of prediction errors can be specified.

- *Train Kohonen:* This node creates a Kohonen network or Knet. This is a variant of neural networks and is referred to as "unsupervised learning". Instead of trying to predict an outcome, Kohonen nets try to uncover patterns in the set of input fields. Records are grouped such that they tend to be similar with the group or cluster. Kohonen networks are also known as self-organising maps.
- *Linear Regression:* This model estimates the best fitting linear formula for predicting the output field based on the input fields. The regression formula represents a straight line or plane that minimises the squared differences between predicted and actual output values. Only numeric fields can be used in a regression model.
- *Generalised Rule Induction (GRI):* This node discovers association rules in the data. Association rules are in the form of "if... then...." The information content is measured using an index that takes both the generality and accuracy of the rules into account.
- *Apriori:* The Apriori node also discovers association rules in the data. It tends to be faster to train than the GRI and has no limit on the number of rules that can be retained. There are five training methods available thus allowing more flexibility in matching the datamining method to the problem.
- *Train K-Means:* This is a further variant of cluster analysis. There is no target output and records are grouped such that records within a group tend to be similar to each other. This node is often the fastest means of clustering large datasets.
- *Logistic Regression:* This node works by building a set of equations that relate the input field values to the probabilities associated with each of the output field categories. Once

the model is generated, it can be used to estimate probabilities for new data. Both symbolic and numeric input fields can be handled.

- *Factor/PCA:* There are two similar but distinct approaches, which are used to reduce the complexity of data. This technique therefore builds more robust models that can execute more quickly than would be possible with raw input fields. Factor analysis attempts to identify underlying concepts or factors that explain the pattern of correlations within a set of observed fields. Factor analysis focuses on shared variance only. Variance that is unique to specific fields is not considered. Principal components analysis (PCA) finds linear combinations of the input fields that are best able to capture the variance in the entire set of fields. Both shared and unique variances are considered. Only numeric fields can be used in this model.
- *Two Step Cluster:* This is another method of clustering data. The first step makes a single pass through the data during which the raw input data is compressed into a manageable set of subclusters. The second step uses hierarchical clustering to merge the subclusters into larger and larger clusters. This has the advantage of not requiring the number of clusters to be selected ahead of time.
- *Classification and Regression (C&R) Trees*: This is very similar to the C5.0 algorithm and can accommodate numeric as well as symbolic values. It also uses recursive partitioning to split the training records into segments with similar output field values. The split defines two subgroups, each of which is subsequently split into two more subgroups, and this continues until one of the stopping criteria is triggered.

5.2.4.2.6 Generated nodes (nuggets)

Generated nodes, described in Clementine as "nuggets", represent the results of modelling and learning and are divided into 2 classes:

1. *Non-executable* models, which are unable to yield answers. These can be browsed and any interesting associations observed.
2. *Executable* models can be used and tested. These are a type of manipulation node which modifies records by adding their predictions as new fields.

5.2.4.2.7 Output nodes

Output nodes provide the means to obtain information about datasets and generated models. These nodes provide a means of exporting data in various formats to interface with other software. There are fourteen output nodes and many offer a choice of output styles. The following options are available: Screen, Printer, Formatted (*.tab), Text file (*.txt), Data file (*.dat), HTML (*. html). The output palette consists of the following nodes:

- *Table node*: This node allows a table to be created from data. It is usually connected to a Source node and the table produced can be displayed on screen or written to a file.
- *Matrix node*: This node displays the relationship between fields. The information is displayed in the form of a grid, for example, with columns for numbers of positive or negative and rows for males and females.
- *Analysis node*: This node performs various comparisons between predicted values and actual values for one or more generated model nodes and allows the analysis of these generated predictive models in order to evaluate their ability to generate accurate predictions.
- *Statistics node*: This node provides basic summary information about numeric fields.
- *Set Globals node*: This node scans data and computes summary values that can be used in CLEM expressions. CLEM is an "in-house" range of expressions that will then enable the average, sum, minimum, maximum and standard deviation of values in a selected field to

be expressed.

- *File node*: This allows data to be written to a delimited text file and is useful for exporting data and also for appending to other files.
- *Report node*: This node allows data from records to be mixed with free-format text output. A template specifies how the report is to be constructed. It can fall into fixed, where text and data are simply reported, or global, where the text refers to specific conditions, for example "for record xx, the value of age is nn, or conditional, where a typical line would be "where sex=M, age=nn".
- *ODBC node*: This node is used to write data to ODBC compliant relational data sources.
- *SPSS Export node*: This node allows data to be exported in the SPSS.sav files. This can then be read by other SPSS products.
- *Quality node*: This node reports on the quality of the dataset by checking for missing values or blanks.
- *SPSS Procedure node*: This node allows an SPSS procedure to analyse Clementine® data and either import the results into Clementine® or save the results in a normal SPSS output file.
- *Excel node*: This node exports data and opens it in the form of a spreadsheet in the Microsoft Excel format.
- *Solution Publisher node*: This allows an entire Clementine® stream to be exported in order that the stream can be embedded in other external applications.
- *SAS file node*: This option writes records to an SAS file.

5.2.4.3 Summary

Clementine® is a datamining software package with embedded machine learning capabilities. It also contains extensive and versatile capabilities to enable the examination and evaluation of data via a friendly user interface. Designed primarily for non-technical users, it contains editable

features and applications in many facets of industry where likely users would be executives and policy decision makers. This study makes use of the machine learning capabilities within this datamining package and the simplicity of this format for application in a clinical context. This looks prospectively toward the possibility of its use in future similar studies.

5.3 An experiment using Clementine® for the identification of high risk groups within a dataset

5.3.1 Introduction

Several aspects of Clementine®'s data manipulation and machine learning capabilities were explored using the 1995 dataset obtained from various sources including a hospital outpatient's department, a general medical practice and executive staff in an industrial company, (Julien *et al* 1995; Downer *et al,* 1995). This was an early experiment to "put Clementine® through its paces" in order to ascertain that such a software package developed primarily for commercial and industrial use contained sufficient features that could be adapted for effective use in a clinical setting. Several features of this package were explored in this experiment, although not all of them were later used. Some instances of these were different methods of arriving at the same goal, for example different methods of random data sampling.

5.3.2 Aims

- To test the efficacy of the machine learning capabilities of Clementine 6® (SPSS Inc), in the identification of sub-populations at higher risk of oral cancer and precancer.
- To find the optimum negative to positive ratio of cases that would yield the best results for the machine learning models.

5.3.3 Materials and methods

A dataset obtained from two previous pilot screening programmes was used. This comprised lifestyle data from 2024 individuals who had been screened for oral cancer and precancer. The available dataset was obtained from a number of studies (Jullien *et al,* 1995; Downer *et al,* 1995), and consisted of 1927 negative cases and 97 cases that had cancerous or precancerous lesions. All the cases had the "gold standard" diagnosis made by a specialist.

The fields present in this demonstration study were:

Rec_no:

Sex

Specialist_diagnosis

Age_years

Male

Female

Smoked_in_last_10_yrs

Smoke_over_20-cigs_pd

Smoking_over_20_yrs

Drinks_beer_or_wine

Drinks_fort_or_spirit

Drinks_over_limit

Dentist_in_last_12_mths

The fields for age, gender and levels of smoking and alcohol consumption were used, as was the consumption of beer or wine and fortified wines or spirits. These were included to investigate the possibility that the consumption of different sorts of alcohol, namely beers, wines, spirits and fortified wines, could influence the outcome. Blot *et al* (1988), and Kulasegaram *et al* (1995) produced evidence that spirits and beer were more important risk factors for oral cancer than wine. Other workers (Franco *et al*,1989; Barra, 1990), found the highest risks to be associated

with wine consumption. However, Mashberg (1993) and Johnson and Warnakulasuriya (1993) take the view that there is no difference in the risk potential between different types of alcohol and the variables were due to the amounts consumed.

The raw data in the form of two data files, an all-negative set and an all-positive set, was imported into Clementine® via a Source node. These Positive and Negative sets were appended together into one file. Some other fields were added, such as "nonsmoker" field, a "heavy smoker and heavy drinker" field, then a selector field to assign a random number to each case in the dataset. All records within this dataset were then randomly assigned a number between one and three. This was done via a Derive field node (Fig 5.2) using the expression "Random(3)". Then for a following node the expression "Random(3)>1" gave the possibility of two answers, "true" or "false". A "true" then grouped the records with numbers 2 and 3; while "false" defined the group with the number 1. In this way, two separate groups of data were randomly generated in the proportion 1:2, one, with the numbers "1" and the other including the numbers "2" and "3". Another method of deriving data samples was by a Sample node shown in Fig 5.3, where the data could be selected on the basis of a percentage, an "1-in n" or a "first n". This node was also used to derive data for training. The purpose was to create two separate datasets, a test set, consisting of one third of the data, and a larger training set, consisting of the other two thirds. After randomisation and separation, the derived training set consisted of 1309 negatives and 64 positives and the test set had 640 negatives and 33 positives. This smaller "test set" was separated from the rest of the experiment to be used later for testing the trained machine learning models

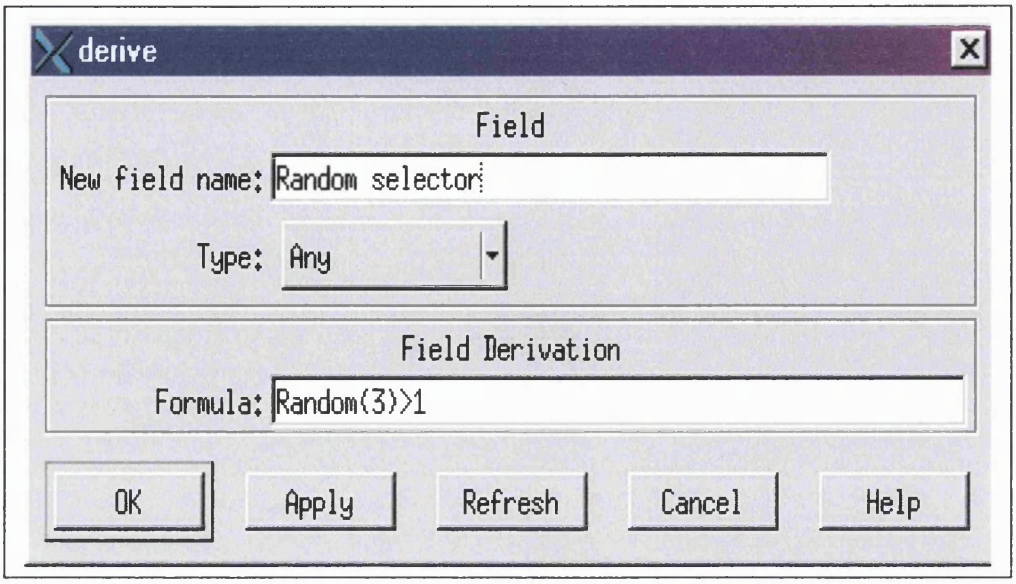


Figure 5.2 A Dialogue box from a Derive node where a new field "Random selector" is produced. The formula expressed partitioned the dataset into cases that were given numbers 2 and 3 (>1) and another set where cases had been given the number 1.

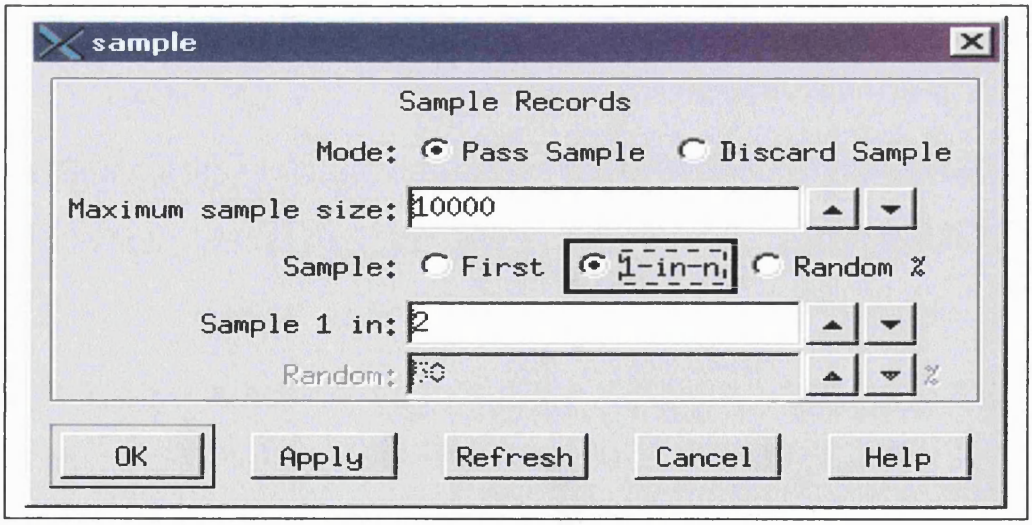


Figure 5.3 A dialogue box from a Sample node illustrating the various means whereby sample records can be passed on or discarded.

The "training set" was used to train the neural network and the C5.0 algorithms. This training set was divided up once again between negative and positive records to create 12 training sets with varying ratios of negative cases. Due to the small number of positive cases, all of the positive cases were used while the number of negatives cases was varied. So proportions of 0.25 negatives to 1 positive, 0.5 negatives to 1, 1:1 and so on to 10:1 were produced. The purpose of this was to determine the performance of the learning algorithms over the varying ratios of negative and positive cases and to find the ratios that would give optimum results. A neural network model was generated then C5 models. These were tested on the unseen test set both singly and in combination. Sensitivities and specificities were calculated for each of the 12 training sets. Each model was given an identifying name according to the sequence trained and then tested singly and in series on the unseen test set. The combination was identified as "$X-Specialist_diagnosis" which is a function recognised by Clementine® to indicate a special category of diagnosis where positives identified by either model counted toward the total of positives.

A representative dialogue box of this function by a derive node is shown in Figure 5.5. There were other methods of drawing data randomly from the pool of negative cases; these were done:

- by the use of a *sample node*: (Fig 5.3) which draws out an exact number of records, which could be the first "n" number of records, a "1-in-n", for example one in every ten cases, or simply a random percentage.
- by the use of a *balance node* (Fig 5.4) where a multiplier can be applied. The positive set would be multiplied by 1 to maintain the number of positive cases at 64, while the negatives would be multiplied by 0.05 to obtain a number of negative records approximating 64 for the 1:1 ratio (1309 x 0.05 = 65) and 0.1 for the 2:1 ratio (1309 x 0.1= 130).

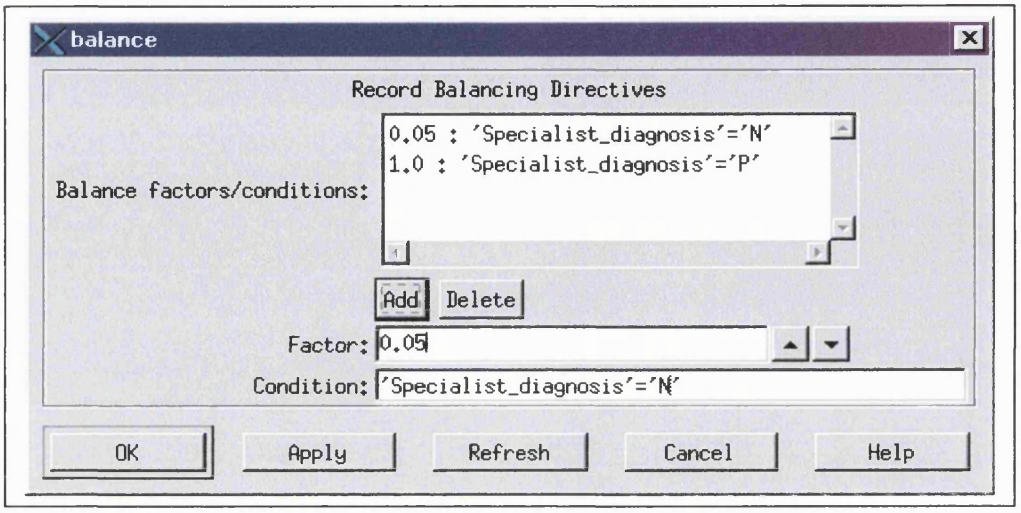


Figure 5.4 The dialogue box from the Balance node showing the balance factors used to return a better proportion of negative and positive data.

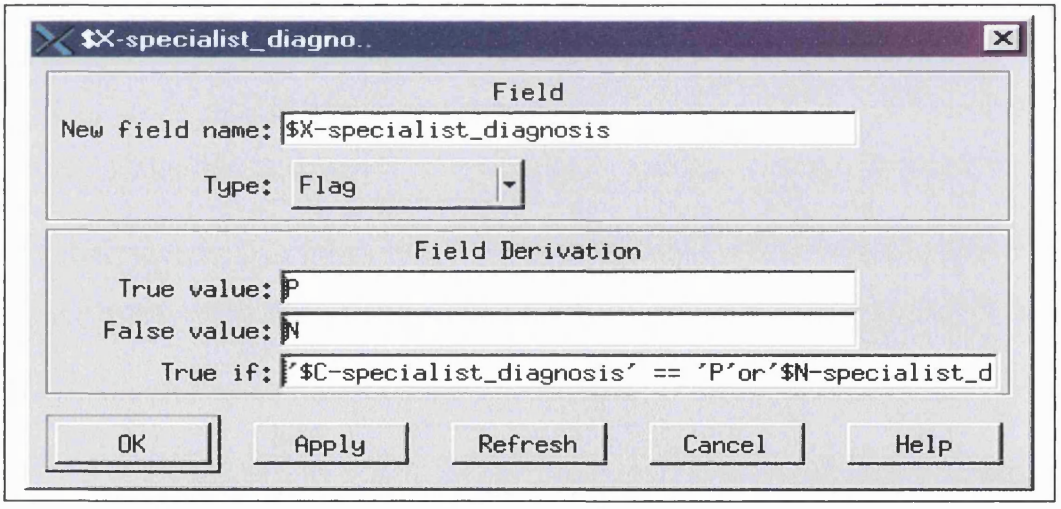


Figure 5.5 The expanded dialogue box for the Derive node labelled as $X-specialist_diagnosis, which combined the performance of a C5.0 and a neural network model with the Clementine® expression: "True" if: [' $C-specialist_diagnosis == 'P' or ' $N-specialist_diagnosis == 'P']. This node causes either a recognition by the neural network model or the C5 model to be logged as a positive.

Both methods were used and the results tabulated in the form of ratios, sample sizes, performances of the C5.0 algorithm and the neural network individually and with the two combined in series against the unseen test set.

A further method of drawing randomised data was the use of the Kohonen Net method of data sampling. This technique was also used in this part of the study. A Kohonen Net is a form of neural network categorised as unsupervised learning, which attempts to cluster data that have similarities. Representative samples of negative and positive data are drawn from these clusters and later recombined as training and test sets. The purpose of this was to eliminate possible bias in data that could have inadvertently resulted from random selection. This makes use of a variant of neural networks described as Self Organising Maps (see Chapter 3) attributed to Kohonen. The data is clustered according to similarities within the set and samples and displayed on a graph in the from of clusters of data; representative cases from each cluster were drawn and recombined into test and training sets for machine learning. A table was then created of the values and results obtained (Table 7.1). This table was used as an illustrative example of the other tables produced as all the data sampling resulted in similar information.

The machine learning models produced were all given alphanumeric identifiers. A further 10 samples of data were drawn in proportions of 1:1, then a further 10 in the positive to negative proportions of 2:1. Graphs showing the performance of these machine learning models were produced, with sensitivity and specificity plotted against varying ratios of negative to positive data caches. Three such graphs were produced and all were alike in shape and trend; Table 7.2 provides a representative example of all three. The best-performing models were combined variously and tested against the test set and results tabulated. The machine learning models were then tested against a "cut-down" sample of the test set to emulate similar numbers of test cases from the earlier work by Speight *et al* (1995) where the test set consisted of 10 positive cases and 355 negative cases. This reduced, or "cut down", test set now consisted of 16 positive cases and 309 negative cases. This table, Table 7.2, was shown as an example of the other similar tables that were produced. The derived machine learning models were also tested in combination by

connecting them up in series via the use of a Derive node, where a new term $X specialist _diagnosis was created to link the neural network and C5.0 algorithms together. The main features of this experiment are shown as a Clementine® onscreen image in Figure 5.6 and a flowchart in Figure 5.7

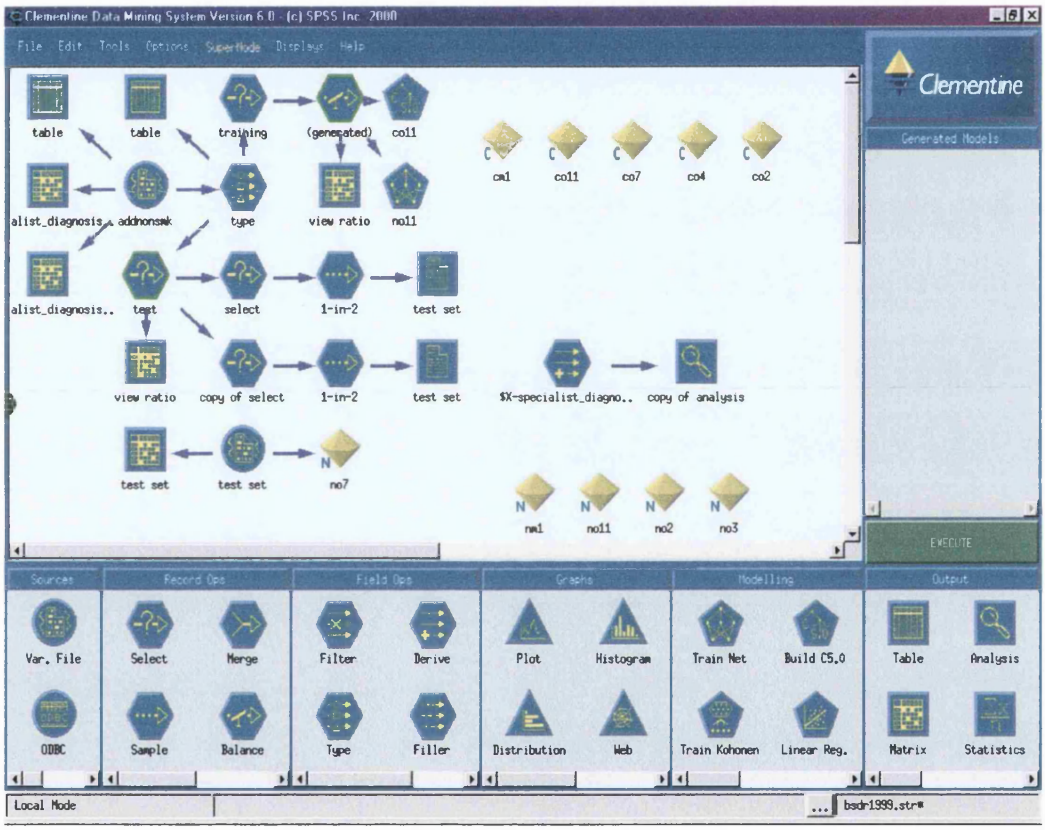


Figure 5.6 A typical Clementine® onscreen image of a working pane displaying the various streams used to obtain training and test sets, and the stream used to produce the machine learning models. The machine models, seen as yellow "nuggets", were then tested on the test set and results analysed via the analysis node.

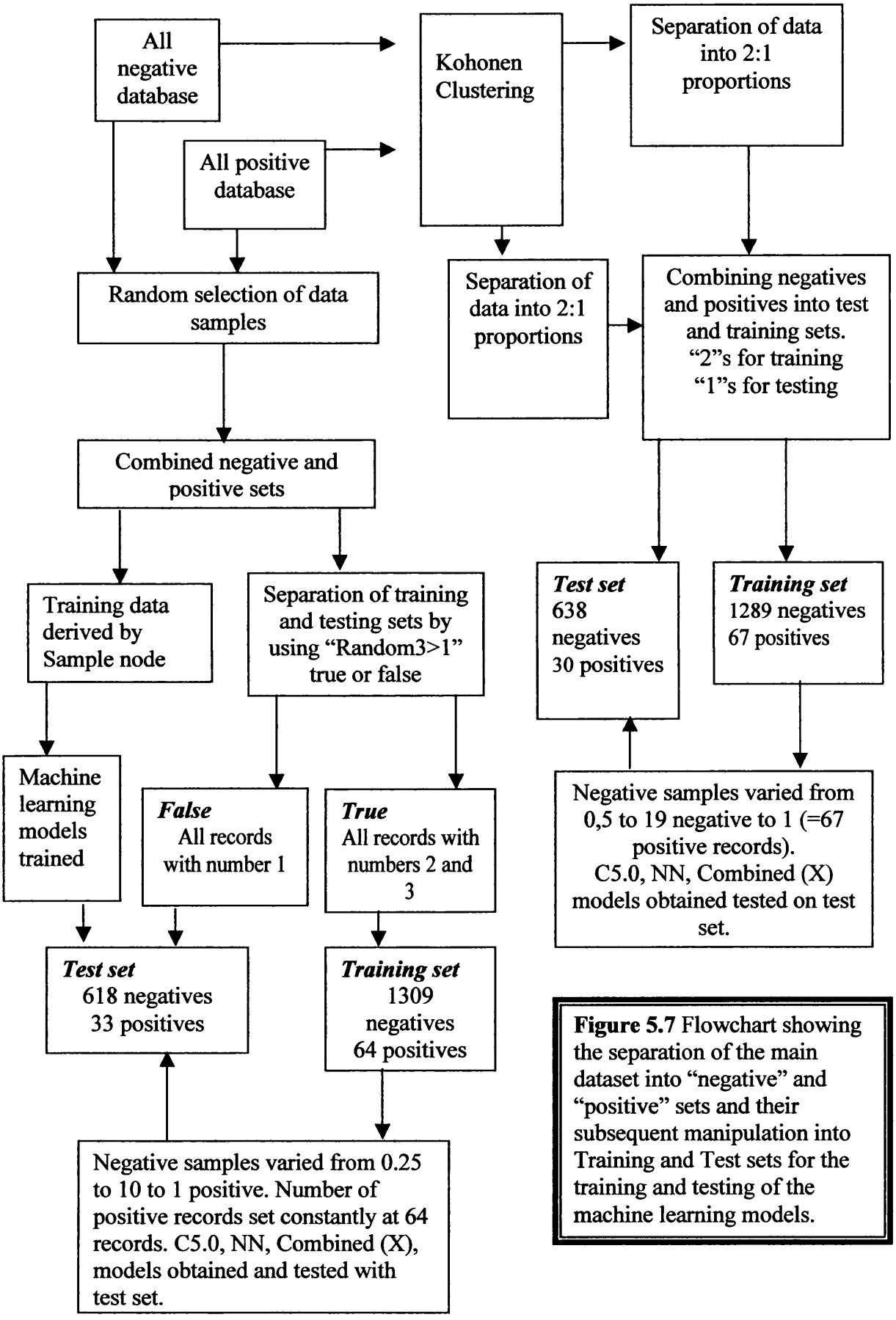


Figure 5.7 Flowchart showing the separation of the main dataset into "negative" and "positive" sets and their subsequent manipulation into Training and Test sets for the training and testing of the machine learning models.

5.4 A comparison of the descriptive epidemiology of an earlier available dataset and one obtained by screening in a general dental practice setting

5.4.1 Introduction

The two datasets described represented a total of over 4000 cases where individuals in the south of England, mainly in London, were screened for oral cancer and precancer. As the sets of data were gathered some five years apart, this section compared the descriptive epidemiology of these two groups. The first group had the outcomes verified by a specialist and hence had a "gold standard" whereas the second group was screened by general dental practitioners. Since this was an observational study of prevalence of cancer and precancerous lesions and conditions, no performance evaluation of the diagnoses was carried out. This section investigated results as entered by general dental practitioners and compared them to the profiles of the earlier dataset. The study also compared the prevalence of positive lesions and conditions with the 1995 dataset where the prevalences could be considered to be benchmarked. The distribution of risk habits and indeed the distribution of positives within the various categories of these two datasets were observed. This comparison of datasets also anticipated the use of one set (the first dataset) as a training set for the machine learning to be tested on the second dataset.

5.4.2 Aim

The aim was to make a comparison of the descriptive epidemiology for age, gender, lifestyle habits of smoking and drinking and dental attendance between a dataset obtained from earlier evaluation studies for oral cancer and precancer screening (Jullien *et al*, 1995; Downer *et al*, 1995) with that obtained from the screening study in general dental practice referred to in Section 5.1. It was to observe differences and similarities in these two datasets obtained some five years apart, the first, from available data from different but closely related sources in the UK with positive cases confirmed by a specialist, hence with a gold standard and the second obtained via

an opportunistic screening study in a general dental practice setting done by general dental practitioners, but without specialist confirmation of the positive cases.

5.4.3 Materials and methods

The first dataset was a composite from a number of screening studies that were carried out mainly in an outpatient department of a dental hospital and by postal invitation at an inner city medical practice (Jullien *et al,* 1995). Individuals over the age of 40 years were examined for lesions and conditions deemed as positive. This set was further augmented by a batch of data derived from a study in an industrial setting (Downer *et al,* 1995). The second set of data was from the multicentre opportunistic study carried out in general dental and described in Section 5.1. The age, gender, smoking and alcohol consumption and dental attendance categories for both datasets were aligned so that both sets had the same headings and criteria for heavy, moderate, non- or light users (smokers and drinkers). Positives were defined by criteria outlined in Table 5.1 and groupings and amounts for smoking and drinking were set according to established guidelines as shown in Table 5.2. Both datasets were modified to make them compatible with each other, aligning the results obtained from each dataset into fields according to age, gender, non- or light, moderate and heavy smoker and drinker categories, and dental attendance. The earlier dataset contained fields on the consumption of beer, wine, spirits and fortified wines and these were removed as they did not form part of this study. Alcohol consumption was measured in units, a unit in the United Kingdom being equal to 10ml or 10 grams of alcohol (ethanol). The units were then converted into three categories: light or non-drinking, moderate and heavy, according to the amount consumed per week, adjusted for gender. The smoking habit was categorised by the number of cigarettes smoked in a day, again, into non- or light, moderate and heavy. As there were insufficient numbers of pipe smokers and users of smokeless tobacco in this study, these groups were not included. Some cases were removed as the respondents gave unclear or incomplete answers to some fields. An example of this was when the "non-smoker" box was

ticked, yet the respondent would indicate that they were at the same time a heavy smoker. In the second (current) study, 236 individuals who were aged between 35 and 39 years had to be removed in order that both sets had a similar age range. By limiting the dataset to the same age range as the earlier study, that is, those aged 40 years and above, four of the positive cases who were aged under 40 years were discarded. The field for ethnic groups was also removed, as comparison was not possible because ethnicity had not been included in the earlier study. Results were shown in the form of tables, both in numbers (Table 8.1) and percentages (Table 8.2), and in bar charts in Appendix 2 comparing values observed in the 1995 and 2000 datasets showing, age (Figs 8.1 and 8.2), gender (Figs 8.3 and 8.4), smoking (Figs 8.5 and 8.6), alcohol consumption (Figs 8.7 and 8.8) and dental attendance (Figs 8.9 and 8.10) for the two datasets. The two datasets were further compared by means of Incidence Rate Ratios using the STATA® statistical software package, Version 7, STATA Corporation, Texas USA, using Poisson Regression.

5.5 Evaluation of machine learning techniques for the detection of groups at high risk of oral cancer and precancer

5.5.1 Introduction

The earlier study referred to in Section 5.3 demonstrated that it was possible for a neural network, written for the purpose, to identify populations who were likely to have a positive lesion by using fields such as age, gender, smoking and drinking habits (Speight *et al*,1995), all information which can be easily gathered from patients in the normal course of dental examination and treatment. The prospect of using such a means of delineating a population at high risk is studied here in greater detail using "off the shelf" neural network programs and also other machine learning paradigms.

5.5.2 Aims

- To generate machine learning models provided in a commercially available datamining software package and evaluate its performance in the identification of sub-populations at high risk of having a cancerous or precancerous lesion or condition.
- To test a hypothesis that individuals who are heavy smokers and heavy drinkers are naturally at higher risk of having positive lesions and conditions.

5.5.3 Materials and methods

The dataset from the earlier 1995 study which was used as the training set comprised 2024 cases, of which 97 were considered to have positive lesions or conditions. The other dataset was that obtained from the screening study in general dental practice described in Section 5.1. This is referred to as the "2000 dataset" and was made up of a total of 2029 cases, 1939 negatives and 90 positives. Although this set originally had 2265 cases, 236 cases were excluded as they were aged under 40 years. This was to make both datasets comparable to each other, as the 1995 dataset consisted of adults aged 40 and above. Clementine®, the datamining package as described in

Section 5.2.4, was used in this experiment. This software package was the source of our machine learning models. From a suite of available models within this package, four different types of machine learning models were selected on the basis of their ability to handle numeric and categorical data. These were the neural networks, C5, Classification and regression trees and Logistic regression models. The experiment was done in two ways.

1. The 1995 dataset was used as the training set and the entire unseen 2000 dataset was used as the test set.
2. The 1995 and 2000 datasets were combined and a randomly drawn set of approximately a third (1365 cases) of the total database was separated to be used as the unseen test set. The machine learning models were trained on the remaining (2688) cases. The justification of pooling the two datasets together was the close similarity of the two datasets in terms of age, gender distribution, distribution of smoking and drinking habits and the amounts of positives within each category (seen in the results described in Chapter 8).

5.5.3.1 Training the machine learning models on the 1995 dataset and testing on the 2000 dataset

The training datasets were entered into Clementine® in the form of Microsoft Excel Comma Separated Values (CSV) files via a Source node (see Section 5.2). This was then connected to a Filter node to filter out redundant fields like "record number", and "dentist" (an identifier for each examining practitioner). As the number of positive cases in the training set was low (97) compared to the numbers of negative cases (1927), all of the positive cases were used in the training. The negative cases however had to be limited to suitable numbers so that the machine learning models would able to "see" them in the presence of 97 positive cases. There were a number of methods available for drawing random numbers of negative cases from the set of negative cases whilst still retaining the same number of positive cases in the batch; these were described in Section 5.3.3. As none was of special benefit over the other, the use of the Balance

node was chosen. Negative cases were selected via the Balance node in the way described in Section 5.2; in this instance, setting the multiplier to 0.05 produced a cache of 93 negative cases to the cache of 97 positive cases. This provided a ratio of approximately 1:1 negative to positive cases. This cache was labelled as 93:97 and saved as a Clementine® file. The multiplier was then set to 0.03 and this drew 64 negative cases. This gave a batch of data having an approximately 2:3 negative to positive ratio. This cache was named 64:97 and also saved as a Clementine® file. Sets of machine learning models were generated by training on the 93:97 cache, then the 64:97 cache. The cached data batches were streamed to a Type node. This is a necessary step before the actual machine learning process as this node specifies the output field and arranges the other fields in ways that the machine learning model can identify and deal with according to the type of information present, for example numerical or categorical. The Type node enabled the selection of the fields required for the training and the choice of the output field. In this case, "diagnosis" was required so "diagnosis" was set as "out" in the "Direction" column. Fields used for training were set as "in" in the "Direction" column. Fields not required in the training process are deselected by choosing the "none" label in the "Direction" column. Figure 5.10 shows the start of the data stream and the Filter, Balance and Type node dialogue boxes superimposed to show the editable parameters. The data was then connected up to the Modelling nodes. The modelling process then produced a generated model which appeared as a golden nugget in the palette on the far right of the screen. This was brought into the working pane for analysis by mouse-clicking on the icon and clicking again on a space on the working screen. The test set in the form of a Source node was connected up to the machine learning model, which was in turn connected to an Analysis node. Fig 5.8 shows this process of data manipulation via the print screen function.

