On the Regional Dose Susceptibility of Parotid Gland Function Loss and Recovery

An Effort Toward Amelioration of Radiotherapy-Induced Xerostomia

by

Haley Clark

B.Sc. Hon. in Physics, University of Alberta, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Physics and Astronomy)

The University Of British Columbia
(Vancouver)

October 2013

Abstract

The introduction of intensity-modulated radiotherapy treatment has produced a small surplus of treatment planning flexibility compared to conventional techniques. Target volumes containing cancerous tumours are given a prescribed dose, but the surrounding normal tissue can sometimes be selectively irradiated. Therefore, as tissue-sparing techniques improve, the knowledge of complication risk in normal tissue becomes increasingly important.

Xerostomia is one of the most common normal tissue complications in head-and-neck cancer patients. It refers to the non-distinct symptom of dry mouth. In the case of radiotherapy-induced xerostomia, it is generally due to the loss of salivary function resulting from radiation damage to the parotid parenchyma. In severe cases it can drastically reduce oral hygiene and is known to strongly detract from a patient's quality of life.

We investigate the regional dose susceptibility of salivary function loss and recovery in the parotid gland with the intent of more precisely quantifying the risk of xerostomia. Reports have indicated regional dose susceptibility of loss in rat parotid glands. Similar results have been seen in human, though they seem to indicate primarily a morphological dependence on the shape of the dose distribution, not specific regional dependence. Further, they consider only subjective xerostomia, not objective salivary function loss nor recovery. The quantification of the regional dependence of loss and recovery of salivary function would substantially benefit our understanding of the complication risk of xerostomia. Immediate improvements in patient outcomes may follow.

To this end, we investigate the functional relation between dose delivered to sub-segments of the parotid and whole mouth saliva measurements. To

enable the investigation, we have developed a contour recognition system which is able to identify embedded planar organ contours with minimal human effort. Additionally, we have developed a sub-segmentation system capable of partitioning organ contours into arbitrarily-complex sub-segments.

Preface

The identification and design of this research program was completed in collaboratively by Vitali Moiseenko, Steven Thomas, Jonn Wu, and Allan Hovan. In particular, the original idea was established, and the collection of patient data began, prior to Hal Clark's involvement in the project.

The collection of patient data used herein was performed by the department of oral oncology at the British Columbia Cancer Agency (BCCA) Vancouver. The anonymization of patient data was performed primarily by Vitali Moiseenko and Hal Clark. Contour verification and quality assurance was performed by Jonn Wu.

Unless otherwise indicated, all analysis presented herein is original work performed by Hal Clark under the supervision and guidance of Vitali Moiseenko, Steven Thomas, and Jonn Wu.

The contents of chapters 5 and 6 are currently in press. Chapter 7 is written in a similar, but extended, style with the intention of future publication. These portions represent research and writing performed by Hal Clark during the present program.

Ethics approval from the UBC BCCA REB was granted for the collection and continued analysis of patient data under the title *A review and comparison of salivary function toxicity from standard techniques in head-and-neck cancer at* BCCA *and calculation of a radiation dose-salivary function response curve* and certificates H01 − 02073 and H07 − 02073.

Table of Contents

List of Tables

List of Figures

Glossary

CT	Computed Tomography
CVLF	Cross-Validation Lexicon Folding
DICOM	Digital Imaging and Communication in Medicine (see [9])
DVH	Dose-Volume Histogram
IMRT	Intensity-Modulated Radiotherapy Treatment
KPS	Karnofsky Performance Status
LENT-SOMA	Late Effects Normal Tissue - Subjective, Objective, Management, Analytic late toxicity grading system
LMS	Least-Median of Squares
LSS	Least-Sum of Squares
MR	Magnetic Resonance
NPLLR	Non-Parametric Local Linear Regression
NTCP	Normal Tissue Complication Probability
PTV	Planning Target Volume
QOL	Quality-of-Life
QUANTEC	Quantitative Analysis of Normal Tissue Effects in the Clinic
RT	Radiotherapy Treatment

Acknowledgments

The author would like to express their gratitude to Vitali Moiseenko, Jonn Wu, Richard Lee, Cheryl Duzenli, and Bradford Gill.

All contributed helpful advice, kindly offered their insight, tolerated extended discussions, and, each, in their own way, extended an invitation to participate with the broader scientific community.

In particular, the author wishes to thank Steven Thomas. Invariably, he was willing to lend a hand, an ear, or offer insight. Sometimes concurrently.

This would have been impossible if not for the love and support of family.

For T.B.B.L., Sarah.

Chapter 1

Introduction

Sometimes a scream is better than a thesis.

Ralph Waldo Emerson

This thesis concerns a radiation-induced toxicity known as *xerostomia*, the onset of which is primarily due to damage to the parotid gland. **Our first goal is to demonstrate that the parotid is inhomogeneously responsible for xerostomia**. In other words, whether damaging certain regions leads to an increase in the onset or severity of xerostomia.

Xerostomia is known, in some cases, to reduce in severity over time, indicating functional recovery. **Our second goal is to quantify the regional dose dependence of recovery**, if any exists.

Investigation will proceed by searching for association between measurements of salivary flow and delivered dose. If such inhomogeneity is found, radiotherapy treatment plans may be able to avoid the regions of increased susceptibility. This would improve patient comfort and quality-of-life both during and after radiotherapy.

1.1 Rationale

Three-dimensional planning has enabled the era of individualized, patient-specific treatment planning. We are now able to weigh estimates of tumour control with estimates of complication risk on a per-patient basis. Through the

use of the Dose-Volume Histogram (DVH), Intensity-Modulated Radiotherapy Treatment (IMRT), and clinically-recommended dose constraints, complex dose distributions can be constructed to deliver tumourcidal doses to diseased regions while simultaneously offering the ability to minimize complication risk in normal tissues. However, estimates of these normal tissue complication risks are not known with certainty, and their determination is a difficult problem. The complexity is demonstrated by sundry reports of recovery and hints of compensatory action. We believe Buckland et al. [11] eloquently summarized the spirit of this thesis.

> *Our philosophy is that truth is high (effectively infinite) dimensional. The more information that is gathered, the greater is the model complexity that the data can support.*

Which is an opinion shared by many in the field (e.g., [31]).

This thesis represents an effort to advance understanding of one such complication: xerostomia. It is hoped that precise knowledge of radiation-induced xerostomia, coupled with the strength of technologies like IMRT, will ultimately lead to its eradication as a substantial complication risk.

1.2 Thesis Organization

Our analysis begins with a brief overview of requisite background material.

Salivary gland function and dysfunction are reviewed in chapter 2. Topics from the literature pertaining to salivary function loss and recovery are presented in chapter 3. Research trends over approximately the last decade are reviewed in section 3.2, while findings of particular consequence for this study are covered in section 3.4.

Chapter 4 gives a precise statement of the questions which this thesis attempts to address. A specific outline of the analysis performed in chapters 5 to 7 is given in section 4.2. At this point, the reader ought to be armed with sufficient knowledge of the topic to proceed to any of the following three chapters.

Finally, concluding remarks and avenues for future research are discussed in chapter 8.

Chapter 2

A Brief Review of Salivary Glands and Xerostomia

My liege, and madam, to expostulate
What majesty should be, what duty is,
Why day is day, night night, and time is time,
Were nothing but to waste night, day and time.
Therefore, since brevity is the soul of wit,
And tediousness the limbs and outward flourishes,
I will be brief.

Hamlet
SHAKESPEARE

2.1 Salivary Glands

Humans have several salivary glands. The parotid, submandibular, and sublingual glands comprise the *major* salivary glands – those responsible for the majority of saliva production. *Accessory* salivary glands are distributed throughout the oral cavity [17]. It is estimated that the major salivary glands supply approximately $60 - 80\%$ of total saliva secreted [67] [72].

Saliva is composed of water (99.5%), with small amounts of amylase and other proteins, inorganic salts, mucins, bicarbonate, and other compounds

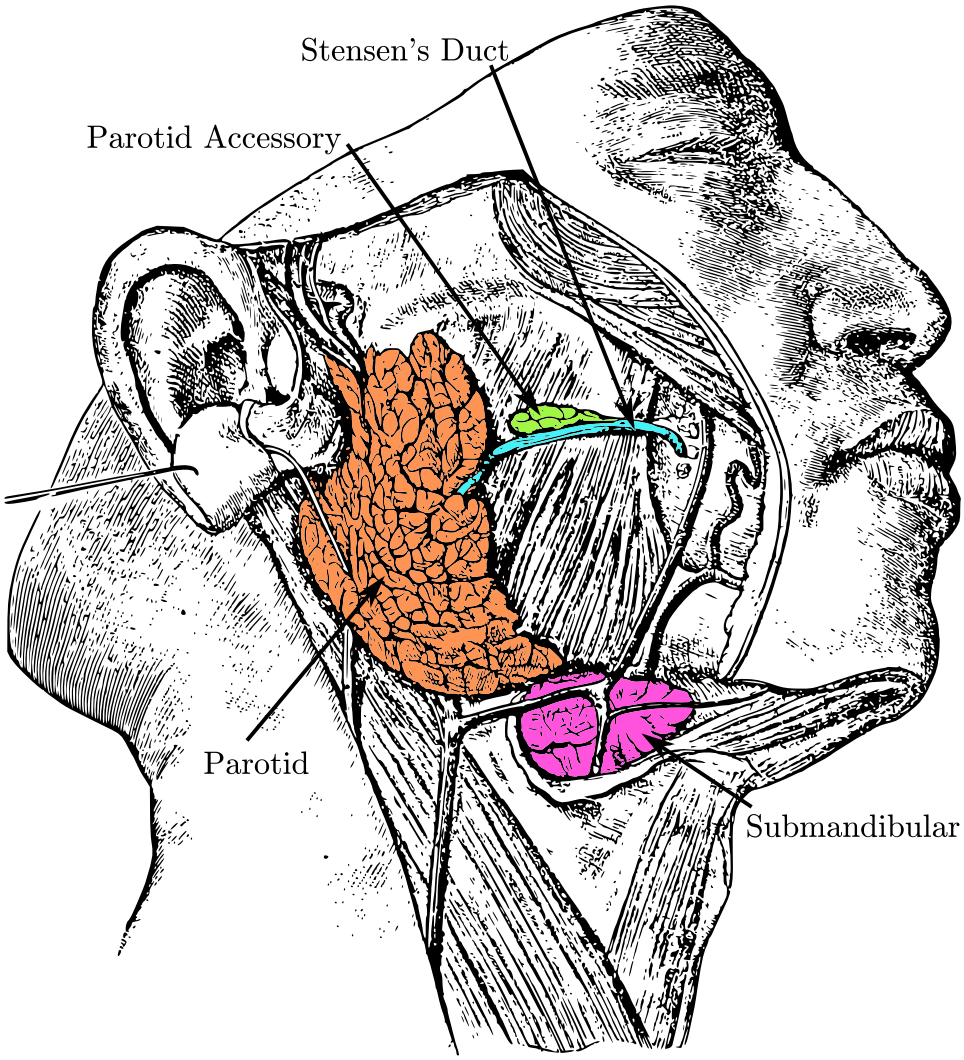

Figure 2.1: Anatomical view of parotid, submandibular, and parotid accessory glands. The ear lobe has been folded so as to not obscure view of the parotid. Image adapted from Toldt and Dalla Rosa [85].

(0.5% altogether) [51]. It is responsible for moistening and softening food, breaking down starch, protecting oral mucosa and teeth, and it fulfills antibacterial functions. Mucins act as mucosal lubricants. Their presence on the mucous membrane surfaces maintains a hydrated state [81]. Total salivary

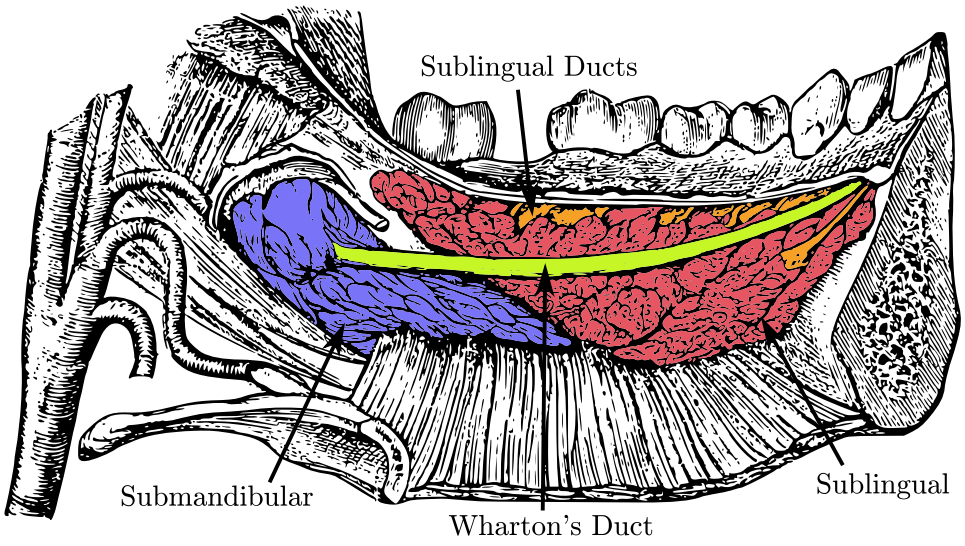

Figure 2.2: Anatomical view of salivary glands medial to the mandible. Submandibular and sublingual glands can be seen, along with sublingual ducts (draining to the oral cavity) and Wharton's duct. Image adapted from Toldt and Dalla Rosa [85].

flow has been estimated to be $1.0 - 1.5\,L$ per day [69].

The composition of the various components of saliva closely follows circadian rhythms. Similarly, the rate of unstimulated whole mouth salivary output varies, with low production occurring during sleep. Peak stimulated whole mouth salivary output occurs daily around 17:00, is lowest at 05:00, and approximately follows a sine wave with a period of one day [22].

2.1.1 Parotid Glands

The parotid glands (figs. 2.1 and 2.3) are the largest salivary glands. They are located in the retromandibular fossa (behind the jaw, below either ear). The parotid is comprised almost exclusively of serous acinar cells, which produce serous saliva. Serous saliva is mainly water in content [17]. It has digestive properties and is not otherwise used for lubrication [22]. The parotids mainly secrete saliva under stimulated conditions, such as chewing, when they supply approximately 60% of total saliva [17].

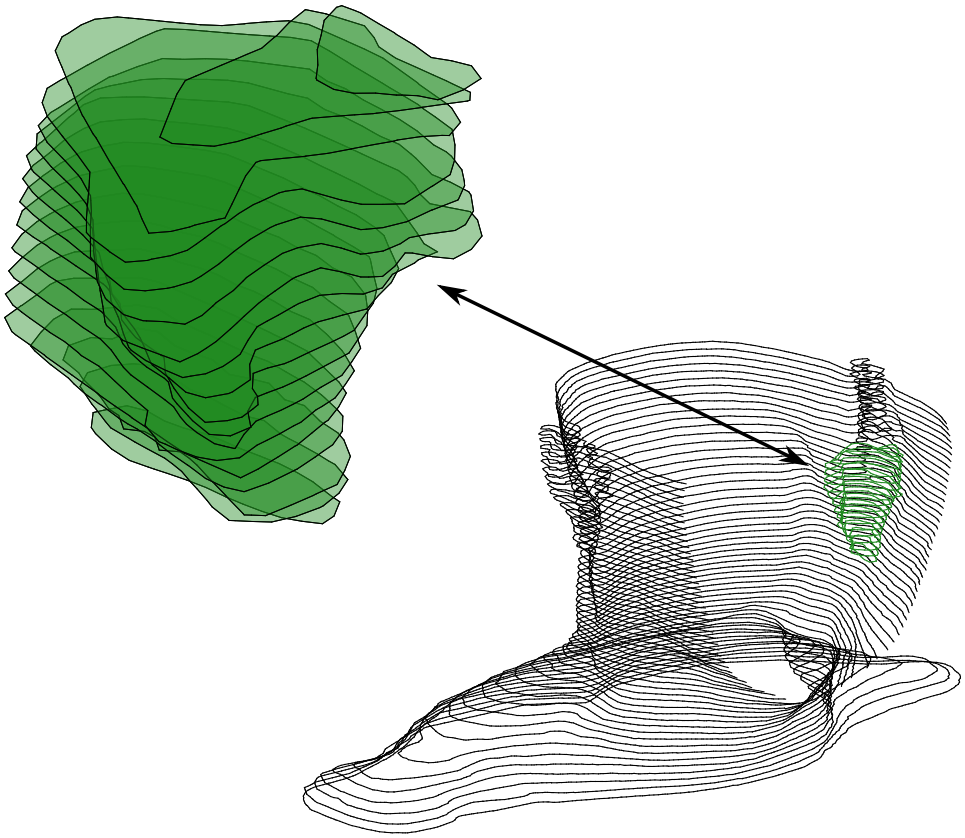

Figure 2.3: Left parotid gland as extracted from patient contours demonstrating location, size, and transversely-inverted pyramid shape.

Anatomically, the parotids house a series of branching ducts through which saliva passes to the oral cavity. The main salivary duct is the Stensen's duct, which drains into the oral cavity at the upper second molar tooth level [17]. The ducts are not visible using conventional Computed Tomography (CT), though they can be imaged using Magnetic Resonance (MR), MR sialography, conventional sialography, scintigraphy, and ultrasonography.

2.1.2 Submandibular Glands

The submandibular glands (figs. 2.1 and 2.2) are the second largest salivary gland pair after the parotids. They are located under the floor of the oral cavity, medial to the mandible [17]. The submandibular is composed of both serous and mucous acini, though mostly serous. The saliva produced is thicker than that of the parotid and is primarily used for lubrication and protection of surfaces [10].

The submandibulars mainly secrete saliva under unstimulated conditions, when they supply $70 - 90\%$ of total saliva. During stimulation, they produce approximately $20 - 40\%$ of total saliva [17].

The main salivary duct is the Wharton's duct, which is about $5\,cm$ in length [69] and drains near the lingual frenula [17] (see fig. 2.2).

2.1.3 Sublingual Glands

The sublingual glands (fig. 2.2) are the smallest of the major salivary glands. They are located in the floor of the oral cavity, above the submandibular glands [17] and between the mandible and genioglossus. Like the submandibulars, they are composed of both serous and mucous acini and produce thicker saliva. Acini are primarily of the mucous type.

Upon stimulation, the sublinguals produce an estimated $2 - 5\%$ of total saliva [17]. Along with the submandibular glands, they synthesize the majority of mucin present in saliva [81].

The anatomical structure is similar to the submandibulars. The ducts drain into the ducts of Rivinus, Bartholin's duct, or directly into the oral cavity [69].

2.1.4 Accessory Glands

The accessory (or *minor*) salivary glands comprise several smaller glands which are distributed throughout the oral cavity. Together, they produce less than 10% of total mucins [17].

2.2 Xerostomia

Worldwide, approximately one half-million new patients are diagnosed with head-and-neck cancer each year [34]. *Xerostomia* is one of the most likely late toxic effects of Radiotherapy Treatment (RT) in patients with head-and-neck cancers [37].

Xerostomia is a symptom of various medical conditions which presents as dry mouth. It is not itself a disease, but results from reduced or absent salivary flow. It can be induced by medication, disease (e.g., Sjögren's syndrome, sarcoidosis, rheumatoid arthritis), or radiotherapy [5].

Salivary dysfunction affects a patient in many ways. Most importantly, basic primary needs such as eating, sleeping, and communication become difficult or painful [62]. Advanced cases are known to severely reduce a patient's perceived Quality-of-Life (QOL) [46].

RT of head-and-neck cancers are complex with multiple organs at risk which should be spared (e.g., brain stem, spinal cord, larynx and pharynx, oral cavity, and parotids). Some structures, such as the spinal cord, *must* be spared, and so irradiating the parotid is often unavoidable. Furthermore, it is common for one or more healthy parotids to obscure a region requiring irradiation – either in the beam foreground or background. Thus, parotids are often irradiated as a geometrical consequence. This is known to cause severe reduction in salivary output [49] [37]. Studies have shown the intensity of salivary gland damage increases in proportion to both the dose received and the volume of gland irradiated [16] [20] [79] [87] [30].

Reports have shown that surgical transplantation of submandibular glands away from the target volume prior to RT can reduce a patient's chance of experiencing xerostomia [73] [2]. Conversely, surgical removal of the submandibular has shown to increase risk of xerostomia [13]. It has also been shown that the mean radiation dose to the accessory salivary glands is a significant predictor of xerostomia [30]. This evidence suggests that xerostomia is not strictly an affliction of the parotid. Nevertheless, the parotid contributes the largest portion of stimulated saliva and therefore imposes considerably more on the patient during salivary dysfunction.

2.2.1 History

Dry mouth was first considered a medical condition in its own right by Bartlet in 1868 [78] [6]. It was given a precise definition and the name *xerostomia* from the Greek 'xeros' (dry) and 'stoma' (mouth) by Hutchinson in 1886 [1]. He described it in detail.

> *The tongue is red, devoid of epithelium, cracked, and absolutely dry, its appearance being like raw beef. The inside of the cheeks, the hard and soft palates, are also dry; the mucous membrane smooth, shiny, and pale. The salivary glands appear normal, and no mechanical obstruction has been detected in their ducts.* [...] *Articulation is difficult in consequence of the absence of moisture, and swallowing has to be assisted by constant sipping.* [...] *The disease reaches its greatest intensity suddenly, and then remains without change for years.*

Radiation-induced xerostomia was described as early as 1938 by Martin. It was likely known prior, though, as it was accompanied by reports of secondary complications and suggested remedies [52].

2.2.2 Endpoints and Grading

Patient-reported xerostomia, by definition, is subjective. Salivary flow can be quantified, though precise quantification is difficult. Although salivary flow is often used as a surrogate measurement of xerostomia, the correlation between patient reported xerostomia and salivary gland function (as measured by salivary flow) has been found to be weak [30] [36] [45] [37] [17] [46]. This could be the result of many factors: large variations in measured salivary flow rates, discrepancy between salivary output and mucosa hydration [30], or changes in patient perceptions of oral dryness following RT [36] being the most intuitive. Others have found significant correlation between salivary gland function and QOL questionnaire scores [10] [3]. Regardless, care should be taken to differentiate xerostomia and salivary function; the distinction is frequently blurred in the literature.

The analytic component of the Late Effects Normal Tissue - Subjective, Objective, Management, Analytic (LENT-SOMA) scale is commonly used to objectively measure reduction of baseline (pre-treatment) function. Severity is divided into four grades, the worst of which (grade IV – "severe xerostomia") comprises reduction of 75% or more of whole mouth baseline salivary output. Many reports do not precisely follow this scale, instead measuring single-organ (or single-organ-pair) output [54]. Thus, reports specifying grade IV xerostomia often differ subtly in conclusion.

Several grading systems exist. The LENT-SOMA, Common Terminology Criteria for Adverse Effects (CTCAE) [86], DAHANCA [41], and Radiation Therapy Oncology Group (RTOG) [21] systems are commonly encountered. Precise translation between systems is not always possible. Jensen et al. [41] is a good reference for learning more about such systems. The LENT-SOMA scale should generally be preferred as reports indicate it is better able to score late radiation toxicity and subjective xerostomia [38] [25] [7].

Imaging endpoints involving scintigraphy [15], radioisotope examination, conventional sialography [4], and MR sialography [91] have been shown to be effective means of quantifying xerostomia. Scintigraphy and radioisotope examination are unsuitable for continued observation, however, due to increased radiation exposure for the patient.

2.2.3 Complications

Xerostomia is known to have a strong, broad impact on a patient's QOL, even imposing on domains not directly related to xerostomia, such as emotion and pain [46] [66].

Vissink et al. [90] notes the onset of secondary complications following xerostomia: mucositis, hyposalivation, loss of taste, osteoradionecrosis, radiation caries, and trismus (spasm of the jaw muscles causing the jaw to remain tightly shut). Hyposalivation will often lead to increased opportunistic infection of the oral cavity.

Martin [52] described the particular severity of dental caries.