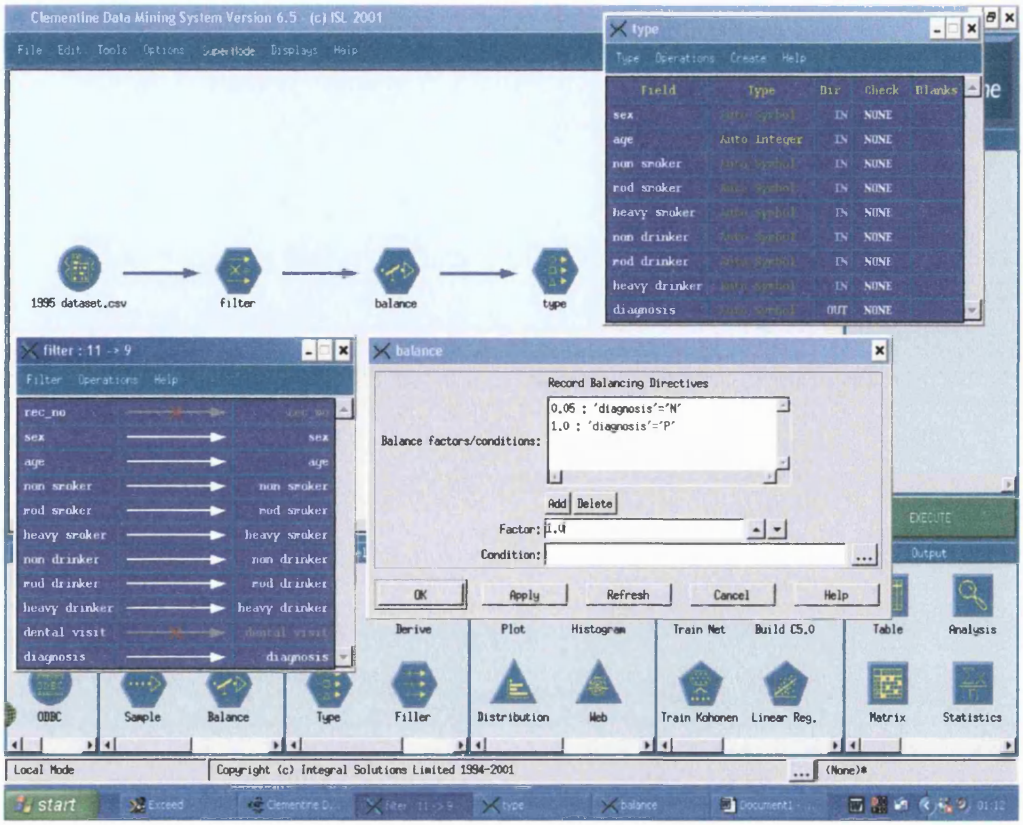


Figure 5.8 A screen shot of a data stream just prior to connection to a machine learning node. The data streams from the Source node (far left) to a Filter node then to a Balance node then to a Type node. Superimposed on this are the three dialogue boxes of the Filter, Balance and Type nodes showing the way that the data were processed.

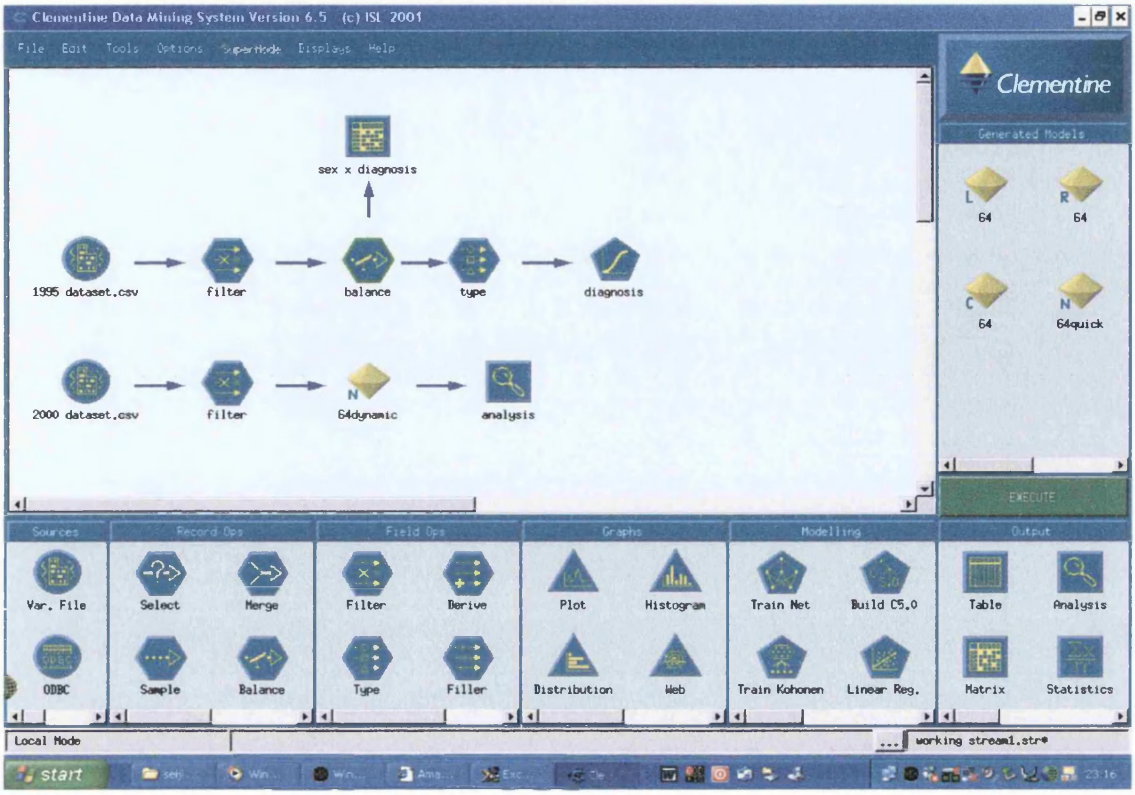


Figure 5.9 The streams for data for training (above) and testing (below). The generated machine learning models can be seen in the palette on the far right of the screen. The first stream, seen in the upper part of the pane, shows data from a Source node filtered; the Filter node is seen as a hexagonal node where non-essential data items, such as dentist's initials, can be filtered out. This is connected to a Balance node thence to a Type node then to the Modelling node. The Balance node, usually in blue, is seen at this time in green; this colour indicates that the draw of cases is cached and can be saved for recall. This is then connected to the Type node. At this node, the output field is selected. The working node with machine learning capabilities is the Modelling node. The modelling node is shown as a pentagonal icon and connected at the end of the data stream and labeled "diagnosis".

5.5.3.1.1 Overview of the machine learning models used

Not all the available machine learning models within Clementine® were relevant for the purposes of this study. Following is a brief overview of the machine learning models used.

i) Neural nets

The *Neural Network* models consisted of a number of variants for the training protocol as described in Section 5.2.4.2.5. The following training variants were used:

Quick: Analyses data to choose the appropriate topology for the network.

Dynamic: Creates an initial topology, and modifies this by adding and/or removing hidden units as training progresses.

Multiple: Creates several networks of different topologies; the model with the lowest error is presented as the final model.

Prune: Starts with a large network and removes (prunes) the weakest units in the hidden and input layers as training proceeds.

RBFN (radial basis function network): Uses a clustering facility to partition data based on values of the output field.

ii) Decision trees and Classification and Regression Trees (CART).

The C5.0 node has two variants, the decision tree and the Ruleset.

The decision tree is a straightforward description of the splits found by the algorithm. This was used on the main default setting with the mode set to generality rather than accuracy.

The Ruleset variant which was featured in the earlier versions of Clementine® is superseded by the Classification and Regression Tree model and was therefore not used.

The Classification and Regression (C&R) Tree node is also a tree-based classification and prediction method. It is similar to the "ruleset" variant of the C5.0 algorithm; this method uses recursive partitioning to split the training records into segments with similar output field values. This node was used on its default settings which had the mode set to "simple" and maximum tree depth set to "7".

iii) Logistic regression

The Logistic regression model is a statistical technique for classifying data based on values of input fields. This works by building a set of equations that relate the input fields to the probabilities associated with the output field category. Training of the Logistic regression model was done on the "simple" mode setting.

All the machine learning models had "expert" options which are editable parameters. These allowed models to be "fine tuned", usually to achieve greater accuracy, for example, in the C5 algorithm it was possible to "penalise" the model for misclassification by setting up misclassifications costs within an editable dialogue box, and cause it to be more biased toward a positive or negative "verdict" as chosen. For the purposes of this study, the training of machine learning models were all done on their default modes. This was to avoid overtraining and overfitting the models, a phenomenon that would tend to give good results within the same dataset but would perform less well on other unseen datasets.

5.5.3.1.2 Training of the machine learning models

Both the previously saved caches of negative and positive cases labelled as 93:97 and 64:97 were used to train each of the neural network variants, then the C5 and Classification and Regression Tree models then the Logistic regression model. The machine learning models derived from the various algorithms were saved in the "generated models palette" on the far right of the screen and tested one by one, on the unseen test set which was the entire 2000 dataset. This was done by linking up a Source node containing the 2000 data file and connecting that up to a Filter node then to the generated machine learning model illustrated in the lower part of the screen in Fig 5.10 and linked to an Analysis node. The Analysis node reported the results in the form of sensitivity and specificity.

5.5.3.2 Training the machine learning models on the pooled 1995 and 2000 datasets and testing them on a previously separated batch of unseen test cases

The methodology was essentially the same as the preceding set and would therefore not need further description except for the points that differ. The two datasets were combined resulting in a dataset consisting of 4053 cases, of which 187 were positives. A Derive node was used to randomly assign a number to each case from 1 to 3. This way, all data assigned the number "1" could be separated from the main set via a Select node and separately formed into a dataset for use later as the unseen test set. Cases with the assigned numbers "2" and "3" became the training set. Machine learning models were generated and results analysed as in the preceding description. This way, a training set of 2568 negatives and 120 positives was obtained with a test set of 1287 negatives and 67 positives. Fig 5.10 shows the process on a Clementine® working pane.

5.5.3.3 The training caches

Two sets of caches, one in the approximate proportion of 1:1, the other in the proportion of 2:3 negatives to positives were generated for each of the two methods of evaluation. In the first method, the 1995 dataset was the source of training data. Two sets of machine learning models (five variants of neural networks, C5, Classification Trees and Logistic regression) were trained on each of the two caches drawn from this set. The first cache consisted of 64 negatives to 97 positives, the second cache consisted of 93 negatives to 97 positives. In the second method, both datasets were pooled and training and test sets were derived from this combined database. The total number of cases in the training set was 2688, made up of 2568 negatives and 120 positives. Two caches were generated from this training set, one consisting of 115 negatives and 120 positives, the other of 70 negatives with the same 120 positives The test set consisted of 1354 cases, 1287 negatives and 67 positives. The same types of machine learning models were trained on the two caches and results analysed and tabulated.

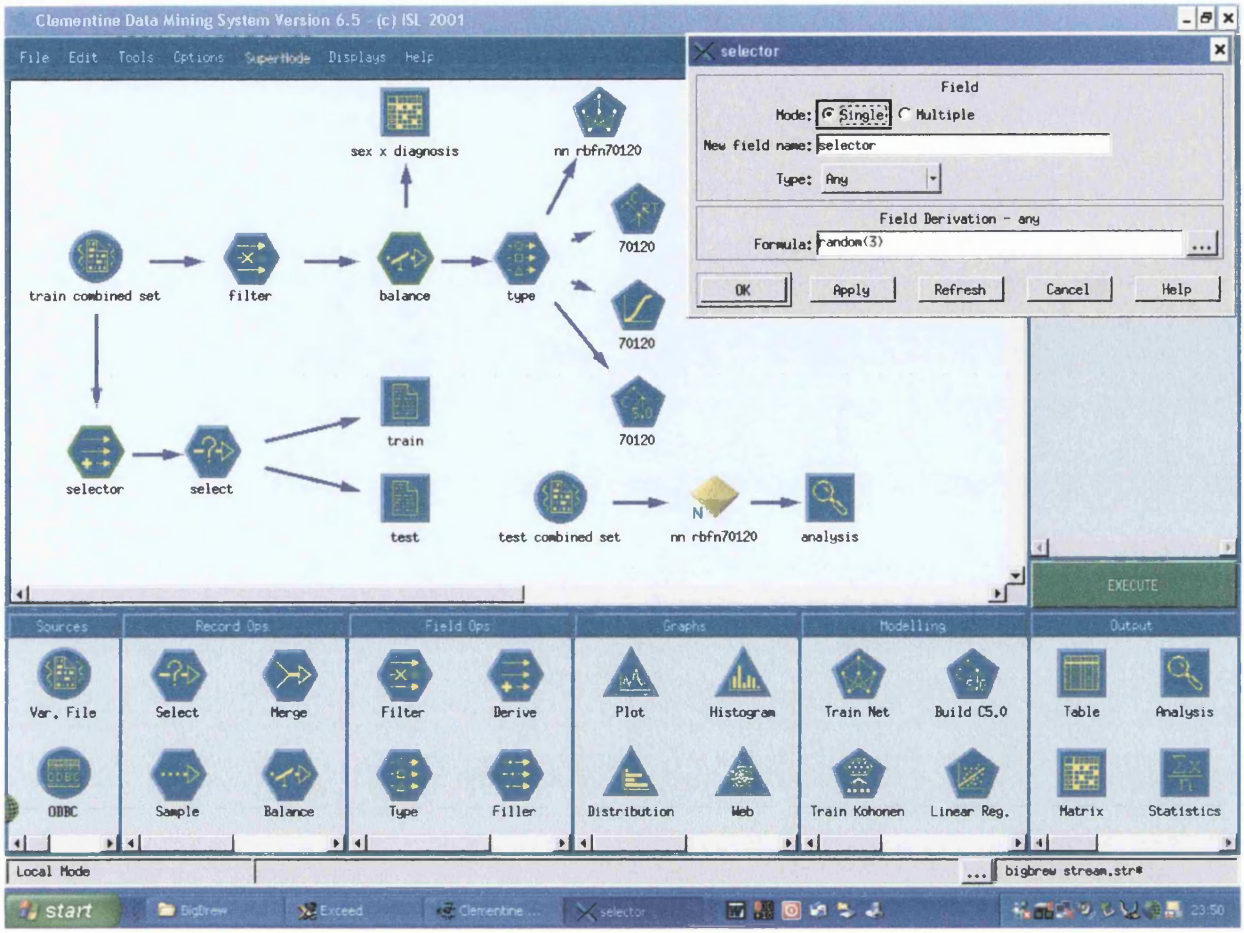


Figure 5.10 The process of generating a training and test set by use of a Derive node (superimposed) which was given the expression "random (3)". This assigns random numbers 1 to 3 to each case, and via a Select node, batches of "1"s or "2s and 3s" can be drawn separately.

5.5.3.4 To test a hypothesis that individuals who are heavy smokers and heavy drinkers are naturally at higher risk of having positive lesions and conditions

It is possible in Clementine® to exclude or include any field of choice. This convenient facility was used to test the widely held assumption that high-risk groups can merely be identified by the fact that they are heavy smokers and/or drinkers. Machine learning models were trained solely on the heavy smoker and heavy drinker fields to assess whether just by using these two fields, positive cases could be identified more efficiently. A filter node was used to remove all fields except "heavy smoker" and "heavy drinker" and "diagnosis" was set as the output. Sets of machine learning models were trained and tested and results tabulated.

Part C

Results

Chapter 6

Results of the Multi-Centre Prevalence Study on Screening in General Dental Practice

Chapter 6 ~ Prevalence of Lesions in General Practice: Results

6.1 General introduction

In all, four experiments were done for this project. The first of these was a field study in a general dental practice environment involving practices situated mainly in the north and south of London, where individuals attending their dentist for routine dental treatment were invited to fill in a health questionnaire and be screened for the presence or absence of oral mucosal lesions. This was not only to observe the prevalence of such positive cases but also to observe the epidemiological profile of this population; their age bands, gender, smoking and drinking habits, frequency of dental attendance and ethnicity.

6.2 Cases collected and distribution of positive cases

2342 sets of forms were returned, 77 (3.3%) patients were excluded due to inconsistencies in their responses or because they were aged less than 35, leaving a total of 2265 patients who were evaluated according to their age, gender, alcohol and tobacco consumption habits, dental attendance, and ethnicity.

There was a variation in the numbers of cases collected by the general dental practitioners. The maximum collected by a practitioner was 423 and the minimum collected was 5 cases. Table 6.1 shows the numbers of cases returned by the participating general dental practitioners.

Table 6.2 gives the chacteristics of the patient population according to gender, age, stratified into 7 bands ranging from 35~39 to 90+ years, and ethnicity. In all, the majority of cases were White (83.5%), Blacks were the next ethnic group constituting 7.6% of the total; South Asians and Chinese constituted 5.3% and 1.4% of this dataset respectively. "Others" refer to population groups such as Middle Eastern or Mediterranean. Information on gender distribution, smoking and alcohol consumption habits, dental attendance and ethnicity are further represented as 100% stacked bar charts in Appendix 2 as Figures 6.1~6.6. The regularity of dental attendance was not used in the following analyses. Table 6.3 show the self reported behavioural chacteristics of

patients and Table 6.4 shows positive lesions categorised under gender, smoking, alcohol consumption and dental attendance. It can be seen that males were more likely to have positive lesions with a prevalence of 5.5% compared with 2.9% for females. There was also an increase in the percentages of positive cases with heavier smoking and increasing alcohol consumption. Heavy smokers constituted 14% of the positives compared to 3.3% of non-smokers. Heavy drinkers accounted for 10% of the positive cases compared to 3.5% for non- or light drinkers. The strengths of the relationships are expressed in Table 6.5 which shows the Incidence Rate Ratios. For both genders heavy smoking greatly increased the risk of having a positive lesion (females IRR 3.58, P=0.010 [CI1.35, 9.50]; males IRR 3.68,P=<0.001 [CI2.10, 6.43]). No significant relationship was between moderate smoking and the risk of having a positive lesion for either gender. Moderate alcohol consumption was associated with nearly twice the risk of having a positive lesion for men (IRR 1.93,P = 0.019 [CI 1.12,3.47]) while heavy drinking was associated with nearly three times the risk (IRR 2.98, P=0.039 [CI 1.06,8.38]). There were no significant associations between alcohol consumption and positive lesions observed in the female group. Overall there were 29% smokers and 16.5% moderate or heavy drinkers. Ethnicity was recorded and is shown as a 100% stacked bar chart in the Appendix as Figure 6.6, however the number of positive lesions occurring in the different groups was too small to submit to statistical analysis.

6.3 Lesions detected

Oral lesions were detected in 319 patients giving a prevalence of 14.1%, with ninety-four (4.2%) lesions considered to be positive. Amongst the positives were 45 cases of white patches (2.0%), 11 cases of red patches (0.5%) and 31 cases of lichen planus (1.4%). Conditions such as Fordyce's spots, furry tongue and "linea alba" were recorded as normal variants and not included. In this study there were two cases of squamous cell carcinoma one of which arose from a case of submucous fibrosis. These are tabulated in Table 6.6.

Table 6.1 The participating general dental practitioners tabulated with the cases returned. The highest number of cases returned by one dentist was 423 (GDP 13) and the lowest was 5 (GDP10)

GDP	Cases returned
1	327
2	222
3	43
4	24
5	198
6	83
7	91
8	173
9	21
10	5
11	301
12	135
13	423
14	39
15	33
16	65
17	69
18	13

Table 6.2 Characteristics of the patient population divided into gender, age bands and ethnic groupings

Factor	Description	Number	%
Gender	Male	1078	47.6
	Female	1187	52.4
Age group	35-39	236	10.4
	40-44	290	12.8
	45-49	323	14.3
	50-54	319	14.1
	55-59	286	12.6
	60-64	229	10.1
	65-69	214	9.4
	70-74	164	7.2
	75-79	109	4.8
	80-84	73	3.2
	85-89	15	0.7
	90+	7	0.3
Ethnic group	White	1892	83.5
	Black	173	7.6
	South Asian	120	5.3
	Chinese	31	1.4
	Other	49	2.1

Table 6.3 Self-reported behavioural characteristics of the patients in regard to smoking, alcohol consumption and dental attendance

Factor	Description	Definition	Number	%
Smoking	Heavy	20 or more cigarettes/day	179	7.9
	Moderate	up to 19 cigarettes/day	474	20.9
	Non	never smoked or not within last 10 years	1612	71.2
Alcohol	Heavy	Units/week: >20 males, >14 females	50	2.2
	Moderate	Units/week: 5-20 males, 5-14 females	323	14.3
	Non or light	up to 4 units/week	1892	83.5
Dental	regular	within the last year	1962	66.6
attendance	irregular	not within the last year	303	13.4

Table 6.4 Distribution of positive lesions grouped under gender, ethnicity, smoking and alcohol consumption

Factor	Description	Number	%
Gender	Male	59	5.5
	Female	35	2.9
Ethnic Group	White	87	4.6
	Black	1	0.6
	South Asian	4	3.3
	Chinese	2	6.5
	Other	0	0.0
Smoking	Heavy	25	14.0
	Moderate	16	3.4
	None	53	3.3
Alcohol	Heavy	5	10.0
Consumption	Moderate	23	7.1
	Non or light	66	3.5

Table 6.5 Incidence Rate Ratios (IRR) of gender, smoking and drinking habits against the prevalence of positive lesions and conditions. Incidence Rate Ratio gave the risk ratios, taking into consideration the ages of the patients.

Factor	**Description**	**IRR (95% CI)**	**P**
Gender	Male	1.86 (1.22, 2.82)	0.004
	Female	1*	
Smoking in females	Heavy	3.58 (1.35, 9.50)	0.010
	Moderate	1.89 (0.87, 4.13)	0.110
	Non	1*	
Smoking in males	Heavy	3.68 (2.10, 6.43)	<0.001
	Moderate	0.57 (0.25, 1.29)	0.179
	Non	1*	
Alcohol consumption in females	Heavy	1.84 (0.25, 13.48)	0.548
	Moderate	2.63 (0.38, 4.08)	0.713
	Non or light	1*	
Alcohol consumption in males	Heavy	2.98 (1.06, 8.38)	0.039
	Moderate	1.93 (1.12, 3.47)	0.019
	Non or light	1*	

* = baseline group for each comparison

Table 6.6 The numbers of lesions and conditions found in this screening study. These were grouped into those determined as positive and those determined as negative.

Lesions determined as positive	**Nos. of patients**
Carcinoma	2
White patch	45
Red patch	11
Lichen planus	31
Ulcer	2
Submucous fibrosis	1
Actinic keratosis	2
	94
Lesions determined as negative	
Non-specific or traumatic ulcers	39
Frictional/traumatic keratosis	30
Fibrous overgrowths	28
Candida/denture stomatitis	14
Smokers keratosis (palate)	12
Aphthous ulcers	10
Amalgam tattoo	9
'Haemangioma'	7
Angular cheilitis	5
Abscess/sinus	5
Mucoceles	4
Geographic tongue	3
Naevus	3
Miscellaneous*	10
	179
Total patients with lesions	**273**

*examples include vesicular lesion, Median rhomboid glossitis, tori, furred tongues

6.4 Discussion

The purpose of this prospective study was to determine the prevalence of relevant lesions and habits in a population visiting their general dentist. There was no specific targeting of high risk groups. The prevalence of mucosal lesions detected was 14.1% which is similar to other studies although comparisons are difficult because there have been few similar studies carried out in non-institutionalised populations. A representative selection of such studies is summarised in Table 6.7. The criterion for lesions varies markedly as do the sites chosen for inclusion in the studies. For example, Reichart (2000) reported that 66.2% of 35-74 year olds had mucosal lesions but this included Fordyce spots (23.7%) and patients with a history of apthous ulceration (18.3%). Martinez-Diaz and Garcia-Pola (2002) reported a prevalence of 58.8% among subjects attending at Dental School for periodontal or prosthodontic treatment; this however included 24.6% cases of oral pigmentation and 10.7% of cases of "linea alba". In another study of a random sample of Sicilian men, mucosal lesions were found in over 81% of this group. 13.8% had leukoplakia which was ascribed to the high use of smoking and alcohol but this study also included 50% of cases with "coated tongue" (Campisi and Marigiotta, 2001). Bouquot (1986) analysed data from a general population based study of 23,616 cases and reported an overall prevalence of 10.3% for mucosal lesions including a 2.9% prevalence of leukoplakia.

In the study reported here, positive lesions were recorded in 4.2% of cases which is more than previously found by Jullien *et al* (1995) in a general medical practice environment (2.2%) and in a dental hospital setting (3%), but less than that reported by Downer *et al* (1995) among subjects screened in a company headquarters (5.5%). The reasons for the disparity observed in the study by Jullien *et al* (1995) and Downer *et al* (1995) are not clear; it could be postulated that in the study by Jullien *et al*, subjects were invited to attend for a screen either directly or via postal invitations and those who smoked heavily or were heavy consumers of alcohol could have refused. The study by Downer *et al* (1995), set in a company headquarters, was opportunistic and thus could be more representative. It is possible that a prevalence of between 4 and 5% is more typical of the population as a whole. The majority of studies in Table 6.7 were observational

prevalence studies with emphasis on various aspects of interest. The screening test used in all the studies was a simple visual examination. In the majority of studies, the screeners were health professionals, mainly dental surgeons but also ENT specialists (Franceschi *et al*, 1990). In some studies such as the Sri Lankan study, primary health care workers were used (Warnakulasuriya *et al,* 1984; Warnakulasuriya and Nanayakkara, 1991). These were trained nurses, health visitors and other hospital staff. The rationale for using health care workers in Asia is that the dentist population ratio in a country like Sri Lanka is about 1:30,000 and screening of such large numbers of individuals in the time would not be otherwise feasible. The ages of the populations screened ranged from 15 yrs and above (Metha *et al* 1972) to 35 and above (Mathew *et al*, 1997). Population groups selected for screening varied from general populations who were invited by post to attend (Julien *et al,* 1995), to house to house case finding (Metha *et al,* 1972), to elective voluntary screening such as that offered to visitors to a state fair (Burzynski, 1997). Targeted populations included those attending clinics for lung disease (Dombi *et al,* 2001), heavy smokers and heavy alcohol consumers where a prevalence of 14.9% was recorded (Talamani *et al*, 1994), elderly institutionalised where a prevalence of 7.4% was recorded (Viglid, 1987) and specific racial groups. Pearson *et al* (2001) in her study targeting a Bangladeshi population in East London showed a prevalence of 25% for leukoplakia and positive association with betel quid or paan chewing. There were variations in the areas examined; some studies included the larynx and oropharynx (Prout *et al,* 1997). Overall, the settings varied but most showed a prevalence of between 1and 6% which is similar to that found in this study.

The proportions of alcohol and tobacco users were also similar to those previously reported. In the previous UK study, 8% were heavy smokers (Julien *et al* 1995), in this study it was 7.9% (Table 6.3). The effects of smoking and alcohol consumption and the likelihood of having a positive lesion were also very similar to earlier UK studies (Jullien *et al* 1995; Downer *et al* 1995). Heavy smokers were more likely to have a positive lesion and men who were heavy drinkers were observed to be about three times more likely to have a positive lesion. The population of smokers (29%) and moderate and heavy drinkers (16.5%) was comparable to the

findings of the ONS omnibus surveys conducted in 2000 and 2001 which show that 26% of the general population are current smokers and 34% of men and 26 % of women in the 45-64 age group drink more than three or four days a week (ONS 2002).

The results of this study have shown that the population attending general dental practices, thought to be self-selecting and therefore not representative of the general population in terms of smoking and alcohol consumption and prevalence of lesions, are representative of the general population and suggests that opportunistic screening in general dental practice could be a viable option.

Table 6.7 A representative selection of screening and prevalence studies spanning the years 1971-2002

Authors	Year	Country	Population/setting	Age	n	cancers	precancers	%+ve
Ross and Gross	1971	USA	gen population	35+	1,2868	1	339	2.6
Mehta *et al.*	1972	India	gen population	15+	101,761	11	1628	1.6
Axell	1976	Sweden	gen population	15+	20,333	1	732	3.6
Warnakulasuriya *et al.*	1984	Sri Lanka	gen population	20+	29,295	4	1230	4.2
Mehta *et al.*	1986	India	gen population	35+	39,331	20	511	1.4
Bouqout *et al.*	1986	USA	gen population	10-90	23,616	22	682	3
Viglid *et al.*	1987	Denmark	institutionalised elderly	65+	285	2	19	7.4
Warnakulasuriya *et al.*	1991	Sri Lanka	gen population	20+	57,124	20	3541	6.2
Ikeda *et al* ,	1991	Japan	gen population	18-63	3,131	0	77	2.5
Banoczy and Rigo	1991	Hungary	lung clinic screening	19-60	7,820	1	104	1.3
Talamini *et al.*	1994	Italy	high risk referrals	35+	627	10	55	14.9
Downer *et al.*	1995	UK	company HQ	40+	292	0	17	5.8
Jullien *et al.*	1995	UK	dental hosp	40+	1,042	1	21	3.0
Jullien *et al.*	1995	UK	medical practice	40+	985	1	31	2.2
Ikeda *et al*	1995	Japan	gen population	60+	802	0	38	4.7
Ikeda *et al.*	1995	Cambodia	gen population	15+	1,319	1	41	3.2
Field *et al*	1995	UK	gen population	<20-> 69	1947	1?	4	0.2
Mathew *et al.*	1997	India	gen population	35-64	2,069	1	212	10.2
Prout *et al*	1997	USA	tobacco users	40+	4,611	1	590	12.8
Szabo *et al.*	1997	Hungary	homeless, alcoholics	Not given	300	8	42	16.7
Zain *et al.*	1997	Malaysia	gen population	25+	11,707	5	165	1.5
Reichart	2000	Germany	aging Germans	35-74	2,022	0	55	2.7
Sankaranarayanan	2000	India	gen population	35+	114,601	63	1310	12
Kovac-Kovacic	2000	Slovenia	perio patients	25-75	1,609	0	106	6.6
Campisi, Margiotta	2001	Sicily	random male	40+	118	1	4	13.9
Dombi *et al.*	2001	Hungary	lung clinic screening	18+	5,034	0	188	3.7
Pearson *et al.*	2001	UK	Bangladeshi	40+	137	0	28	20.4
Current study	2002	UK	dental practices	35+	2,265	2	92	4.2

Part C

Results

Chapter 7

Use of Clementine® for the Detection of Groups at High Risk of Having Positive Lesions and Conditions

Chapter 7~Use of Clementine® Datamining Software

7.1 Introduction

As this experiment was done to verify that a datamining package with machine learning capabilities could be adapted for use in this area of research, several techniques of data manipulation were carried out, many of which achieved similar results in different ways. These included the exploration of all the available methods for random selection of data, the linking up of derived machine learning models in series to achieve better results and testing the models against a test set where the numbers of negatives and positives were reduced (this was done as a comparison with the results of the earlier demonstration study by Speight *et al* (1995), where the test set consisted of 10 positive cases and 355 negative cases).

7.2 Results

A series of neural network and C5 machine learning models was produced by training them on varying proportions of negative to positive cases. A summary of the salient points of this experiment is shown as a flowchart in Fig. 5.7. This shows the flow of data, which were divided into randomly selected sub-groups for the purpose of training and testing. The data were drawn via different functions available in Clementine® such as the Balance node, Sample node, and Kohonen clustering. Batches of data for testing were first taken off from the main dataset, with the remainder being the training set. Multiple sets of data were then produced from this training set and drawn in varying proportions of negatives to the set number of positives. Neural networks and C5 models were trained from these batches of data containing varying proportions of negative cases. The models produced were given alphanumeric identifiers. The results were expressed as sensitivity and specificity and entered in tables and graphs. Altogether, five tables showing the performances of the machine learning models over different data ratios, two tables of summaries and three graphs were produced; Table 7.1 is an illustrative example of one such table, and Fig

7.1 is an example of a graph. As the other tables and graphs essentially conveyed the same information on trends and similar performances, they are not included. It was observed that ratios of between 1:1 and 2:1 negatives to positives tended to give the best results, and also that there was an inverse relationship between sensitivity and specificity. Models that were linked in series and tested appeared to give better results although this facility tended to be over-contrived and could be difficult to adapt for future clinical use.

From the tables, there was a range of sensitivities from the nineties to zero that was achievable with the corresponding specificities ranging from zero to the nineties as the proportions of negative cases increased.

The single best performance was a neural network (no11) C5 (cm1) combination testing on a cut-down test set consisting of 16 positives and 309 negatives, and this is shown in Table 7.2. This gave a sensitivity of 0.88 with a specificity of 0.64.

Table 7.1 An illustrative example of a series of such tables, showing the performance of the C5.0 and Neural Network (NN) models trained in the varying ratios from 0.05:1 through to 19:1 negative to positive cases, tested against an unseen test set of 638 negatives and 30 positives. Data was sampled in several ways. In this instance, the numbers of negative cases were drawn from clusters created using a Kohonen network, hence the title "K Clusters". Sensitivity and Specificity values are recorded and X-sensitivity and X-specificity refers to the performances when the C5 and neural network sets were combined.

Ratios (Neg:Pos)	0.5:1	1:1	2:1	3:1	4:1	5:1	6:1	7:1	8:1	9:1	10:1	11:1	12: 1	13: 1	14: 1	15:1	16:1	17:1	18:1	19:1
Sample size to 30 pos	33	67	207	134	268	335	402	469	536	603	670	737	804	871	938	1005	1072	1139	1206	1273
C5.0 models	kc1	kc2	kc17	kc3	kc4	kc5	kc6	kc7	kc8	kc9	kc10	kc11	kc12	kc13	kc14	kc15	kc16	kc17	kc18	kc19
C5-n score/638	90	109	183	128	214	266	266	272	335	338	432	427	431	488	528	536	621	630	630	637
C5-p score/30	26	25	21	25	19	18	18	19	17	15	10	12	12	6	5	5	1	1	1	0
C5 sensitivity	0.87	0.83	0.70	0.83	0.63	0.6	0.6	0.63	0.57	0.5	0.33	0.4	0.4	0.77	0.17	0.17	0.03	0.03	0.03	0
C5 specificity	0.14	0.17	0.27	0.2	0.34	0.42	0.42	0.43	0.52	0.53	0.68	0.67	0.68	0.2	0.83	0.84	0.97	0.99	0.99	1
NN models	kn1	kn2	kn17	kn3	kn4	kn5	kn6	kn7	kn8	kn9	kn10	kn11	kn12	kn13	kn14	kn15	kn16	kn17	kn18	kn19
NN-n score/638	54	259	173	361	184	258	228	350	351	435	431	432	428	504	524	538	538	609	626	638
NN-p score/30	28	19	22	15	21	19	21	15	16	13	12	12	12	6	5	5	5	3	1	0
NN sensitivity	0.93	0.63	0.73	0.5	0.7	0.63	0.7	0.5	0.53	0.43	0.4	0.4	0.4	0.2	0.11	0.17	0.17	0.1	0.03	0
NN specificity	0.09	0.41	0.27	0.57	0.29	0.4	0.36	0.54	0.55	0.68	0.68	0.68	0.67	0.79	0.82	0.84	0.84	0.95	0.98	1
X-n score/638	46	102	102	128	177	236	233	260	295	357	424	427	427	469	511	536	536	608	620	637
X-p score/30	28	25	25	25	22	20	21	20	19	16	12	12	12	7	6	5	5	3	2	0
X-sensitivity	0.93	0.83	0.83	0.83	0.73	0.67	0.7	0.67	0.63	0.53	0.4	0.4	0.4	0.23	0.2	0.17	0.17	0.1	0.06	0
X-specificity	0.07	0.16	0.16	0.2	0.28	0.37	0.37	0.41	0.46	0.56	0.67	0.67	0.67	0.74	0.81	0.84	0.84	0.95	0.97	1

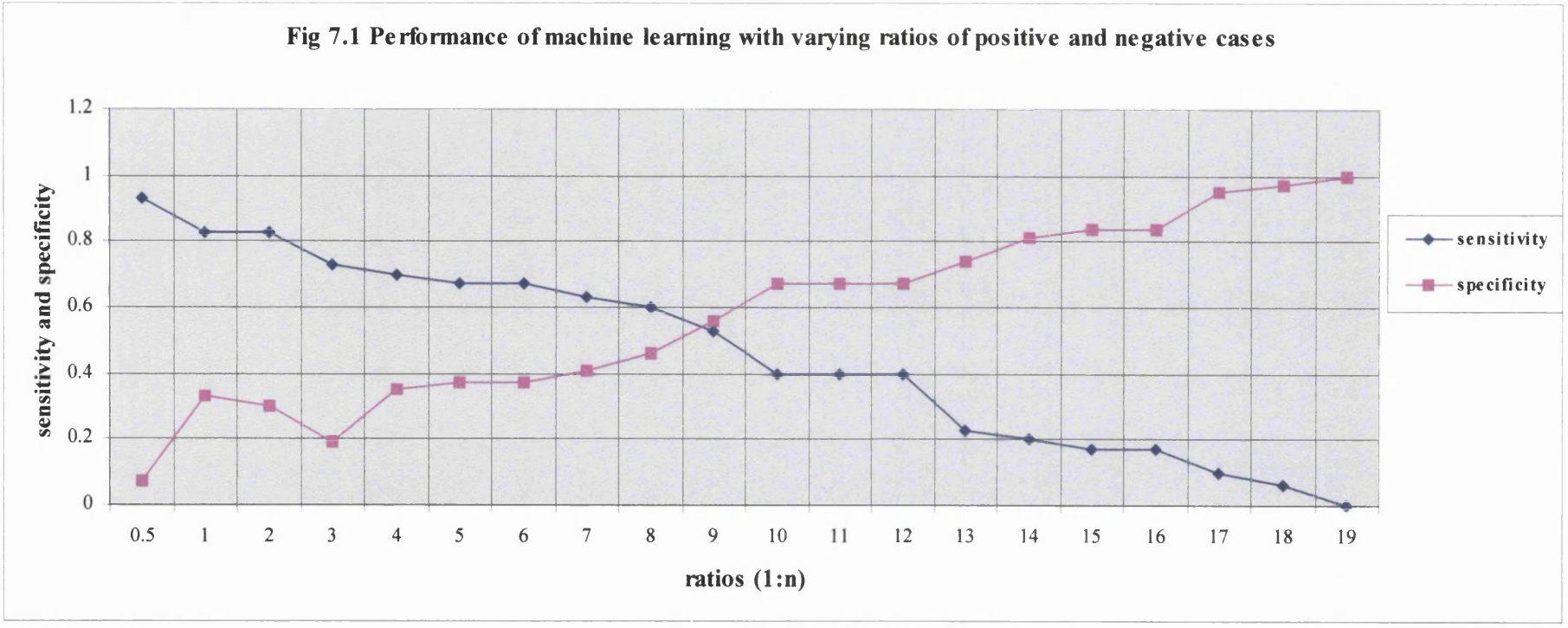


Figure 7.1 The performance of an illustrative combined C5.0 and neural network model, expressed as sensitivity and specificity in the Y-axis plotted against increasing ratios of negative cases to positive cases from 0.5 negatives to 19 negatives. It is acknowledged that points on the graph that are not directly related to each other should not be joined together. However, this was done in this graph for the sake of clarity as linking the points enabled the trends to be seen more clearly.

Table 7.2 The results of combining various neural network models with C5.0 models against the "cut down" unseen test set of 16 positive and 309 negative cases.

Models	no2+	co2	co4	co7	co11	cm1
n-score/309		99	127	133	122	131
p-score/16		13	12	12	12	13
Sensitivity		0.81	0.75	0.75	0.75	0.81
Specificity		0.32	0.41	0.43	0.4	0.42
Models	no3+	co2	co4	co7	co11	cm1
n-score/309		110	174	186	178	188
p-score/16		12	13	13	12	12
Sensitivity		0.75	0.81	0.81	0.75	0.75
Specificity		0.36	0.56	0.6	0.58	0.61
Models	no7+	co2	co4	co7	co11	cm1
n-score/309		91	137	142	128	143
p-score/16		12	11	11	11	12
Sensitivity		0.75	0.69	0.69	0.69	0.75
Specificity		0.29	0.44	0.46	0.41	0.46
Models	no11+	co2	co4	co7	co11	cm1
n-score/309		113	190	197	185	197
p-score/16		13	13	13	13	14
Sensitivity		0.81	0.81	0.81	0.81	0.88
Specificity		0.37	0.62	0.64	0.6	0.64
Models	nm1+	co2	co4	co7	co11	cm1
n-score/309		101	152	158	149	158
p-score/16		13	14	12	14	13
Sensitivity		0.81	0.88	0.75	0.88	0.81
Specificity		0.33	0.49	0.51	0.48	0.51

7.3 Discussion

This experiment evaluated the performance of machine learning models in Clementine®; initial results suggest that such a datamining package could be adapted for use in screening studies. Initially earlier versions of Clementine® only had the neural network and C5 algorithms which were suitable for this application, however later versions of Clementine® include more candidates such as a version of the Ruleset variant of C5 called the Classification and Regression Tree and a Logistic regression node. The clear advantage of this over bespoke machine learning software was that there would be a suite of machine learning algorithms available. In addition, technical support was readily available and glitches encountered in the software were quickly remedied. A further benefit was that these machine learning models tended to be set in a way that a wide range of applications could be catered for; as such, there was less danger of overfitting data than there would be in a purpose written software package. The best result achieved was a sensitivity of 0.88 with a specificity of 0.64 using a selected combination of neural network and C5 models. This however was achieved by training on the *complete set* of fields provided (and testing on a previously separated unseen test set from the same database). Some of these fields such as "drinks fortified spirits or wine" would not be included in the future work which would limit focus on age, gender, smoking, drinking habits and dental attendance. The machine learning models behaved predictably when presented with varying proportions of negative to positive cases, giving high sensitivities when more positives were presented and high specificities when more negatives were presented. The optimum training ratio was in the region of 1:1 and 2:1 negative to positive cases. All methods of data sampling were tried and these did not show any method to be better than another. This project suggested that the use of machine learning models in Clementine® for the forthcoming screening project was a viable proposition.

Part C

Results

Chapter 8

A Comparison of the Descriptive Epidemiology of Data from Two Oral Cancer and Precancer Screening Studies

Chapter 8~ Comparative Description of Two Datasets

8.1 Introduction

Two main datasets are examined in this study; the first was from a collection of screening studies for oral cancer and precancer that took place in1995 at a dental hospital outpatients clinic, an inner city medical practice and an industrial setting (Jullien *et al,* 1995, Downer *et al,* 1995). Risk habits of smoking, alcohol consumption and also dental attendance were recorded, as was the prevalence of positive lesions and conditions. The results of the screening done by qualified dental practitioners had the "gold standard" of the diagnoses confirmed by an oral medicine specialist of consultant status. The second dataset is described in Chapter 6

8.2 Analysis of the numbers and distribution of cases in the two datasets

As the 1995 dataset consisted of cases aged 40 years and above and the 2000 dataset had cases ranging from age 35 years, 236 cases below the age of 40 years were removed to enable equivalent comparisons. The dataset was reduced from 2265 cases to 2029 cases. Data from the 1995 set and the 2000 set are tabulated under age, gender, smoking and drinking habits and dental attendance in Tables 8.1 and 8.2 first, showing the numbers of cases within the categories then expressed as percentages. This same information is further presented as histograms (in Appendix 2) which show the distribution of positive and negative cases over ten-year age bands in numbers and percentages (Figures 8.1, and 8.2), the prevalence of negative cases over positive cases according to gender, (Figures 8.3 and 8.4), smoking habits (Figures 8.5 and 8.6), alcohol consumption (Figures 8.7 and 8.8) and dental attendance (Figures 8.9 and 8.10).

The distribution of cases in the different age bands showed remarkable similarity, for example, in the 1995 dataset, there were 412 negative and 22 positive cases; the similar age group in the 2000 dataset consisted of 422 negative and 21 positive cases. The largest sub-population of cases for

both datasets was in the 40-49-age band, 34.5% in the 1995 set and 30.2% in the 2000 dataset. The oldest age band (90+), constituted 0.3% of cases in both datasets. The largest proportion of positive cases was found in the 50-59 year age band in both datasets and was 1.5% of the total positives. The positive prevalence for the 1995 dataset was 4.7% and that for the 2000 dataset was 4.4%. There were greater numbers of females in both datasets with similar percentages, 55.3% (1120 cases) for the 1995 dataset and 53% (1066 cases) for the 2000 dataset. There were 289 heavy smokers (14.2%) in the 1995 dataset and 152 (7.5%) in the 2000 dataset. There were 103 heavy drinkers (5.1%) in the 1995 dataset and 40 heavy drinkers (2.0%) in the 2000 dataset. In both datasets there was an increase in the proportion of positive cases with increasing smoking and alcohol consumption, for example 3%, 3.5% and 16.5% positives for non-, moderate and heavy smokers respectively in the 2000 dataset and 4%, 6.2% and 8.7% positives for non- or light, moderate and heavy drinkers in the 1995 dataset; the only exception to this trend was in the 1995 dataset for moderate smokers where the moderate smoker category was lower than the non-smokers (4.8%, 4% and 7.6% for non-, moderate and heavy smokers respectively).

8.3 IRR Values, Poisson regression

The two datasets were further compared using the STATA® statistical software package (Version 7.0, STATA Corporation, Texas, USA) to produce Incidence Rate Ratios. These were calculated using Poisson regression and tabulated with P values and Confidence Intervals. Several sets of tables are presented showing results of Poisson regression. The term "adjusting for all categories" refers to the process of taking into account the influence that each factor has on the others and the final values obtained in the computations. The first table, Table 8.3, shows results that take into account the influence that all fields or factors have on each other. Table 8.4 and 8.5 show this same information separated by gender. Table 8.6 shows the influence of just one factor, age, on the prevalence of positive lesions and conditions. Table 8.7 shows the Incidence Rate Ratios for

the 1995 and 2000 datasets for age and smoking *omitting* gender and drinking factors against prevalence of positive lesions and conditions with P values and 95% Confidence Intervals. Table 8.8 shows the Incidence Rate Ratios of the 1995 and 2000 datasets for age and drinking *omitting* gender and smoking factors against prevalence of positive lesions and conditions with P values and 95% Confidence Intervals. Other combinations such as, adjusting for age, gender and smoking habit without the influence of drinking, then adjusting for age, gender and drinking habit without the influence of smoking were tabulated for the sake of completion but not included here as they tended to show similar results and it was not usual to have smokers who were not alcohol consumers or alcohol consumers, especially heavy consumers, who were not smokers. Incidence Rate Ratios were used to establish P values (which is the chance of getting the result if the null hypothesis is true). Significant relationships were seen in heavy smokers in the 2000 dataset (Table 8.3): after adjusting for the influence of all fields, a heavy smoker was 4.15 times more likely to have a positive lesion than one who did not smoke (P= >0.001 [CI 2.46-7.00]). In this category a heavy drinker in the 1995 dataset would be 2.1 times more likely to have a positive lesion than one who was a non- or light drinker (P=0.05 [CI 1.01-4.40]). A similar significance was seen in females in the 2000 dataset where a heavy smoker was 4.47 times more likely to have a positive lesion than one who did not smoke (P= 0.005 [CI1.58-12.68]). Age was not a significant factor for both sets with P values of 0.175 and 0.414 for the 1995 and 2000 set respectively. The following are further interesting examples where the P values were less than 0.05. Male heavy smokers in the 1995 set were 2.48 times more at risk than non smokers (P = 0.022 [CI 1.42-5.39]) and in the 2000, this figure was 3.79 (P =<0.001 [CI 0.24-1.40). For females, heavy drinkers had an IRR of 4.48 (P = 0.008 [CI 1.478-13.589]) in the 1995 set.

Table 8.1 A numerical analysis of the two datasets grouped under age bands, gender, smoking and drinking habits and dental attendance.

		1995 Dataset		2000 Dataset	
		Negative	**Positive**	**Negative**	**Positive**
Age	40-49	674	25	592	21
	50-59	512	30	574	31
	60-69	412	22	422	21
	70-79	254	15	261	12
	80-89	69	5	83	5
	90+s	6	0	7	0
Gender	Male	848	56	906	57
	Female	1079	41	1033	33
Smoking	Heavy	267	22	127	25
	Moderate	867	37	384	14
	Non or light	793	38	1428	51
Drinking	Heavy	94	9	35	5
	Moderate	496	33	265	22
	Non or light	1337	55	1639	63
Dental attendance	*yes*	1363	77	1697	78
	no	564	20	242	12

Table 8.2 Comparison of the two datasets grouped under age bands, gender, smoking and drinking habits and dental attendance expressed in percentages.

		1995 Dataset %		2000 Dataset %	
		Negative	**Positive**	**Negative**	**Positive**
Age	40-49	33.3	1.2	29.2	1
	50-59	25.3	1.5	28.3	1.5
	60-69	20.4	1	20.8	1
	70-79	12.6	0.7	12.9	0.6
	80-89	3.4	0.3	4.1	0.3
	90+s	0.3	0	0.3	0
Gender	Male	41.9	2.8	44.7	2.8
	Female	53.3	2	50.9	1.6
Smoking	Heavy	93.1	7.6	83.5	16.5
	Moderate	96	4.1	96.5	3.5
	Non or light	95.3	4.8	97	3
Drinking	Heavy	91.3	8.7	87.5	12.5
	Moderate	93.8	6.2	91	9
	Non or light	96	4	96.3	3.7
Dental	*yes*	94.7	5.4	83.6	3.8
attendance	*no*	96.6	3.4	11.9	0.6

Table 8.3 The Incidence Rate Ratios for the 1995 and 2000 datasets with P values and 95% Confidence Intervals for each dataset adjusting for *all* fields.

* Refers to the baseline for each comparison.

1995 dataset					**2000 dataset**				
Factor	**Description**	**IRR**	**P value**	**95%CI**	**Factor**	**Description**	**IRR**	**P value**	**95%CI**
Age		1.04	0.104	1.0-1.03	**Age**		1.02	0.079	0.99-1.034
Gender	Male	1.47	0.077	0.96-2.25	**Gender**	Male	1.53	0.062	0.98-2.391
	Female	1*				Female	1*		
Smoking	Heavy	1.26	0.423	0.72-2.19	**Smoking**	Heavy	4.15	0.001	2.46-7.00
	Moderate	0.78	0.294	0.49-1.24		Moderate	1.00	0.984	0.55-1.84
	Non or light	1*				Non or light	1*		
Alcohol	Heavy	2.1	0.047	1.01-4.40	**Alcohol**	Heavy	1.92	0.171	0.75-4.94
Consumption	Moderate	1.47	0.098	0.93-2.33	**Consumption**	Moderate	1.47	0.140	0.88-2.47
	Non or light	1*				Non or light	1*		

Table 8.4 The Incidence Rate Ratios for the 1995 and 2000 datasets for *males*, of age, smoking and drinking habits against prevalence of positive lesions and conditions adjusted for all fields with P values and 95% Confidence Intervals. * Refers to the baseline for each comparison.

1995 dataset					**2000 dataset**				
Factor	**Description**	**IRR**	**P value**	**95%CI**	**Factor**	**Description**	**IRR**	**P value**	**95%CI**
Age		1	0.850	0.979-1.025	**Age**		1.01	0.099	0.99-1.04
Smoking	Heavy	2.48	0.022	1.142-5.39	**Smoking**	Heavy	3.79	0.000	2.06-6.97
	Moderate	1.64	0.187	0.788-3.393		Moderate	0.58	0.225	0.24-1.40
	Non or light	1*				Non or light	1*		
Alcohol	Heavy	1.4	0.500	0.526-3.729	**Alcohol**	Heavy	2.38	0.110	0.82-6.92
Consumption	Moderate	1.48	0.173	0.843-2.585	**Consumption**	Moderate	1.79	0.047	1.00-3.19
	Non or light	1*				Non or light	1*		

Table 8.5 The Incidence Rate Ratios for the 1995 and 2000 datasets for *females*, of age, smoking and drinking habits against prevalence of positive lesions and conditions adjusted for all fields with P values and 95% Confidence Intervals. * Refers to the baseline for each comparison.

1995 dataset					**2000 dataset**				
Factor	**Description**	**IRR**	**P value**	**95%CI**	**Factor**	**Description**	**IRR**	**P value**	**95%CI**
Age		1.02	0.062	1.0-1.04	**Age**		1.01	0.476	0.98-1.04
Smoking	Heavy	0.51	0.304	0.146-1.820	**Smoking**	Heavy	4.47	0.005	1.58-12.68
	Moderate	0.38	0.011	0.183-0.805		Moderate	2.01	0.100	0.87-4.63
	Non or light	1*				Non or light	1*		
Alcohol	Heavy	4.48	0.008	1.478-13.589	**Alcohol**	Heavy	1.33	0.790	0.17-10.57
Consumption	Moderate	1.39	0.440	0.597-3.272	**Consumption**	Moderate	0.71	0.636	0.17-2.99
	Non or light	1*				Non or light	1*		

Table 8.6 The Incidence Rate Ratios for the 1995 and 2000 datasets for *age*, without gender, smoking and drinking factors against prevalence of positive lesions and conditions with P values and 95% Confidence Intervals. It can be seen from the IRRs and the P values that the influence of age differences was not significant in both datasets.

1995 dataset	IRR	P value	95%CI		2000 dataset	IRR	P value	95%CI
Age	1.01	0.175	0.99-1.03		**Age**	1.01	0.414	0.98-1.02

Table 8.7 The Incidence Rate Ratios for the 1995 and 2000 datasets for *age* and *smoking* omitting gender and drinking factors against prevalence of positive lesions and conditions with P values and 95% Confidence Intervals.

1995 dataset		IRR	P value	95%CI	2000 dataset		IRR	P value	95%CI
Age		1.01	0.177	0.99-1.03	**Age**		1.01	0.091	0.99-1.03
Smoking	Heavy	1.67	0.055	0.99-2.83	**Smoking**	Heavy	5.25	0.000	3.21-8.59
	Moderate	0.90	0.685	0.57-1.42		Moderate	1.11	0.736	0.61-2.02
	Non or light	1*				Non or light	1*		

Table 8.8 shows the Incidence Rate Ratios for the 1995 and 2000 datasets for *age* and *drinking* omitting gender and smoking factors against prevalence of positive lesions and conditions with P values and 95% Confidence Intervals.

1995 dataset		IRR	P	95%CI	2000 dataset		IRR	P value	95%CI
Age		1.02	0.060	1.00-1.03	**Age**		1.01	0.265	0.99-1.03
Drinking	Heavy	2.40	0.016	1.78-4.90	**Drinking**	Heavy	3.51	0.007	1.41-8.74
	Moderate	1.63	0.031	1.04-2.54		Moderate	2.13	0.002	1.31-3.46
	Non or light	1*				Non or light	1*		

8.4 Discussion

Two datasets from separate and unrelated studies five years apart showed remarkable similarities both for the prevalence of positive cases and the distribution of numbers in the age, gender, smoking and alcohol consumption categories. The distribution of positive cases within those categories was also similar. Dental attendance, used as a surrogate measure of dental awareness in the first dataset, was left out because of the way the data were gathered in the second screening study, that is, from a general dental practice environment which naturally biases the population to those attending more regularly. In the event, the trend observed was the opposite from what was expected, in that regular attendees were more likely to have positive lesions in the second set. The results of this category were also not statistically significant. Poisson regression was done on the data and Incidence Rate Ratios were produced. These were the correlations that had strong statistical significance. Although this was not the situation in every case, the trend of increasing smoking or drinking habits and a concomitant increase in numbers of positives was consistent.

8.5 Conclusion

The similarity of these two datasets supports the view that populations are similar in epidemiological profile and that it could be possible to use these datasets to train and evaluate machine leaning models for the detection of sub-groups at high risk of oral cancer and precancer.

Part C

Results

Chapter 9

Evaluation of Machine Learning Techniques for the Detection of Groups at High Risk of Oral Cancer and Precancer

Chapter 9~ Results of Machine Learning in High Risk Identification

9.1 Introduction

This chapter reports the results of the performance of machine learning models that were trained with the view to comparing their performances with those obtained from the early pilot study previously referred to (Speight *et al,* 1995), and from the study by dental surgeons and other health professionals (Moles *et al,* 2000). Machine learning models within a commercially available datamining package were used. In all, four main types of machine learning models were evaluated, one of them, the neural networks, consisting of five variants. The models were trained then tested on an unseen test set. This was done by two main methods: in the first, the 1995 dataset was the training set from which batches of training data were drawn, testing on the 2000 dataset. In the second, both datasets were pooled and training and test sets separated for training. Training was done on two caches of data in 1:1 then 2:1 proportions.

9.2 Performance of the neural networks

9.2.1 Brief review of salient points

Clementine® contains five variants of neural networks: *quick,* where the program selects the best topology and working parameters, usually time prioritised; *dynamic*, where the program starts with a topology and modifies to achieve less error; *multiple,* much like the dynamic model except that the program starts with different topologies, trains them simultaneously and picks and presents the one with the least error; *prune,* where the program starts with a large network and reduces the hidden connections until the "best" configuration is found based on error feedback; and *RBFN*, radial basis function network which uses a clustering technique for training.

9.2.2 Tabulation of results

Table 9.1 Representative samples of reports of the trained neural network models. The left column identifies the training caches and the type of neural network and the right column shows the topology used and the relative importance of the training fields.

Neural network architecture"64 quick"	**Topology** Input Layer : 8 neurons Hidden Layer #1 : 4 neurons Output Layer : 1 neuron Predicted Accuracy : 65.84% **Relative Importance of Inputs** age : 0.29965 heavy drinker : 0.23531 non drinker : 0.21885 sex : 0.19851 heavy smoker : 0.16524 non smoker : 0.14355 mod smoker : 0.12042 mod drinker : 0.03436
Neural network architecture "93 Multiple"	**Topology** Input Layer : 8 neurons Hidden Layer #1 : 19 neurons Hidden Layer #2 : 17 neurons Output Layer : 1 neuron Predicted Accuracy : 72.63% **Relative Importance of Inputs** age : 0.51405 sex : 0.29915 non smoker : 0.20638 non drinker : 0.20563 mod smoker : 0.16537 heavy drinker : 0.14503 heavy smoker : 0.12319 mod drinker : 0.08744

Table 9.2 The results of the neural network variants, quick (quick), dynamic (dyn), multiple (multp), prune (prun), and radial basis function network (rbfn) for the 93 negatives to 97 positives and 64 negatives to 97 positives in descending order with results expressed in sensitivity and specificity. These were trained on two separate caches of data using the 1995 set as a training set and the 2000 set as the test set. TN = true negative, FN =false negative, TP = true positive, FP = false negative, PPV = positive predictive value, NPV = negative predictive value.

Model	Cache n: 97pos	TN	FN	TP	FP	Sensitivity	Specificity	PPV	NPV
nn rbfn	93	1147	44	46	792	51	59	0.05	0.96
nn quick	93	686	24	66	1253	73	35	0.05	0.97
nn prun	93	1334	49	41	605	46	69	0.06	0.96
nn multp	93	1297	49	41	642	46	67	0.06	0.96
nn dyn	93	1338	53	37	601	41	69	0.06	0.96
nn rbfn	64	885	37	53	1054	59	46	0.05	0.96
nn quick	64	371	14	76	1568	84	19	0.05	0.96
nn prun	64	593	22	68	1346	76	31	0.05	0.96
nn multp	64	584	22	68	1355	76	30	0.05	0.96
nn dyn	64	831	22	68	1108	76	43	0.06	0.97

Table 9.3 The results of the neural network variants, quick (quick), dynamic (dyn), multiple (multp), prune (prun), and radial basis function network (rbfn) for the 115 negatives to 97 positives and 70 negatives to 97 positives in descending order with results expressed in sensitivity and specificity. These were trained on the pooled data of the 1995 and 2000 sets A proportion of data was previously isolated and used as the test set. TN = true negative, FN =false negative, TP = true positive, FP = false negative, PPV = positive predictive value, NPV = negative predictive value.

Model	Cache n: 120pos	TN	FN	TP	FP	Sensitivity	Specificity	PPV	NPV
nn rbfn	115	739	19	48	559	72	57	0.08	0.97
nn quick	115	580	16	51	718	76	45	0.07	0.97
nn prun	115	553	29	38	745	57	43	0.05	0.95
nn multp	115	620	26	41	678	61	48	0.06	0.96
nn dyn	115	334	7	60	964	90	26	0.06	0.98
nn rbfn	70	253	16	51	1045	76	20	0.05	0.94
nn quick	70	184	13	54	1114	81	14	0.05	0.93
nn prun	70	155	21	46	1143	69	12	0.04	0.88
nn mulp	70	411	24	43	887	64	32	0.05	0.94
nn dyn	70	344	15	52	954	78	27	0.05	0.96

9.3 C5 and Classification and Regression Trees (CARTs)

9.3.1 Brief review of salient points

Sets of C5 and Classification and Regression Tree machine learning models are considered together as they are essentially similar paradigms; these were trained first on data from the 1995 set then on the pooled 1995 and 2000 sets using two separate caches per series, in the 1:1 and 2:1 positive to negative ratios.

9.3.2 Tabulation of results

Table 9.4 Typical decision sequences for C5 algorithms trained on the 64:97, 93:97 and 115:120 caches showing the internal workings of the C5 algorithm for the different data caches. These are less cryptic that those found for the neural networks.

C5 64:97 Cache	heavy drinker T [Mode: P] -> P heavy drinker F [Mode: P] mod smoker T [Mode: P] sex M [Mode: P] -> P sex F [Mode: N] -> N mod smoker F [Mode: P] age =< 53 [Mode: N] non smoker T [Mode: N] -> N non smoker F [Mode: P] -> P age > 53 [Mode: P] -> P
C5 93:97 Cache	sex M [Mode: P] -> P sex F [Mode: N] age =< 68 [Mode: N] heavy drinker T [Mode: P] -> P heavy drinker F [Mode: N] -> N age > 68 [Mode: P] -> P
C5 115:120 Cache	heavy smoker T [Mode: P] -> P heavy smoker F [Mode: N] age =< 49 [Mode: N] -> N age > 49 [Mode: P] -> P

Table 9.5 Performances of the C5 and CART trained models on the 1995 dataset showing the training caches with the 93 negative to 97 positive caches and the 64 to 97 caches set in descending order.

Model	Cache n: 97pos	TN	FN	TP	FP	Sensitivity	Specificity	PPV	NPV
C5	93	815	26	64	1124	71	42	0.05	0.97
CART	93	1057	42	48	882	53	55	0.05	0.96
C5	64	720	23	67	1219	74	37	0.05	0.97
CART	64	867	26	64	1072	71	45	0.06	0.97

Table 9.6 Performances of the C5 and CART trained models on data from the pooled 1995 and 2000 datasets showing the training caches, with the 115 negatives to 97 positives and 70 negatives to 97 positives set in descending order.

Model	Cache n: 120 pos	TN	FN	TP	FP	Sensitivity	Specificity	PPV	NPV
C5	115	356	10	57	942	85	27	0.06	0.97
CART	115	308	10	57	990	85	24	0.05	0.97
C5	70	249	8	59	1049	88	19	0.05	0.97
CART	70	372	26	41	926	61	29	0.04	0.93

CART model, 64:97 Cache	CART model, 115:120 Cache	CART model, 93:97 Cache
non drinker [T] [Mode: P] age < 52.5 [Mode: N] age < 45.5 [Mode: N] sex [M] [Mode: P] non smoker [T] -> N non smoker [F] [Mode: P] age < 41.5 [Mode: P] age < 40.5 -> P age >= 40.5 -> P age >= 41.5 -> P sex [F] -> N age >= 45.5 [Mode: N] sex [M] -> N sex [F] [Mode: N] age < 50 [Mode: N] mod smoker [T] -> N mod smoker [F] -> N age >= 50 -> P age >= 52.5 [Mode: P] mod smoker [T] [Mode: N] age < 84 [Mode: N] age < 75 [Mode: N] age < 67.5 [Mode: P] age < 66.5 -> N age >= 66.5 -> P age >= 67.5 [Mode: N] age < 72.5 -> N age >= 72.5 -> P age >= 75 -> P age >= 84 -> N mod smoker [F] [Mode: P] age < 63.5 [Mode: P] age < 57.5 -> P age >= 57.5 -> P age >= 63.5 [Mode: P] age < 70.5 [Mode: N] sex [M] -> P sex [F] [Mode: N] age < 65.5 -> P age >= 65.5 -> N age >= 70.5 [Mode: P] age < 73.5 -> P age >= 73.5 [Mode: P] age < 78 -> N age >= 78 -> P non drinker [F] [Mode: P] sex [M] [Mode: P] age < 62 [Mode: P] heavy smoker [T] [Mode: P] age < 57.5 -> P age >= 57.5 -> P heavy smoker [F] -> P age >= 62 -> P sex [F] [Mode: P] mod drinker [T] [Mode: N] age < 65.5 [Mode: N] age < 44.5 -> P age >= 44.5 [Mode: N] age < 54.5 -> N age >= 54.5 -> P age >= 65.5 -> P mod drinker [F] -> P	age < 49.5 [Mode: N] non smoker [T] [Mode: N] age < 40.5 -> P age >= 40.5 [Mode: N] age < 48.5 [Mode: N] non drinker [T] [Mode: N] sex [M] -> N sex [F] -> N non drinker [F] -> N age >= 48.5 -> N non smoker [F] [Mode: P] age < 48.5 [Mode: P] age < 40.5 -> P age >= 40.5 [Mode: P] non drinker [T] [Mode: N] age < 42.5 -> N age >= 42.5 [Mode: P] age < 43.5 -> P age >= 43.5 -> N non drinker [F] [Mode: P] age < 46.5 [Mode: P] age < 45.5 -> P age >= 45.5 -> P age >= 46.5 -> N age >= 48.5 -> N age >= 49.5 [Mode: P] non drinker [T] -> P non drinker [F] [Mode: P] mod smoker [T] -> P mod smoker [F] [Mode: P] age < 58.5 -> P age >= 58.5 [Mode: P] age < 61.5 -> P age >= 61.5 -> P	sex [M] [Mode: P] age < 44.5 -> P age >= 44.5 [Mode: P] age < 52.5 [Mode: N] mod drinker [T] [Mode: N] heavy smoker [T] -> P heavy smoker [F] [Mode: N] age < 49.5 [Mode: N] age < 46 -> N age >= 46 -> N age >= 49.5 -> P mod drinker [F] -> N age >= 52.5 [Mode: P] age < 57.5 [Mode: P] non smoker [T] [Mode: P] age < 55 -> N age >= 55 -> P non smoker [F] [Mode: P] mod drinker [T] -> P mod drinker [F] -> P age >= 57.5 [Mode: P] mod drinker [T] [Mode: N] non smoker [T] -> P non smoker [F] [Mode: N] age < 65.5 -> N age >= 65.5 -> N mod drinker [F] [Mode: P] mod smoker [T] [Mode: P] age < 67.5 -> P age >= 67.5 -> P mod smoker [F] [Mode: P] age < 66.5 -> N age >= 66.5 -> P sex [F] [Mode: N] age < 68 [Mode: N] non drinker [T] [Mode: N] non smoker [T] [Mode: N] age < 65.5 [Mode: N] age < 59.5 [Mode: N] age < 55.5 -> N age >= 55.5 -> N age >= 59.5 [Mode: P] age < 63.5 -> P age >= 63.5 -> P age >= 65.5 -> N non smoker [F] [Mode: N] age < 59.5 [Mode: N] age < 57 [Mode: N] age < 43.5 -> P age >= 43.5 -> N age >= 57 -> P age >= 59.5 -> N non drinker [F] [Mode: P] age < 54.5 [Mode: N] age < 48 -> P age >= 48 -> N age >= 54.5 [Mode: P] mod smoker [T] -> P mod smoker [F] -> P age >= 68 [Mode: P] non smoker [T] [Mode: P] age < 71.5 -> P age >= 71.5 [Mode: P] age < 81 -> P age >= 81 -> P non smoker [F] -> N

Table 9.7 The analyses of the CART algorithm for the different data caches.

9.4 The Logistic regression node

9.4.1 Brief review of salient points

The Logistic regression models were derived by training on the two methods as described in the previous sections. The performances of these machine learning models are tabulated below.

Table 9.8 The fields used in the workings of the Logistic regression model. This same format which was set at default model limiting the tree depth to 7, was used for all the data caches in the training process.

Fields used by the Logistic regression model for every data cache	Predicted by the model diagnosis Numeric variables used age Symbolic variables used sex non smoker mod smoker non drinker mod drinker

9.4.2 Tabulation of results

Results from both the training methods are tabulated as Tables 9.9 and 9.10 which show the true negatives (TN), false negatives (FN), true positives (TP) and false positives (FP) along with sensitivities, specificities, positive predictive values and negative predictive values.

Table 9.9 The performances of the Logistic regression models derived by training on the 1995 dataset and testing on the 2000 dataset. Two caches were used for the training.

Model	Cache n: 97pos	TN	FN	TP	FP	Sensitivity	Specificity	PPV	NPV
LR	64	598	18	72	1341	80	31	0.05	0.97
LR	93	873	26	64	1066	71	45	0.06	0.97

Table 9.10 The performances of the Logistic regression models derived by training on data from the pooled 1995, 2000 dataset and testing on a separate and unseen portion of this set. Two caches were used for the training

Model	Cache n:120 pos	TN	FN	TP	FP	Sensitivity	Specificity	PPV	NPV
LR	70	147	3	64	1151	96	11	0.05	0.98
LR	115	732	19	48	566	72	56	0.08	0.97

9.5 Results

The 1995 and 2000 datasets showed much similarity, as observed in Chapter 8; this gave rise to two possible methods of training the machine learning models. First, using the entire 1995 dataset as the training set for the machine learning models and testing the trained models on the unseen 2000 dataset and second, pooling the two datasets together, dividing the data into a training and testing set and training the machine learning models on the larger training set and testing them on the unseen test set.

The performances of all the machine learning models are summarised again in two tables, Table 9.12 showing the results of models trained on data from the 1995 dataset and tested on the 2000 dataset, and Table 9.13 trained on data drawn from the pooled 1995 and 2000 datasets and tested on a portion of data previously isolated from this set. These are arranged in descending order of sensitivities. There was no machine learning model type that performed consistently better than another. Models trained on the pooled data tended to perform better than those trained on just the 1995 data. The sensitivities obtained ranged from 96% to 57% using the former method and 84% to 41% in the latter. This was possibly due firstly to the fact that pooling two datasets together tended to "iron out" minor differences within the two sets giving a more homogenous dataset for the machine learning models to train and test on, secondly there was more data available, 2024 from the 1995 set, 2029 from the 2000 dataset making 4053 cases in the pooled dataset. A larger training set provided the machine learning models with more instances of positive profiles, as the pooled training set consisted of 187 positive cases as opposed to 97 in the 1995 dataset. There was an inevitable inverse relationship between sensitivity and specificity. The positive predictive values were low, as was expected in diseases of low prevalence such as oral cancer and precancer in the United Kingdom and the negative predictive value was conversely high. The best sensitivity was achieved by a Logistic regression model trained on the 70:120 negative to positive cache from the combined dataset. This model returned a

sensitivity of 96% missing only 3 positives out of a test set consisting of 67 positives. The specificity unfortunately was low at 11% and it returned a large number of false positives (1151 cases). Most machine learning models were able to return high sensitivities; a neural network trained on the 1995 dataset achieved a sensitivity of 84% with a specificity of 19%, in the same batch a Logistic regression model achieved 80% sensitivity with a higher specificity (31%). A Classification and Regression Tree model trained on the pooled data gave a sensitivity of 85% with 24% specificity. A good performer would need to also have reasonable specificity. One such was a C5 model trained on the pooled dataset using the 115:120 negative: positive cache which gave a sensitivity of 85% and a specificity of 27%. A Logistic regression model trained on the 1995 dataset using the 64:97 negative: positive cache gave a sensitivity of 80% and a specificity of 31%. Still higher specificities could be achieved but always at the expense of sensitivity. For example, a neural network model trained on the 1995 dataset achieved a specificity of 69% but with a sensitivity of 41%.