> *A complication occasionally associated with radiation xerostomia is a peculiar form of dental caries.* [...] *Beginning about two or three months after irradiation of the pharynx or of the oral cavity, metal fillings and inlays tend to loosen and fall out. In the cases of greater involvement the teeth may lose their natural glistening appearance and assume a dull, chalky hue. The substance becomes rather brittle and may wear away at the occlusal surfaces. Numerous cavities develop, especially near the gingival margin, so that the teeth tend to crumble or break off, leaving the root exposed at the gum level. Toothache is a prominent symptom. The direct cause of these dental complications is somewhat obscure.*

The modern belief is that radiation caries are caused by hyposalivation, though it has been suggested that diffuse radiation on the bones of the jaws and a reduction of blood supply through apical arteries could be to blame. Due to the severe impact on general health, early recommendations included extraction of teeth prior to irradiation. This was generally to the detriment of the patient as osteoradionecrosis and osteomyelitis of the mandible often followed extraction [24].

Modern preventative measures include heightened oral hygiene practices, frequent application of fluoride solutions, limitation of cariogenic foods, and application of artificial salivary agents. Though a patient may be able to avoid such complication through diligence, effectiveness is limited by the patient's tolerance and rigor [90]. Especially so for those patients in which xerostomia persists indefinitely.

The issue of whether to extract teeth prior to treatment is still debated. It is generally accepted that tooth extraction is warranted in cases of teeth with a questionable prognosis or patients with questionable motivation [71] [84] [90].

2.2.4 Function Loss and Recovery

The reduction of salivary flow occurs rapidly after RT begins [10]. Conventional fractionated RT of head-and-neck cancers often results in a marked decline of

function, even if the patient perceives few symptoms.

Xerostomia can be induced by medication. When it is, it typically subsides when medication is halted. Disease- and radiation-induced xerostomia may persist briefly, for years, or indefinitely [5].

Though the nature of functional loss is not understood, its occurrence is well-known. Reports of recovery are mixed – indeed, individual accounts of recovery are often mixed. Mossman et al. [58] remarked on the matter.

> *Results of studies of the time and extent of recovery of normal taste and salivary function in man following radiotherapy are contradictory. Several investigators have observed complete recovery of taste and salivary function in patients 1-3 months following treatment, whereas others have not. Although little or no improvement in salivary function has been observed in some patients at least two years following curative courses of radiotherapy, partial recovery 8 months after radiotherapy has also been reported.*

Eneroth et al. [32] reported the case of a patient irradiated with $65\,Gy$ whose parotid tissue remained functioning nine years after RT, suggesting a recovery mechanism which can withstand high doses.

The mechanism of radiation damage in the parotid is poorly understood. Studies show that the magnitude of salivary gland damage in pig, rat, and human increases in proportion with dose and irradiated volume [16]. Damage to the parenchyma is thought to reduce salivary function, though it is unclear what portion of loss is due to serous or mucinous acini, intercalated or striated duct, or excretory duct damage. Indeed, vascular changes have been observed, beginning with periarteritis (inflammation of the outer arterial coat) and endarteritis (inflammation of the inner arterial lining) which eventually progresses to fibrosis of the tunica intima. Eventual destruction of the lumena of blood vessels results, to the detriment of nearby bone [76].

Both serous and mucinous acinar cell atrophy occurs at a discernible level, though serous cells appear to be less capable of survival during RT. Saliva in patients with salivary dysfunction is generally thick and discolored [78] – supporting the hypothesis that the loss of serous acini (with or without

survival of mucinous acini) is key to xerostomia. Finally, a reduction in gland volume generally occurs over the course of RT.

It is most likely that xerostomia is initiated by several factors, some objective (e.g., salivary flow, mucous content) and some subjective (e.g., patient perception of oral dryness).

Chapter 3

Relevant Literature

> *The first process* [...] *in the effectual study of sciences must be one of simplification and reduction of the results of previous investigations to a form in which the mind can grasp them.*
>
> James Clerk Maxwell

3.1 Review of Relevant Topics from the Literature

3.1.1 Noise in Salivary Measurements

Consistently, studies investigating salivary function report the presence of non-negligible noise.

In a comprehensive study of the variability of salivary flow, Burlage et al. [14] noted that standard deviations of approximately 24% should be expected for measurement of whole mouth stimulated flow. Blanco et al. [10] performed a small experiment on five healthy volunteers with the intent of evaluating the reproducibility of measurement of baseline output. They encountered a standard deviation of 27% and suggested that variability in patients may be higher due to disease or other comorbid conditions. A study by Bergdahl and Bergdahl [8] involving 1427 healthy volunteers of varying sociodemographic

backgrounds found 44%. Other reports generally indicate standard deviations between $30 - 40\%$.

Stimulated salivary flow is reasonably stable over the span of two hours. One would hope to exploit this stability to pool measurements and thus reduce variability. Unfortunately, it appears that increasing the number of collections has a negligible effect [14].

Though noise appears to be inherent to the domain of salivary measurement, reports have shown that the variability of the stimulated saliva production rate does not increase appreciably with age. Unstimulated flow rates generally decrease, however, and this contributes negatively to oral health [35].

3.1.2 Choice of Salivary Flow Model

The choice of salivary flow model and the factors considered are tightly coupled – but nevertheless distinct – issues. Many models have been considered in the literature.

Early reports used the Lyman-Kutcher-Burman model, which uses a power law with an exponent that controls the volume effect and was applicable only for single-organ contributions [31]. Many have investigated the use of equivalent uniform dose (i.e., generalized mean) with reasonable success.

Blanco et al. [10] considered six distinct whole mouth saliva models, each motivated by physical arguments. All involved an exponential suppression of salivary flow with increased dose. The data was best described by a parallel-exponential model (which considered the parotid to be composed of independent functional subunits), though a simplistic mean dose exponential was only slightly worse. All models considered were found to exhibit highly significant correlations with measured saliva flow rates. The use of the simple models were advocated.

El Naqa et al. [31] compared several techniques for determining model complexity. Except in the case of overfitting, multivariate models were found to be more predictive than the univariate variety. Though, overall, multivariate models performed only slightly better than univariate. Houweling

et al. [39] found univariate models to be preferable.

The use of univariate models has been suggested for simplicity arguments. Without additional information or substantially more patient measurements, it seems the inherent noise in whole mouth salivary measurement will continue to blur the distinction between models [10].

3.1.3 Relevant Clinical Factors

Many factors have been conflated in salivary output prediction models. The use of multivariate logistic regression and other model parameter selection techniques are widely used.

Using multivariate analysis, Chao et al. [16] investigated correlation between salivary function and irradiation dose to the parotid glands. They found that the rate reduction of stimulated salivary flow at six months post-RT was not significantly influenced by a patient's gender, age, tumour stage, radiation technique (IMRT vs. non-IMRT), or concurrency of chemotherapy. A derived toxicity score from dose to the parotid was found to be the sole significant factor for xerostomia.

A similar analysis involved patient age, gender, ethnicity, date of treatment start, treatment technique (IMRT alone vs. otherwise), treatment aim (definitive vs. postoperative RT), Karnofsky Performance Status (KPS), chemotherapy, tumour stage, treatment duration, histologic features (squamous vs. other), tumour subsite (oropharyngeal vs. non-oropharyngeal primary), and the logarithm of the mean dose-exponential model prediction. The mean dose-exponential model was the most independently significant factor, followed by considerably less significant gender and KPS factors. If the dose-exponential model was replaced with other, closely related, models for regression, the KPS factor was replaced by the chemotherapy factor as third-most significant. Overall, the predictive power of the model was predominantly due to the dose-volume term [10].

In a multivariate study, El Naqa et al. [31] demonstrated a consistent preference of variables when constrained to five factors (only): mean dose, gender, KPS (the three most significant), and technique and treatment aim

(the two of considerably lower significance than the previous three). They therefore recommend using a simplified model with fewer factors.

Teshima et al. [83] found a correlation between decreased parotid gland volume and decreased saliva production following RT. They noted that no such correlation existed between *total* volume and *total* salivary output following RT.

More recently, using Bayesian multivariate logistic regression Buettner et al. [13] considered volume of the parotids, mean dose to submandibulars, surgical removal of the ipsi-lateral submandibular, gender, tumour site (hypopharynx or oropharynx), age, use of neo-adjuvant chemotherapy, and the presence of hypertension. Factors relating to dose were found to be most significant, though removal of the submandibular was found to significantly increase the risk of developing xerostomia. Morphological dose susceptibility was found to play a significant role in the onset of xerostomia.

3.1.4 Contributions from Accessory Glands

Dose to the oral cavity, and thus to the accessory salivary glands, has been found to be both significant [30] and insignificant [10] as a predictor for xerostomia.

Most reports differ subtly in terms of the dose delivered to salivary organs. Blanco et al. [10] made no attempt to spare submandibular or oral cavity glands. Patients for which glands could be spared were excluded from their study. Unsurprisingly, they found little link between non-parotid salivary gland dose and xerostomia. Similarly, Chao et al. [16] intentionally only considered patients whom had received more than $50\,Gy$ to the submandibulars. Findings were similar. Nishimura et al. [62] found no statistically significant mean dose dependence, though no mention of submandibular sparing was given.

Studies in non-human salivary glands have had higher control over non-parotid dose [89], though it is uncertain to what degree the results are applicable to humans.

Attempts to model non-parotid salivary contributions to stimulated flow,

including compensation, have generally been inconclusive. It has been suggested that even though compensation has not been found to be a significant factor for xerostomia, it has not yet been ruled out completely [10].

3.1.5 Quality of Life Questionnaires

Whether salivary output correlates with xerostomia is somewhat controversial (see section 2.2.2), most likely due to the subjective, symptomatic nature. However, many reports indicate a strong correlation between QOL questionnaires and salivary measurements [16] [46] [65] (cf. section 2.2.2). Meirovitz et al. [53] advocates the use of patient-reported xerostomia over direct measurement. Still, others report the converse [42].

Most report a rapid decrease in patient QOL shortly after RT and a gradual, but steady, improvement over the following year. It is unclear if this is related to salivary output recovery or is due to alteration of patient perceptions of oral dryness. If the latter, it would seem that salivary flow rate alone would be insufficient for adequate prediction of xerostomia.

3.1.6 Collection Technique

Early salivary measurement techniques collected whole mouth salivary output over the span of several minutes. Saliva was directed into a collection cup under stimulated or unstimulated conditions. Stimulated conditions were induced by chewing a small paraffin block. This basic procedure has been used by many [16] [10] [68].

Whole mouth saliva is contributed from both major and accessory salivary glands. In particular, both left and right variants of organs contribute. Disease, damage, or other factors may lead to an unbalanced contribution from individual organs, and thus whole mouth saliva measurements convolute the signal from each.

Attempts have been made to measure individual parotid gland salivary flow using small suctioned collection cups (e.g., Lashley cups). Compared with whole mouth salivary measurement, discrepancies in dose response were noted [30] [16]. It is unknown to what degree such discrepancies are due

to differences in collection technique [59]. Indeed, it is unclear if individual parotid flow measurement is more or less relevant for objectively measuring xerostomia. Further research is needed, though it is likely that data from many patients will be required due to the inherent noise in salivary output measurements.

3.2 Research Trends

Complication risk is an active field of research. Three-dimensional planning has enabled the era of individualized, patient-specific treatment planning. At the British Columbia Cancer Agency (BCCA), within the last ten years, IMRT has gradually supplanted previously-common RT techniques. Compared with traditional techniques, IMRT allows the delivery of both more complex and more precise dose distributions. The surplus of flexibilty in planning allows complication risk to assume a more critical role in the treatment planning process; planning target volumes are still given the prescription dose, but the surrounding normal tissue, in some cases, can now be selectively spared.

In the early years of the past decade, research tended to focus on the recognition of factors which could be used to more precisely assess the risk of xerostomia. Chao et al. [16] and Blanco et al. [10] performed multivariate analysis to search for such factors. These attempts mostly investigated clinically available information such as gender, concurrent chemotherapy, and predictions from plausible dose models. Mixed success was interposed amongst general failure to identify such factors (see section 3.1.3), though mean dose appears to have nearly unanimously been found to strongly predict salivary output loss.

Later, El Naqa et al. [31] took a more high-level approach by evaluating the process by which models are synthesized from the data. Emphasizing the underlying complexity and apparent non-linearity of salivary function, they used information-theoretic techniques to provide guidelines for model generation. By focusing not on the specific values of parameters, but on the problem of determining *how many* parameters a good model would have, they transitioned away from the parameter-determination mentality. Others

would follow suit throughout the decade.