9.6 Discussion

The best performing model was not necessarily the one returning the highest sensitivity as that was invariably accompanied by a low specificity which tended to disqualify its use. Table 9.11~ 13 list the models that performed with high sensitivities and reasonable specificities according to the type of algorithm (Table 9.11) and the different training methods used (Tables 9.12,13).

Even considering its role as a pre-screen filter, a very sensitive model would have the tendency to identify an unacceptably large number of cases as being at high risk (false positives). As a means of comparison, Table 9.14 shows the performances of a number of screening studies where outcome measures were evaluated (after Moles *et al,* 2000). These are tabulated with the results from studies using machine learning software. It can be seen that levels of sensitivity achievable by machine learning models were similar or better than

results returned by some studies, although in this current study the specificities are low. It may be argued that there may be a role for such models as a pre-screening tool for identifying those at high risk of oral cancer and precancer. False positives could be considered as those being at higher risk of having the disease. Unfortunately the observation from Tables 9.11 ~ 9.13 is that whilst achieving high sensitivities and identifying about 60% to 70% of a population as being at high risk (the false positives), the machine learning models were still under-diagnosing. The best performer was a Logistic regression model trained on pooled data that correctly identified 64 out of 67 positive cases in the test set of 1365 cases. This was an exception and the average number of true positives missed in the first training method was 31 out of 97 (32%) and the second was 16 out of 67 (24%). Table 9.14 compares the sensitivity and specificity of an illustrative sample of screening studies with the performances achieved by representative machine learning models from all four main types used, which were the Neural networks, C5, Classification and Regression Trees and Logistic regression. These models, whilst achieving good sensitivities, would have missed a number of positives and this is shown in Table 9.11 where it can be seen that the models were under-diagnosing by 10 to 20 %.

Table 9.11 A representative selection of each of the main machine learning model types used, sensitivities and specificities achieved and the numbers and percentages of true positives missed.

Machine Learning Model	Sensitivity	Specificity	Positives correctly identified (%)	Positives missed (%)
C5 algorithm	85	27	57 (85)	10 (15)
Logistic Regression	80	31	72 (80)	18 (20)
Neural Network	90	26	60 (90)	7 (10)
Classification and Regression Tree	85	24	57 (85)	10 (15)

Table 9.12 The performances of the various machine learning models trained on the 64:97 and 93:97 negative to positive caches from the 1995 dataset.

Model	Cache n:97pos	TN	FN	TP	FP	Sensitivity	Specificity	PPV	NPV
nn quick	64	371	14	76	1568	84	19	0.05	0.96
LR	64	598	18	72	1341	80	31	0.05	0.97
nn prun	64	593	22	68	1346	76	31	0.05	0.96
nn multp	64	584	22	68	1355	76	30	0.05	0.96
nn dyn	64	831	22	68	1108	76	43	0.06	0.97
C5	64	720	23	67	1219	74	37	0.05	0.97
nn quick	93	686	24	66	1253	73	35	0.05	0.97
LR	93	873	26	64	1066	71	45	0.06	0.97
CART	64	867	26	64	1072	71	45	0.06	0.97
C5	93	815	26	64	1124	71	42	0.05	0.97
nn rbfn	64	885	37	53	1054	59	46	0.05	0.96
CART	93	1057	42	48	882	53	55	0.05	0.96
nn rbfn	93	1147	44	46	792	51	59	0.05	0.96
nn prun	93	1334	49	41	605	46	69	0.06	0.96
nn multp	93	1297	49	41	642	46	67	0.06	0.96
nn dyn	93	1338	53	37	601	41	69	0.06	0.96

Table 9.13 The performances of the various machine learning models trained on the 70:120 and 115:120 negative to positive caches from the pooled 1995/2000 dataset

Model	Cache n:120pos	TN	FN	TP	FP	Sensitivity	Specificity	PPV	NPV
LR	70	147	3	64	1151	96	11	0.05	0.98
nn dym	115	334	7	60	964	90	26	0.06	0.98
C5	70	249	8	59	1049	88	19	0.05	0.97
CART	115	308	10	57	990	85	24	0.05	0.97
C5	115	356	10	57	942	85	27	0.06	0.97
nn quick	70	184	13	54	1114	81	14	0.05	0.93
nn dym	70	344	15	52	954	78	27	0.05	0.96
nn rbfn	70	253	16	51	1045	76	20	0.05	0.94
nn quick	115	580	16	51	718	76	45	0.07	0.97
nn rbfn	115	739	19	48	559	72	57	0.08	0.97
LR	115	732	19	48	566	72	56	0.08	0.97
nn prun	70	155	21	46	1143	69	12	0.04	0.88
nn mulp	70	411	24	43	887	64	32	0.05	0.94
nn mulp	115	620	26	41	678	61	48	0.06	0.96
CART	70	372	26	41	926	61	29	0.04	0.93
nn prun	115	553	29	38	745	57	43	0.05	0.95

Table 9.14 The range of sensitivities and specificities achieved in screening studies along with results obtained from machine learning algorithms. It can be observed that machine learning models can achieve comparable sensitivities. (*) refers to a neural network program that was written for the purpose, and (**) refers to machine learning models that were incorporated in commercially available datamining packages.

Authors	Year	Screeners	Sensitivity	Specificity
Metha *et al*	1986	Primary Health Care Workers	59	98
Warnakulasuriya and Pindborg	1990	Primary Health Care Workers	95	81
Mathew *et al*	1997	Primary Health Care Workers	94	98
Ikeda *et al*	1995	General Dental Practitioners	60	94
Downer *et al*	1995	Dental Surgeons	71	99
Julien *et al*	1995	Dental Surgeons	64	99
Julien *et al*	1995	Dental Surgeons	81	99
Speight *et al*	1995	Neural Network*	80	77
current study	2003	C5 algorithm**	85	27
current study	2003	Logistic Regression**	80	31
current study	2003	Neural Network**	90	26
current study	2003	Classification and Regression Tree**	85	24

9.7 The heavy smokers and heavy drinkers

There is a widely held assumption that if an individual is a heavy smoker and a heavy drinker, such a person would inevitably be at high risk. As a demonstration study, this axiom is investigated.

9.7.1 Results

Table 9.15 shows the numbers of heavy smokers and drinkers combined and the numbers of positives within that category for the 1995 and 2000 datasets, with a column for the cases that would be missed if only this axiom was used. Using the machine learning software, models were trained only on the heavy smoker and heavy drinker fields using the 93 and 64 negative to 97 positive caches and results are tabulated in table 9.16. It can be seen that there are three broad outcomes. Either the models would default to assigning the entire dataset as being positive (100% sensitivity with 0% specificity) or veer heavily on the side of the positives and leave a smaller group that would be very good at identifying the negatives but not the positives. These were the ones trained on the cache where there were more negative cases.

Table 9.15 Positive cases in the combined heavy smoker and heavy drinker categories, showing that 86 to 92% of positives would be missed if just the criterion that combined heavy smokers and heavy drinkers to signal a positive case was applied.

Dataset	Combined heavy smoker, heavy drinkers	Positives	% cases that would be missed if this axiom was applied
1995	35	5	86
2000	13	1	92

Table 9.16 The performances of machine learning models trained only on the heavy smoker and heavy drinker fields, grouped in descending order of the training data caches. Only the 64:97 and the 93:97 caches were used in this part of the experiments.

Model	Cache n:97	TN	FN	TP	FP	Sensitivity	Specificity	PPV	NPV
nn quick	93	1789	61	29	150	0.32	0.92	0.16	0.97
nn dyn	93	0	0	90	1939	1.00	0	0.04	0.00
nn multp	93	1789	61	29	150	0.32	0.92	0.16	0.97
nn prun	93	1789	61	29	150	0.32	0.92	0.16	0.97
nn rbfn	93	1789	61	29	150	0.32	0.92	0.16	0.97
C5	93	0	0	90	1939	1.00	0	0.04	0.00
CART	93	1812	65	25	127	0.28	0.94	0.16	0.97
LR	93	1789	61	29	150	0.32	0.92	0.16	0.97
nn quick	64	0	0	90	1939	1.00	0	0.04	0.00
nn dyn	64	0	0	90	1939	1.00	0	0.04	0.00
nn multp	64	0	0	90	1939	1.00	0	0.04	0.00
nn prun	64	0	0	90	1939	1.00	0	0.04	0.00
nn rbfn	64	0	0	90	1939	1.00	0	0.04	0.00
C5	64	0	0	90	1939	1.00	0	0.04	0.00
CART	64	0	0	90	1939	1.00	0	0.04	0.00
LR	64	0	0	90	1939	1.00	0	0.04	0.00

9.7.2 Discussion

The machine learning models obtained by using this method of training were considered unusable because they would either ascribe every case as a positive or perform with a high specificity and a low sensitivity and this would negate the purpose.

Part D

General Discussion

Chapter 10

Discussion, Conclusions and Future Work

Chapter 10~General Discussion, Conclusions, and Future Work

10.1 General discussion

Oral cancer is an important health problem worldwide and appears to be a growing problem especially in developing countries. Studies of birth cohorts in the United Kingdom indicate a possible increase in the numbers of cases in the coming decades. There is also an increase of cases in young males in the 35-40 age group (Hindle *et al,* 1996). Whilst surgical and rehabilitation techniques have improved over the years, the 5 year survival rate is still less than 50% and increasing numbers of patients are presenting at T_3 and T_4 stages (Stell and McCormack, 1985). Yet oral cancer is a disease that is very amenable to treatment if discovered early in its natural history. Disease prevention appears to be a useful strategy. Patient education could be an effective means of prevention as oral cancer appears to be associated with heavy smoking, alcohol consumption, habits such as areca nut chewing and dietary deficiencies, so health counselling could raise patient awareness and compliance to screening initiatives. Currently, however, despite well publicised campaigns such as "Oral Health Week" and media publicity, there appears to be widespread ignorance or indifference to oral cancer. Warnakulasuriya *et al* (1999), in a random public survey of 1,894 members aged over 16 years, found that whilst 96% had heard of skin cancer and 97% of lung cancer and 86% of cervical cancer, only 56% were aware that cancer could occur in the mouth. Oral cancer fulfils most of the criteria for a screenable disease as determined by Wilson and Junger (1968) (Chapter 2, Tables 2a, b and c). Much can be gained by screening for oral cancer and precancer as prognosis decreases markedly if lesions are over 2.0cms when detected (Speight and Morgan, 1993). Caught when small, management involves simple excision. Management of advanced lesions, however, is often expensive and involves the team work of many disciplines such as surgeons, radiologists, physiotherapists and speech therapists, and the aftermath, even with the most conscientious

treatment usually leaves considerable disfigurement and loss of function. If discovered and treated early, survivors will have a better quality of life in terms of function, eating and social life. They will be able to re-enter the work force and not be dependent on special care.

The screen for oral cancer is a simple visual examination. This is atraumatic and well accepted by almost all patients. It has been shown that a simple visual examination of the oral mucosa for oral cancer and precancer (Pindborg, 1984; Mock, 1985; Julien *et al,* 1995) gave sensitivity and specificity values comparable to prostate cancer (Mettlin *et al* 1991; Babaian *et al,* 1992), cervical cancer (Soost *et al* 1991), breast cancer (Haiart and Henderson, 1991; Hakama *et al,* 1991) and colorectal screening (Favennac *et al,* 1992). A recent meta-analysis of seven investigations in oral cancer and precancer indicated that overall the studies showed a high discriminatory ability, suggesting that screening is feasible (Moles *et al*, 2002). As a cancer of relatively low incidence in the United Kingdom, (CancerStats 2003) it is generally accepted that population screening cannot be recommended in terms of cost effectiveness (Rodrigues *et al,* 1998). This tends to be the problem in developed countries therefore different alternative strategies have been tried such as targeting high-risk groups, where screening is carried out in hospital centres treating lung cancer (Dombi *et al,* 2000), or in an ENT centre examining heavy smokers and drinkers (Talamani *et al,* 1994). Examinations have also been carried out within some "captive or convenience populations" such as in industrial settings, where there are likely to be populations of middle-aged males who tend to be stressed and could have smoking and drinking habits that render them at higher risk of having positive lesions or conditions (Downer *et al,* 1995). The possibility of establishing "outreach posts" for screening was suggested but a difficulty encountered here could be finding the location of such venues, as well as the compliance of the targeted individuals (Warnakulasuriya, 2001).

Opportunistic screening in general dental practice may provide a close approximation to an unbiased sample of the general population in terms of age range, gender distribution, and risk habits as related to the likelihood of having positive lesions and conditions. With the exception of dental emergencies coming in "off the street", this population can be said to be self-selecting consisting of individuals who are more conscientious about dental health and therefore less likely to have lesions and conditions seen to be positive. The findings of this study, however (Chapter 8), suggest that risk habits of smoking and alcohol consumption and the overall prevalence of positive lesions are similar to those seen in the general population. Opportunistic screening or case finding within this population group could therefore be beneficial. 50% of the UK population visit their dentist regularly and these patients can be screened for oral cancer when they attend for routine treatment. It is implied under the terms of service of the National Health Service that a soft tissue inspection is part of the examination process for which a fee is allocated. In a recent postal survey of dentists, Warnakulasuriya and Johnson (1999) noted that of the 2519 respondents, 84% reported that they routinely perform oral mucosal screening. Whilst this was encouraging, it was noted in this paper that this number represented only 16% of the 15,836 dentists who were sent the questionnaire. The actual numbers of dentists who regularly and routinely screen for oral cancer and precancer is not verified and could be a subject for further research.

The use of artificial intelligence or machine learning software in medicine, especially as an aid to diagnosis or decision-making, is increasing in popularity. The use of a neural network to identify sub-populations at high risk of having oral cancer and precancer just on information such as age, gender, and smoking and drinking habits was initially studied by Speight *et al* (1995) and returned encouraging results. As it has the potential of serving as a useful tool for a clinician to be made aware of an individual's risk status of oral cancer and precancer, it was developed further in this project. The pattern or profile for a positive case

appears more subtle than just one who smokes heavily and drinks heavily. Using this axiom in our two separate datasets would have resulted in 86 to 92 % of positive cases being missed; an experiment in Chapter 9 (Section 9.7) derived some machine learning models that achieved a low sensitivity and a high specificity, suggesting that this axiom could work better at identifying negatives, in that if an individual was not a heavy smoker and heavy drinker, he or she would be likely to be a negative case, but if one was a heavy smoker and a heavy drinker, one could not say for certain that they were likely to be a positive case.

Machine learning models trained in identifying positive cases could also be developed as part of an interactive patient health information and motivation package. As a "spin off" from this research, an early prototype of such a model is described in Section 10.2. Our machine learning models did not compare very well with applications in other fields of medicine, possibly because there were no strongly distinguishing features for this disease as there is in, for example, the case of electrocardiogram patterns in ischaemic cardiac disease where ECG patterns are strongly pathognomonic. Machine learning models need to be given strong diagnostic cues to return good results. This was not the case for oral cancer and precancer as there was not a sufficiently strong pattern of smoking and alcohol consumption habits to definitely signal a positive case. Good sensitivities, comparable with "human" screeners, were achieved in this study. However the accompanying low specificities are a cause for concern and can be considered as a disadvantage. Moreover, there was still a tendency for the models to underdiagnose, missing 24 to 32% of the positive cases whilst identifying an inordinately large population of negatives as being at high risk, that is, the false positives, as reflected in the low specificities. For these reasons, it remains debatable if this device, at this stage of development, is appropriate for use as a diagnostic aid. Its role as a pre-screen filter could be possible, but the problem of under-diagnosing, in other words, missing positives, still has to be addressed. The ideal situation would be that a machine

learning model could correctly identify all the positives as well as a number of negatives with a similar profile to a positive. The machine learning model would thus ascribe a risk status to the patient in order that the examining practitioner could be alerted and be able to more closely monitor this patient. At this stage of development therefore, because of the tendency of the learning models to underdiagnose the positives while still picking up large numbers of negatives, this method cannot be recommended for field use. At this point, it is worth considering the high sensitivities and specificities achieved by the neural network in the earlier study (Speight *et al* 1995). This was a bespoke program written for the purpose and could have been vulnerable to overfitting and overtraining, a well-known phenomenon that results in a program performing very well within a set but less well if tested on prospective datasets. The original program unfortunately was not available for use in these current studies. The author of the program also was not available, further highlighting the disadvantages of bespoke machine learning programs.

10.2 The scoring model

A result of current studies in machine learning algorithms was the development of a prototype Scoring model which can either be used as part of a patient education package, or as a means of data gathering. The interactive screen is shown in Figure 10.1. There is an underlying Clementine® stream shown in Figure 10.2. Information given by the subject about their age, gender, and smoking and drinking habits is entered into a Microsoft Excel CVS file and also passed "downstream" onto a machine learning model which scores the entry with a numerical value and a categorical value, for example high, medium or low risk, by a "publisher" node. This Scoring Model is highly adaptable and different fields such as Betel nut users can be included in future screening studies; however this would necessitate the training of new machine learning models. The underlying machine learning model can

also be further refined with new data and even completely replaced as new data improves its accuracy. The cut-off points for low, medium or high risk can be set by the investigator who decides on the cut-off points of the range of results expressed numerically from 0 to 1.

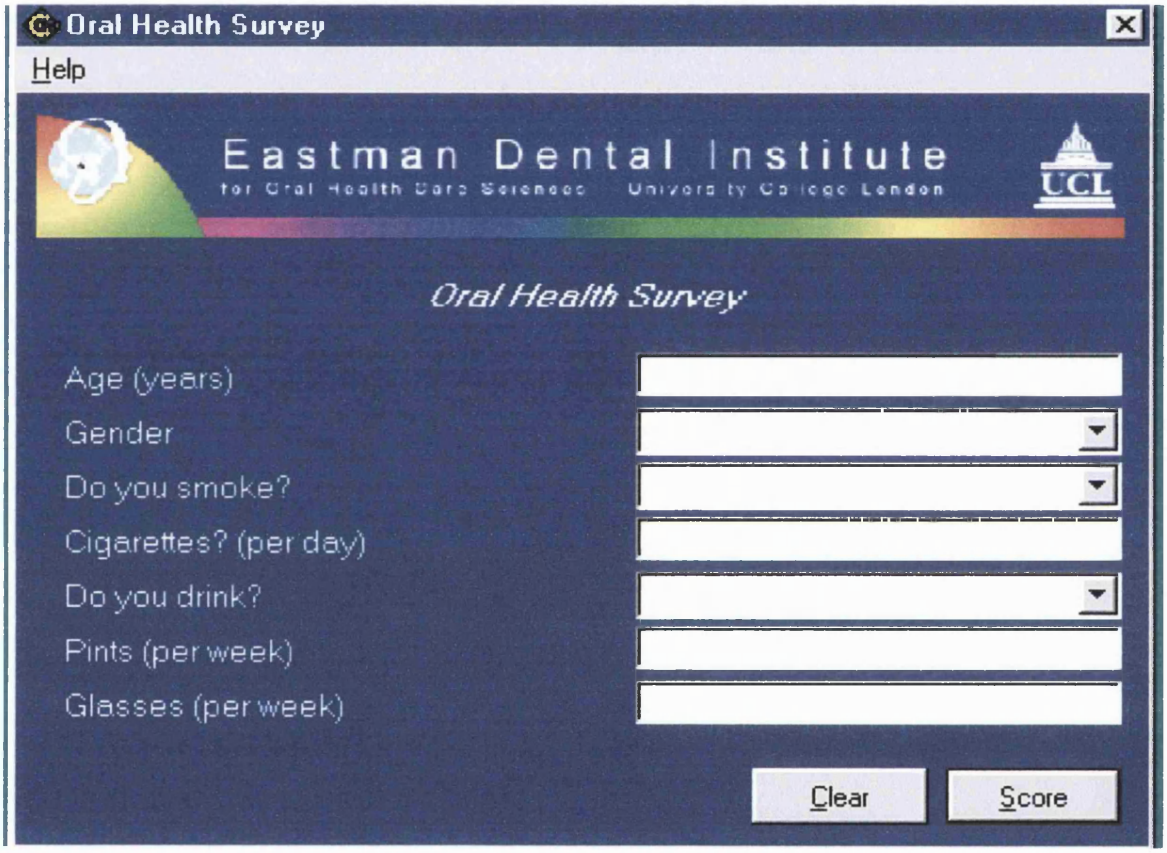


Figure 10.1 An interactive pane from the Scoring model. This could be presented to the patient as an Oral Health awareness package.

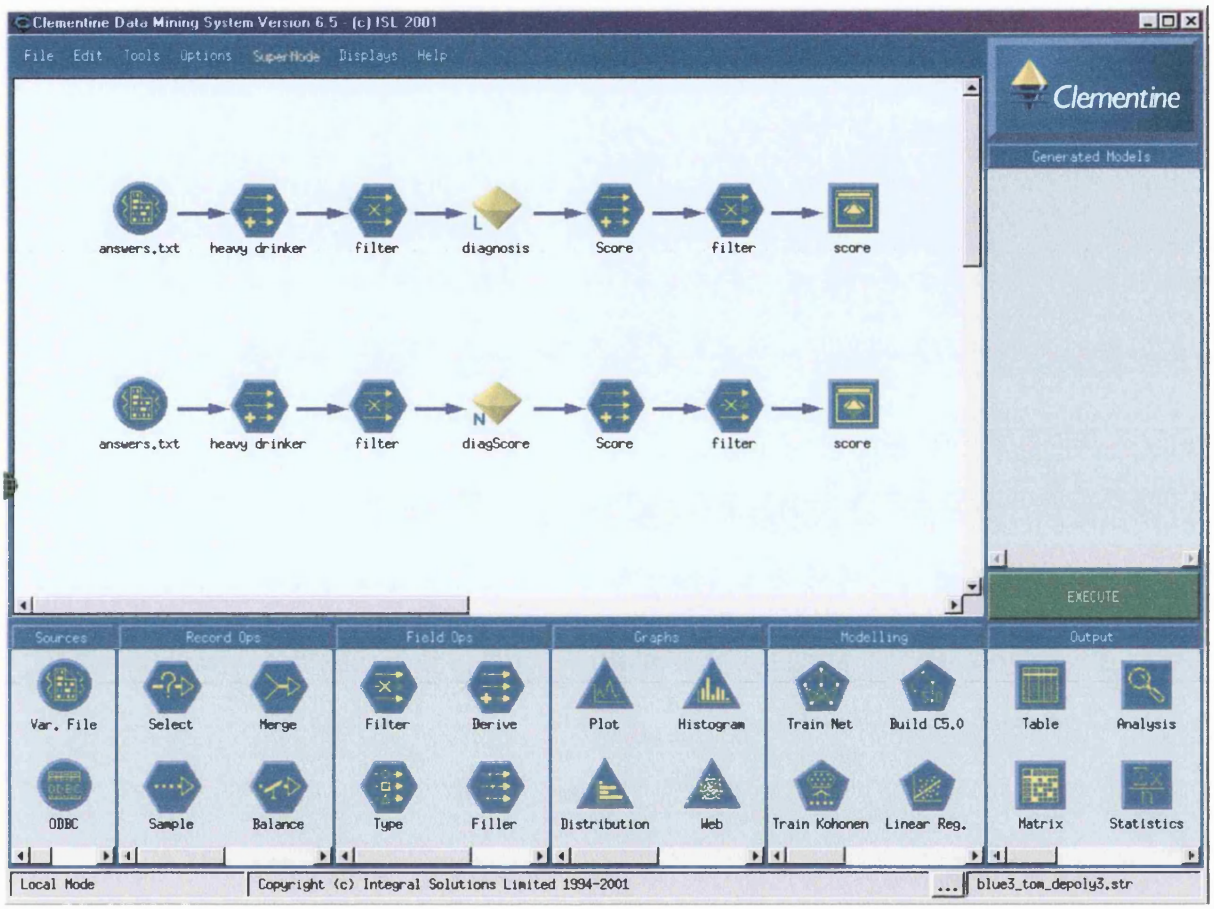


Figure 10.2 A screen image taken from a Clementine® stream to demonstrate two data streams, one using a neural network, the other using the Logistic regression model, where data entered in on the left of the screen via a Source node is streamed toward a trained machine learning model seen as a yellow "nugget". Just prior to this, a Derive node is used to translate glasses and pints of alcohol to units.

10.3 Observations and experiences gained from this study

10.3.1 The patients

- This screening study was well accepted by the patients. When the nature and purpose of the study were explained to them, some were actually glad and grateful to have it. This is in accordance with the findings of a recent study by Humphris (2001), which concluded that patients were less anxious (Mann Whitney U test: $z=-2.07$, $P<0.05$) and more willing to have a screen (Mann Whitney U test: $z=-3.67$, $P<0.001$) if sufficient knowledge was available to them for example in the form of patient information leaflets
- The form was designed by a professional graphic designer with an emphasis on simplicity but a number of individuals still tended to have problems giving the required responses. For example, some would indicate that they were non-smokers but also ticked the box for smoking more than 20 a day. This could possibly be interpreted as inferring that they were presently non-smokers, but when they did smoke, they smoked more than 20 cigarettes a day. Forms with such conflicts were not included in the study. It is difficult to further simplify these forms. Failures in this area probably reflect on the patients and dentists rather than the forms.
- Some patients may have tended to poorly recall or downplay their smoking and drinking habits. This appears to be a well recognised phenomenon and has been described in many studies (Muir *et al* 1998; MacDonald 1999; Watt *et al* 2000; Warnakulasuriya 2002)
- Patients also tended, though not invariably, to have difficulty recounting their alcohol consumption in units. They tended to relate better to "pints" and "glasses".
- Alcohol consumption was most difficult to reliably quantify. Measures, for example a glass of wine, when poured at home tend to be larger in quantity than in pubs. It may be worth considering a further question as to whether most of the alcohol consumption is

done at home or socially in pubs and similar venues. There was a group of occasional but heavy drinkers, the "bingers", who usually consume little or no alcohol on a daily basis but have large quantities at the weekend or on social outings. This group is difficult to classify; the significance and effects of "bingers" and "bingeing" could be worth further study.

- In both alcohol and tobacco consumption data, the amounts used were self-reported and difficult to verify without the use of salivary cotinine assay for tobacco consumption or markers such as carbohydrate-deficient transferrin (CDT) and gamma-glutamyl transferase (GGT) for alcohol consumption (Helander *et al,* 1996). This study is based on self-reported values.

10.3.2 The screeners

- This study was done on a voluntary basis. General dental practitioners were invited to take part in this study without financial or other incentives other than as fulfilment of Continuing Professional Development units or payment for Section 63 courses during the lectures. As such, it was difficult to impose time, quota and other demands on them.
- Examining dentists varied considerably in their responses, with some exceeding their screening quota and others barely returning any forms. Each dentist was requested to complete about 200 forms. The highest number of forms returned by any dentist was 423 and the lowest was 5. This is an area that needs further study, as the co-operation of dentists and receptionists in the field is pivotal to the success of this or any similar study. To this end, financial incentives could be helpful. This task could be better accepted if it was remunerated as an item of service within the National Health Service.

- To ensure the smooth progress of a screening study, it is necessary to have at least one person directly responsible for the day-to-day running of the study. This person's duties would include the collection of data, checking to see if any logistical problems are encountered in the field, and generally informing the screeners of the state of progress and any interesting conditions and lesions found by other participants. Such feedback is usually very well received by participating examiners and could serve to further motivate their efforts.
- For this study, the initiative for all cases identified as positive and needing onward referral was with the dentist. It was not in the scope of this study to follow up the cases which were referred, but future studies could include the follow-up of referrals.
- There is need for a simple, standardised, robust screening protocol which could include:

i) Publicising the advantages of oral cancer screening by media announcements or leaflets in the waiting rooms of dental and medical practices which target the general public

ii) Consensus on the clinical criteria for lesion identification.

iii) Agreement as to which mucosal lesions are to be noted.

iv) Calibration and, in the case of studies taking more than a year, recalibration of the screeners.

v) Involvement of the key members of the practice staff such as receptionists and head nurses so that they know what is going on and can meaningfully co-operate with the study

vi) A set and regular data collection procedure within the participating practices

vii) A financial incentive for the staff. This could be advantageous because whilst the

actual screening examination does not usually take more than 3-5 minutes, explanations to the patient about the project for their cooperation and informed consent, and processing the paperwork does take considerable time, usually up to 15 minutes; this is generally seen as disruptive for most busy dental practices

viii) Regular feedback from the screening dentists. This is helpful as they are usually able to provide useful practical insights regarding problems encountered. This is best done in the form of evening dinners accompanied by a talk or update on the progress of the study.

10.3.3 The machine learning software

- Machine learning software can be used to identify subsets of populations at higher risk of oral cancer or dysplastic lesions with good levels of sensitivity
- This can be in the form of "bespoke" programs or commercially available software, as found, for example, in datamining software packages

Advantages of specially written software:

i) Exactly fits the purpose

ii) More likely to perform well as the parameters can be adapted and the models can be guided to give the desired outputs with domain knowledge.

iii) High sensitivities and specificities can be achieved

Disadvantages of specially written software:

i) Expensive to set up and maintain

ii) Difficult to modify especially when the original author of the software has moved away

iii) Very dependent on the person who wrote the program for maintenance and modifications

iv) Risk of overtraining and overfitting data possibly resulting in poor performance in prospective new datasets

Advantages of commercially available software:

i) Readily available "off the shelf"

ii) Technical support more readily available

iii) Initial costs lower, generally less expensive to purchase

iv) Packages can be updated and supported more easily

v) Easier to adapt to changing requirements, for example the inclusion of new fields

vi) Still possible to have individual requirements catered to via expert support

vii) Possible deals on related products

Disadvantages of commercially available software:

i) Selection of machine learning models limited to that which is packaged in the software

ii) Not always easy to adapt to exact clinical requirements

iii) Maintenance or service agreements are mandatory for continued newer versions

Skills learned in the data manipulation of machine learning software within these packages can be adapted for use in other fields.

- The accuracy of machine learning models can be influenced by adjusting internal bias and function, also by the proportion of positive to negative cases used in the training.
- Neural networks, Logistic regression, C5 and Classification and Regression Trees tended to perform equally well or equally badly.

- The results can be expressed as categorical data, as in negatives or positives, or in a range where cut-off points can be chosen, for example low, moderate, or high risk. This application can be the basis of an interactive health education tool for patients to assess their risk status via a user-friendly visual interface such as a computer touch-screen. A prototype "Scoring model" was developed.
- The combination of Heavy Smoker and Heavy Drinker alone does not necessarily signal an individual at high risk. Applying just these guidelines for this study would have resulted in 86 to 92% of positive cases being missed.

10.4 Conclusions

10.4.1 Major conclusions of this study

1. The population attending the dentist for routine treatment, once considered as biased and self-selecting, is indeed representative of the general population in terms of the risk habits of smoking and alcohol consumption and the distribution of positive lesions within the non-, moderate and heavy user categories (Chapter 6). The prevalence of oral cancer and precancerous lesions and conditions within this population group is comparable with published national figures and similar to those generally found in developed Western countries (Table 6.7). This suggests that screening in a general dental practice setting can be a viable prospect. It is possible that the population screened in this study could still be atypical of the general population and that the two separate studies done five years apart in different populations are not representative of the

general population. To prove or disprove this, mass population screening needs to be carried out; unfortunately there appears to be no economic justification for this at the present time (Rodrigues *et al* 1998).

2. General dental practitioners are correctly finding positive and negative lesions and conditions in accordance with the criteria that they have been taught (Jullien *et al* 1995)
3. Screening or case finding in a primary care environment such as a general dental practice is feasible and could result in the diagnosis of more cases of oral cancer and precancer at an earlier stage (downstaging).
4. Machine learning software as applied to the identification of groups at high risk of oral cancer and precancer can achieve sensitivities comparable to a previous study using an artificial neural network as well as screeners who were dental surgeons and other health professionals (Speight *et al* 1995, Jullien *et al* 1995). High sensitivities however, were always associated with low specificities and there was a tendency for the models to miss up to 20% of positive cases (Chapter 9,Table 9.11). The ideal performance for our machine learning models would be the ability to identify all positive cases as well as some who were negative but have the profile of a positive case, which therefore would be designated as being at high risk. This unfortunately had not been achieved in this study and as such the use of machine learning models cannot be recommended at this stage of development.
5. If more clear cut criteria associated with oral dysplasia can be found, such as gene mutations, machine learning based case finding might be more reliable. Current diagnostic tools include the use of tolonium blue

staining, and Brush biopsy techniques, but present costs would be a disincentive for their widespread use in general practice. A possible means of improving the performance could be the exposure and training on much larger sets of data.

10.4.2 Suggestions for future work

More data will be helpful in giving a better understanding of the populations visiting their dentist. Building on the premise that general dental practitioners can reliably identify positive lesions and conditions and that 59% of the general population visit their dentist regularly (Nuttall *et al* 1998), screening studies can be extended to include more general dental practices in London and the provinces as well as other regions such as Northern Ireland and Scotland, where the prevalences of oral cancer and precancer are reported to be higher (CRC, 2000).

The use of Patient Information Leaflets has been shown to result in better patient acceptance and significantly increased awareness of correct signs and risk factors associate with oral cancer (Humphris *et al* 2001) Appropriately written information leaflets could be made available in the waiting room or even in non clinical settings such as public libraries and health clubs.

Screening could be extended to medical practices. Pearson *et al* (2001), in their study of Bangladeshis living in the Tower Hamlets area of London, screened a population of 137 Bangladeshis including women and found 28 cases (20.4%) considered to be positive. These individuals would not normally have attended at dental practices. The availability of access centres and walk-in dental clinics in the UK presents further opportunities for screening. Other health workers such as nurses and health visitors could also be involved in screening for oral cancer and precancer. It is widely accepted that suitably trained health

workers can reliably detect positive lesions (Warnakulasuriya and Pindborg 1990, Mathew 1997).

Sub-populations seen to be at a higher risk for oral cancer and precancer such as the homeless, heavy smokers and alcoholics could also be targeted. A representative study is one by Dombi *et al* (2001), where screening was done in a lung cancer clinic. Similar local studies could be carried out in locations where individuals at high risk can be targeted, such as food kitchens for the homeless, rehabilitation centres for drug and alcohol abusers, and medical oncological units.

Any increase in numbers of cases referred to tertiary care as a result of more widespread screening could potentially overload hospital resources. Toward avoiding this, the concept of sentinel practices (Tickle *et al,* 2000) could be adopted and adapted to screening for oral cancer and precancer in primary health care. Sentinel practices or individual general dental practitioners with a special interest in oral medicine could be employed to initially deal with, or triage cases referred. Facilities could be set up to provide on-going training to such practitioners.

With the increase of data, machine learning software can be re-evaluated as a potential tool for use as a pre-screen filter. At this stage of development, machine learning software can have a role as an interactive health awareness package which could be integrated into a patient information module for use in dental and medical practices.

There is a possibility that the populations screened are still not representative of the general population; however for the sake of pragmatism, opportunistic screening for oral cancer and precancer in general dental practices in the United Kingdom appears presently to be the most efficient compromise between mass population screening and not screening at all.

References

References

Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, Wright GL Jr. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res* 2002 Jul 1; **62**(13): 3609-14

Adami HO, Ponten J, Sparen P, Bergstrom R, Gustafsson L, Friberg LG. Survival trend after invasive cervical cancer diagnosis in Sweden before and after cytologic screening. 1960-1984. *Cancer* 1994 Jan 1; **73**(1): 140-7

Alhbom HE. Simple achlorhydric anaemia. Plummer Vinson syndrome and carcinoma of the mouth, pharynx and oesophagus in women. *BMJ* 1936; **2**:331-33.

Allison P, Locker D, Feine JS. The role of diagnostic delays in the prognosis of oral cancer: a review of the literature. *Oral Oncol* 1998; **34**: 161-170.

Almadori G, Palidetti G, Cerullo M, Ottaviani F, D'Alatri L Marijuana smoking as a possible cause of tongue carcinoma in young patients. *J Laryngol Otol.* 1990 Nov; **104**(11): 896-9.

Almog DM, Heisler EM. Computer intuition: guiding scientific research in imaging and oral implantology. *J Dent Res* 1997 Oct; **76**(10): 1684-9

Andreasen J. Oral manifestations in discoid and systemic lupus erythematous. *Acta Odontol*

Scand 1964; **22**: 295- 310.

Austoker J, Reducing alcohol intake. *BMJ* 1994; **308**: 1549-52. (a)

Axell T. A prevalence study of oral mucosal lesions in an adult Swedish population *Odontol Revy* 1976; **27**: 1-103.

Axell, T, Pindborg, JJ., Smith, CJ., Waal van der, I. And an International Collaborative Group on Oral White Lesions, Oral white lesions with special reference to precancerous and tobacco-related lesions: conclusions of an international symposium held in Uppsala, Sweden, May 18-21, 1994. *J Oral Pathol Med* 1996, **25**, 49-54.

Bagnardi V, Blangiardo M, La Vecchia C, Corrao G. Alcohol consumption and the risk of cancer: a meta-analysis. *Alcohol Res Health* 2001; **25**(4):263-70

Balaji KC, Fair WR, Feleppa EJ, Porter CR, Tsai H, Liu T, Kalisz A, Urban S, Gillespie J. Role of advanced 2 and 3-dimensional ultrasound for detecting prostate cancer. *J Urol* 2002 Dec; **168**(6): 2422-5

Balaram P, Sridhar H, Rajkumar T, Vaccarella R, Samkaranarayanan RH, Gajalakshimi V, Munoz N, Franceschi S. Oral cancer in southern India: The influence of smoking, drinking, paan chewing and oral hygiene. *International Journal of Cancer* 2002 Jan; Vol **98** (3): 440-445

Banoczy J, Csiba A. Occurrence of epithelial dysplasia in oral leukoplakia: analysis and

follow-up of 12 cases. *Oral Surg* 1976; **42**: 6: 766-774.

Banoczy J. Follow-up studies in oral leukoplakia. *J Maxillofac Surg* 1977; **5**:69-75.

Banoczy J, Rigo O. Prevalence study of oral precancerous lesions within a complex screening system in Hungary. *Community Dent Oral Epidemiol* 1991; **19:265**-7.

Barnard NA, Scully C, Eveson JW, Cunningham S, Porter SR. Oral cancer development in patients with oral lichen planus. *J Oral Pathol Med* 1993; **22**:421-4

Barrett AW, Kingsmill VJ, Speight PM. The frequency of fungal infection in biopsies of oral mucosal lesions. *Oral Dis* 1998 Mar; 4(1): 26-31

Batsakis J. Histology of Oral Cancer-Chapter 4. Oral Cancer by Shah J, Johnson NW, Batsakis J. 2003; MartinDiaz Pub, ISDN 1-899666 87X

Baum M "Screening – A cruel deception" *The Practitioner*, May 2002 Vol **26**: 293

Baxt WG, Shofer FS, Sites FD, Hollander JE. A neural computational aid to the diagnosis of acute myocardial infarction. *Ann Emerg Med* 2002 Apr; **39**(4): 366-73

Berg JW, Schotlenfeld D, Ritter F. Incidence of multiple primary cancers. 3: cancers of the respiratory and upper digestive system as multiple primary cancers. *J Natl Cancer Inst* 1976:**44**:263-74.

Bessho F. Effects of mass screening on age-specific incidence of neuroblastoma. : *Int J Cancer* 1996 Aug 7;67(4):520-2

Billroth T. Die allgemeine chirurgishe Pathologie und Therapie. In 51 Vorlesungen. In: Reimer G. cd. *Handbuch fur Studierende and Arzie.* 14th ed. Berlin. 1889:908.

Binnie WH, Rankin KV, Mackenzie IC. Etiology of squamous cell carcinoma. *J Oral Pathol* 1983 Feb: **12** (1): 11-29

Blamey RW, Wilson AR, Patnick J, Dixon JM. ABC of breast diseases. Screening for breast cancer. *BMJ* 1994 Oct 22; **309**(6961):1076-9

Blinkhorn AS, Jones JH. Behavioural aspects of oral cancer screening. *Community Dent Health* 1993 Jun; **10** Suppl 1:63-9

Bouqout JE. Common oral lesions found during a mass screening examination. *JADA*, Jan 1986; **Vol 112**:50-7.

Bouquot JE. Epidemiology. In: Gnepp DG (ed.) Pathology of the head and neck. New York: Churchill Livingstone; 1987:263-314.

Bouquot JE, Gorlin RJ. Leukoplakia, lichen planus, and other oral keratoses in 23,616 white Americans over the age of 35 years. *Oral Surg Oral Med Oral Pathol* 1986; **61**: 373-381

Boyle P, Macfarlane GJ, Scully C. Oral Cancer: necessity for prevention strategies. *Lancet* 1993; **342**:1129.

Brandberg Y, Bolund C. Michelson H, Mansson-Brahme E, Ringborg U, Sjoden P. Psychological reactions in public melanoma screening. *Eur J Cancer* 1993; **29**:860-863.

Breiman L, Friedman J, Olshen R , Stone C. 1984 CART: Classification and Regression Trees. Belmont, Calif. Wadsworth.

Brickley MR, Cowpe JG, Shepherd JP. Performance of a computer simulated neural network trained to categorise normal, premalignant and malignant oral smears. *J Oral Pathol Med* 1996 Sep; **25**(8): 424-8

Brugere I, Guenel P, Leclerc A, Rodriguez 1. Differential effects of tobacco and alcohol in cancer of the larynx, pharynx and mouth. *Cancer* 1986; **57**: 391-391.

Burzynski NJ, Firriolo FJ, Butters JM, Sorrell CL. Evaluation of oral cancer screening. J Cancer Educ. 1997 Summer;12(2):95-9.

Butler IA. Prevention may be more expensive than cure. *J R Soc Med* 1993; **86**: 341-344.

Campisi G, Margiotta V Oral mucosal lesions and risk habits among men in an Italian study population *J Oral Pathol Med* 2001;Jan;(**1**): 22-8.

Cancer Research Campaign –Oral Cancer 2000.

CancerStats-Incidence UK April 2003

Canniff J. Oral submucous fibrosis: its pathogenesis and management. *Br Dent J* 1986 Jun **21**: 160 (12) :429-34

Castellaro C, Favaro G, Castellaro A, Casagrande A, Castellaro S, Puthenparampil DV, Salimbeni CF. An artificial intelligence approach to classify and analyse EEG traces. *Neurophysiol Clin* 2002 Jun; **32** (3): 193-214

Cenci M, Nagar C, Vecchione A. PAPNET-assisted primary screening of conventional cervical smears. *Anticancer Res* 2000 Sep-Oct; 20(5C): 3887-9

Chamberlain J. Evaluation of screening for cancer. *Community Dental Health* 1993; **1O**,Supplement 1:5-11.

Chen JK, Katz RV, Krutchkoff DJ. Intraoral squamous cell carcinoma. Epidemiologic patterns in Connecticut from 1935 to 1985. *Cancer* 1990 Sep 15; **66**(6):1288-96

Cianfriglia F, Di Gregorio DA , Maneri A. Multiple primary tumours in patients with squamous cell carcinoma. *Oral Oncology* 1999 Mar **35** (2) 157-63.

Cross SS, Hamson RP, Kennedy RL. Introduction to neural networks. *Lancet* 1995; **346**: 1075-79.

Day GL, Blot WJ, Austin DF, Bernstein L, Greenberg RS, Preston Martin S, Schoenberg JB, Winn DM, Mc Laughlin JK, Fraumeni JF Jr. Racial differences in risk of oral and

pharyngeal cancer: alcohol, tobacco and other determinants. *J Natl Cancer Inst* 1993; **85**:465-73.

Djavan B, Remzi M, Zlotta A, Seitz C, Snow P, Marberger M. Novel artificial neural network for early detection of prostate cancer. *J Clin Oncol* 2002 Feb 15; **20**(4): 921-9

.

Dombi C, Voros-Balog T, Czegledy A, Hermann P, Vincze N, Banoczy J. Risk group assessment of oral precancer attached to X-ray lung screening examinations. *Community Dent Oral Epidemiol* 2001; **29**: 9-13.

Donald PJ. Marijuana smoking-possible cause of head and neck carcinoma in young patients. *Otolaryngol Head Neck Surg* 1986:**94**, 517-521.

Downer MC, Speight PM. Cost and value considerations in screening for oral cancer and precancer. *Community Dent Health* 1993; **1O**: Suppl 1:71-78.

Downer MC. Today's proposal, tomorrow's answer? In: Introduction to dental public health (eds) Downer MC, Gelbier S, Gibbons DE. FDI World Press; London, 1994: lO6-26.

Downer MC, Evans AW, Hughes-Hallett CM, Jullien JA, Speight PM, Zakrzewska JM. Evaluation of screening for oral cancer and precancer in a company headquarters. *Community Dent Oral Epidemiol* 1995; **23**: 84-88.

Downer MC, Jullien JA, Speight PM. An interim determination of health gain from oral cancer and precancer screening: 1. Obtaining health state utilities. *Comm Dent Health,*

1997a: **14**, 139-142.

Downer MC, Jullien JA, Speight PM. An interim determination of health gain from oral cancer and precancer screening: 2. Developing a model of population screening. *Community Dental Health* 1997b; **14**: 227-232.

Downer MC, Jullien JA, Speight PM. An interim determination of health gain from oral cancer and precancer screening: 3. Preselecting high risk individuals. *Community Dental Health* 1998; **15**: 72-76.

Drew PJ, Monson JR. Artificial intelligence for clinicians. *J R Soc Med* 1999 Mar; **92**(3): 108-9

Dyer CR. Machine Learning CS 540 Lecture notes (Chapter 18.1-18.4) University of Wisconsin-Madison

Dyer O. "News" reporting on the WHO Multicentric Cervical Cancer Study Group report. Lancet 2002; 359:1085-92 .*BMJ* April 2002 Vol **324**: 808

Eddy DM. Screening for cancer. Theory, analysis and design. Englewood Cliffs, New Jersey: Prentice Hall (1980).

Eddy DM. Setting priorities for cancer control programs. *J Natl Cancer Inst* 1986; **76** (2): 187-199.

Edwards PJ, Hall DMB. Screening, ethics and the law. *BMJ* 1992; **305**:267-8.

Eisenberg E, Krutchkoff DJ. Lichenoid lesions in oral mucosa; diagnostic criteria and their importance in the alleged relationship to oral cancer. *Oral surgery, Oral Medicine, Oral Pathology* 1992;**73**:699-704.

Ercal F, Chawla A, Stoecker WV, Lee HC, Moss RH. Neural network diagnosis of malignant melanoma from color images. *IEEE Trans Biomed Eng* 1994 Sep; **41**(9): 837-45

Eversole LR. Diseases should not be considered entities unto themselves. *Oral Surg Oral Med Oral Pathol* 1992 Jun; **73** (6): 707.

Farnsworth A, Chambers FM, Goldschmidt CS. Evaluation of the PAPNET system in a general pathology service *MJA* 1996; 165: 429-431

Fearon E, Vogelstein B A genetic model for colorectal tumorigenesis. *Cell* 1990:**61**, 757-767

Ferlay J, Bray F, Pisani P *et al.,* GLOBOCAN 2000: cancer incidence, mortality and prevalence worldwide, version 1.0. IARC Cancer Base No.5. Lyon: IARC, 2002

Field EA, Darling AE, Zakrzewska JM. Oral mucosal screening as an integral part of routine dental care *Br Dent J* 1995; 179:262-266

Field EA, Ellis A, Friedmann PS, Leigh IM, Field JK. Oral tylosis: a re-appraisal. *Oral Oncol* 1997 Jan;33(1):55-7

Finlay J, Dix A. An Introduction to Artificial Intelligence 1999 UCL Press.

Flaitz CM, Nichols CM, Walling DM, Hicks MJ. Plasmablastic lymphoma: an HIV-associated entity with primary oral manifestations. *Oral Oncol* 2002 Jan; **38**(1) 96-102

Floyd CE Jr, Lo JY, Tourassi GD. Case-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions. *Am J Roentgenol* 2000 Nov; **175**(5): 1347-52

Flynn FV. Screening for presymptomatic disease. *J C/in Pathol* 1991; 44:529-538.

Fontana S. Screening for lung cancer. In Miller AB. ed. *Screening for cancer.* 1985: 375-94. Orlando:Academic press

Fowler G, Mant D. Health checks for adults. *BMJ* 1990; **300**: 13 18-20.

Franceschi S, Levi F, La Vecchia C, Conti E, Dal Maso L, Barzan L, Talamini R. Comparison of the effect of smoking and alcohol drinking between oral and pharyngeal cancer. *Int J Cancer* 1999 Sep 24; **83**(1): 1-4

Franceschi S, Barzan L, Talamini R. Screening of the Head and Neck: if not now, when? *Oral Oncology*, 1997; Vol **33** No.5, 313-316

Fung M, Gallagher C, Machtay M. Lung and aero-digestive cancers in young marijuana smokers. *Tumori* 1999 Mar-Apr; **85**(2):140-2

Gallagher B, Wang Z, Schymura MJ, Kahn A, Fordyce EJ. Cancer incidence in New York State acquired immunodeficiency syndrome patients. *Am J Epidemiol* 2001 Sep 15; **154**(6): 544-56

Gangal SG, Tatake RJ. Immune status in head and neck cancer with special reference to cancer of the oral cavity.1991 In Risk Markers for Oral Diseases. Vol 2. Oral Cancer. Detection of Patients and Lesions at risk. (ed Johnson NW); Cambridge University Press, Cambridge, pp114-54.

Gordis L. The scope of screening. *J Med Screening* 1994; **1**: 98-100.

Gorsky M, Silverman S Jr.Tobacco use in patients with head and neck carcinomas: habit chages and second primary oral/ oropharyngeal cancers in patients from San Francisco. *Cancer J.* 1994: **7**: 78-80.

Gross G, Gundlach K, Reichart PA. Virus infections and tumors of the oral mucosae. Symposium of the "Arbeitskreis Oralpathologie und Oralmedizin" and the "Arbeitsgemeinschaft Dermatologische Infektiologie der Deutscher Dermatologischen Gesellschaft", Rostock, July 6-7, 2001 *Eur J Med Res* 2001 Oct ; **6**(10):440-58

Guggenheimer J, Verbin RS, Johnson JT, Horkowitz CA, Myers EN. Factors delaying the diagnosis of oral and orophryangeal carcinomas. *Cancer* 1994; **64:** 932-935.

Gupta, PC, Mehta, FS, Daftary DK., Pindborg JJ, Bhonsle RB, Jalnawalla PN, Sinor PN,

Pitkar VK, Murti PR, Irani RR, Shah HT, Kadam PM, Iyer KS, Iyer HM, Hedge AK, Chandrashekar GK, Shiroff BC, Sahiar BF, Mehta MN. Incidence rates of oral cancer and natural history of oral precancerous lesions in a 10-vear follow-up study of Indian villagers. *Community Dental and Oral Epidemiology*, 1980; **8**: 287-333.

Gupta PC, Mehta FS, Pindborg JJ, Aghi MB, Bhonsle RB, Daftary DK, Murti PR, Shah HT. Intervention study for primary prevention of oral cancer among 36 000 Indian tobacco users. *Lancet* 1986; **1**:1235-9.

Gupta PC, Mehta FS, Pindborg JJ, Bhonsle, Murti PR, Daftary DK, Aghi MB. Primary prevention trial of oral cancer in India: a 10-year follow-up study. *J Oral Pathol Med.* 1992;(**2**):433-9.

Gupta PC, Murti PR, Bhonsle RB, Mehta FS, Pindborg JJ. Effect of cessation of tobacco use on the incidence of oral mucosal lesions in a 10 year study of 12212 users. *Oral Diseases* 1995; **1:54**-58.

Gupta PC, Nandakumar A. Oral cancer scene in India. *Oral Diseases* 1999; **5**, 1-2

Gurney K, An Introduction to Neural Networks1997 UCL press

Hakama M. A screening programme that worked: discussion paper. *J R Soc Med* 1990; **83**: 322-324.

Halloppeau H. Sur un cas de lichen de Wilson gingival avec neoplasie voisine dans la

region maxillaire. *Bull Soc Fr Dennatol Syphgr* 1910; 17:32.

Haque MF, Harris M, Meghji S, Speight PM. An immunohistochemical study of oral submucous fibrosis. *J Oral Pathol Med* 1997 Feb; **26**(2): 75-82

Hardcastle JD, Chamberlain JO, Robinson MH, Moss SM, Amar SS, Balfour TW, James PD, Mangham CM.Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. *Lancet,* 1996; **348:**1472-147

Helander A, Carlsson AV, Borg S. Longitudinal comparison of carbohydrate-deficient transferrin and gamma-glutamyl transferase: complementary markers of excessive alcohol consumption. Alcohol Alcohol. 1996 Jan; 31(1): 101-7.

Hindle I, Nally F. Oral cancer: a comparative study between 1962-1967 and 1980-1984 in England and Wales. *Br Dent J* 1991; **170**: 15-19.

Hindle I, Downer MC, Speight PM. Necessity for preventative strategies in oral cancer. *Lancet* 1994; **343**: 178-9.

Hindle I, Downer M C, Speight P M. The epidemiology of oral cancer. *Brit J Oral Max Fac Surg* 1996; **34**: 471-476.

Hobaek A. Leukoplakia oris. *Acta Odontol Scand* 1946; 7:61-91.

Hogewind WFC, van Der Waal I. Prevalence study of oral leukoplakia in a selected

population of 1000 patients from the Netherlands. *Community Dent Oral Epidemiology* 1988; **16**:302-5.

Holmstrup P, Thorn JJ, Rindum J, Pindborg JJ. Malignant development of lichen planus-affected oral mucosa. *J Oral Pathol* 1988; **17**:219-25.

Hopfield JJ. Neural Networks and physical systems with emergent colletive computational properties. 1982 Proceedings of the National Academy of Sciences of the USA 79,2554-88

Howie NM, Trigkas TK, Cruchley AT, Wertz PW, Squier CA, Williams DM. Short-term exposure to alcohol increases the permeability of human oral mucosa. *Oral Dis* 2001 Nov; 7(6): 349-54

Humphris GM, Duncalf M, Holt D, Field EA. The experimental evaluation of an oral cancer information leaflet. Oral Oncol. 1999 Nov;35(6):575-82.

Humphris GM, Ireland RS, Field EA. Randomised trial of the psychological effect of information about oral cancer in primary care settings. Oral Oncol. 2001 Oct;37(7):548-52.

Hun HH, Broders AC. Multiple primary malignant neoplasms. *J Lab Clin Med* 1933; **I8**: 765-77.

Ikeda N, Ishii T, Kawai T. Epidemiologic study of oral leukoplakia based on mass screening for oral mucosal disease in a selected Japanese population. *Community Dent Oral epidemio* 1991; **19**(3): 160-163.

Ikeda N, Downer MC, Ishii T, Fukano H, Nagao T, Inoue K. Annual screening for oral cancer and precancer by invitation to 60-year-old residents of a city in Japan. *Community Dent Health* 1995 Sep;12(3):133-7

Ikeda N, Handa Y, So Po Khim, Durward C, Axell T, Mizuno T, Fukano H, Kawai T.. Prevalence study of oral mucosal lesions in a selected Cambodian population, *Community dentistry and Oral Epidemiology* 1995; **23:49**-54.

Jacobs IJ, Skates SJ, MacDonald N. Screening for ovarian cancer: a pilot randomised controlled trial. *Lancet* 1999; **353** (9160): 1207-1210.

Jando G, Siegel RM, Horvath Z, Buzsaki G. Pattern recognition of the electroencephalogram by artificial neural networks. *Electoencephalogr Clin Neurophysiol* 1993; **86** (2): 100-09.

Johnson NW. Tobacco use and oral cancer: a global perspective. *J Dent Educ* 2001 Apr; **65** (4): 328-39

Johnson NW. Orofacial neoplasms: global epidemiology, risk factors and recommendations for research *International Dental Journal* 1991; **41**: 365-375

Johnson NW, Warnakulasuriya KA. Epidemiology and aetiology of oral cancer in the United Kingdom. *Community Dent Health* 1993 Jun;10 Suppl 1:13-29

Johnson NW, Warnakulasuriya S, Tavassoli M. Hereditary and environmental risk factors:

clinical and laboratory risk markers for head and neck, especially oral cancer and precancer. *European Journal of Cancer Prevention* 1996; **5**:5-17

Jullien JA, Downer MC, Zakrzewska JM, Speight PM. Evaluation of a screening test for the early detection of oral cancer and precancer. *Community Dental Health* 1995; **12**: 3-7.

Jullien JA, Zakrzewska JM, Downer MC, Speight PM. Attendance and compliance at an oral cancer-screening programme in a general medical practice. *Oral Oncol, Eur J Cancer* 1995; **31B**: 202-206

Kampschoer GH, Fujii A, Masuda Y: Gastric cancer detected by mass survey: comparison between mass survey and outpatient detection. *Scandinavian Journal of Gastroenterology* 1989; **24**(7): 813-817.

Kaplan BR. Oral Lichen planus and Squamous carcinoma: case report and update of the literature. *Rhode Island Dent J* 1991; **24:**405-14.

Kaufman S, Grabau JC, Lore JM. Symptomatology in head and neck cancer. *Am J Public Health,* 1980; **70,** 520-522.

Kay E, Locker D. A systematic review of the effectiveness of health promotion aimed at improving oral health. *Comm Dent Health* 1998 Sep; **15**:132-144.

Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. *JAMA* 1995; **273**: 149-154.

Kerlikowske K. Efficacy of screening mammography among women aged 40-49 years and 50-59 years: comparison of relative and absolute benefit. *Natl Cancer Inst Monogr* 1997; (22): 79-86.

Kernick D. Costing Interventions in primary care *Family Practice* Vol 17, No 1, 66-70 Oxford University Press 2000

Keyhani-Rofagha S, Schreiber K, Valente PT. Significant reduction in the rate of false-negative cervical smears with neural network-based technology (PAPNET Testing System). *Human Pathology.* 1997 Oct. (10): 1196-203,

King GN, Thornhill MH. Dental attendance patterns in renal transplant recipients. *Oral Dis* 1996 Jun; 2(2): 145-7

Kleinman DV, Swango PA, Niessen LC, *et al.* Epidemiologic studies of oral mucosal conditions-methodologic issues. *Comm Dent Oral Epidemiol* 1991; **19**:129-40.

Kleinman DV, Swango PA, Pindborg JJ, Gupta P. Toward assessing trends in oral mucosal lesions: lessons learned from oral cancer *Adv Dent Res* 1993 Jul; **7** (1): 32-41.

Koss LG, Sherman ME, Cohen MB, Anes AR, Darragh TM, Lemos LB, McClellan BJ, Rosenthal DL, Keyhani-Rofagha S, Schreiber K, Valente PT. Significant reduction in the rate of false-negative cervical smears with neural network-based technology (PAPNET Testing System). *Hum Pathol* 1997 Oct; 28(**10**): 1196-203

Kovac-Kovacic M, Skaleric U. The prevalence of oral mucosal lesions in a population in Ljubljana, Slovenia. *J Oral Pathol Med* 2000 Aug;(7): 331-5.

Kowalski LP, Franco EL, Torloni H, Fava AS, de Andrade Sobrinho J, Oliveira BV, Curado MP. Lateness of diagnosis of oral and oropharyngeal carcinoma; factors related to the tumour, patient and health professionals. *Oral Oncol, Eur J Cancer* 1994; **30B,** 167-173

Kramer IRH, El-Labban N, Lee KW. The clinical features and risk of malignant transformation in sub-lingual keratosis. *Bri Dent J* 1978; **144**:6: 171-180

Krogh P, Hald B, Holmstrup P. Possible mycological etiology of oral mucosal carcinoma: catalytic potential of infecting Candida albicans and other yeasts in production of N-nitrosobenzylmethylamine. *Carcinogenis* 1987; **8**: 1543-8

Kulasegaram R, Downer MC, Julien JA, Zakrzewska J, Speight PM. Case-control study of oral dysplasia and risk habits among patients of a Dental Hospital. *Oral Oncol, Eur J Cancer* , 1995; Vol **31B** No.4: 227-231.

Lader D, Meltzer H. Smoking related behaviour and attitudes 2001, London ONS 2002

Lader D, Meltzer H, Drinking related behaviour and attitudes 2001, London ONS 2002

Lerman CE, Rimer BK. Psychosocial impact of cancer screening. *Oncology* 1993; 7:67-72.

Lewis PC, Harrell JS, Deng S, Bradley C. Smokeless tobacco use in adolescents: the Cardiovascular Health in Children (CHIC II) Study. J Sch Health 1999 Oct;69(8):320-5

Lewis S, Menon U Screening for ovarian cancer. *Expert Rev Anticancer Ther* 2003 Feb; **3**(1): 55-62

Lillehaug IS, Lajoie SP. AI in medical education—another grand challenge for medical informatics. *Artificial Intelligence in Medicine.* 1998 Mar; **12**(3): 197-225,

Llewellyn C, Linklater K, Bell J, Johnson NW, Warnakulasuriya KA. Squamous cell carcinoma of the oral cavity in patients aged 45 years and under; a descriptive analysis of 116 casses diagnosed in the South East of England from 1990-1997. *Oral Oncol* 2003 Feb; **39** (2):106-14

Lucht R, Delorme S, Brix G. Neural network-based segmentation of dynamic MR mammographic images. *Magn Reson Imaging* 2002 Feb; **20**(2): 147-54

Lynch MA, Brightman VJ, Greenberg MS. Burkets Oral Medicine, diagnosis and treatment 8th edition. Lippincott, Philadelphia 1984.

Macdonald I (ed). "Health Issues Related to Alcohol Consumption". 1999 September; ILSI Europe Alcohol Task Force (second edition)

Macfarlane GJ, Boyle P, Evstifeeva TV, Robertson C, Scully C. Rising trends of oral cancer mortality among males worldwide: the return of an old public health problem. *Cancer*

Causes and Control 1994a May; **5**(3): 259-65

Macfarlane GJ, Evstifeeva TV, Robertson C, Boyle P, Scully C. Trends of oral cancer mortality among females worldwide. *Cancer Causes and Control* 1994b, **5**; 255-258

Maier H, Zoller J, Herrmann A Keriss nbM, Heller WD. Dental status and oral hygiene in patients with head and neck cancer. *Otolaryngol Head Neck Surg* 1993 Jun; **108**: 655-61

Mandel JS, Bond JH, Church TR, Snover DC, Bradley GM, Schuman LM, Ederer F. Reducing mortality from colorectal cancer by screening for fecal occult blood. *New England Journal of Medicine* 1993; **328**(19): 1365-1371.

Mandel JS, Church TR, Ederer F, Bond JH. Colorectal cancer mortality: effectiveness of biennial screening for fecal occult blood. *Journal of the National Cancer Institute* 1999; **91**(5): 434-437.

Mant D, Fowler G. Mass screening : theory and ethics. *BMJ* 1990; **300**: 916-8

Marshall JR, Boyle P. Nutrition and oral cancer. *Cancer Causes Control* 1996 Jan; **7** (1): 101-11.

Marteau T. Reducing the psychological costs. *BMJ* 1990; **301:26**-28.

Martinez-Diaz AI, Garcia-Pola MJ. Epidemiological study of oral mucosal pathology in patients of the Oviedo School of Stomatology. *Med Oral* 2002; **7**: 4-16

Mashberg A, Meyers H. Anatomical site and size of 222 early asymptomatic oral squamous cell carcinomas: a continuing prospective study of oral cancer II. *Cancer* 1976 May: **37** (5): 2149-57.

Mashberg A, Boffetta P, Winkelman R., Garfinkel L. Tobacco smoking, alcohol drinking and cancer of the oral cavity and oropharynx among US veterans. *Cancer*.1993; **72**(4): 1369-1375.

Mathew B, Sankaranarayanan R, Sunilkumar KB, Kuruvilla B, Pisani P, Nair MK. Reproducibility and validity of oral visual inspection by trained health workers in the detection of oral precancer and cancer. *British J Cancer* 1997; **76** (3): 390-394.

Mayor S, Introduce the single European diet. *Health and Ageing* 2002 May: 58-9.

Metha FS, Gupta PC, Daftary DK, Pindborg JJ, Choksi SK. An epidemiologic study in oral cancer and precancerous conditions among 101761 villagers in Maharashtra, India. *Int J Cancer* 1972; **10**:134-41.

Metha FS, Gupta PC, Bhonsle RB, Murti PR, Daftary DK, Pindborg JJ. Reproducibility of oral cancer using basic health workers in an area of high oral cancer incidence in India. *Cancer Detection and Prevention* 1986; **9**:**219**-225.

Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC. Abstract. In: Cancer screening. International Union against Cancer, Cambridge University Press 1991.

Miller CS, Johnstone BM. Human papillomavirus as a risk factor for oral squamous cell carcinoma: a meta-analysis 1982-1997 *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 2001 Jun; **91**:622-35.

Mitchell TM, Machine Learning 1997 McGraw-Hill Companies

Moles DM, Downer MC, Speight PM. Meta-analysis of measures of performance reported in oral cancer and precancer screening studies. *British Dental Journal* 2002 March; Vol **192**; No 6:340-343

Murakami R, Tsukuma H, Ubukata T, Nakanishi K, Fujimoto I, Kawashima T, Yamazaki H, Muir J, Lancaster T, Fowler G, Neil A. Self reporting overestimates smoking cessation rates. *BMJ* 1998;**316**:704

Murti PR, Daftary DK, Bhonsle RB, Gupta PC, Mehta FS, Pindborg JJ. Malignant potential of oral lichen planus: observations in 722 patients from India. *J Oral Pathol* 1986; **15**:71-7.

Nunes FD, Loducca SV, de Olivera EM, de Araujo VC. Well differentiated liposarcoma of the tongue. *Oral Oncol* 2002 Jan; **38**(1): 117-9

Office of National Statistics. Cancer statistics registrations, Registrations of cancer diagnosed in 1998, England. Series MB1. No 29. London ONS, 2002

Office of National Statistics. Cancer statistics registrations, Registrations of cancer

diagnosed in 1998, England. Series MB1. No 30. London ONS, 2003

Ogden GR. Second malignant tumours in head and neck cancer. *BMJ* 1991 Jan 26; **302**(6770): 193-4

Ogden GR. Recurrence of malignant tumours in the head and neck. *BMJ* 1996; **313**:700-01.

O'Grady JF, Reade PC. Candida albicans as a promoter of oral mucosal neoplasia.*Carcinogenesis* 1992 May;**13**(5):783-6

Oshima A.: Estimation of validity of mass screening program for gastric cancer in Osaka, Japan. *Cancer* 1990 **65** (5): 1255-1260

Oshima A, Hirata N, Ubukata T, Umeda K, Fujimoto I. Evaluation of a mass screening program for stomach cancer with a case-control study design. International Journal of *Cancer* 1986; **38(**6): 829-833

Parkin DM, Pisani P, and Ferlay J. Estimates of worldwide incidence of eighteen major cancers in 1985. *Int J Cancer* 1993; **54**: 594-606

Parkin DM, Pisani P, Ferlay J. Estimates of worldwide incidence of twenty-five major cancers in 1990. *Int J Cancer* 1999; **80**: 827-841.

Patnick J (Ed.) Cervical screening programme review 1999; Muir Gray, Oxford

Pearson N, Croucher R, Marcenes W, O'Farrell M. Prevalence of oral lesions among a sample of Bangladeshi medical users aged 40 years and over living in Tower Hamlets, UK. *International Dental Journal* 2001 **51**: 30-34.

Peters ES. Epidemiology of oral cancer. 2002, Harvard School of Dental Medicine Website: http://www.hsph.harvard.edu/cohands/home/oralcancer_files/frame.htm

Pindborg JJ, Poulsen HE, Zachariah J. Oral epithelial changes in 30 Indians with oral cancer and submucous fibrosis. *Cancer* 1967; **20**: 1141-46.

Pindborg JJ, Roed-Peterson B, Renstrup G. Role of smoking in floor of the mouth leukoplakias. *J Oral Pathol* 1972; **1**(1): 22-9

Pindborg JJ. Oral cancer and precancer. Bristol, Wright. 1980

Pindborg JJ, Daftary DK, Gupta P, *et al.* Public health aspects of oral cancer: implications for cancer prevention in the community. In: Johnson NW, ed. Risk markers for Oral diseases Vol. 2. Oral cancer: Detection of patients and lesions at risk. Cambridge: Cambridge University Press, 1991.

Platz H, Fries R, Hudec M. Prognosis of oral cavity carcinomas, Results of a multi centric retrospective observational study. 1986, Hanser Munich.

Prout MN, Sidari J, Witzburg A *et al.* Head and neck cancer screening among 4611 tobacco

users older than 40 years. *Otolaryngol Head Neck Surg* 1997; **116**: 201-8.

Pukkala E, Soderholm AL, Lindqvist C. Cancers of the lip and oropharynx in different social and occupational groups in Finland. *Eur J Cancer. Oral Oncol* 1994; **30B:** 209-15.

Quinlan J. 1988, C4.5 Programs for Machine Learning. Morgan Kaufmann Publishers, Inc.

Reichart PA. Oral mucosal lesions in a representative cross-sectional study of ageing Germans. *Community Dent Oral Epidemiol* 2000; **28**:390-8.

Renson CE. Oral Cancer: a growing problem. *Dent Update* 1990; **17**: 399-401.

Rickart RM. A clinical staining test for the in vivo delineation of dysplasia and carcinoma in situ. *Amer J Obstret Gynae* 1963; **86**(6): 703-712.

Risk JM, Mills HS, Garde J, Dunn JR, Evans KE, Hollstein M, Field JK. The tylosis oesophageal cancer (TOC) locus: more than just a familial cancer gene. *Dis Esophagus* 1999: **12** (3) 173-6

Rodrigues VC, Moss SM, Tuomainen H. Oral cancer in the UK: to screen or not to screen. *Oral Oncol.* 1998 Nov; **34**(6): 454-65.

Roed-Petersen B. Effect on oral leukoplakia of replacing or ceasing tobacco smoking. *Acta Dermato- Venereol* 1982; **62**:164-167.

Ross NM, Gross E. Oral findings based on an automated multiphasic health screening program. *J Oral Med* 1971;**26**:21-26.

Sample PA, Goldbaum MH, Chan K, Boden C, Lee TW, Vasile C, Boehm AG, Sejnowski T, Johnson CA, Weinreb RN. Using machine learning classifiers to identify glaucomatous change earlier in standard visual fields. Invest Ophthalmol *Vis Sci* 2002 Aug; **43**(8): 2660-5

Sankaranarayanan R, Mathew B, Jacob BJ, Thomas G, Somanathan T, Pisani P, Pandey M, Ramadas K, Najeeb K, Abraham E. Early findings from a community-based cluster-randomised, controlled oral cancer screening trial in Kerala, India. The Trivandrum Oral Cancer Screening Study Group. *Cancer* 2000 Feb; **88** (3): 664-73.

Santana JC, Delgado L, Miranda L: Oral cancer case finding programme (OCCFP). 1997 Oral *Oncol*; **33**:10-12.

Schantz SP, Byers RM, Goepfert H, Shallenberger RC, Beddingfield N. The implication of tobacco use in the young adult with head and neck cancer. *Cancer* 1988; **62**:1374-80.

Schwartz LH, Ozsahin M, Zhang GN *et al* Synchronous and metachronous head and neck carcinomas *Cancer* 1994: **74**: 1933-8

Sciubba J. Improving the detection of precancerous and cancerous oral lesions. *JADA* Oct 1999; Vol **130**: 1445-1457.

Scully C, El-Kom M. Lichen planus: a review and update on pathogenesis. *J Oral Path*

1985; **14**: 431-458.