Amidst these efforts, many more direct xerostomia incidence-reduction strategies were investigated. Chemotherapy agents (e.g., amifostine, pilocarpine) were found to benefit patients. Likewise for surgical transfer of submandibular glands (see section 2.2). Improvements in patient immobilization, demanded by the higher precision of dose delivery, allowed for more precise dose sculpting. All the while, improvements in computation dramatically followed Moore's law, improving planning software, analysis capabilities, and, ultimately, insight.

Gradually, as computation became more powerful, cheaper, and more prevalent, researchers recognized the importance of taking a data-driven approach to investigating xerostomia, with some claiming that the construction of "potentially increasingly complicated models" would be an effective way to approach outcomes model building. In particular, it was recognized that a data-driven approach appears necessary for the prediction of multi-factor complications like xerostomia [31].

In the latter part of the decade, the shortcomings of various models describing dose response became highly scrutinized. The failure of Normal Tissue Complication Probability (NTCP) models in animals were investigated by van Luijk et al. [88]. Dijkema et al. [27], using ten years of patient data, reported failure of mean dose based models to fully describe the effects of radiotherapy on parotid glands. Later, dose volume dependence in the rat parotid gland was found [89].

Recently, researchers have attempted to transplant stem cells into dysfunctional parotid glands following RT with some success [34] [47] [80]. Better pathological understanding of the underlying nature of xerostomia could lead to identification of more relevant factors pertinent to xerostomia. Alternatively, identification of the regions of the parotid particularly susceptible to radiation damage may provide more accurate targets for implantation and could therefore improve regrowth efforts.

Modern research trends tend to focus on either analyzing large pools of data or investigating specific factors. In both cases, the level of sophistication of analysis tools has increased. Houweling et al. [39] pooled data

from multiple centres for a study involving 384 patients. Their aim was to validate a collection of models. Using sophisticated analysis techniques, including Bayesian multivariate logistic regression and statistical moments, Buettner et al. [13] found an increase in predictive power using morphological factors which had been previously opaque to those considering only mean dose. Without the use of sophisticated statistical techniques and ubiquitus computational power, the discovery of such dependence seems tedious at best and, at worst, perhaps improbable.

3.3 Clinical Recommendations

Central to the search for predictive factors and relevant models was the need to establish practical clinical recommendations.

In 1999, Eisbruch et al. [29] recommended a mean dose $\leq 26\,Gy$ in $30-35$ fractions to permit substantial sparing of the parotid. In 2003 Amosson et al. [3], based on patient-reported QOL questionnaires, found that patients felt they had 'too little' saliva when the contralateral parotid received a mean dose $\geq 22.5\,Gy$. Six years after Eisbruch et al. [29], in 2005, Blanco et al. [10] recommended a mean dose $\leq 25.8\,Gy$. Variation in recommendations have since appeared to be minimal, though many have been given [54].

In 2010 a joint effort by many researchers, authors, reviewers, and support personnel provided comprehensive summaries of the available dose-volume and outcomes data and accompanying clinical recommendations referred to as the Quantitative Analysis of Normal Tissue Effects in the Clinic (QUANTEC) guidelines. For parotids, $< 20\,Gy$ mean dose to the contralateral gland or $< 25\,Gy$ mean dose to both is advised. It is currently recommended to follow the QUANTEC guidelines [23].

3.4 Notable Findings

In this section we detail results which are of particular relevance to the present report.

Perhaps the most important finding is the apparently unavoidable noise in salivary output measurements (see section 3.1.1). In a comprehensive study

on the variability of salivary measurements, Burlage et al. [14] found that increasing the number of collections has a negligible effect on the reliability of baseline measurements. Further, repeated collections did not result in a significant reduction of intrapatient variation. This unfortunate fact appears unavoidable, and so it is with importance that the number of patients enrolled in studies requiring salivary function measurements be sufficiently numerous to overcome the variability.

In spite of the noise, mean dose models have been found to be most practical for predicting the onset of xerostomia. Still, the failures of the model are slowly becoming apparent. The subject of recovery from xerostomia is substantially less investigated than its onset, most likely due to historically mixed reports and limited recovery (see section 2.2.4), especially with conventional RT. After the introduction of IMRT, increased recovery was noted by many but no apparent functional or dose dependence has been described [46] [10]. However, studies with rat parotids have shown the regional dose dependence of loss of salivary output [89]. Further, the findings of Buettner et al. [13] of the morphological dependence of loss hold promise that the same mechanisms may provide a similar dependence in recovery.

Finally, one must question the utility of the basic premise: the characterization of complication risk. The introduction of IMRT has provided the impetus for more precise determination of complication risk – but has it led to an improvement in patient outcomes? A recent study by Mortensen et al. [57] showed that incidences of both xerostomia and dysphagia have been reduced following the introduction of IMRT. Another report by Rathod et al. [66] found that by using IMRT, the magnitude of QOL impairment in most domains was lesser, and recovery was both more rapid and more complete, compared with previous RT techniques. Furthermore, a *tripling* of cancer survivorship occured in the United States from 1970 to 2001 [7]. Causality aside, complications are playing an increasingly important role in patient's lives; the continued research effort appears to be warranted.

Chapter 4

Statement of Research Questions

"Let us descend now into the blind world,"
Began the Poet, pallid utterly;
"I will be first, and thou shalt second be."

The Divine Comedy
Dante Alighieri

4.1 Statement

The goals of this thesis are to address the following questions.

1. Is the parotid inhomogeneously responsible for xerostomia? Will delivering dose of equal magnitude to different regions of the parotid lead to an appreciable difference in risk of xerostomia?

2. Is recovery of salivary output dose-dependent? If so, is it likely to play a significant role in subjective xerostomia?

If dose-dependence is found in either case, an additional goal is to quantify it.

Studies with rat parotids have shown regional dose dependence of salivary output loss [89]. Similarly, morphological dose dependence for grades II to

IV xerostomia has been demonstrated in humans Buettner et al. [13]. It is presently unknown to what degree these results indicate regional dose dependence of salivary output loss and recovery in humans.

Statistically significant mean recovery has been observed by many. Dose-dependence of recovery has, to our knowledge, not been reported. Nor have the techniques we propose to use; sub-segmentation of organ contours. We believe such a technique is the most appropriate device for conclusively examining regional dose-volume dependence and quantifying salivary output.

Answering the stated goals will lead to a stronger estimate of the risk of xerostomia in head-and-neck cancer patients. Application of the findings, via IMRT, should lead to immediate improvement in outcomes if regional dependence is found.

4.2 Outline of Approach

Due to the nature of the proposed analysis, a requirement for conducting our investigation was the development of a computational system (`DICOMautomaton`) for the automated examination of dose and outcomes data. Novel aspects of its operation are described in chapters 5 and 6.

Chapter 5 demonstrates semi-automated contour recognition, which is necessary for combing through patient data without having to perform manual identification on each contour. Chapter 6 demonstrates the sub-segmentation capabilities available in `DICOMautomaton` which are used to regionally demarcate organ contours into sub-volumes. An overview of available dose-volume methods is also given.

With the requisite machinery in place, we involve patient outcomes data and perform analysis of salivary output loss and recovery in chapter 7. The unknown nature of dose response in (sub-)segments and the expectation of significant noise warrant the use of statistically robust non-parametrical, resampling, and random sampling techniques, which are discussed in detail in chapter 7.

Chapter 5

Contour Recognition[1]

Obscure, profound it was, and nebulous,
So that by fixing on its depths my sight,
Nothing whatever I discerned therein.

The Divine Comedy
Dante Alighieri

The search for regional organ dose susceptibility demands robust, flexible, and composable contour sub-segmentation methods. If such methods are able to be driven in an automated fashion, feedback-driven sub-segmentation can be employed. Unfortunately, the ever-growing volume of dosimetric patient data accumulated by medical centres is varied not only in terms of quality, scope, and format, but also in a more subtle way; each center has a unique naming and contouring *dialect*. This variety makes contour recognition difficult to automate.

While the Digital Imaging and Communication in Medicine (DICOM) standard addresses the problem of interoperability between various types of hardware and software, it implicitly encourages the development of local vernacular. In particular, the flexibility of contour generation, specification, and labeling presents a challenge for systematic identification. For example,

[1]The content of this chapter is currently in press with the title *Semi-automated contour recognition using* `DICOMautomaton` [18].

60 patient contour sets from one centre over one year contained more than a dozen labels indicating the left parotid. In some cases both parotids were contoured into a single 'parotids' structure while in others the parotids had not been contoured - sometimes because they had been surgically removed.

An experienced researcher could reliably identify a left parotid contour, even if it were misshapen or mislabeled. However, it is not conceivable to perform manual identification in analyses involving hundreds of patients. In order to surmount this difficulty and enable automated recognition of contour data, a set of practical lexicographical and geometrical techniques have been developed for use in `DICOMautomaton`.

5.1 Methods

We consider the situation whereby a researcher has a collection of existing contour data of mixed origin and a set of structures (e.g., organs) which are to be identified. An example might be locating all left parotids, taking into account that not every structure set is required to contain a left parotid and that unknown aliases may be used.

5.1.1 Lexicographical Measures

We define a *lexicon* l to be the collection of exact mappings from a set of raw labels d_i to a unique label c_i such that $l(s) = c_i$ *if and only if* s is a label in d_i. We denote the set of all d_i as d and the set of all c_i as c. The lexicon represents the researcher's domain knowledge and is similar to a thesaurus. For instance a limited, toy lexicon for head-and-neck cancers might be

$$\begin{aligned} d_1 &= \{\text{l parotid}, \text{lt\_par}, \text{Left Parotid}\} & c_1 &= \text{Left Parotid} \\ d_2 &= \{\text{r partd}, \text{r\_par}, \text{Right Parotid}\} & c_2 &= \text{Right Parotid} \\ d_3 &= \{\text{chiasm}, \text{opt\_chiasm}\} & c_3 &= \text{Optic Chiasm.} \end{aligned}$$

Using this lexicon, one would be unable to identify a structure named "left par" though it is clear that it most closely relates to d_1. Natural language requires interpolation to determine the meaning of a word in a given context. To capture this intuitive notion of similarity we model group membership

using string similarity measures, which consider two inputs similar if they share specific features.

The similarity between strings s_1 and s_2 for measure J is written as a score $S_J(s_1, s_2) \in [0, 1]$ where $S_J = 0$ denotes no similarity and $S_J = 1$ perfect similarity. Scoring allows us to incorporate the similarity of strings from many measures. It is tempting to treat S_J as the probability that $s_1 = s_2$ or that they are logically linked, but to do so is erroneous; Perfect similarity scores do not necessarily indicate perfect matches. At best, similarity scores can be treated roughly as estimators of the probability of a match. Given a lexicon l and input s which refers to an unknown member of c, we choose a combining function F to weight the similarity scores with the elements of d_i into a total score for c_i

$$T_{c_i}(s) = F\left[\left\{S_J(s, \hat{s}) \,\middle|\, \hat{s} \in d_i\right\}; l\right] \in [0, 1].$$

Typically, F is chosen to help reduce statistical uncertainty by producing a high $T_{c_i}(s)$ when many measures produce high scores and a low $T_{c_i}(s)$ otherwise. For well behaved general measures, F could be a simple mean. Denoting the set of all $T_{c_i}(s)$ as $T(s)$, the suspected translation $\tilde{c}$ is determined by choosing a mapping function $M(T(s)) = \tilde{c} \;\middle|\; \tilde{c} \in c$. Both F and M are subject to the desirable property that if $s \in d_i$ then $\tilde{c} = c_i$ as one would find using the lexicon without any consideration of similarity. We can judge the validity of $\tilde{c}$ using $T(s)$.

Due to wide diversity of contour labeling encountered, F and M were equipped with tunable minimum thresholds to help reliably control type I and II errors. For this study, F was implemented as a weighted-average while M was taken to be a `maximum` function with additional degeneracy-handling logic. Further discussion regarding the recognition system is provided in appendix A.

Well-known lexicographic measures implemented in `DICOMautomaton` include Levenshtein-Damerau [44] and Jaro-Winkler [93] edit measures, Soundex [43], Double Metaphone [64], and Match Rating Approach [56] phonetic measures, statistical measures including longest common substrings/sequences,

N-grams of user-specified order N (character or word based in a variety of flavours, including Dice [26], Jaccard [40], and cosine metrics), and a generic bag-of-characters measure. The Levenshtein-Damerau measure is popular due to its speed and simplicity. It assumes similarity can be modeled by counting the minimal edits required to transform one string to another, and therefore works particularly well dealing with typographical errors and some abbreviations. The Jaro-Winkler measure performs similarly but somewhat complementary [94]. Alternative lexicographical measures implemented include an artificial neural network-based measure, a self-orthogonalizing measure which ignores all common N-grams by elements of d, a custom measure (which we refer to as DICOMhash) which differentiates labels which are liable to confuse other techniques (e.g., 'CTV+3mm' vs. 'CTV+5mm'), and a domain-specific head-and-neck measure which hard-codes an individual centre's naming conventions.

5.1.2 Geometrical Measures

Geometrical methods considered are similar to the aforementioned lexicographic methods; instead of a string input, they take a set of ordered two-dimensional contours embedded in $\mathbb{R}^3$. They rely on geometrical features, but otherwise serve an identical purpose to the lexicographic measures. Geometric measures implemented include probability spheres and boolean overlap relative-position measures, Fourier descriptor and eccentricity shape-based measures, a domain-specific measure, and simple feature measures involving volume, perimeters, planar surface area, and centroids [63].

The most intuitive measures involve the spatial location of objects in $\mathbb{R}^3$, so we describe here probability spheres to demonstrate the basic approach: in normalized coordinates, each target structure is given a centre, an effective radius, and a radially-dependent normalized weighting $W(\vec{r})$. The purpose of weighting is to prescribe inhomogeneous regional similarity. The similarity score between structures is evaluated as $\int W_1(\vec{r})W_2(\vec{r})d\vec{r}$. The more structures overlap in $\mathbb{R}^3$, the more likely we are to think of them as similar.

5.1.3 Cross-Validation Lexicon Folding

Cross-Validation Lexicon Folding (CVLF) was used to estimate the overall effectiveness of the system. CVLF involves randomly choosing a fraction f of d and measuring the ability of the system to correctly translate the entirety of the complete lexicon. This is loosely analogous to measuring the ability of an individual to reconstruct a complete foreign vocabulary after being given only a sample portion.

Random selection of a portion of the lexicon may result in omission of c_i, yielding a system artificially unable to recognize any of the corresponding d_i. Where applicable we correct for this effect. Denoting the number of elements in d as N_d, omission of fN_d elements reduces the maximum CVLF recognition rate, on average, to $1 - \sum_i P(c_i) N_{d_i}/N_d$ where

$$P(c_i) = \frac{\Gamma(N_d - N_{d_i} + 1)\,\Gamma(N_d - fN_{d_i} + 1)}{\Gamma(N_d - (f+1)N_{d_i} + 1)\,\Gamma(N_d + 1)} \quad \Big| \quad fN_d \leq N_d - N_{d_i}$$

and is otherwise zero. To evaluate system performance, contour label data from 60 head-and-neck cancer patients was used for CVLF, producing a lexicon composed of $N_d = 325$ and $N_c = 18$ elements denoting 16 unique structures and two honey-pots (i.e., for interception of artifacts). The honey-pots comprised 146 raw labels and the remaining 179 were distributed with an average N_{d_i} of 11.2 ($\sigma = 5.1$).

5.2 Results

M and F thresholds were found to greatly affect the recognition and error rates. Both recognition and type I error rates increased when decreasing thresholds. Type II error rates increased when increasing thresholds. Reasonable default values were found to be 0.3 for F and 0.5 for M.

Results of CVLF for various retention fractions f are shown in figure 5.1. We estimate that in typical circumstances a user would possess between $50 - 80\%$ of d and $75 - 100\%$ of c from the complete lexicon. Focusing on an information-deficient situation where $f = 0.3$ (i.e., 30% of d are known), the Levenshtein-Damerau measure performed successful recognition 76% of the

time with type I errors occurring 23% of the time. The use of several (mixed) lexicographic measures improved the raw recognition rate 7% and reduced type I errors 6%, representing an effective overall improvement in more than 10% of recognition attempts.

Geometrical techniques performed similar to lexicographic techniques. In general, they produced both lower recognition and error rates compared with lexicographic techniques. Remarkably, the type I error rate did not increase above 7% at any f using mixed geometrical measures (probability spheres, perimeter length, a lateral-position discriminator, and centroid comparison). At $f = 0.3$, mixed geometrical measures performed successful recognition 71% of the time. For probability spheres, it was found that an appropriate effective structure radius was $r_{eff} = 3\,(3V/4\pi)^{1/3}$ where V was the structure's volume. The preceding factor of 3 was chosen to help handle variations in position due to patient geometry and orientation.

Domain-specific lexicographic measures which hard-coded naming conventions performed dramatically better, achieving an optimal recognition rate of 97% when $f = 0.3$.

5.3 Discussion

The use of mixed lexicographic measures resulted in an effective overall improvement in more than 10% of recognition attempts compared with the common Levenshtein-Damerau approach.

Although geometrical techniques generally showed reduced error rates compared with lexicographical techniques, they required considerably more processing power, memory, and storage to use. Furthermore, they introduced subtle dependencies on data fidelity and were somewhat unwieldy. Notwithstanding, when f is low (i.e., the domain is not well-known), the reliably low type I error rates make geometrical techniques a powerful tool to have in the arsenal.

Domain-specific measures performed optimally in recognition. Unfortunately, they are by nature often incapable of handling previously-unseen input, and tended to produce the highest possible type I error rates at a

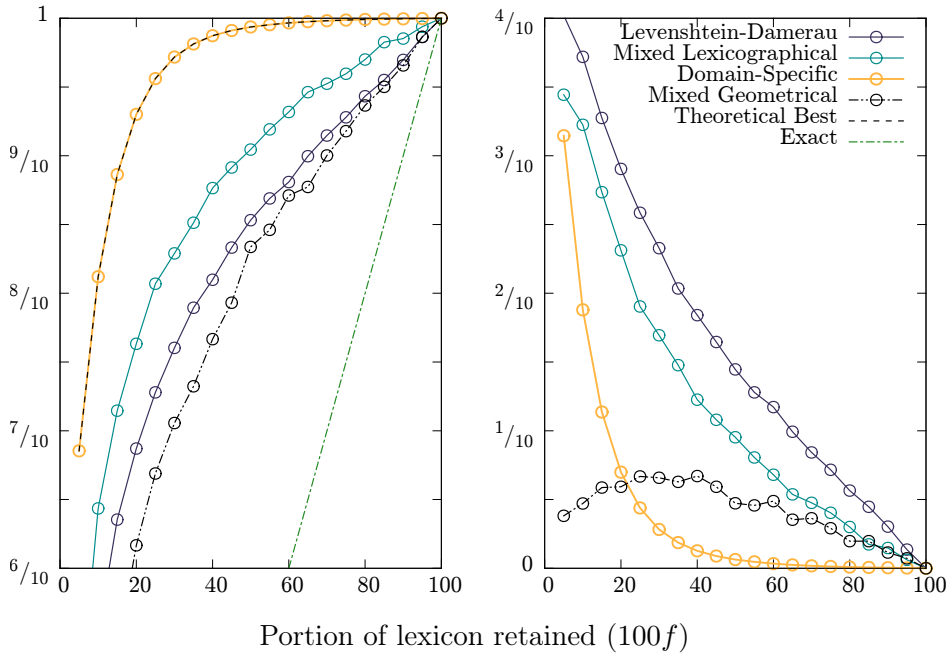

Figure 5.1: Comparison of select measures. Recognition (left) and type I error (right) rates versus lexicon retention during CVLF. Also shown are the optimal recognition and exact match rates. The latter refers to application of l without the use of similarity measures. The mixed lexicographical measure contains Levenshtein-Damerau, Jaro-Winkler, DICOMhash, N-gram, and bag-of-characters measures. The mixed geometrical measure contains probability spheres, perimeter length, a lateral-position discriminator, and centroid comparison measures. The domain-specific (lexicographic) measure achieves optimal recognition.

fixed recognition rate. If required, they could be made more selective, but the effort to do so may eclipse the overall utility of contour recognition in cases. It is unclear how to appropriately gauge their performance during real usage. A balance between the uncertainty of domain-specific measures and the less-performant, but more robust, lexicographic and geometrical measures can be achieved with appropriate mixing. However, an estimate of the portion of the lexicon which is known (f) is required for reliable weighting. This may

be difficult to assess in an unknown or broad domain.

Hybrid techniques further improve system efficiency by involving meta-information. For example, simultaneous recognition on a collection of mutually exclusive input can ensure that the system does not erroneously detect two spinal cords in a single patient. Another useful technique is realtime lexicon building; similar to a spell-checker, if the system is unable to produce a prediction with strong confidence, the user is queried and the input is inserted into the lexicon permanently. This is most useful when f is low (i.e., the domain is not well-known). Such techniques were tested but are not described here.

Actual usage indicates performance higher than that suggested by CVLF. We believe this is due to CVLF randomly removing elements of d; in reality some labels are encountered frequently (e.g., 'body') while others are rare. This regularity provides increased recognition power.

Finally, we hope to have demonstrated that recognition is, in some cases, a flexible way to handle contour data. It rejects the rigidly-defined database paradigm and encourages data sharing across labeling and contouring conventions; instead of prospectively requiring radiotherapy staff to use a fixed lexicon or certain phrases, our approach encourages the use of local naming conventions, descriptive labels or phrases, extra notes or modifiers (e.g., '+3mm'), and abbreviations.

5.4 Conclusions

`DICOMautomaton` can be used to perform accurate, semi-autonomous contour recognition. Lexicographic methods are generally suitable to this end when the domain is well-known, while more computationally-burdensome geometrical methods are available for information-deficient situations.

Mixing lexicographic measures produces an effective recognition improvement of more than 10% over a pure-Levenshtein-Damerau approach, while domain-specific measures can achieve optimal recognition.

Increasing domain knowledge in the lexicon increases performance, and so continued usage will tend to increase the successful recognition rate.

Chapter 6

Segmentation and Dose-Volume Analysis[1]

Divide et Impera – Divide and Conquer.

PHILIP II, KING OF MACEDON

Dose-volume-response for normal tissue complications has been extensively studied. Typically, organs at risk are demarcated on CT as a whole organ and dose-volume analysis is based on DVH constraints, which ignore positional dose dependence. The assumption underlying this treatment is that function is homogeneous throughout the organ volume. Evidence to the contrary has been demonstrated. Combining mean lung dose with positional dependence leads to improved predictive power for incidence of radiation pneumonitis [50]. Incidence of subjective xerostomia in head-and-neck cancer patients correlates more strongly with dose to the lateral cranial regions of the parotid compared with mean dose to the whole parotid [13]. The search for regional effects requires sub-segmenting normal tissue contours (i.e., dividing them into regional components) based on anatomical (e.g., liver lobes) or geometrical features. This search cannot conceivably be performed manually. Herein we present a computational system, `DICOMautomaton`, which can rapidly perform

[1] The content of this chapter is currently in press with the title *Automated Segmentation and Dose-Volume Analysis with* `DICOMautomaton` [19].

autonomous contour sub-segmentation and generic dose-volume computations, substantially reducing the human effort required for exploratory dose-volume analyses.

6.1 Methods

6.1.1 Sub-Segmentation Techniques

`DICOMautomaton` offers a variety of segmentation routines which can be applied identically from single contour loops to entire patient organ sets. The most intuitive are those which split data along an embedded $\mathbb{R}^2$ plane. For example, a researcher requiring equal-volume lateral halves of a given organ need only specify that a lateral plane be used. Not all segmentation is limited to organs, as per-contour or per-patient segmentation is available.

Alternatively, a user might not want a flush, planar boundary between sub-segments, instead preferring to sub-segment in terms of distance from an organ's surface. `DICOMautomaton` can perform this *ray-casting* segmentation along a specified direction. Planar and casting segmentation are sufficient for producing sub-segment of equal volume, contour (planar) area, or number in a variety of orientations and configurations. Each technique can be chained together sequentially or mixed selectively on a per-segment basis.