Scully C, Malamos D, Levers BG, Porter SR, *et al.* Sources and patterns of referrals of oral cancer: role of general practitioners *BMJ.* 1986; **293**:599-601.

Scully C. Clinical diagnostic methods for the detection of premaiignant and early malignant oral lesions. *Community Dent Health* 1993; **1O**: Suppl 1:43-52.

Scully C. Oral Precancer:Preventive and medical approaches to management. *Oral Oncol, Eur J Cancer.* 1995: Vol **31B** No.1:16-26

Shafer WG, Waldron CA. Erythroplakia of the oral cavity. *Cancer* 1975; **36**: 1021-8

Shah JP, Candela FC, Poddar AK. The patterns of cervical lymph node metastases from squamous carcinoma of the oral cavity. *Cancer* 1990 Jul 1; 66(1): 109-13

Shapiro S. Goals of screening. *Cancer* 1992; **70**: 1252-1258.

Shimakage M, Horii K, Tempaku A, Kakudo K, Shirasaka T, Sasagawa T. Association of Epstein-Barr virus with oral cancers. *Hum Pathol* 2002 Jun; **33**(6): 608-14

Siegel RS, Gartenhaus RB, Kuzel TM. Human T-cell lymphotropic-I-associated leukemia/lymphoma. *Curr Treat Options Oncol* 2001 Aug; **2**(4): 291-300

Silverman S, Bhargava K, Mani, NJ, Smith LW. Malaowalla AM. Malignant

transformation and natural history of oral leukoplakia in 57,518 industrial workers of Gujarat, India. *Cancer*, 1976, **38** :1790-1795.

Silverman S, Gorsky M, Lozada-Nur F. A prospective follow-up study of 570 patients with oral lichen planus: persistence, remission and malignant association. *Oral Surg Oral Med Oral Pathol* 1985; **60**:30-4.

Silverman S. Early diagnosis of oral cancer. *Cancer* 1988; **62**: 1796-9.

Slaughter DP, Southwick HW, Smejkal W. Field cancerisation in oral stratified squamous epithelieum. *Cancer* 1953; **6**: 963-8

Smith EM, Hoffman HT, Summersgill KS, Kirchner HL, Turek LP, Haugen TH. Human papillomavirus and risk of oral cancer. *Laryngoscope* 1998 Jul; **108**: 1098-103

Smukler AJ, Ratner L. Hepatitis viruses and hepatocellular carcinoma in HIV-infected patients. *Curr Opin Oncol* 2002 Sep; **14**(5): 538-42

Snow PB, Smith DS, Catalona WJ. Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study. *J Urol* 1994 Nov; **152** (5 Pt 2):1922

Somoza E, Somoza JR. A neural-network approach to predicting admission decisions in a psychiatric emergency room. *Med Decis Making* 1993; **13**:273-80.

Speight PM, Downer MC, Zakrzewska JM. Screening for Oral Cancer and Precancer. Oral

Oncology, *European Journal of Cancer* 1992; **28B**: 45-48.

Speight PM, Downer MC, Zakrzewska JM eds. 1993, Screening for oral cancer and precancer: Report of a UK working group. *Comm Dent Health*; **10**: (Suppl 1): 1-89.

Speight PM, Morgan PR. The natural history and pathology of oral cancer and precancer. *Comm Dent Health* 1993; **10**(Suppl.1): 31-41

Speight PM, Elliott A, Jullien JA, Downer MC, Zakzrewska JM. The use of artificial intelligence to identify people at risk of oral cancer and precancer. *Brit Dent J.* 1995; **179**:382-387.

Stell PM, McCormick MS. Cancer of the head and neck: are we doing any better. *Lancet* 1985; ii: 1127.

Stitch HF, Rosin MP, Hornby AP, Mathew B, Sankaranarayanan R, Nair MK. Remission of oral leucoplakias and micronuclei in tobacco/quid chewers treated with beta-carotene, and with beta-carotene plus vitamin A. *Int J Cancer* 1988 **421**:195-9

Sugerman PB. The role of oncogenes, tumour supressor genes and growth factors in oral squamous cell carcinoma: a case of aptosis versus proliferation. *Oral Dis* 1995; **1**: 172-88

Sundstrum B, Mornstad H, Axell T. Oral carcinomas associated with snuff dipping. Some clinical and histological characteristics of 23 tumors in Swedish males. *Journal of Oral Pathology* 1982 Jun (3): 245-51.

Svedlow AJ, Marmot MG, Grundig AE, Head J. Cancer mortality in Indian and British ethnic immigrants from the Indian sub-continent to England and Wales. *Br J Cancer* 1995; 72: 1312-1319

Szabo G, Klenk G, Veer A. Correlation between the combination of alcohol consumption and smoking and oral cancer (screening of the population at risk) *Orvsoi Hetilap.* 1997 Dec; **138** (52): 3297-9.

Talamini R, Barzan L, Franceschi S, Caruso G, Gasparin A, Comoretto R. Determinants of compliance with an early detection programme for cancer of the head and neck in north eastern Italy. *European J Cancer Part B Oral Oncology* 1994; **30B**: 415-8.

Thomas DW, Seddon SV, Shepherd JP. Systemic immunosuppression and oral malignancy: a report of a case and review of the literature. *Br J Oral Maxillofac Surg* 1993 Dec;**31** (6):391-3

Thomas S, Kearsley J. Betel nut and oral cancer: a review. *Eur J Cancer Oral Oncol* 1993; **29B**; 251-5.

Tickle M, Williams M, Jenner T, Blinkhorn A. Sentinel practices in dentistry: a preliminary evaluation. Prim Dent Care. 2000 Jan; 7(1): 39-43.

Tominaga S, Kuroishi T, Aoki K. Cancer mortality statistics in 33 countries 1953-1992. 1998 UICC Prevention and Epidemiology Program. Nagoya. International Union Against

Cancer.

Townsend P, Phillmore P, Beattie A. Health and Deprivation. Inequality and the North London: Croon Holm. 1988.

Van der Waal,, Schepman KP, van der Meij EH. Smeele LE Oral Leukoplakia: a Clinicopathological Review. *Oral Oncology* 1997 Vol. **33**, No. 5, pp. 291-301.

Van Wyk CW, Stander I, Padayachee A, Grobler-Rabie AF. The areca nut chewing habit and oral squamous cell carcinoma in South African Indians. *S Afr Med J* 1993; **31**:517-21.

Vedtofte P, Holmstrup P, Hjôorting-Hansen, Pindborg JJ. Surgical treatment of premalignant lesions of the oral mucosa. *Int J Oral Maxillofac. Surg* 1987; **16**: 656-664

Vigild M. Oral Mucosal lesions among the institutionalised elderly in Denmark. *Dent Oral Epidemiol* 1987; **15**:309-13.

Warnakulasuriya S, Pindborg J. Reliability of oral cancer and precancer by primary health care workers in Sri Lanka. *Community Dental Health* 1990; **7**: 73-79

Warnakulasuriya KA, Johnson NW. Dentists and oral cancer prevention in the UK: opinions, attitudes and practices to screening for mucosal lesions and to counselling patients on tobacco and alcohol use: baseline data from 1991. *Oral Dis* 1999 Jan; **5**:10-4

Warnakulasuriya KA, Johnson NW, Linklater KM, Bell J. Cancer of the mouth, pharynx and nasopharynx in Asian and Chinese immigrants resident in Thames regions. *Oral Oncol* 1999Sept; **35** (5); 471-5

Warnakulasuriya KA, Harris CK, Scarrott DM, Watt R, Gelbier S, Peters TJ, Johnson NW.An alarming lack of public awareness towards oral cancer. *Br Dent J* 1999 Sep; **187** (6): 319-322.

Warnakulasuriya KAAS, Ekanayake AN, Sivayoham S, Stjernsward J, Pindborg JJ, Sobin LH, Perera KS. Utilisation of primary health care workers for early detection of oral cancer and precancer cases in Sri Lanka. *Bulletin of the World Health Organisation* 1984; **62**:243-250.

Warnakulasuriya KAAS, Johnson NW. Strengths and weaknesses of screening programmes for oral malignancies and potentially malignant lesions. European *Journal of Cancer Prevention* 1996; **5**: 93-98.

Warnakulasuriya KAAS Johnson NW. Sensitivity and specificity of OraScan toluidine blue mouthrinse in the detection of oral cancer and precancer. *J Oral Pathol Med* 1996; **25**: 97-103.

Warnakulasuriya KAAS, Nanayakkara, BG. Reproducibility of an oral cancer and precancer detection program using a primary health care model in Sri Lanka. *Cancer Detection and Prevention* 1991; **15**: 331-334.

Warnakulasuriya S. Analytical studies of the evaluation of the role of betel-quid in carcinogenesis. In: Betel-quid and tobacco smoking. Usage and health issues, chapter 6 London: *Transcultural Health*; 1995, 61-70.

Warnakulasuriya S. Global aspects of oral cancer; Reported screening programmes and outcomes. *CDP Dentistry* 2001; **2**(1): 5-8

Warnakulasuriya S. (Guest Editor), Biology, medical and socio-economic aspects of areca nut use: Proceedings of a symposium. *Addiction Biology, 2002* Jan: vol 7, No 1.

Warnakulasuriya S Effectiveness of tobacco counselling in the dental office. *J Dent Educ* 2002 Sep; **66**(9): 1079-87

Warren S. Gates O. Multiple primary malignant tumours: a survey of the literature and a statistical study. *Am J Cancer* 1932: **51**:1358-1403.

Watt RG, Johnson NW, Warnakulasuriya KA. Action on smoking--opportunities for the dental team. *Br Dent J* 2000 Oct **14**; 189(7): 357-60

Wiering M. Machine Learning Website 2002 Institute of Informaton and Computing Sciences, University of Utrecht.

Wight AJ, Ogden GR. Possible mechanisms by which alcohol may influence the development of oral cancer-a review. *Oral Oncology* 1998 Nov; **38** (6): 441-447.

Wilkinson CE, Peters TI, Stott NC, Harvey IM. Prospective evaluation of a risk scoring system for cervical neoplasia in primary care. *Brit J General Practice*1994; **44**: 341-4.

Williams C. Ovarian and cervical cancer. In: Mead GM, ed. Current Issues in Cancer. London: BMJ publishing group, 1992:42-53.

Wilson JMG, Junger G. Principles and practice of screening for disease. Public Health Papers no. 34. Geneva: World Health Organization, 1968.

Wilson E. On Lichen planus. *J Cutan Med. Dis. Skin* 1896; **3**:117

Winn DM, Blot WJ, McLaughlin JK *et al* Mouthwash use and oral conditions in the risk of oral and pharyngeal cancer. *Cancer Res* 1991; **51**:3044-47.

Winn DM, Diehl SR, Brown LM, Harty LC, Bravo-Otero E, Fraumeni JF Jr, Kleinman DV, Hayes RB. Mouthwash in the etiology of oral cancer in Puerto Rico. *Cancer Causes Control* 2001 Jun; **12**(5): 419-29

World Health Organisation-International Statistical Classification of Diseases and Related Health Problems ICD-10 Tenth Revision. with User's Guide, English 1995 World Health Organization Marketing and Dissemination CH-1211 Geneva 27 Switzerland

World Health Organisation- Application of the International Diseases classification to Dentistry and stomatology. Geneva WHO 2nd Edition. 1978.

World Health Organisation- Multicentric Cervical Study Group report 2002

Wynder EL, Hultberg S, Jaconsson F. Environmental factors in cancer of the upper alimentary tract: a swedish study with specific reference to Plummer Vinson syndrome. *Cancer* 1957; **10**:470-482.

Yamamoto K, Hayashi Y, Hanada R, Kikuchi A, Ichikawa M, Tanimura M, Yoshioka S. Mass screening and age-specific incidence of neuroblastoma in Saitama Prefecture, Japan. *J Clin Oncol* 1995 Aug; **13**(8): 2033-8

Zain PB, Ikeda N, Razak IA, Axell T, Majid ZA, Gupta PC, Yaacob M. A National epidemiological survey of oral mucosal lesions in Malaysia. *Community Dent Oral Epidemiol* 1997 Oct; **25**:377-83.

Zakrzewska JM, Hindle I, Speight PM. Practical considerations for the establishment of oral cancer screening programme. *Community Dental Health,* 1993; **10:** Supplement.

Appendix 1

Opportunistic cancer screening in general dental practice: results of a pilot study.

Eastman Dental Institute for Oral Health Care Sciences http://www.eastman.ucl.ac.uk

K Lim*, P Hammond, DR Moles, MC Downer,and PM Speight.

Department of Oral Pathology, Eastman Dental Institute, University College London. London, UK.

University College London

Introduction

Studies have shown that population screening for oral cancer may not be feasible in the UK because of its relatively low prevalence [1,2]. However, there is evidence of a rising incidence and mortality in Western industrialised countries [3] with the increase in the younger age groups of males and females. This is thought to be due to the effects of increased alcohol consumption and the synergistic, multiplicative effects of tobacco use.

The 5 year relative survival rate has remained at less than 50% with survivors likely to suffer devastating morbidity and loss of function.

The detection of potentially malignant lesions and conditions in the mouth is worthwhile because early intervention can lead to their resolution [4]. Unfortunately, there is a lack of public awareness of the disease and patients are slow to seek attention [5]. Over 60% of lesions present greater than 2 cms when the prognosis is significantly worse [6].

It is widely agreed that mass population screening for oral cancer is not a practical proposition in terms of health economics. Invitational screening has been studied but the response rate has been typically disappointing. In a recent pilot invitational study, the response was about 25% [7].

Opportunistic targeted screening in a primary care setting may be a viable option [8]. However, it has been suggested that patient populations in such a setting are not representative of the general population and those at high risk are less likely to attend and therefore many lesions will be missed.

Aim

The aim of this study was to determine the prevalence of relevant lesions in patients screened opportunistically in general dental practice. Data relating to age, gender, ethnicity, tobacco and alcohol consumption were also collected.

Materials and Methods

18 general dental practitioners took part in this pilot oral cancer screening programme. Each attended training sessions to standardise criteria for positive and negative lesions and conditions. The criteria for a positive screen were the presence of a red, white or red and white lesion, or of an ulcer that had not healed for more than 2 weeks. Lesions regarded as positive included leucoplakia, erythroplakia (and speckled variants), lichen planus, submucous fibrosis, actinic keratosis and squamous cell carcinoma or basal cell carcinoma or orofacial skin.

A patient information sheet for informed consent and 2 other forms were used. The first, a Health Questionnaire, was given to the patient to complete and related to personal details and tobacco and alcohol consumption.

The second form was for the dentist to record the presence or absence of lesions. This was filled in by the dentist and both forms were collected and the information collated at the Eastman Dental Institute.

The dentists were blinded to the Health Questionnaire filled in by the patients.

Results

Table 1. Shows a profile of the population studied according to age, gender, smoking and drinking habits, ethnicity and regularity of dental attendance. Non-smokers were defined as those who did not smoke or have not smoked for the last 10 years. Moderate smokers smoked less than 20 cigarettes a day and heavy smokers smoked more than 20 cigarettes a day.Non drinkers were defined as those who did not drink alcohol or who drank less than 5 units a week. Moderate drinkers, if female, were those who drank between 5 to 14 units a week. If male, 5 to 21 units a week. Heavy drinkers were considered as females who drank 15 or more units a week and males who drank 22 or more units a week. Regular attendance at the dentist was taken as a surrogate measure of the level of dental motivation and oral care. Table 2. Shows the prevalence of lesions and conditions reported in this study determined as being positive or negative.

Table 1. Profile of dataset

Age	40s	613
	50s	605
	60s	443
	70s	273
	80s	88
	90+s	7
Gender	Male	963
	Female	1066
Non smoker		1480
Moderate smoker		397
Heavy smoker		147
Non or light drinker		627
Moderate drinker		1466
Heavy drinker		188
Regular Dental Attender		1774
Ethnic Grouping	Asian	101
	Black	135
	Chinese	27
	Other	44
	White	1721

Table2. Lesions determined as positive — **No of patients**

Carcinoma	2
White patch	45
Red patch	11
Lichen planus	29
Ulcer	2
Submucous fibrosis	1
Actinic keratosis	2
	92

Lesions determined as negative

Non-specific or traumatic ulcers	39
Frictional/traumatic keratosis	30
Fibrous overgrowths	28
Candida/denture stomatitis	14
Smokers keratosis (palate)	12
Aphthous ulcers	10
Amalgam tattoo	9
'Haemangioma'	7
Angular cheilitis	5
Abscess/sinus`	5
Mucoceles	4
Geographic tongue	3
Naevus	3
Miscellaneous	10
	179
Total patients with lesions	271

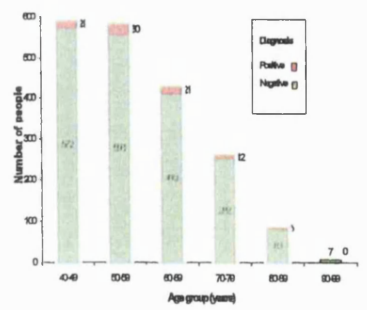
Table 3. is a regression analysis table highlighting gender and risk habit associations with the prevalence of positive lesions. The rate ratio (IRR) of each is shown with 95% Confidence Intervals. Figure1. A Bar chart demonstrating the occurrence of positive lesions in the various age groups.

Table 3. Factors associated with the presence of oral cancer and pre-malignant lesions

Factors	IRR (95% CI)	P
Gender		
Male	1.92 (1.25, 2.95)	0.003
Female	1*	
Smoking in females		
Heavy smoker	4.79 (1.80, 12.77)	0.002
Moderate smoker	2.05 (0.90, 4.66)	0.086
Non-smoker	1*	
Smoking in males		
Heavy smoker	4.69 (2.67, 8.22)	<0.001
Moderate smoker	0.63 (0.26, 1.50)	0.293
Non-smoker	1*	
Drinking in females		
Heavy drinker	2.53 (0.83, 10.12)	0.189
Moderate drinker	2.83 (1.08, 6.45)	0.034
Non- or light drinker	1*	
Drinking in males		
Heavy drinker	3.97 (1.74, 9.06)	0.001
Moderate drinker	1.23 (0.61, 2.50)	0.560
Non- or light drinker	1*	

* = baseline group for each comparison

Figure 1: Profile of the sample by age group and diagnosis for people aged 40 and over



Summary

The overall prevalence of positive lesions was 4.17%. Males were more likely to have a positive lesion with an IRR of 1.92 [95% CI 1.25, 2.95]. Heavy smokers were 4 times more likely to have a positive lesion (Females: IRR 4.79 [95% CI 1.80, 12.77]; Males: IRR 4.69 [95% CI 2.67, 8.22]). Heavy drinkers were 2 to 3 times more likely to have a positive lesion than non or light drinkers (Females, IRR 2.53 [95%CI 0.63, 10.12]; Males IRR 3.97 [95%CI 1.74, 9.06]).

Conclusions

The results of this study show a prevalence of lesions and habits comparable to similar studies in an industrial setting [9] showing a 5% prevalence of positive lesions but greater than one done in a hospital and medical practice setting [10] which showed a prevalence of 2.7%. This compares well with other international studies done in Sri Lanka, giving a prevalence of 6.2%, USA (4.2%) and Hungary (3.7%).

It has been suggested that screening programmes for oral cancer and precancer in general dental practice may not be useful as the groups considered to be most at risk, that is, the lower socio-economic groups, heavy users of tobacco and alcohol and the elderly are less likely to attend the dentist on a regular basis. In this pilot study, these high risk groups were not specifically targeted. Indeed, that would negate the purpose of this study.

However, it was encouraging to observe a prevalence rate that compares well and is consistent with known rates in the UK and other parts of the world. This suggests that the population attending general dental practice, usually regarded as self selecting and therefore biased, are representative of the general population and supports the concept that opportunistic screening in a primary dental care setting may be feasible. The added advantage is the opportunity to do repeated screening of the same patients as they present for their routine dental check-ups.

References

1. Warnakulasuriya KA., Johnson NW. Strengths and weaknesses of screening programmes for oral malignancies and potentially malignant lesions. European Journal of Cancer Prevention, 1996 5(2): 93-8.
2. Boyle,P., Macfarlane,G.J., Blot, W.J., *et al.* European School of Oncology advisory report to the European Commission for the Europe against cancer Programme: Oral carcinogenesis in Europe. Oral Oncology, European Journal of Cancer, 1995 31B: 75-86.
3. Hindle, I., Nally, F. Oral cancer; a comparative study between 1962-67, and 1980-84 in England and Wales. British Dental Journal, 1991 170: 15-20.
4. Gupta, P.C., *et al.* Effect of cessation of tobacco use on the incidence of oral mucosal lesions in a 10-year follow up study of 12,212 users. 1995 Oral Diseases, 1: 54-8.
5. Warnakulasuriya, K.A., *et al.* An alarming lack of public awareness towards oral cancer British Dental Journal, 1999 187(6): 319-22.
6. Speight,P.M., Morgan,P.R. The natural history and pathology of oral cancer and precancer. Community Dental Health, 1993 10(Supplement 1): 31-41.
7. Julien, J.A., *et al.* Attendance and compliance at an oral cancer screening programme in a general medical practice. European Journal of Cancer Part B, Oral Oncology, 1995 31B(3): 202-6.
8. Speight ,P.M., *et al.* Screening for Oral Cancer and Precancer. Oral Oncology, European Journal of Cancer, 1992 28B(1): 45-48.
9. Downer, M.C., *et al.* Evaluation of screening for oral cancer and precancer in a company headquarters. Community Dental Health, 1995 23: 84-88.
10. Julien, J.A., *et al.* Evaluation of a screening test for the early detection of oral cancer and precancer. Community Dental Health, 1995 12: 3-7.

Eastman Dental Institute

For Oral Health Care Sciences
http://www.eastman.ucl.ac.uk

Optimising artificial intelligence for the identification of individuals at high-risk of oral cancer.

Ken Lim , Peter Hammond, Martin C Downer, Paul M Speight,
Department of Oral Pathology, Eastman Dental Institute ,University of London, UK

UCL
Affiliate of University College London

Introduction

The incidence of oral cancer in the UK is 4.5/100,000 or equivalent to about 2,600 cases a year. The 5 year survival rate is low at 40%. There has been no improvement in survival over the decades and recent studies show that the number of cases is actually increasing.
The risk of having advanced disease does not necessarily depend on patient or professional delay, since symptomology appears to be independent of tumour stage.

Although small lesions of oral cancer and precancerous lesions are easy to detect, our previous studies have shown that attendance at oral cancer screening programmes is low, making population screening non-viable. Targeting of high-risk groups for case finding and primary prevention is generally seen as beneficial. We have shown how computerised artificial intelligence can be used to identify a high risk subset of the population.

Aims

The purpose of this study was to determine the optimal parameters for training artificial intelligence models for the identification of individuals at high-risk of oral cancer

Materials and Methods

The basic dataset contained information on the smoking and drinking habits of 2024 individuals of whom 97 were known to have oral cancer or pre-cancer. A test set was separated from the data, the rest was used for training.

Training sets consisted of 64 positives and various numbers of negatives to give negative to positive ratios from 0.25:1 to 10:1.

The test set consisted of 309 negative and 16 positive cases.

A commercially available data-mining package (Clementine®) was used.

"Data mining" refers to the use of a variety of techniques to identify decision making knowledge in bodies of data. These can then be extracted for use in areas such as decision support, prediction, forecasting and estimation.

Clementine has basically two machine learning models, neural networks and rule induction. Also facilities for data manipulation and visualisation of data.

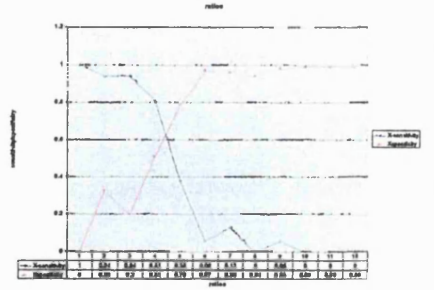
A total of 36 training sets were used to train several neural networks, and C5 algorithms (decision trees). For all models, the best performance was obtained with negative to positive ratios of 1:1 or 2:1. Further models were generated using these ratios. Further sets within these ratios were generated and the learned models produced were then tested singly and in combinations against the test set for accuracy.

Results

Three separate tables were generated. One using exact numbers of negative cases to positives, one using an approximate ratio as determined by a random selector function and a third using unsupervised Kohonen network clustering to eliminate any possible hidden bias within the data. Below is the table for the approximate ratios. Each learnt model was given an identifier. The various models were also linked in series to from a combined learning unit consisting of a neural network and a C5 algorithm.

aprox ratios	0.25:1	0.5:1	1:01	2;1	3:01	4:01	5:01	6:01	7:01	8:01	9:01	10:01
sample size to64pos	17	36	72	126	191	282	321	394	463	514	587	668
C5	cm12	cm7	cm6	cm1	cm2	cm3	cm4	cm5	cm8	cm9	cm10	cm11
C5-n score/309	0	139	77	245	266	301	298	290	302	305	307	307
C5-p score/16	16	15	15	7	5	1	2	0	1	0	0	0
C5sensitivity	1	0.94	0.94	0.44	0.31	0.06	0.13	0.94	0.98	0.99	0.99	0.99
C5specificity												
NN	nm12	nm7	nm6	nm1	nm2	nm3	nm4	nm5	nm8	nm9	nm10	nm11
NN-n score/309	0	175	184	184	261	309	309	309	309	309	309	309
NN-p score/16	16	8	10	10	2	0	0	0	0	0	0	0
NNsensitivity	1	0.5	0.63	0.63	0.13	0	0	0	0	0	0	0
NNspecificity	0	0.57	0.6	0.6	0.85	1	1	1	1	1	1	1
X-n score/309	0	103	61	158	244	301	298	290	302	305	307	307
X-p score/16	16	15	15	13	6	1	2	0	1	0	0	0
X-sensitivity	1	0.94	0.94	0.81	0.38	0.06	0.13	0	0.06	0	0	0
Xspecificity	0	0.33	0.2	0.51	0.79	0.97	0.96	0.94	0.98	0.99	0.99	0.99

Below is a graphic representation of the relationship between sensitivity and specificity over the various ratios.
Also a screenshot of a data manipulation stream in Clementine ®



A table of combinations of well performing models against the test set.

	no3	oo2	oo4	oo7	oo11	om1
n-score/309		110	174	186	178	188
p-score/16		12	13	13	12	12
sensitivity		0.75	0.81	0.81	0.75	0.75
specificity		0.36	0.56	0.6	0.58	0.61
	no7+	oo2	oo4	oo7	oo11	om1
n-score/309		91	137	142	128	143
p-score/16		12	11	11	11	12
sensitivity		0.75	0.69	0.69	0.69	0.75
specificity		0.29	0.44	0.46	0.41	0.46
	no11	oo2	oo4	oo7	oo11	om1
n-score/309		113	190	197	185	197
p-score/16		13	13	13	13	14
sensitivity		0.81	0.81	0.81	0.81	0.88
specificity		0.37	0.62	0.64	0.6	0.64
	nm1	oo2	oo4	oo7	oo11	om1
n-score/309		101	152	158	149	158
p-score/16		13	14	12	14	13
sensitivity		0.81	0.88	0.75	0.88	0.81
specificity		0.33	0.49	0.51	0.48	0.51

Summary

Two machine learning models, a neural network and a C5 algorithm were trained on life-style risk habits of a varying number of negative and positive individuals. The models were then tested on their ability to predict positive and negative cases in a separate unseen test set.
The results were expressed as sensitivity and specificity.

It was found that the machine learning models trained best in the proportion of 1:1 and 2:1 , negatives to positives. They also tended to perform better when linked in combination.

It was also possible to train to a higher level of accuracy but at the expense of a significantly raised number of false negatives.

The sensitivity ranged from 0.73 to 0.81 and the specificity from 0.53 to 0.63. An optimum performance of 0.8 and 0.6 was obtained.

Conclusions

The results show that this artificial intelligence programme can be trained to identify 80% of subjects with cancer or precancer. There were 40% false positives but this may be desirable for a programme designed as a "pre-screen" filter whose purpose is to select high risk subjects for detailed clinical examinations.
Further studies will test the performance prospectively in general dental practice.

References

Julien JA *et al.* Comm Dent Health 1995;**12**,3-7
Julien JA *et al.* Oral Oncol, 1995,**31B**,202
Speight *et al* BDJ,1995 **179**,382.
Speight PM, Morgan PR, Comm Dent Health 1993;**10** (Suppl)31-41.
Hindle I, Nally F. BDJ 1991;**170**:15
Hindle I, Speight PM. Downer MC. Lancet 1994;**343**,178
Warnakulasurya WAAS,Johnson NW, European Jn of Cancer Prevention 1996,**5**,93-98
LaVecchia *et al* Oral Oncology,1997 **Vol 33**, No5, 302-307

4630

Form A

Patient Number ☐ Dentist Number ☐

Health questionnaire

Please answer a number of simple questions about yourself and about your smoking and drinking habits. Your answers will help us in a study comparing the use of tobacco and alcohol to oral health. Your answers will be regarded as confidential and you will only be identified by the anonymous number at the top of this sheet. Completion of this form will not affect your treatment in any way.
Please circle the correct response, or tick the boxes.

1. Age ☐

2. Sex: Male ☐ Female ☐

3. Do you smoke cigarettes, a pipe or cigars? Yes ☐ No ☐

4. If No to 3, have you smoked in the last 10 years? Yes ☐ No ☐

5. If you smoke or have smoked in the last 10 years please indicate the amount:

   Less than 20 cigarettes, pipes or cigars per day ☐

   20 or more cigarettes, pipes or cigars per day ☐

6. Do you drink alcohol? Yes ☐ No ☐

7. If you drink alcohol, please indicate the amount (please see bottom of page):

   14 or fewer units per week ☐

   Between 15 and 21 units per week ☐

   Between 22 and 28 units per week ☐

   29 or more units per week ☐

8. Have you you visited a dentist in the last 12 months : Yes ☐ No ☐

9. Please indicate your ethnic group : ☐

   White ☐ Black ☐ Asian ☐ Chinese ☐ Other ☐

Units of alcohol are used as a rough guide to determine the amount we drink.
1 unit of alcohol = half a pint of beer or lager;
or = one glass of wine
or = one measure of spirits or port or sherry

Office use only

Study Number 4630

Form B

Soft tissue examination form

Patient Number _____ Dentist Number _____

Record the presence of soft tissue lesions in the table below. If there is no abnormality tick the first column for each site. If a lesion is present, indicate its code letter in the appropriate box.

Site	No Abnormality (Tick box)	Benign Lesion **U** for Ulcer **F** for Fibrous overgrowth **O** for Other	Suspect Lesion **W** for White patch **R** for Red patch **U** for Ulcer **O** for Other	Suspect Condition **L** for Lichen planus **S** for submucous fibrosis **A** for Actinic lesion **O** for Other
Lips and commisures				
Buccal mucosa				
Tongue, dorsum				
Tongue, lateral				
Floor of mouth/ventral tongue				
Gingiva				
Lower alveolus/retromolar				
Hard palate/upper alveolus				
Soft palate/fauces				
Buccal and lingual sulci				

If **O**ther, specify the nature of the lesion _____

_____

_____

_____

This form is to be returned to the Eastman Dental Institute with the Habits Questionnaire Form (Form A) for the same patient. Please ensure that the patient number is entered at the top of the form and that the numbers on Forms A and B match

The evaluation of artificial intelligence to identify individuals at high risk of oral cancer and precancer in a general dental practice

Patient Information Sheet

The mouth, like anywhere in the body is at risk of cancer. Oral cancer in the UK however, is still comparatively rare.

It is nonetheless a potentially serious condition, which can be disfiguring or even fatal. Dental Surgeons can very easily identify oral cancer and in its early stages is very easily cured and controlled.

Because of its rarity however, it is uneconomic to run large scale screening programmes such as in those for Breast and Cervical cancer. A better way to try to detect the disease would be to identify people who are at risk and then ask them to have their mouths examined. People at most risk are those who smoke cigarettes and drink alcohol.

The Eastman Dental Institute in London has recently developed a Computer based program that was proven effective in identifying individuals who could be at high risk of getting oral cancer. This is done by asking a few simple questions on life-style habits such as smoking and alcohol consumption.

The Computer program would then work out a risk factor and only those at high risk could be screened and supervised more closely.

The purpose of this study is to obtain information from people who go to the dentist so that we can evaluate the computer program and find out if it really works.

Your Dentist will ask you to fill in a Health Questionnaire, and your Dentist will also complete a form after he or she has examined your mouth during the course of your normal treatment. The Health Questionnaire filled in by you would be sent to the research team at the Eastman Dental Hospital for evaluation. Because the information is anonymous, you will not receive any results from this study. Your treatment with your Dentist will continue in the normal way. The information you provide will be used to help others in the future.

If you do not wish to participate in this study you do not have to – this will not affect your treatment in any way. If you would like further information you may ask your dentist, or you can contact the head of the research team at the Eastman Dental Institute- Professor Paul Speight, Tel no. 0207 915 1055

Thank you for your help

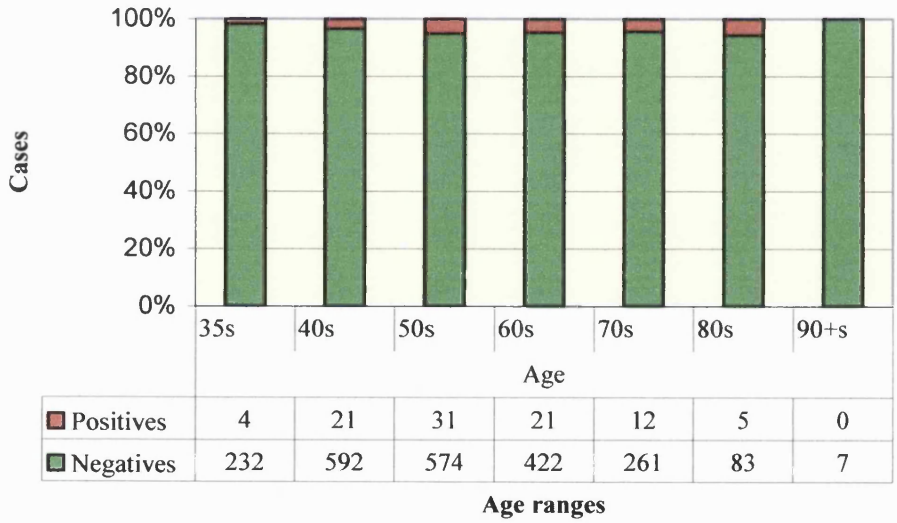
Appendix 2

Appendix 2

Chapter 6 Histograms and bar charts

Figures 6.1-6.6 present this data in the form of a histogram for age bands and bar charts for gender, smoking, alcohol consumption, dental attendance and ethnicity. These show the numbers and percentages of individuals with negative or positive screens within each group and also their distributions within the increasing scales in the smoking and alcohol consumption categories.