More complex segmentation is possible including fractal-like branching, which is useful for imitating branching vessels, core-and-peel, for separating inner and outer portions, and constrained segmentation, which can partition voxels by means of a user-provided heuristic. A canonical use for the latter would be locating all voxels within an organ receiving dose above a certain threshold. The heuristic function is a stateful closure which can be passed in and altered at run-time, admitting highly complex selection heuristics to be constructed.

6.1.2 Sub-Segmentation Scheduling

It is often desirable to test a variety of segmentations for some given suitability. This feedback allows for automatic searching of regional organ dose-sensitivity.

DICOMautomaton provides rudimentary facilities for the automation of such searching: given a set of segmentation rules (i.e., a "schedule"), the system will attempt to optimize a user-provided cost function by traversing a sub-segmentation graph. The schedule can be provided at run-time.

6.1.3 Dosimetric Facilities

To provide a useful platform for automation of dosimetric analyses, DICOMautomaton offers the ability to compute standard dosimetric quantities on any set of contour sub-segment. DVH, minimum/mean/median/maximum voxel dose, and statistical moment routines exist, while the implementation of additional routines is straightforward. All dosimetric routines avoid resampling and interpolation where possible by working directly with DICOM-style contour and voxel dose data.

To encourage interoperability, contours which have been segmented can be injected back into anonymized patient data permanently in a DICOM-conforming manner. Lossless segmentation persists across such injection, and remains fully reversible.

6.2 Results

To demonstrate the segmentation capabilities, we have produced a variety of images using patient contours. In figure 6.1 we demonstrate simple planar segmentation of a whole left parotid. Figure 6.2 shows mixed ray-casting and planar segmentations proceeding in sequence. Finally, more complex segmentations are shown in figure 6.3 including random-orientation sequential planar, core-and-peel, and dose-threshold heuristic segmentations.

System accuracy was measured by comparing minimum, mean, and maximum doses from 60 head-and-neck cancer dosimetric data sets to Varian's Eclipse™ (Varian Medical Systems, Inc.). The mean dose to the left parotid varied from 2 to 65 Gy. The average absolute percent discrepancy was 0.8% ($\sigma = 1.2\%$) with the largest occurring at low mean doses ($< 10\,Gy$). Total computation time, including file input and output, was 80 seconds on an Intel® Pentium® E2200. DVHs produced by DICOMautomaton were found

to strongly agree with those produced by Varian's Eclipse™. An example comparison is shown in figure 6.4.

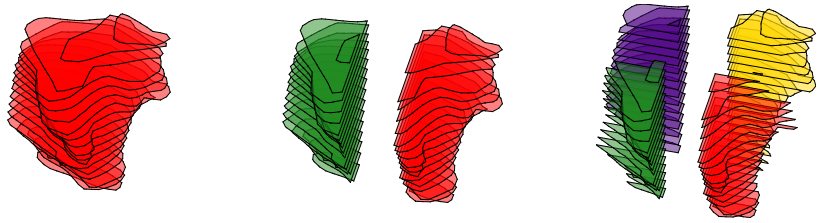

Figure 6.1: Sequentially performed planar segmentation on a whole left parotid. Per-organ (per-patient) planar boundaries are identical for all individual contours within an organ (patient) whereas per-contour planar boundaries vary from contour to contour.

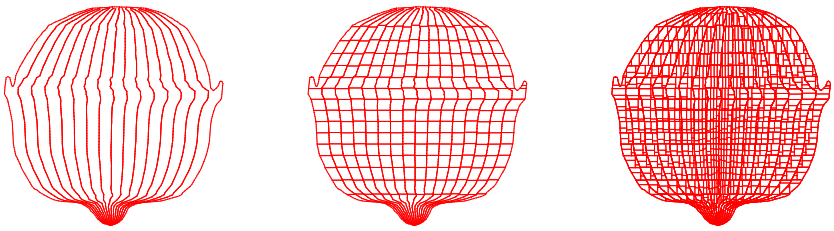

Figure 6.2: Demonstration of mixed contour segmentations of a body contour at the facial level in sequence. From left to right: simple lateral ray-casting, planar (coronal) per-contour, and several mixed planar per-organ and per-contour segmentations. This sequence demonstrates the composability of segmentation operations, which can easily produce highly complex subdivisions.

A typical segmentation workflow using popular analysis software takes a voxel approach whereby contours are resampled onto legitimately voxel (e.g., image or dosimetric) data for the convenience of having a single, unified structure or to simply fit into a voxel paradigm. This approach requires irreversible resampling and interpolation to project the contours onto a voxel grid and is wasteful in the sense that algorithms working on contour lines embedded in planes decrease considerably in computational efficiency when adapted to run on voxels in $\mathbb{R}^3$. Efficient voxel image segmentation algorithms

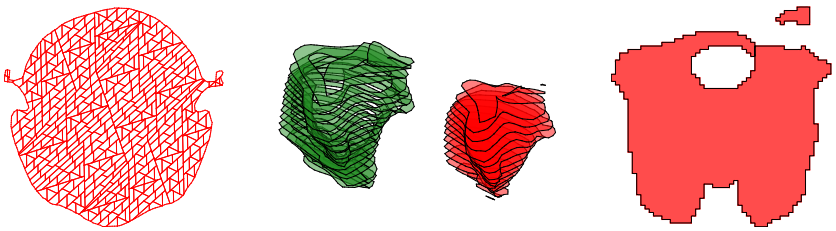

Figure 6.3: Demonstration of complex segmentations. From left to right: random-orientation sequential planar segmentation on a body contour at the upper neck level, core-and-peel segmentation on a whole left parotid, and dose-threshold heuristic segmentation of a body contour at the shoulder level. In the latter, voxels with high dose are exclusively encircled with a contour; spinal cord sparing is evidenced by the omitted (low-dose) circular region. Sub-segments can be further sub-segmented well below the typical voxel size, if desired.

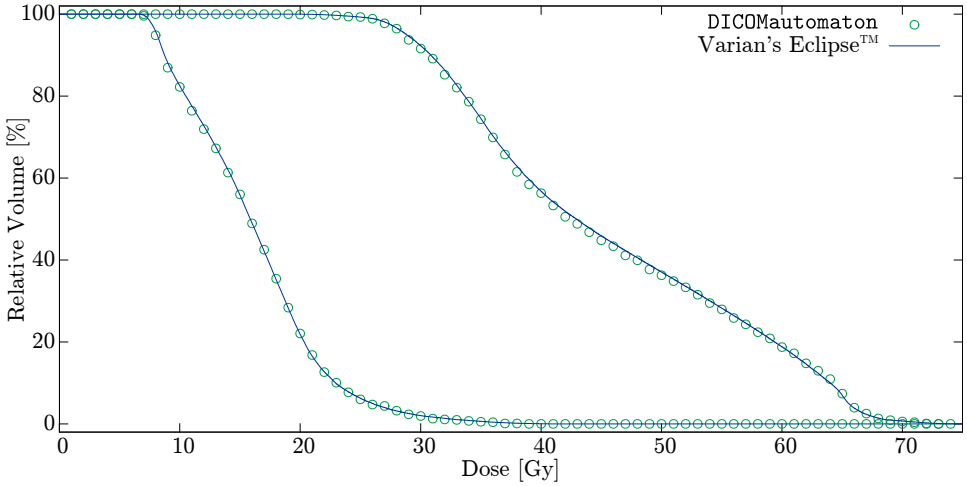

Figure 6.4: Example DVHs as produced by `DICOMautomaton` and Varian's Eclipse™ for a typical left (lower) and right (upper) parotid. Strong agreement is observed.

run in $O(n \log n)$ time where n is the number of voxels [33], which, due to the dimensionality, makes n unfortunately large at even modest resolutions (e.g., 256^3 voxels). The popular marching cubes algorithm [48], which is more directly adaptable to voxel-based contour segmentation, not only suffers from the aforementioned dimensionality issue but also notoriously exhibits topological inconsistency and produces fragmented volume boundaries [55]. Attempts to address these limitations are numerous [61]. To avoid such problems `DICOMautomaton` forgoes the voxel paradigm, instead working directly with DICOM-style contours. This enables lossless, reversible segmentation, reduces resampling and truncation errors to near-machine precision, and dramatically reduces computation time.

6.3 Conclusions

`DICOMautomaton` allows users to rapidly, accurately, and autonomously subsegment large amounts of contour data into intricate structures suitable for analyses of regional organ dose-sensitivity. Dosimetric routines are robust; quantities computed with `DICOMautomaton` strongly agree with those produced by Varian's Eclipse™.

Chapter 7

Analysis of Salivary Measurements

I would have written a shorter letter,
but I did not have the time.

Blaise Pascal

Xerostomia is the subjective symptom of chronically dry mouth. It is one of the most common late toxicities in patients receiving RT for head-and-neck cancers. In the severe case, it is a debilitating condition, affecting a patient's ability to eat, sleep, and communicate [37]. It is known to have a strong, broad impact on a patient's QOL, even imposing on domains not directly related to xerostomia, such as emotion and pain [46] [66].

Much effort has been spent identifying patient-, clinical-, and disease-factors which could be used to predict the onset of xerostomia. Historically, the strongest predictors were based on mean dose to whole parotid glands [10]. Recently, predictors based on morphological features have been found to accurately predict the complication risk of xerostomia [13].

Reports describing late recovery are less numerous and mixed in conclusion. Generally, parotid glands exhibit (limited) late recovery in function so long as the mean dose is below threshold values [10] [30]. It is otherwise unclear how the magnitude of recovery is associated with the dose distribution a patient

receives.

The aim of this report was to establish whether regional geometric sub-segmentation is able to better predict the loss or recovery of salivary output following RT when compared to standard mean dose models. Our technique geometrically sub-segments organs into sub-volumes. Using a variety of such sub-segments, we examine the relationship between regional dose and functional loss and recovery.

7.1 Methods

7.1.1 Patients

> *"Data! Data! Data!" he cried impatiently.*
> *"I can't make bricks without clay."*
>
> Sherlock Holmes in *The Adventure of the Copper Beeches*
> Sir Arthur Conan Doyle

Between February 2008 and May 2013, 102 patients undergoing RT for head-and-neck cancers were evaluated for xerostomia.

Patients were required to give informed consent before pariticipating in this study. They underwent saliva collection under stimulated and unstimulated conditions prior to RT and both three months and one year after the RT-completion date. The population admitted for study was inhomogeneous; both the RT aim (i.e., curative vs. palliative) and technique were varied. Pertinent demographic information is supplied in table 7.1.

Patients were deemed unsuitable if they satisfied any of the following.

- They were unable to produce more than one gram of whole mouth stimulated saliva at baseline over a span of five minutes[1].
- They took unusual or uncommon chemotherapy agents or medication.
- They received electron therapy to the primary tumour.

[1]The saliva collection protocol required a five minute collection period. The procedure is described in appendix B.

- Multiple, inconsistent CT sets were used for planning or during RT.
- They underwent modified radical neck dissections that involved removing one or more salivary glands.
- They had had previous RT which may have affected salivary function.
- They had incomplete salivary measurement data due to missed follow-up.
- They failed to comply with the collection protocol (which is described in appendix B).

Data collection is ongoing. Of the approximately 750 patients for which salivary function has been measured, 102 were found to be suitable for inclusion in this study.

Table 7.1: Demographics for the 102 patients considered in this study.

Quantity		$n^{\ddagger}$ or Time
Gender	male	69
	female	33
Grade IV Xerostomia[§]	3 months*	44
	1 year*	27
RT Technique	IMRT*	65
	unilateral	21
	BLPOP*	16
Age	mean	$57.7\,y$
	range	$19.0 - 82.7\,y$

Dose $[Gy]$	$n_F^{\dagger}$	$n^{\ddagger}$
50	25	3
52.5	20	1
55	20	1
55	25	1
56	28	1
60	25	16
60	30	7
60	33	1
60	35	1
63	30	1
66	33	5
70	34	1
70	35	63

Tumour subsite	$n^{\ddagger}$
oropharynx – tonsil	27
nasopharynx	26
oropharynx – tongue	18
other	15
hypopharynx	6
larynx	5
oral cavity	5

$n^{\ddagger}$ – Number of patients.
$n_F^{\dagger}$ – Number of fractions.
3 months* – Nominal amount of time following RT completion. Median = 91 days, $\sigma = 23$ days.
1 year* – Nominal amount of time following RT completion. Median = 379 days, $\sigma = 56$ days.
IMRT* – Fixed-gantry IMRT or volumetric modulated arc therapy.
BLPOP* – Bilateral parallel-opposed fields.
§ – As objectively defined by salivary flow being $\leq 25\%$ of baseline.

7.1.2 Contour Generation

Using CT images, contours were generated for the purpose of RT by an expert oncologist specializing in contouring at the BCCA using Varian's Eclipse™. Parotid contours were later examined by a single senior head-and-neck oncologist for quality assurance (using the same software) for the purposes of this study. Re-contouring was performed if deemed necessary.

7.1.3 Measurement of Saliva Output

Xerostomia is a subjective symptom. Several studies have shown a weak correlation between subjective xerostomia reported by patients and salivary function (as measured by salivary output – see section 2.2.2).

Subjectivity notwithstanding, the widely used LENT-SOMA classification scheme advocates objectively measuring xerostomia via salivary output. Therefore, we wish to stress that within this report we focus not on subjective xerostomia but rather on directly quantifiable salivary output.

The collection procedure is described in appendix B. It is commonly used in the literature (e.g., in [10]).

7.1.4 Sub-Segmentation and Dose Computation

Our choice of contour sub-segmentation was motivated by both the basic anatomical structure of the parotid and geometrical constraints typical during treatment.

Both parotids were considered separate entities – insofar as the whole mouth collection of saliva permitted such distinction. In particular, we focus on whole parotid (no sub-segmentation), lateral halves (medial and lateral) of equal volume, and core-and-peel (inner core and outer shell) volumes.

To facilitate rapid, flexible contour recognition and dose-volume computations, a computer system, `DICOMautomaton`, was created for the purposes of the present study. The relevant capabilities of `DICOMautomaton` are described in chapters 5 and 6.

7.1.5 Non-Parametric Regression

We make use of non-parametric regression techniques for viewing and evaluating the overall trend of salivary output loss and recovery. Specifically, we use the Non-Parametric Local Linear Regression (NPLLR) described by Wasserman [92] with a tricube kernel and derive our confidence intervals from the method described by Wasserman [92] and the report by Yu and Jones [95]. In light of the (surprisingly) well-behaved weighting parameters over bandwidth space, we make use of an unconstrained Nelder-Mead simplex [60] to locate the bandwidth parameter which optimizes overall smoothness.

Assumptions required by non-parametric methods are taken as weak as possible and are generally fewer in number than for parametric methods. The benefit of such methods include the ability to homogeneously model generic functional relations (viz. non-linear or non-monotonic) and the computability of local confidence intervals. Such techniques ensure we minimally bias our results with inappropriate salivary models.

7.1.6 Measuring Association

To evaluate the association between salivary output and delivered dose, we use Spearman's non-parametric rank correlation coefficient (R_s) [77]. It was chosen as it is able to measure correlation in any non-linear monotonic function and is asymptotically ignorant of the normality of errors in the data [75]. Compared with alternative correlation measures, R_s is robust to outliers due to its use of ranking. Compared with alternative rank correlation measures, it is generally of equivalent power [82] [96].

The use of non-parametric measures of association come at a price: reduced statistical power. Conversely, if an association is found it is more likely to exist in the data instead of being imposed by a chosen model. We expect salivary output loss to be appropriately described by R_s [75].

However – recovery is thought to have an arched dose-dependence, which defies the monotonicity demanded by correlation. At low dose when the parotid parenchyma is damaged little, we expect little recovery. Likewise, at high doses, reports have indicated that chances of recovery are weak and

we therefore expect little, if any, recovery. In between these extremes we expect enough damage will be inflicted that recovery may be initiated, but no so much so that it would be impossible. Therefore, the use of correlation (rank or otherwise) to measure association is inappropriate, so we rely on parametric techniques to measure the dose-dependence of recovery.

7.1.7 Modeling and Fitting

A smart model is a good model.

Tyra Banks

Salivary output over a fixed time can be effectively modeled by a simple exponential model [10] [31]. Therefore, to model contributions to total salivary output S from parotid sub-segments we use

$$\frac{\Delta S\left(D_L, D_R; A\right)}{S_b} \equiv \frac{S_t - S_b}{S_b} = \frac{\exp\left(-AD_L\right) + \exp\left(-AD_R\right)}{2} - 1, \qquad (7.1)$$

where S_b is a pre-treatment salivary output (i.e., $t < 0$), D_L (D_R) is the dose delivered to the left (right) parotid (at time of RT, $t \approx 0$), and A is a free parameter. We make use of this model for all parametric analysis. As we are investigating an arbitrary number of parotid sub-segments, it was prudent to limit the possibility of overfitting due to an excess of free parameters. Therefore, we only consider a particular sub-segment (in each parotid) for each fit and disregard the contributions from remaining organs and sub-segments.

S is, in reality, the sum of contributions from various organs and all sub-segments, which may contribute unequally. To account for this and other possible functional dependence, we have non-parametrically investigated a class of models we call *combined dose* which satisfy

$$\psi\left(S_b, S_{3m}, S_{1y}\right) \propto \phi\left(D_L, D_R\right) \qquad (7.2)$$

for some salivary signal ψ and some combining function ϕ. We defer discussion of ψ and ϕ to section 7.1.9.

We perform non-linear parametric regression using an unconstrained Nelder-Mead simplex. As our objective function, we try both weighted Least-Sum of Squares (LSS) and weighted Least-Median of Squares (LMS). The latter helps reduce the impact of outliers and is considered more statistically robust than the former, although it is of lesser statistical power. We do not wholly rely on LMS because, although it is of reduced sensitivity to outliers, it is highly unstable [74] and is for this reason considered to be an unsuitable replacement for LSS [70]. Therefore, for each parametric fit we compute both LSS and LMS best-fit parameters. If the agreement between the two sets is strong, we can be reasonably satisfied with the LSS parameters. Otherwise, disagreement will indicate the presence of outliers or noise which have skewed the fit. We use such discrepancy to help assess the reliability of our fits.

For estimates of uncertainty and the production of confidence intervals for parameters derived using both LSS and LMS, we use the bootstrap method [28]. The bootstrap is a data resampling technique whereby a statistical population is simulated using a limited number of samples. The procedure is simple and is described here in brief. A *resampled* distribution is constructed from the measured distribution. Elements of the measured distribution are chosen at random (including elements which have already been chosen) until the distributions have the same number of elements N. The estimator of interest is applied to the resampled distribution, yielding an estimate of the quantity resulting from, on average,

$$N\left(1-\left(\frac{N-1}{N}\right)^N\right) \xrightarrow[N\gg 1]{} \frac{e-1}{e}N$$

unique samples. If this procedure if performed M times, we have M estimates of the desired quantity. The resulting distribution of estimates simulates the spread of the quantity of the true population. Further discussion and a more complete introduction can be found in [28].

To ensure the quality of bootstrap computation, no less than 50000 resamples are performed. Confidence intervals are estimated by sorting the bootstrap estimates and finding the corresponding percentile (i.e., 95%

confidence intervals can be found by considering the 2500^{th} and 47500^{th} ordered estimates).

7.1.8 Measuring Goodness-of-Fit

In the case of LSS, we use χ^2 and t-tests to measure goodness-of-fit. Whenever possible, p-values are computed by direct computation or enumeration of permutations. We cannot directly compute significance of LMS fits or our choice of ϕ, but we compare with LSS using t-tests.

For the purposes of χ^2 fitting, an estimate of the per-measurement (pointwise) variance is constructed using reported values of variance for similar saliva collection procedures, an assumed fractional uncertainty, and standard error propagation techniques. In contrast, NPLLR trends are shown in comparison with standard errors derived directly from the abscissæ and ordinate. Deriving standard errors from binned data gives a clearer picture of population variance, and it is believed to be most relevant for the purposes of viewing the NPLLR trends. Conversely, the pointwise variance estimates are more appropriate for χ^2 fitting, where an estimate of the uncertainty of a particular datum is most appropriate.

7.1.9 Monte Carlo Model Sampling

> *Do not quench your inspiration and your imagination;*
> *do not become the slave of your model.*
>
> VINCENT VAN GOGH

To choose the non-parametric salivary signal ψ, combining function ϕ, and evaluate our choice of loss and recovery models, we have implemented a novel Monte Carlo computation technique which is described here. The purpose of the technique is to non-parametrically locate functional response by sampling many possible models.

1. We produced a small list of reasonable candidate dose combination functions along with an estimate of how each might fail to describe the data or be generalized. These represent the dose combining function ϕ.

2. Using the measured primitives: baseline (S_b), three months (S_{3m}), and one year (S_{1y}), we constructed a handful of salivary signals which the data might be well-suited for describing (e.g., early loss $(S_{3m} - S_b)/S_b$, recovery $(S_{1y} - S_{3m})/S_b$, overall loss $(S_{1y} - S_b)/S_b$, etc.) and performed a similar breakdown analysis. These represent ψ.

3. We identified the combinations we thought would elicit the largest monotonic response. All expressions were broken into primitive components and operations; arithmetical and functional.

4. Using a symbolic computation system developed by the author, expressions for ψ and ϕ were randomly sampled using components of the candidate expressions, arithmetical operators, functions, and constants (integers). Maximum equation order and depth was specified and uniqueness of samplings was ensured.

5. Each generated expression (eq. (7.2) ; $\psi \propto \phi$) was populated with required measurements and dose data. R_s was used to evaluate monotonic association non-parametrically.

Examining the cluster of expressions with highest R_s, we are able to estimate the appropriateness of our candidate ψ and ϕ, and their relevance to the data (i.e., whether the data shows the capacity to describe the desired relationship). Further, we can search the generated expressions for hints of dimorphism between various sub-segments and can provide an estimate of the upper limit of monotonic association which the data (in the desired functional form) can support.

We wish to emphasize that the aforementioned technique is not 'data dredging' (i.e., looking for statistically significant results in many potential models). A fundamental first step involves specification of particular models, features, and operations. It is more in the spirit of *classifying* the *types* of models which exhibit strong monotonic associativity, and ensuring our choice of ϕ is amongst them. In summary, as a sort of dual to resampling techniques, which validate models by random choice of data, we simulate model permutation in the functional neighbourhood of the specified models.

7.2 Results

7.2.1 Output Loss and Recovery Were Significant

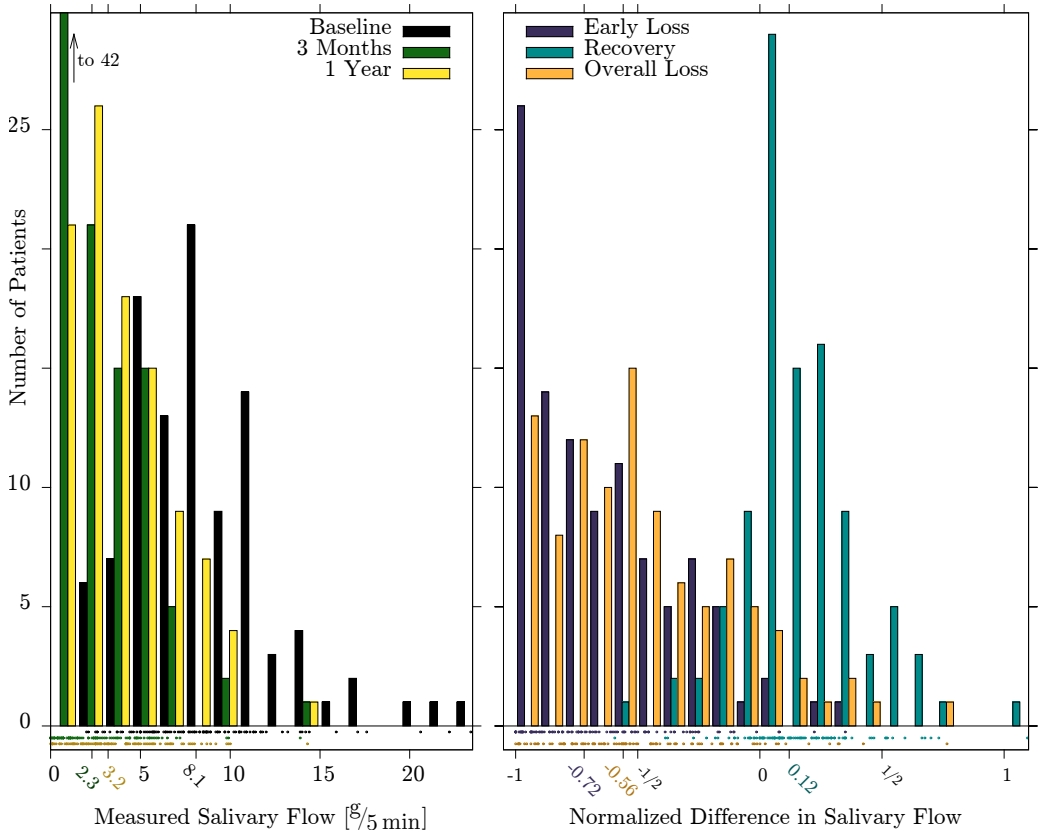

Figure 7.1: Histograms showing various distributions. At left: measured salivary function at baseline, three months, and one year. At right: salivary function loss and recovery. Below each plot are the raw data. The medians for each distribution are indicated along the abscissæ.