Fig 6.1 Negative and positive cases within the age bands

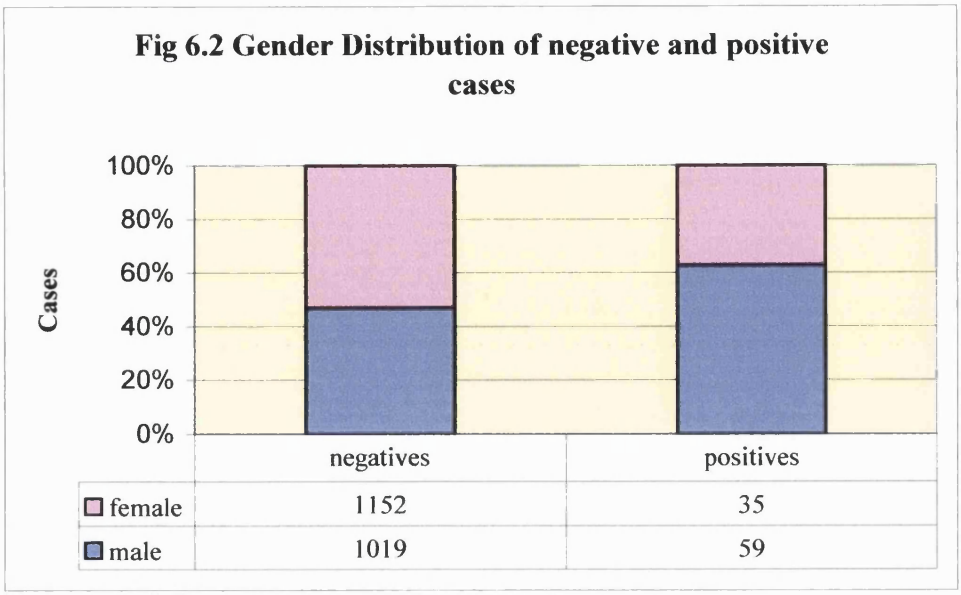


	35s	40s	50s	60s	70s	80s	90+s
Positives	4	21	31	21	12	5	0
Negatives	232	592	574	422	261	83	7

Age ranges

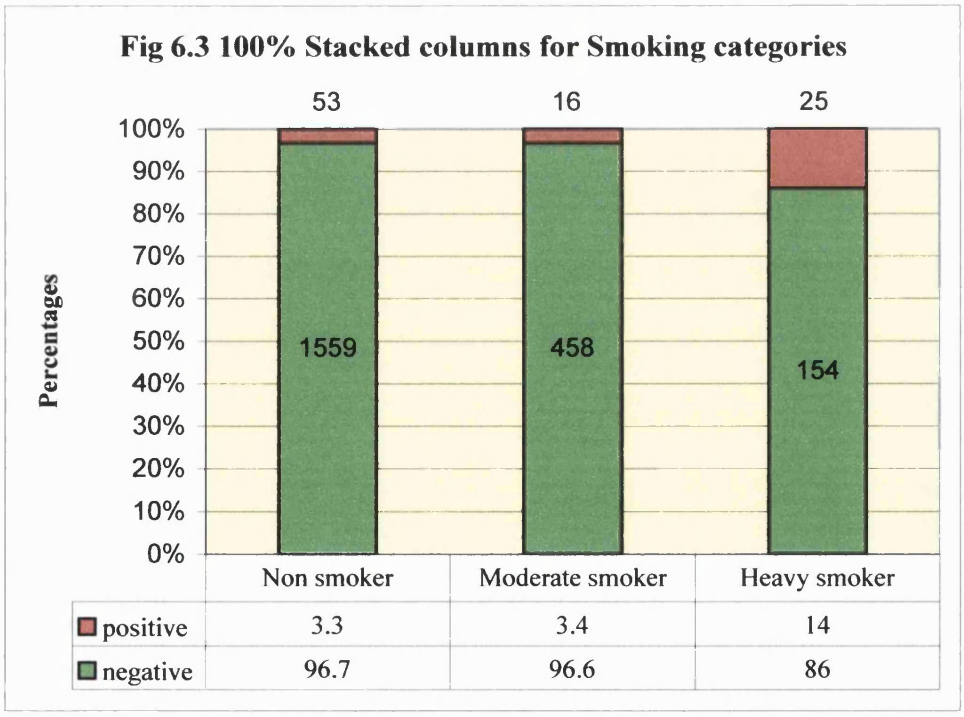
chi-squared = 6.55, P = 0.26.

Figure 6.1 The age distribution of the patients examined set in age bands with the number of positive cases found for each age band shown in red at the top end of the histogram.



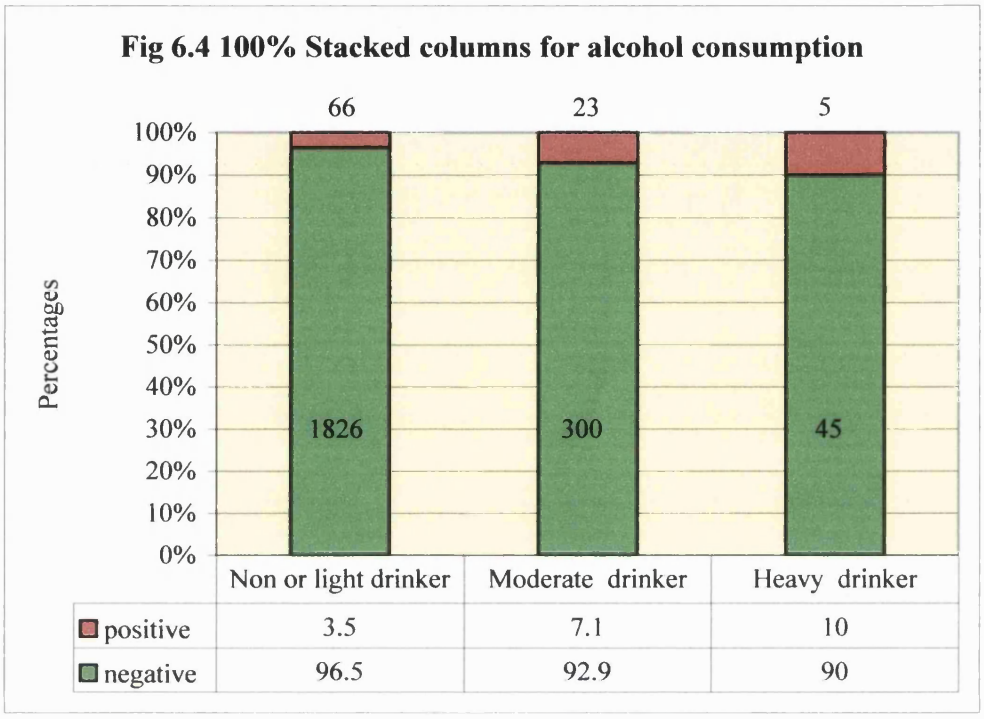
chi-squared = 8.43, P = 0.004.

Figure 6.2 The gender distribution of positive and negative cases with number and percentage shown in 100% Stacked columns.



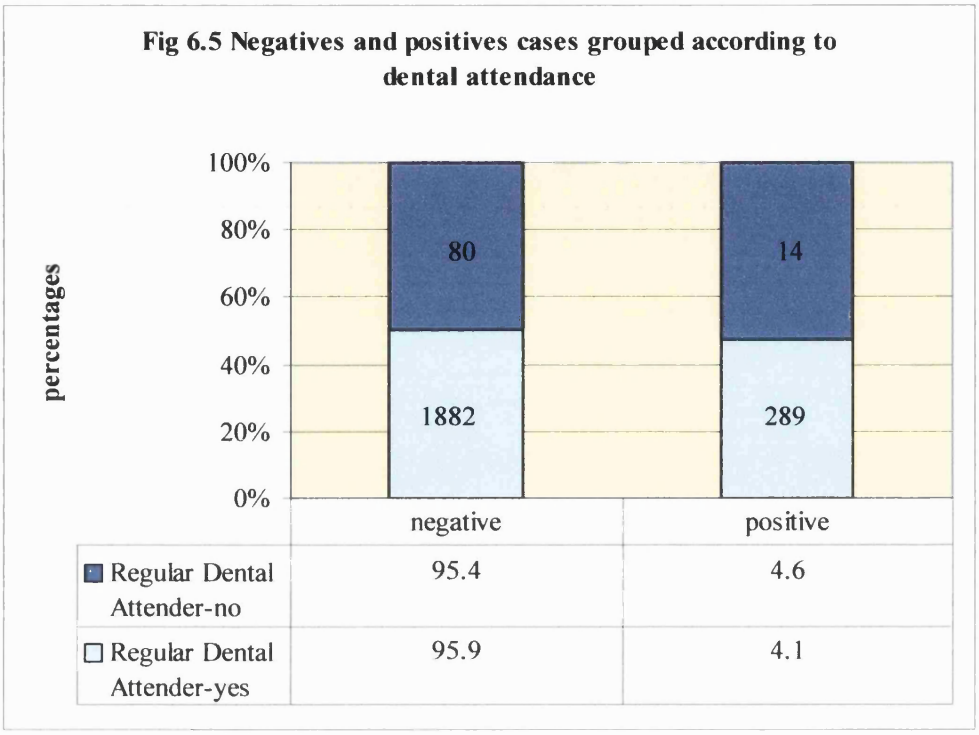
chi-squared = 47.09, P = <0.001.

Figure 6.3 Three 100% stacked columns representing non-, moderate and heavy smoker categories and the distribution of positive and negative cases within each category, their respective numbers and percentages. Percentages for negative and positive cases are shown in the legend and numbers in the columns.



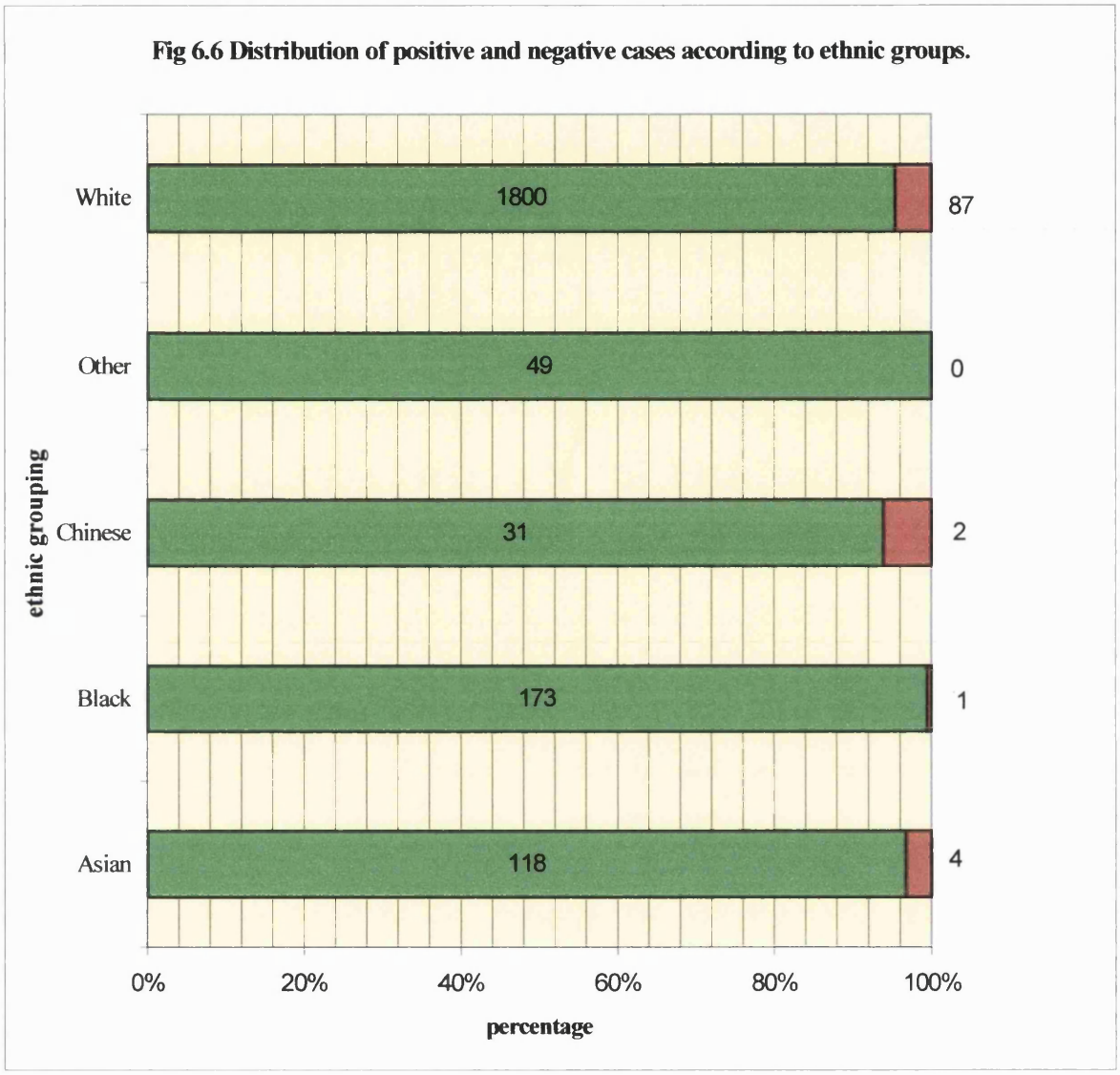
chi-squared = 13.55, P = 0.001.

Figure 6.4 Three 100% stacked columns representing proportions of negative and positive cases for non-, moderate and heavy alcohol consumption. The legend shows the percentage of cases. The number of cases is shown in the column for the negative cases and above the column for the positive cases.



chi-squared = 0.08, P = 0.78.

Figure 6.5 100% stacked columns for negative and positive cases in dental attendees. The legend shows the percentage. The number of cases is superimposed on the columns.



chi-squared=9.25, P=0.10

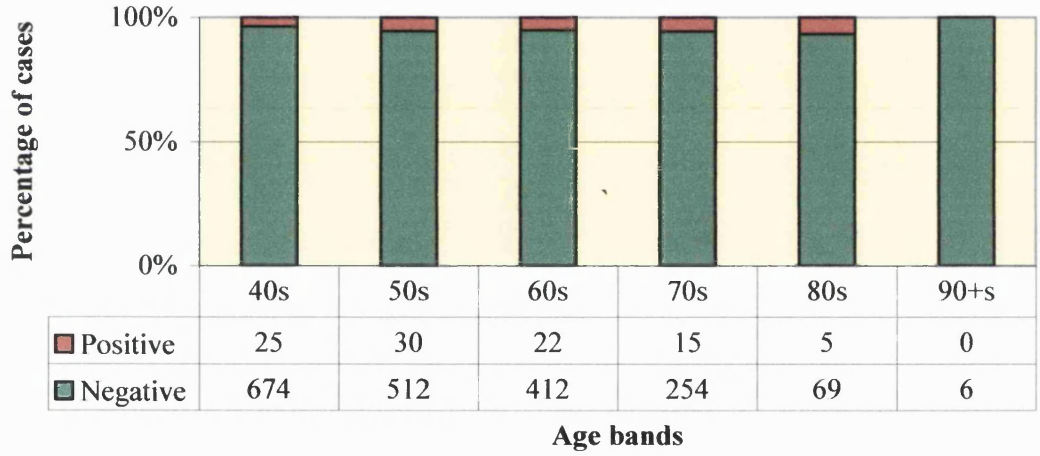
Figure 6.6 100% stacked bars showing the distribution of the positive and negative cases within the different ethnic groups screened. Green represents the negatives and red the positives. The percentages of all cases are on the x –axis. The number of cases is superimposed on the bars for the negatives and immediately adjacent the bars for the positives

Chapter 8 Histograms and bar charts

8.1 Age groups within the two sets

There was an interesting similarity in the numbers of cases in the various age groups in the two datasets. Figures 8.1and 8.2 show 100% stacked bar charts of the various age groups and the numbers of cases within the different age bands for the two datasets. It can be observed that both the numbers of individuals in the respective age bands and the distribution of negative and positive cases are very similar.

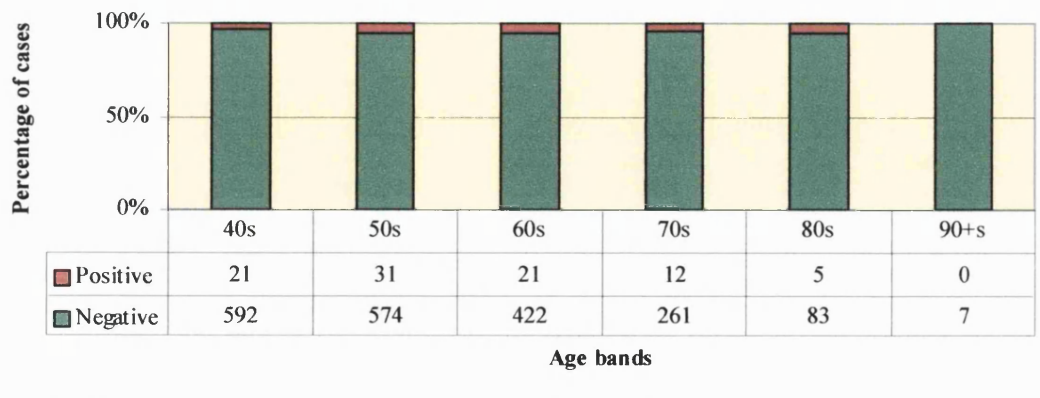
Fig 8.1 Age distribution in the 1995 set and prevalence of positive cases



	40s	50s	60s	70s	80s	90+s
Positive	25	30	22	15	5	0
Negative	674	512	412	254	69	6

chi squared = 3.76, P = 0.044.

Fig 8.2 Age distribution in the 2000 set and prevalence of positive cases

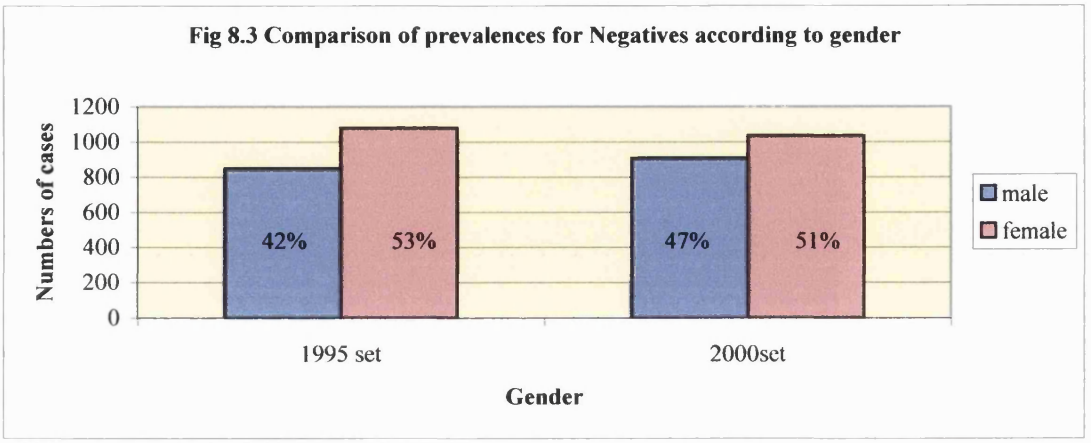


	40s	50s	60s	70s	80s	90+s
Positive	21	31	21	12	5	0
Negative	592	574	422	261	83	7

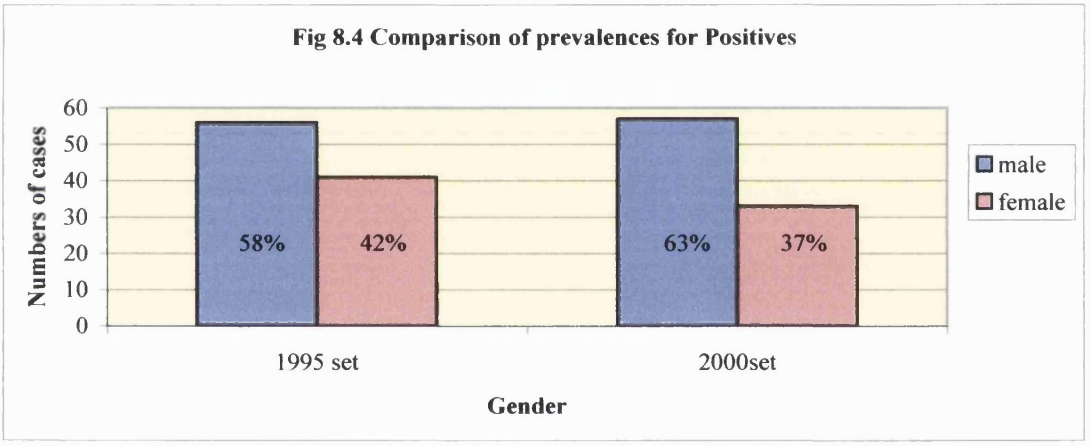
chi squared = 2.40, P = 0.66.

Figures 8.1 and 8.2 are histograms showing the negative and positive cases over the

different age bands for the 1995 and 2000 datasets respectively.

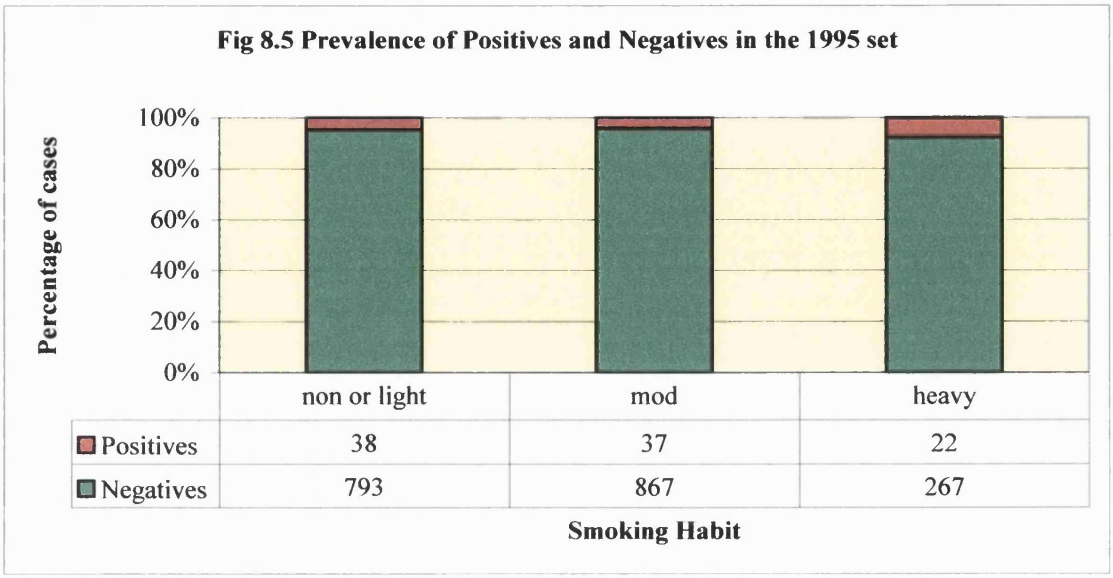


chi squared = 6.50, P = 0.01.

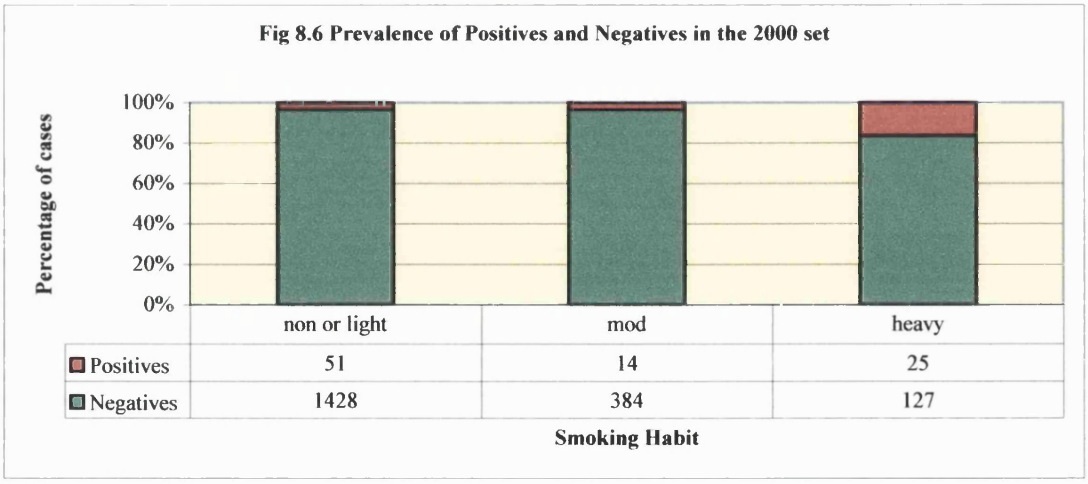


chi squared = 8.86, P = 0.003.

Figures 8.3 Bar charts that show the gender distribution for negatives and **Fig 8.4** for positives in the1995 and 2000 datasets.

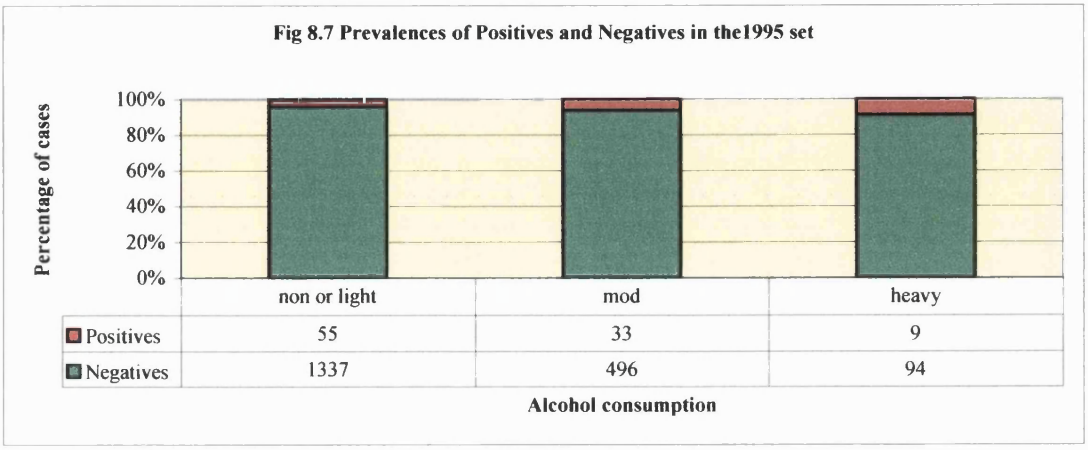


chi squared = 6.09, P = 0.047.

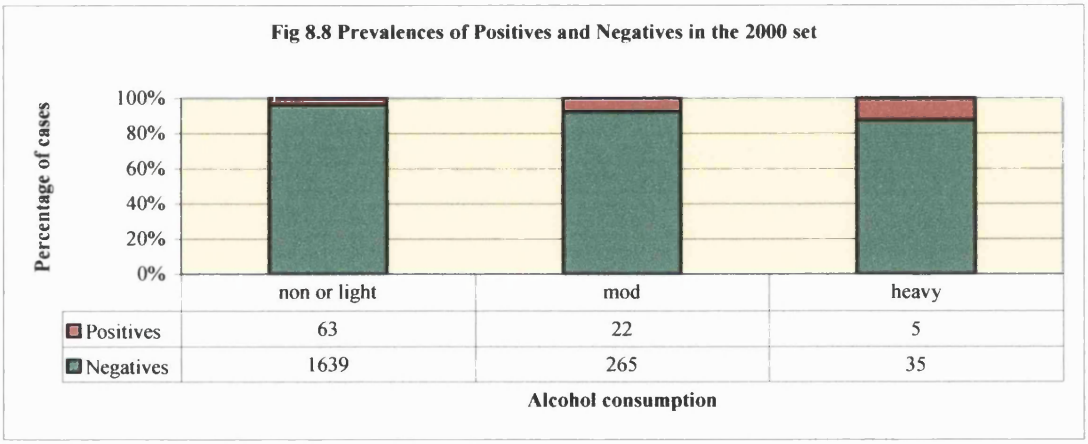


chi squared = 55.93, P = <0.001.

Fig 8.5 and **8.6** The distribution of positive and negative cases according to smoking habits over the different use categories in the 1995 and 2000 datasets respectively.



chi squared = 8.10, P = 0.02.



chi squared = 15.36, P = <0.001.

Figures 8.7 and **8. 8** The non-or light, moderate and heavy consumers of alcohol for both datasets both in numbers of cases and in percentages also expressed as 100% stacked columns. It can be observed that increasing levels of consumption resulted in an increase of positive cases.

Dental attendance is included here for the sake of completion. This field was not used as the 2000 series was collected in a general dental practice environment and as such most of the subjects would have been regular attendees and this would have caused a bias toward dental attendees.

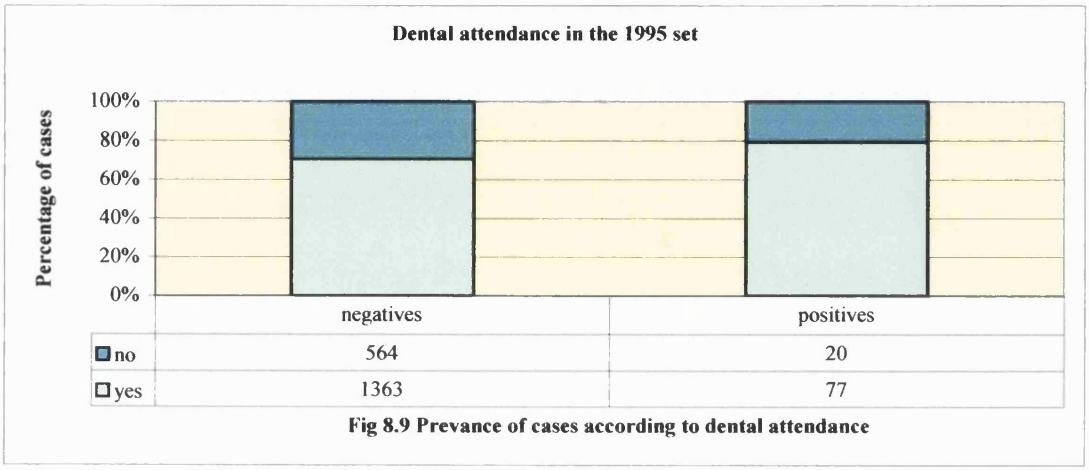


Fig 8.9 Prevance of cases according to dental attendance

chi squared = 2.96, P = 0.09

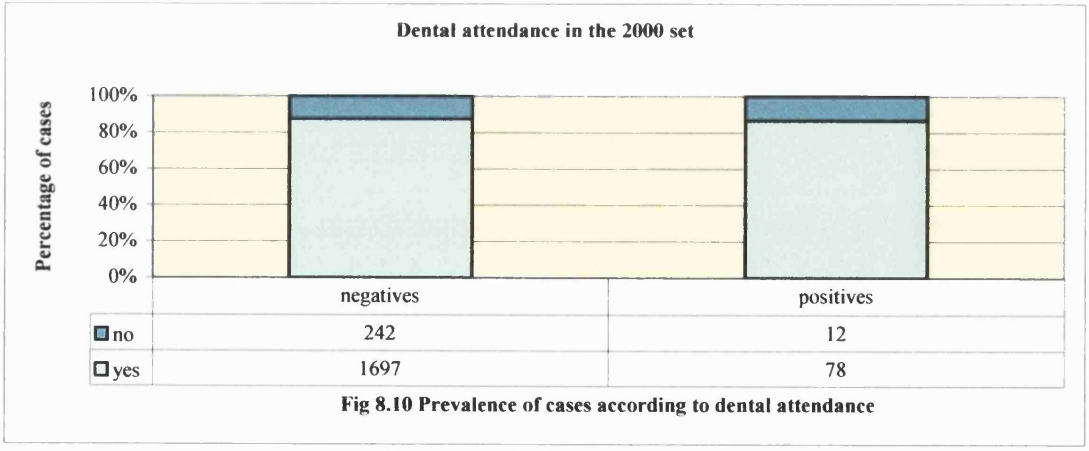


Fig 8.10 Prevalence of cases according to dental attendance

chi squared = 0.01, P = 0.94

Dental attendance was the last field examined, here; this is shown as 100% stacked columns in **Figures 8.9** for the 1995 and **8.10** for the 2000 datasets with the percentages on the-axis and the numbers of cases grouped in the legend.

Publications

RESEARCH SUMMARY

Has the time come for opportunistic oral cancer screening?

Opportunistic screening for oral cancer and precancer in general dental practice: results of a demonstration study
K. Lim, D. R. Moles, M. C. Downer and P. M. Speight Br Dent J 2003; 194: 497–502

Objectives
To demonstrate the feasibility of opportunistic oral cancer and precancer screening in general dental practice and to determine the prevalence of relevant lesions and risk habits in a population of general dental practice attenders.

Design, setting and methods
A prospective demonstration study, recruiting patients opportunistically in general dental practices. Eighteen general dental practitioners took part in this study. Each attended training sessions to be advised of the study protocol and the criteria of a positive and negative screen. Patients over the age of 35 years were prospectively and opportunistically recruited. Each patient was asked to complete a health questionnaire concerning age, gender, ethnicity, smoking and drinking habits. The dentist then examined the soft tissues and recorded the presence or absence of lesions independently on a second form. The forms were collated and data were analysed to determine prevalence of lesions and associations with risk habits.

Results
Data on 2,265 patients were available for analysis. Oral lesions were detected in 319 patients (14.1%). Ninety-four patients (4.2%) had lesions considered to be either malignant or potentially malignant. There was a significant association between positive lesions and male gender (IRR 1.86, 95% CI 1.22-2.82), heavy smoking (males: IRR 3.68, 95% CI 2.10-6.43: female; IRR 3.58, 95% CI 1.35–9.50) and heavy alcohol use in males (IRR 2.98, 95% CI 1.06–3.47).

Conclusions
The results suggest that patients attending general dental practices are representative of the general population both in terms of lesion prevalence and high risk habits such as smoking and drinking. This supports the view that opportunistic screening in a general dental practice setting may be a realistic alternative to population screening. Further research is needed to determine the cost effectiveness of this approach and to investigate the value of targeting high risk groups within this population. General dental practice is ideal for the evaluation of such systems prior to extending these studies to other healthcare settings.

IN BRIEF

- Oral cancer is an important disease that is easily identified in a dental practice setting.
- This study shows that GDPs can easily detect relevant lesions, using a simple systematic mucosal examination, with little disruption to a normal practice routine. The prevalence of relevant lesions and of risk habits among patients who attend general dental practices appears to be representative of the general population.
- The data suggest that opportunistic screening in a general dental practice setting, particularly if high risk groups can be targeted, might be a realistic option.

COMMENT
Of all oral diseases, oral cancers are the most life threatening. Further, the primary risk factors for these cancers — use of tobacco and alcohol products — have been known for decades. Yet dental schools have placed little emphasis on oral cancer prevention and early detection, especially compared with other content areas such as restorative and prosthetic dentistry. Not surprisingly, general dentists have essentially ignored detection of these cancers. Equally unsurprising, most oral cancers are detected at late stage and, as a result, five-year survival rates are among the lowest for all major cancers. Of all health providers, general dentists are the most logical group to screen for these cancers. But some argue that those persons at highest risk for oral cancers do not seek dental services. Moreover, the evidence to demonstrate the benefits of early cancer detection has been scant. As a result, few leading authorities suggest population screening for oral cancers. Others have suggested screening for oral cancers opportunistically especially when high-risk groups can be identified. But, here too, little evidence has been available to support the recommendation.

The study by Lim *et al.* is important because it demonstrates that opportunistic screening for oral cancers by general dentists is feasible and worthwhile. Here, opportunistic screening means examining patients who are at high risk for oral cancer because of risky behaviours, for soft tissue lesions. Moreover, this study demonstrates that the prevalence of positive lesions is consistent with the general population prevalence and that the dental attendees mimicked the general population in their use of tobacco and alcohol.

Eighteen general dentists were trained and standardized to identify erythroplakia (red lesions) and leukoplakia (white lesions), the primary lesions associated with being potentially malignant or malignant. This examination was in addition to their regular procedures and required about three minutes. Prior to the examination, patients completed a short survey regarding their age, sex, ethnicity and smoking and drinking practices.

This pivotal study determined, for the first time, that opportunistic screening in general dental practice is the place to screen for oral cancer because these investigators showed that the prevalence of lesions associated with oral cancers and the prevalence of risky behaviours (use of tobacco and/or alcohol) among dental attenders are similar to the general UK population. Thus, opportunistic screening by general dentists is a significant step forward in our efforts to decrease morbidity and mortality resulting from oral cancers.

A. M. Horowitz, Senior Scientist,
National Institute of Dental and Craniofacial Research, National Institutes of Health, Department of Health and Human Services
Bethesda, MD USA

IN BRIEF

- Oral cancer is an important disease that is easily identified in a dental practice setting.
- This study shows that GDPs can easily detect relevant lesions using a simple systematic mucosal examination, with little disruption to a normal practice routine.
- The prevalence of relevant lesions and of risk habits among patients who attend general dental practices appears to be representative of the general population.
- The data suggest that opportunistic screening in a general dental practice setting, particularly if high risk groups can be targeted, might be a realistic option.

Opportunistic screening for oral cancer and precancer in general dental practice: results of a demonstration study

K. Lim,[1] D. R. Moles,[2] M. C. Downer[3] and P. M. Speight[4]

Objectives To demonstrate the feasibility of opportunistic oral cancer and precancer screening in general dental practice and to determine the prevalence of relevant lesions and risk habits in a population of general dental practice attenders.

Design A prospective demonstration study, recruiting patients opportunistically.

Setting General dental practices.

Methods Eighteen general dental practitioners took part in this study. Each attended training sessions to be advised of the study protocol and the criteria of a positive and negative screen. Patients over the age of 35 years were prospectively and opportunistically recruited. Each patient was asked to complete a health questionnaire concerning age, gender, ethnicity, smoking and drinking habits. The dentist then examined the soft tissues and recorded the presence or absence of lesions independently on a second form. The forms were collated and data were analysed to determine prevalence of lesions and associations with risk habits.

Results Data on 2,265 patients were available for analysis. Oral lesions were detected in 319 patients (14.1%). Ninety-four patients (4.2%) had lesions considered to be either malignant or potentially malignant. There was a significant association between positive lesions and male gender (IRR 1.86, 95% CI 1.22-2.82), heavy smoking (males: IRR 3.68, 95% CI 2.10-6.43: female; IRR 3.58, 95% CI 1.35-9.50) and heavy alcohol use in males (IRR 2.98, 95% CI 1.06-3.47).

Conclusions The results suggest that patients attending general dental practices are representative of the general population both in terms of lesion prevalence and high risk habits such as smoking and drinking. This supports the view that opportunistic screening in a general dental practice setting may be a realistic alternative to population screening. Further research is needed to determine the cost effectiveness of this approach and to investigate the value of targeting high risk groups within this population. General dental practice is ideal for the evaluation of such systems prior to extending these studies to other healthcare settings.

[1]GDP, Research Fellow [2]Clinical Lecturer, [3]Emeritus Professor, [4]Professor, Head of Department, Department of Oral and Maxillofacial Pathology, Eastman Dental Institute
Correspondence to: Prof. Paul M. Speight, Department of Oral and Maxillofacial Pathology, Eastman Dental Institute for Oral Health Care Sciences, University College London, 256 Gray's Inn Road, London WC1X 8LD
E-mail: p.speight@eastman.ucl.ac.uk

Refereed paper
Received 11.07.02; Accepted 15.01.03

Oral cancer is a major global health problem and is the sixth most common cancer worldwide[1] and according to the International Agency for Research on Cancer (IARC) there were over 266,000 new cases of intra-oral cancer in the year 2000, with the majority (64%) occurring in males.[2] The world age-standardised incidence rate (the number of new cases per 100,000 per year) was 6.42 for males and 3.27 for females. The estimated number of deaths from mouth cancer in the year 2000 was nearly 128,000.

In England the latest data show 2,870 new cases of oral cancer in 1998[3] with an overall incidence of 7.6 for males and 3.6 for females. In recent years the number of cases has increased steadily, for example from 2,377 in 1995, and from a total for England and Wales of less than 2,000 per year in 1985.[4] This trend is reflected in increasing incidence and mortality rates that have particularly affected younger males in the 35–64 age group.[5]

For the United Kingdom as a whole, the IARC estimated that registrations of oral cancer for the year 2000 were 4,459 of which 2,923 occurred in males and 1,536 in females. The estimated number of deaths from oral cancer in the UK was 1,334 for males and 717 for females.[2] The figures for incidence and deaths do not fully indicate the magnitude of the burden of the disease in the population. Another indicator is prevalence. The UK 5-year prevalence (number of people alive with oral cancer diagnosed in the preceding 5-years) in the year 2000 was believed to be 12,740 (8,186 male, 4,554 female). Thus at any time nearly 13,000 people in the UK are living under the shadow of oral cancer.[2]

In addition to increasing incidence, mortality remains high with over 50% of patients dying of their disease. Most studies have shown no improvement in survival for decades,[5] but a recent, more detailed examination of the data suggests small improvements for most oral sites, which are most evident in more affluent socio-economic groups.[6] These figures are despite the fact that the oral cavity is easily accessible and about half the population receives regular oral examinations during routine dental treat-

ment. Unfortunately, there is a lack of public awareness of the disease and patients are slow to seek attention.[7–9] The simple fact is that over 60% of all patients present with late lesions when the prognosis is already poor and metastatic spread has already occurred.[6,10,11]

Although primary prevention in the form of advice and education about risk factors is important, this is largely ineffectual as evidenced by increased tobacco use in the UK despite knowledge of the risks, and the lack of evidence of effectiveness of mass education programmes in improving oral health.[12] There is clearly a large gap between knowledge and behaviour change.

One way forward is to improve early detection of oral cancer, either by case finding or by organised screening programmes. The rationale for this is that oral cancer may be preceded by a clinically detectable potentially malignant lesion (leukoplakia or erythroplakia) or that it may begin as a small, localised, often asymptomatic lesion in the early part of its natural history.[11] If detected when small, these lesions can be treated thus avoiding the notoriously high mortality and morbidity associated with this cancer. It is unlikely however that population screening for oral cancer will ever be instituted. The UK Working Group on screening for oral cancer and precancer reported in 1993 that there was insufficient evidence to support population screening and laid out an agenda for further research.[13]

Recent pilot studies of oral cancer screening programmes using a simple oral examination showed that dental screeners in a hospital, medical practice or industrial setting could detect relevant lesions with a sensitivity and specificity equivalent to that achieved in other screening programmes.[14,15] However the compliance for screening following a postal invitation was only about 25%,[16] leading the researchers to conclude that such a programme may not be cost effective. A viable alternative might be to carry out screening opportunistically, especially if high-risk groups could be targeted. Using the data from the pilot programmes[14,15] an artificial neural network was trained which was shown to be able to detect relevant oral lesions with a sensitivity of 0.80 and specificity of 0.77.[17] By means of simulation modelling techniques it was subsequently shown that this computer system could preselect high-risk individuals and identify 80% of all lesions by screening only 25% of the population.[18,19]

Initially, the most obvious place to evaluate opportunistic screening or case finding is in a general dental practice environment. At present however, there are no data on the prevalence of lesions, or of high risk habits in a population of dental attenders, and there have been no attempts to evaluate oral cancer screening in primary dental care.

The purpose of the present investigation was to establish a demonstration study of opportunistic oral cancer screening in a dental primary care environment and to determine the characteristics of routine general dental practice attenders in terms of the prevalence of relevant habits and of malignant or potentially malignant lesions.

MATERIALS AND METHODS

For the purposes of this study, oral cancer was defined as squamous cell carcinoma of the lip (ICD-9, 140), tongue (ICD-9, 141), gum, floor of mouth, 'mouth', oropharynx (ICD-9, 143-146), hypopharynx and 'other sites' (ICD-9 148, 149).[4,20] (The equivalent codes in the current ICD-10 are: C00 – C06, C09, C10 and C14.) The setting for this study was general dental practice. Eighteen general dental practitioners took part, 16 were located in Greater London, one in Nottingham and one in Aldershot. The practices predominantly offered a mix of National Health Service and private dentistry. Each dentist attended two training sessions to standardize criteria for the identification of lesions and to receive instruction on the protocol of the study. In the context of screening for potentially malignant or malignant lesions, the dentists were advised of the criteria for a positive and negative screen. A positive screen was defined as the presence of a white patch or a red patch, or of an ulcer of longer than 2 weeks duration.[14] These basic criteria were modified by defining a number of well known clinical entities which might have these appearances but should be included as positive or negative lesions.[21] For example, ulcers and white lesions with an obvious traumatic aetiology and recurrent ulcers (aphthae) were regarded as negative, whereas lichen planus, actinic keratosis and submucous fibrosis were defined as positive.

Table 1 Characteristics of the patient population

Factor	Description	Number	(%)
Gender	Male	1,078	(47.6)
	Female	1,187	(52.4)
Age group	35-39	236	(10.4)
	40-44	290	(12.8)
	45-49	323	(14.3)
	50-54	319	(14.1)
	55-59	286	(12.6)
	60-64	229	(10.1)
	65-69	214	(9.4)
	70-74	164	(7.2)
	75-79	109	(4.8)
	80-84	73	(3.2)
	85-89	15	(0.7)
	90+	7	(0.3)
Ethnic group	White	1,892	(83.5)
	Black	173	(7.6)
	South Asian	120	(5.3)
	Chinese	31	(1.4)
	Other	49	(2.2)

Patients aged over 35 were prospectively and opportunistically recruited. It was not possible to prescribe set criteria for the method of recruitment and each dentist selected the most appropriate method to suit their practice routine. Some recruited sequentially, while others recruited randomly or on fixed days per week. The aim was for each dentist to attempt to recruit around 200 patients into the study. Each patient was first asked to complete a questionnaire concerning his or her age, gender, ethnicity, and smoking and drinking habits. This was normally completed in the waiting room under the guidance of a practice nurse or receptionist.

Before commencing routine treatment the dentist then recorded the presence or absence of lesions independently on a second form. The site and appearance of all lesions, whether positive or negative, were recorded. The nature of each lesion or a diagnosis, when obvious, was also entered on the form. The dentists were blinded to the results of the questionnaire at the time of examination. The two forms were later collated and sent in batches to the research team at the Eastman Dental Institute.

Patients with lesions were referred as appropriate for secondary care, according to the normal protocols of each practice. There was no attempt, in this study, to record the results of secondary care.

Data were analysed using the *STATA* statistical software package (Version 5.0, Stata Corporation, Texas, USA). Incidence rate ratios (relative risk for having a positive lesion) were calculated for for smoking and alcohol consumption using Poisson regression adjusted for the age of the patient.

RESULTS

The 18 participating practitioners returned a total of 2,342 forms (minimum 5, maximum 423). One-hundred-and-twelve forms (4.8%) were excluded due to either the patient being less than 35-years-old (77 forms) or as a result of illogical combinations of responses (35 forms). The remaining forms, relating to 2,265 patients, were evaluated. Tables 1 and 2 summarise the character-

Table 2 Self-reported behavioural characteristics of the patients

Factor	Description	Definition	Number	(%)
Smoking	Heavy	20 or more cigarettes/day	179	(7.9)
	Moderate	up to 19 cigarettes/day	474	(20.9)
	Non	never smoked or not within last 10 years	1,612	(71.2)
Alcohol	Heavy	Units/week: > 20 males, >14 females	50	(2.2)
	Moderate	Units/week: 5-20 males, 5-14 females	323	(14.3)
	Non or light	Up to 4 units/week)	1,892	(83.5)
Dental attendance	'Regular'	(within the last year)	1,962	(86.6)
	'Irregular'	(not within the last year)	303	(13.4)

Table 3 Lesions detected

		Number	(%)
Positive lesions	Carcinoma	2	(0.1)
	White patch	45	(2.0)
	Red patch	11	(0.5)
	Lichen planus	31	(1.4)
	Persistent ulcer	2	(0.1)
	Submucous fibrosis	1	(<0.1)
	Actinic keratosis	2	(0.1)
	Total positive	94	(4.2)
Negative lesions	Non-specific or traumatic ulcers	39	(1.7)
	Frictional/traumatic keratosis	30	(1.3)
	Fibrous overgrowths	28	(1.2)
	Candida/denture stomatitis	14	(0.6)
	Smokers keratosis (palate)	12	(0.5)
	Aphthous ulcers	10	(0.4)
	Amalgam tattoo	9	(0.4)
	'Haemangioma'	7	(0.3)
	Angular cheilitis	5	(0.2)
	Abscess/sinus	5	(0.2)
	Mucoceles	4	(0.2)
	Geographic tongue	3	(0.1)
	Naevus	3	(0.1)
	Miscellaneous	56	(2.5)
	Total negative	225	(9.9)
	Total lesions	319	(14.1)

istics of the sample of patients examined. Table 1 indicates the physical characteristics of the sample and Table 2 indicates their self-reported smoking and drinking habits.

Oral lesions were detected in 319 patients, giving an overall lesion prevalence of 14.1% (Table 3). Ninety-four patients (4.2%) had lesions that were considered to be either malignant or potentially malignant (positive lesions). This included two squamous cell carcinomas one of which had developed from submucous fibrosis. Table 4 gives the distribution of positive lesions according to gender, ethnicity, smoking habits, and alcohol consumption. The number of positive lesions occurring in the different ethnic groups was too small to submit to formal statistical analysis.

The associations between smoking and alcohol consumption and the prevalence of positive lesions are indicated in Table 5. The strengths of the associations were quantified by the incidence rate ratio (IRR), which is a measure of relative risk adjusted for the patient's age. The results are stratified by gender since there is evidence that males have nearly twice the risk of a positive lesion compared with females and that the risk factors may not have exactly the same effect for both males and females. For both genders, heavy smoking significantly increased the risk of a positive lesion by over 350%. No significant relationship could be detected between moderate smoking and the risk of having a positive lesion for either males or females. The effect of alcohol consumption differed between men and women. For men, there was a significant association between alcohol and risk. Moderate alcohol consumption was associated with nearly twice the risk of having a positive lesion, while heavy drinking was associated with nearly three times the risk when compared with non/light drinkers. No significant relationships between alcohol consumption and positive lesion prevalence were detected for women.

DISCUSSION

This was a prospective study carried out opportunistically in typical general practice settings. The purpose was to determine the prevalence of relevant lesions and habits in a dental practice population and therefore there was no specific targeting of high-risk groups; indeed doing so would have negated the purpose of this study.

The prevalence of mucosal lesions detected in this study was 14.1%, which is similar to other studies. However comparisons are difficult, because there have been few similar studies in non-institutionalised populations and the criterion for lesions varies markedly. For example, in a recent study in Germany, Reichart[22] reported that 66.2% of 35–74-year-olds had mucosal lesions, but this included Fordyce spots (23.7%) and patients with a history of aphthous ulceration (18.3%). In Spain, Martinez-Diaz and Garcia-Pola[23] reported a prevalence of 58.75% among subjects attending a dental school for periodontal or prosthodontic treatment. However this included pigmentation (24.6%) and 'linea alba' (10.7%). In another study of a random sample of Sicilian men mucosal lesions were found in 81% of the group,[24] but this included over 50% with 'coated tongue'. Nevertheless the prevalence of leukoplakia in this group was 13.8%, which was ascribed to the high numbers who used alcohol and tobacco. It should be noted that in the present study variants of normal were not included as mucosal lesions, thus the prevalence of 'furry tongue', Fordyce spots, varices and 'leukoedema' have not been recorded.

Table 4 Distribution of positive lesions

Factor	Description	Number	%
Gender	Male	59	(5.5)
	Female	35	(2.9)
Ethnic group	White	87	(4.6)
	Black	1	(0.6)
	South Asian	4	(3.3)
	Chinese	2	(6.5)
	Other	0	(0.0)
Smoking	Heavy	25	(14.0)
	Moderate	16	(3.4)
	None	53	(3.3)
Alcohol consumption	Heavy	5	(10.0)
	Moderate	23	(7.1)
	Non or light	66	(3.5)

Table 5 Factors associated with the presence of a positive lesion

Factor	Description	IRR	(95% CI)	P
Gender	Male	1.86	(1.22, 2.82)	0.004
	Female	1*		
Smoking in females	Heavy	3.58	(1.35, 9.50)	0.010
	Moderate	1.89	(0.87, 4.13)	0.110
	Non	1*		
Smoking in males	Heavy	3.68	(2.10, 6.43)	<0.001
	Moderate	0.57	(0.25 1.29)	0.179
	Non	1*		
Alcohol in females	Heavy	1.84	(0.25, 13.48)	0.548
	Moderate	2.63	(0.36, 4.08)	0.713
	Non or light	1*		
Alcohol in males	Heavy	2.98	(1.06, 8.38)	0.039
	Moderate	1.93	(1.12, 3.47)	0.019
	Non or light	1*		

* = Baseline group for each comparison

In a more comparative, population based, study of subjects over 35 in the USA, Bouquot[25] analysed data from 23,616 oral examinations and reported 3,783 mucosal lesions in 2,824 people – giving an overall prevalence of 10.3% of subjects. This included 1% with Fordyce spots. In this same study the prevalence of white keratotic lesions was 3.4% and of leukoplakia 2.9%.[26]

In the present study, positive lesions were recorded in 4.2% of subjects. This is more than previously found in a general medical practice environment (2.2%) or among patients of a dental hospital (3.0%),[14] but is slightly less than that recorded among subjects screened in a company headquarters (5.5%).[15] Since the age groups in these studies were similar, the reasons for these differences are not clear. In the previous studies, subjects were invited to have an oral examination, either directly or by letter and it is possible that those who abused alcohol, or were heavy smokers refused. In the company headquarters however, and in the present study, recruitment was opportunistic. It is possible therefore that a prevalence of 4–5% is more representative of the population as a whole.

There have been a number of studies which have determined the prevalence of potentially malignant and malignant lesions and some of these are summarised in Table 6.[14,15,22,24,25,27–47] Although overt cancers were recorded, the most common relevant lesion was 'leukoplakia' and therefore most of these studies may be regarded as demonstrating the prevalence of this entity. However leukoplakia is poorly defined and, similar to the current study, most can be regarded as recording mainly the prevalence of persistent white or keratotic lesions. The setting of these studies varies, but overall, most show a prevalence of between 1 and 6%, which is similar to that found in this study and our other UK studies.[14,15] Most studies are of specific age groups, but those which show a high prevalence have usually been conducted in high risk groups and have been carried out with some element of targeting, either in countries where the prevalence of oral cancer and precancer is high,[28,30,31,33,40] in high-risk populations such as heavy smokers and drinkers or betel quid users,[24,36,42,47] as part of multiphasic screening such as those attending lung cancer clinics[46] or in convenience populations like the institutionalised elderly, forces personnel and attenders of hospital outpatients clinics or medical practices.[14,15,32,39]

An informative UK study was that carried out by Pearson *et al.*[47] among the Bangladeshi population of East London. This showed a prevalence of leukoplakia of 25%, with a positive association with betel quid or paan chewing. This highlights the potential dangers of extrapolating data from one population to another without taking account of the nature of our multicultural society.

The proportions of alcohol and tobacco users were also similar to those previously reported. In our previous study 8% were heavy smokers (current study; 7.9%, Table 2) and 3% were heavy drinkers (current study; 2.2%). Overall, 29% were smokers and 16.5% were moderate or heavy drinkers. Data from the ONS omnibus surveys conducted in 2000 and 2001 show that 26% of the population are current smokers[48] and 34% of men and 26% of women in the 45–64 year age group drink more than three or four days per week.[49]

The effects of smoking and alcohol consumption and the likelihood of having a positive lesion or condition were also clearly seen and are similar to our previous study where heavy smoking showed a significant relationship.[14] In the current study, heavy smokers were more likely to have a positive lesion by a factor of 3.58 for females and 3.68 for males. This study also showed that men who drank heavily were about three times more likely to have a positive lesion.

Because of the relatively low prevalence of the disease, and a lack of adequate knowledge of the natural history, it is generally agreed that mass population screening for oral cancer and precancer may not be cost-effective and cannot be recommended.[13,50,51] Opportunistic screening, undertaken when patients attend a healthcare professional for some other purpose, may however be beneficial,[13,20,50] especially if high risk groups can easily be identified and targeted for primary preventive advice and a mucosal examination.[13,19] Opportunistic case finding is already an important component of a routine dental check-up, but further research is needed to determine the prevalence in the general population as well as in different subpopulations and ethnic groups, so that appropriate high risk groups can be targeted.[52,53]

The results of this study suggest that the population attending general dental practices, who are self selecting and therefore thought to be not representative, is representative of the general population both in terms of lesion prevalence and high risk habits. This suggests that opportunistic screening in a general dental practice setting may be a realistic option. Careful targeting of high risk individuals within this group may make this more cost effective and on-going studies are investigating methods of pre-selection and targeting within primary care environments. It is recognized that, at any one time, only 50% of the population are registered with a dentist, but general dental practice is an ideal environment to evaluate these systems prior to investigating applications in other healthcare settings.

Table 6 Representative screening and prevalence studies for oral cancer and precancer from 1971 to 2002

Authors [reference]		Year	Country	Population/setting	Age	*n*	Cancers	Precancers	%+ve
Ross and Gross	[27]	1971	USA	Gen. population	35+	12,868	1	339	2.6
Mehta *et al.*	[28]	1972	India	Gen. population	15+	101,761	11	1,628	1.6
Axell	[29]	1976	Sweden	Gen. population	15+	20,333	1	732	3.6
Warnakulasuriya	[30]	1984	Sri Lanka	Gen. population	20+	29,295	4	1,230	4.2
Mehta *et al.*	[31]	1986	India	Gen. population	35+	39,331	20	511	1.4
Bouqout	[25]	1986	USA	Gen. population	10-90	23,616	22	682	3
Viglid	[32]	1987	Denmark	Institutionalised elderly	65+	285	2	19	7.4
Warnakulasuriya	[33]	1991	Sri Lanka	Gen. population	20+	57,124	20	3,541	6.2
Ikeda *et al.*	[34]	1991	Japan	Gen. population	18-63	3,131	0	77	2.5
Banoczy and Rigo	[35]	1991	Hungary	Lung clinic screening	19-60	7,820	1	104	1.3
Talamini *et al.*	[36]	1994	Italy	High risk referrals	35+	627	10	55	14.9
Downer *et al.*	[15]	1995	UK	Company HQ	40+	292	0	17	5.8
Jullien *et al.*	[14]	1995	UK	Dental hosp,	40+	1,042	1	21	3.0
Jullien *et al.*	[14]	1995	UK	Medical practice	40+	985	1	31	2.2
Ikeda *et al*	[37]	1995	Japan	Gen. population	60+	802	0	38	4.7
Ikeda *et al.*	[38]	1995	Cambodia	Gen. population	15+	1,319	1	41	3.2
Field *et al.*	[39]	1995	UK	Gen. population	<20->69	1,947	1?	4	0.2
Mathew *et al.*	[40]	1997	India	Gen. population	35-64	2,069	1	212	10.2
Prout *et al.*	[41]	1997	USA	Tobacco users	40+	4,611	1	590	12.8
Szabo *et al.*	[42]	1997	Hungary	Homeless, alcoholics	Not given	300	8	42	16.7
Zain *et al.*	[43]	1997	Malaysia	Gen. population	25+	11,707	5	165	1.5
Reichart	[22]	2000	Germany	Aging Germans	35-74	2,022	0	55	2.7
Sankaranarayanan	[44]	2000	India	Gen. population	35+	114,601	63	1,310	12
Kovac-Kovacic	[45]	2000	Slovenia	Perio patients	25-75	1,609	0	106	6.6
Campisi, Margiotta	[24]	2001	Sicily	Random male	40+	118	1	4	13.9
Domoi *et al.*	[46]	2001	Hungary	Lung clinic screening	18+	5,034	0	188	3.7
Pearson *et al.*	[47]	2001	UK	Bangladeshi	40+	137	0	28	20.4
Current study		2002	UK	Dental practices	35+	2,265	2	92	4.2

There are few studies worldwide where oral cancer screening has been thoroughly evaluated with appropriate endpoints. Two studies however, from Cuba and India, provide encouraging results with evidence that morbidity and mortality can be reduced.[44,54] Santana *et al.*[54] report a study from Cuba where an oral cancer case finding programme was carried out between 1983 and 1990, this was shown to be effective in reducing the number of oral cancers presenting at a late stage. Patients presenting with stage I lesions rose from 22.8% to 48.2% and stage II, III, and IV fell from 77.25 to 51.8%. In the Indian study,[44] one of the largest in the world, the intervention group showed 72.3% of early stage lesions compared with 12.5% in the control group. Furthermore, this study was able to demonstrate a reduction in mortality, with 3 year fatalities of 14.9% in the intervention group compared with 56.3% in the control group. Neoplasms discovered when small, could result in substantial cost savings for the Health Service. Early diagnosis of lesions that are smaller than 2.0 cm (Stage I) have a good prognosis[10,11] and since 20% of oral cancer patients get another cancer within a 5-year period[55,56] regular screening in a general practice setting could also detect second, metachronous lesions while they are small.

SUMMARY AND CONCLUSIONS

The prevalence of mucosal lesions in a population attending typical general dental practices was 14.1%, with 4.2% of lesions being regarded as malignant or potentially malignant. This, and the proportions of smokers and drinkers are comparable with data for the general population as a whole and to other prevalence studies in similar settings. The results suggest that dental attendees are quite representative of the general population and support the view that opportunistic oral cancer screening in general dental practice may be a realistic option. Further studies are needed to determine the potential of targeting high risk groups within these populations, to evaluate similar approaches in other healthcare settings and to determine the cost-effectiveness.

K Lim was funded by the Oral and Dental Research Trust as the recipient of a Dunhill Trust Fellowship.
The authors are grateful to the following general dental practitioners who so kindly gave of their time in their busy daily schedule to make this study possible. Peter Fine, Nick Unsworth, Richard Pinns, Malcolm Alexander, Malcolm Gald, Rohit Mistry, Colin Gaunt, Douglas Miller, Nigel Monkman, Tushar Patel, Sandip Patel, Sangeeta Patel, Raj Parma, Peter Jenkinson, Rebecca Roberts, Darren Kowlessar, and Harris Yee-Chong.

1. Parkin D M, Pisani P, Ferlay J. Estimates of worldwide incidence of twenty-five major cancers in 1990. *Int J Cancer* 1999; **80:** 827-841.
2. Ferlay J, Bray F, Pisani P *et al. GLOBOCAN 2000: cancer incidence, mortality and prevalence worldwide, version 1.0.* IARC Cancer Base No.5. Lyon: IARC, 2002.
3. Office of National Statistics. *Cancer statistics registrations, Registrations of cancer diagnosed in 1998, England.* Series MB1. No 29. London ONS, 2002.
4. Office of Population Census and Surveys. *Cancer statistics, registrations 1985.* Series MB1. Nó 18. London OPCS, 1990.
5. Hindle I, Downer M C, Speight P M. The epidemiology of oral cancer. *Br J Oral Max Fac Surg* 1996; **34:** 471-476.
6. Coleman M P, Babb P, Damiecki P, *et al. Cancer survival trends in England and Wales 1971-1995: depravation and NHS region.* London: TSO, 1999.
7. Scully C, Malamos D, Levers B G, Porter S R, *et al.* Sources and patterns of referrals of oral cancer: role of general practitioners. *Br Med J* 1986; **293:** 599-601.
8. Warnakulasuriya K A A S. An alarming lack of public awareness towards oral cancer. *Br Dent J* 1999 Sep 25; **187:** 319-322.
9. Allison P, Locker D, Feine J S. The role of diagnostic delays in the prognosis of oral cancer: a review of the literature. *Oral Oncol* 1998; **34:** 161-170.
10. Platz H, Fries R, Hudec M. *Prognosis if oral cavity carcinomas, results of a multi centric retrospective observational study.* Munich: Hanser 1986.
11. Speight P M, Morgan P R. The natural history and pathology of oral cancer and precancer. *Comm Dent Health* 1993; **10**(Suppl. 1): 31-41.
12. Kay E, Locker D. A systematic review of the effectiveness of health promotion aimed at improving oral health. *Comm Dent Health* 1998; **15:** 132-144.
13. Speight P M, Downer M C, Zakrzewska J M eds. Screening for oral cancer and precancer:

report of a UK working group. *Comm Dent Health* 1993; **10:** (Suppl 1): 1-89.
14. Jullien J A, Downer M C, Zakrzewska J M, Speight P M. Evaluation of a screening test for the early detection of oral cancer and precancer. *Community Dent Health* 1995; **12:** 3-7.
15. Downer M C, Evans A W, Hughes-Hallett C M, Jullien J A, Speight P M, Zakrzewska J M. Evaluation of screening for oral cancer and precancer in a company headquarters *Community Dent Oral Epidemiol* 1995; **23:** 84-88.
16. Jullien J A, Zakrzewska J M, Downer M C, Speight P M. Attendance and compliance at an oral cancer screening programme in a general medical practice. *Oral Oncol, Eur J Cancer* 1995; **31B:** 202-206.
17. Speight P M, Elliott A, Jullien J A, Downer M C, Zakzrewska J M. The use of artificial intelligence to identify people at risk of oral cancer and precancer. *Br Dent J* 1995; **179:** 382-387.
18. Downer M C, Jullien J A, Speight P M. An interim determination of health gain from oral cancer and precancer screening: 2. Developing a model of population screening. *Community Dental Health* 1997; **14:** 227-232.
19. Downer M C, Jullien J A, Speight P M. An interim determination of health gain from oral cancer and precancer screening: 3. Preselecting high risk individuals. *Community Dental Health* 1998; **15:** 72-76.
20. British Dental Association. *Opportunistic oral cancer screening.* Occasional paper No 6. April 2000.
21. Speight P M, Downer M C, Zakrzewska. Screening for Oral Cancer and Precancer. *Oral Oncology, European J Cancer* 1992; **28B:** 45-48.
22. Reichart P A. Oral mucosal lesions in a representatives cross-sectional study of ageing Germans. *Community Dent Oral Epidemiol* 2000; **28:** 390-398.
23. Martinez-Diaz A I, Garcia-Pola M J. Epidemiological study of oral mucosal pathology in patients of the Oviedo School of Stomatology. *Med Oral* 2002; **7:** 4-16.
24. Campisi G, Margiotta V. Oral mucosal lesions and risk habits among men in an Italian study population. *J Oral Pathol Med* 2001; **1**: 22-28.
25. Bouqout J E. Common oral lesions found during a mass screening examination. *J Am Dent Ass* 1986; **112:** 50-57.
26. Bouquot J E, Gorlin R J. Leukoplakia, lichen planus, and other oral keratoses in 23,616 white Americans over the age of 35 years. *Oral Surg Oral Med Oral Pathol* 1986; **61:** 373-381.
27. Ross N M, Gross E. Oral findings based on an automated multiphasic health screening program. *J Oral Med* 1971; **26:** 21-26.
28. Metha F S, Gupta P C, Daftary D K, Pindborg J J, Choksi S K. An epidemiologic study in oral cancer and precancerous conditions among 101761 villagers in Maharashtra, India. *Int J Cancer* 1972; **10:** 134-141.
29. Axell T A. A prevalence study of oral mucosal lesions in an adult Swedish population *Odontol Revy* 1976; **27:** 1-103.
30. Warnakulasuriya K A A S, Ekanayaki A N I, Sivayoham S, *et al.* Utilisation of primary health care workers for early detection of oral cancer and precancer cases in Sri Lanka. *Bulletin of the World Health Organisation* 1984; **62:** 243-250.
31. Metha F S, Gupta P C, Bhonsle R B *et al.* Reproducibility of oral cancer using basic health workers in an area of high oral cancer incidence in India. *Cancer Detection and Prevention* 1986; **9:** 219-225.
32. Vigild M, Oral Mucosal lesions among the instituitionalised elderly in Denmark. *Dent Oral Epidemiol* 1987; **15:** 309-313.
33. Warnakulasuriya K A A S, Nanayakkara B G, Reproducibility of an oral cancer and precancer detection program using a primary health care model in Sri Lanka. *Cancer Detection Prevention* 1991; **15:** 331-334.
34. Ikeda N, Ishii T, Kawai T. Epidemiologic study of oral leukoplakia based on mass screening for oral mucosal disease in a selected Japanese population. *Community Dent Oral Epidemiol* 1991; **19:** 160-163.
35. Banoczy J, Rigo O, Prevalence study of oral precancerous lesions within a complex screening system in Hungary. *Community Dent Oral Epidemiol* 1991; **19:** 265-267.
36. Talamini R, Barzan L, Franceschi S, Caruso G, Gasparin A, Comoretto R. Determinants of compliance with an early detection programme for cancer of the head and neck in north eastern Italy. *Eur J Cancer, Part B Oral Oncol* 1994; **30B:** 415-418.
37. Ikeda N, Downer M C, Ishii T, Fukano H, Nagao T, Inoue K. Annual screening for oral cancer and precancer by invitation to 60-year-old residents of a city in Japan. *Comm Dent Health* 1995; **12:** 133-137.
38. Ikeda N, Handa Y, So Po Khim, *et al.* Prevalence study of oral mucosal lesions in a selected Cambodian population. *Comm Dent Oral Epidemiol* 1995; **23:** 49-54.
39. Field E A, Morrison T, Darling A E, Parr T A, Zakrzewska J M. Oral mucosal screening as an integral part of routine dental care. *Br Dent J* 1995; **179:** 262-266.
40. Mathew B, Sankaranarayanan R, Sunilkumar KB, *et al.* Reproducibility and validity of oral visual inspection by trained health workers in the detection of oral precancer and cancer. *Br J Cancer* 1997; **76:** 390-394.
41. Prout M N, Sidari J, Witzburg A, *et al.* Head and neck cancer screening among 4611 tobacco users older than 40 years. *Otolaryngol Head Neck Surg* 1997; **116:** 201-208.
42. Szabo G, Klenk G, Veer A. Correlation between the combination of alcohol consumption and smoking and oral cancer (screening of the population at risk) *Orvsoi Hetilap* 1997; **138:** 3297-3299.
43. Zain P B, Ikeda N, Razak I A, Axell T, Majid Z A, Gupta P C, Yaacob M. A National epidemiological survey of oral mucosal lesions in Malaysia. *Community Dent Oral Epidemiol* 1997 Oct; **25:** 377-383.
44. Sankaranarayanan R, Mathew B, Jacob B J, *et al.* Early findings from a community-based cluster-randomised, controlled oral cancer screening trial in Kerala, India. The Trivandrum Oral Cancer Screening Study Group. *Cancer* 2000 Feb 1; **88:** 664-673.
45. Kovac-Kovacic M, Skaleric U. The prevalence of oral mucosal lesions in a population in Ljubljana, Slovenia. *J Oral Pathol Med* 2000 Aug; 331-335.
46. Dombi C, Voros-Balog T, Czegledy A, *et al.* Risk group assessment of oral precancer attached to X-ray lung screening examinations. *Community Dent Oral Epidemiol* 2001; **29:** 9-13.
47. Pearson N, Croucher R, Marcenes W, O'Farrell M. Prevalence of oral lesions among a sample of Bangladeshi medical users aged 40 years and over living in Tower Hamlets, UK. *Int Dent J* 2001; **51:** 30-34.
48. Lader D, Meltzer H. *Smoking related behaviour and attitudes 2001.* London: ONS, 2002.
49. Lader D, Meltzer H. *Drinking related behaviour and attitudes 2001.* London: ONS, 2002.
50. K A A S Warnakulasuriya, Johnson N W. Strengths and weaknesses of screening programmes for oral malignancies and potentially malignant lesions. *Eur J Cancer Prevention* 1996; **5:** 93-98.
51. Rodrigues V C, Moss S M, Tuomainen. Oral cancer in the UK: to screen or not to screen. *Oral Oncol* 1998; **34:** 454-465.
52. Klienman D V, Swango P A, Niessen L C, *et al.* Epidemiologic studies of oral mucosal conditions-methodologic issues. *Comm Dent Oral Epidemiol* 1991; **19:** 129-140.
53. Klienman D V, Swango P A, Pindborg J J. Toward assessing trends in oral mucosal lesions: lessons learned from oral cancer. *Adv Dent Res* 1993; **7:** 32-41.
54. Santana J C, Delgado L, Miranda J, *et al.* Oral cancer case finding programme (OCCFP). *Oral Oncol* 1997; **33:** 10-12.
55. Franceschi S, Barzan L, Talamini R. Screening of the head and neck: if not now, when? *Oral Oncol* 1997; **33:** 313-316.
56. Johnson N W. *Oral cancer screening- part 5.* FDI World, 1/98. 14-19.