Considerable noise was encountered in the salivary flow measurements, consistent in magnitude with previous reports (see section 3.1.1).

The distribution of measured stimulated salivary output at baseline (S_b), three months (S_{3m}), and one year (S_{1y}) are shown in fig. 7.1. Also shown are

the baseline-normalized early loss $(S_{3m} - S_b)/S_b$, recovery $(S_{1y} - S_{3m})/S_b$, and overall loss $(S_{1y} - S_b)/S_b$ constructed from salivary output measurements. The medians for each distribution are indicated along the abscissæ. For the purposes of this study, *early loss* should be taken to be synonymous with "baseline-normalized loss of salivary output at three months" and *overall loss* with "baseline-normalized loss of salivary output at on year."

Recovery was seen in 79 of 102 patients ($p < 0.0001$, Wilcoxon sign rank test). The median magnitude of recovery was 12% of baseline output. Unfortunately, for most patients loss was considerably higher in magnitude than recovery. The median magnitude of early loss was 72% of baseline output. Overall loss was slightly less severe at 56%.

An increase in mean salivary output was seen (mean $\pm\sigma$: $2.7 \pm 0.1\,{}^{g}/_{5\,min}$ at three months vs. $3.8 \pm 0.1\,{}^{g}/_{5\,min}$ at one year; $p < 0.000001$, two-tailed t test). This increase, although small, represents 13% of the baseline output and 42% of the three month output. Barring outliers, maximum recovery appears to be $\approx$ 80% of baseline output. A summary of these results is tabulated in table 7.2.

Salivary output loss and recovery exhibited no discernible laterality.

7.2.2 Output Loss was Strongly Correlated with Mean Dose

Whole organ and sub-segmented volumes exhibited significant correlation between salivary output loss and mean dose delivered (within the volume).

The Monte Carlo model sampling technique of section 7.1.9 indicated that the proposed combined dose $\phi = min(D_L, 45) + min(D_R, 45)$, chosen to explicitly limit the amount of dose damage each parotid can contribute to the combined dose, elicited strong salivary loss response[2]. Figure 7.2 displays early loss vs. combined dose for whole parotid. The NPLLR trend and its 95% confidence interval are compared with the salivary measurements. The same is shown for combined dose to laterally sub-segmented parotid in fig. 7.3. Both trends depart from linearity. The presence of a 'shoulder' in the lateral sub-segment is prominent, as it is for several other tested sub-segments and

[2]Results of the Monte Carlo model sampling technique are given in section 7.2.5.

Table 7.2: Stimulated salivary flow measurement statistical summary.

Quantity	Range	Median	Mean $\pm\sigma$, 95% c.i.†
Baseline S_b	$[2.0, 23.4]$	8.1	8.5 ± 0.2, $[8.1, 8.8]$
Three Months S_{3m}	$[0.0, 13.9]$	2.3	2.7 ± 0.1, $[2.5, 2.9]$
One Year S_{1y}	$[0.0, 14.3]$	3.2	3.8 ± 0.1, $[3.6, 4.1]$
Early Loss $(S_{3m} - S_b)/S_b$	$[-1.00, 0.35]$	-0.72	-0.64 ± 0.02, $[-0.68, -0.61]$
Recovery $(S_{1y} - S_{3m})/S_b$	$[-0.59, 1.10]$	0.12	0.16 ± 0.03, $[\ 0.10,\ \ 0.21]$
Overall Loss $(S_{1y} - S_b)/S_b$	$[-1.00, 0.77]$	-0.56	-0.49 ± 0.02, $[-0.53, -0.44]$

All measurement are in units of $^{g}/_{5\,min}$.
† *Confidence interval* of the mean.

a wide variety of Monte Carlo sampled models. We defer discussion of the shoulder to section 7.2.4. Sub-segmentating into the lateral halves reduced the cluster of combined doses around $40 - 60\,Gy$, qualitatively improving congruence of the NPLLR trend with combined dose in both cases.

Spearman's rank correlation coefficients have been computed (table 7.3) for a variety of sub-segments. All demonstrate strong significance and, therefore, dose dependence.

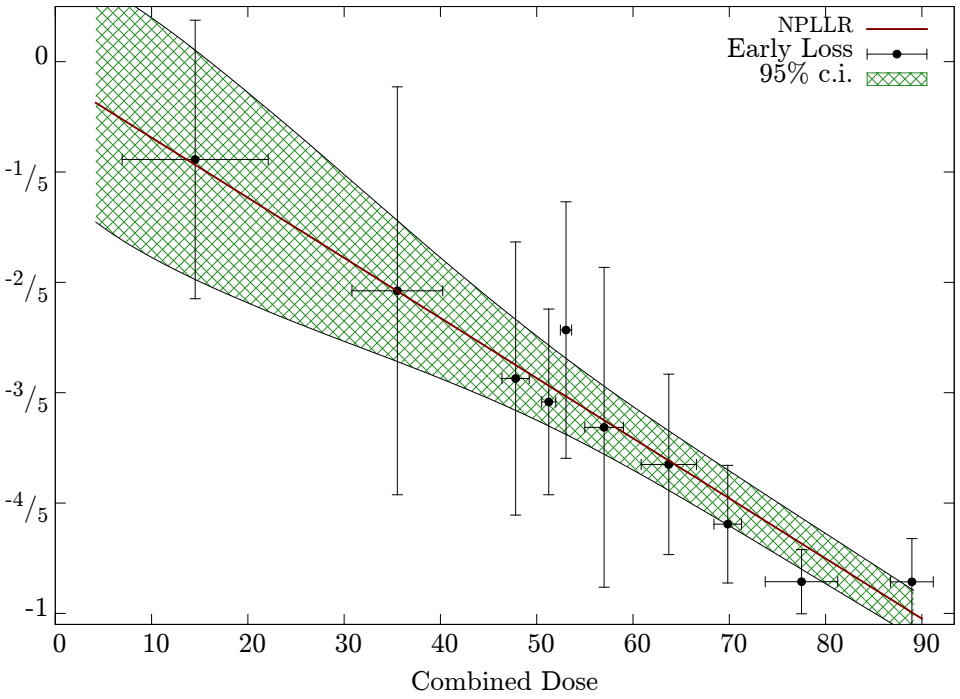

Figure 7.2: Early loss $(S_{3m} - S_b)/S_b$ vs. combined dose $\phi = min(D_L, 45) + min(D_R, 45)$ for whole parotid with standard errors. The NPLLR trend and its 95% confidence interval are given. Note the clustering between $40 - 60\,Gy$. An identical plot with non-binned data, which shows the clustering in more detail, is shown in fig. C.1.

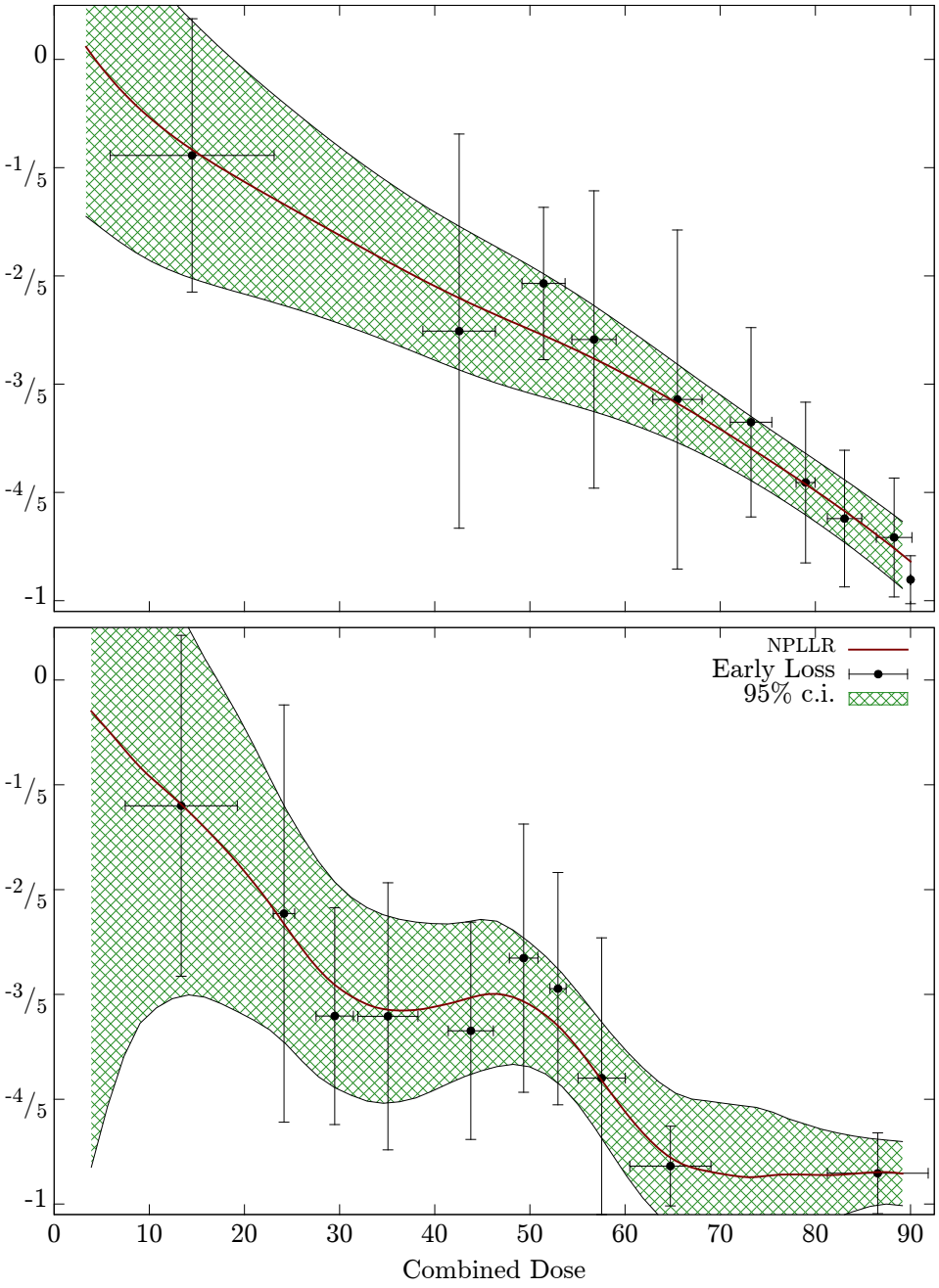

Figure 7.3: Early loss $(S_{3m} - S_b)/S_b$ vs. combined dose $\phi = min(D_L, 45) + min(D_R, 45)$ with standard errors for laterally-sub-segmented halves (of equal volume) of the parotid. The medial half is on top and the lateral on bottom. The NPLLR trends appear, qualitatively, to better suit the data compared to the whole parotid of fig. 7.2. An identical plot with non-binned data is shown in fig. C.2.

Table 7.3: Spearman's rank correlation coefficients (R_s) and two-tailed p values describing salivary output loss vs. combined dose for a variety of sub-segments.

Quantity	Sub-segment	R_s	p
Early Loss	Whole	-0.729	< 0.000001
	Medial half	-0.754	< 0.000001
	Lateral half	-0.608	< 0.000001
	Core	-0.671	< 0.000001
	Peel	-0.704	< 0.000001
Overall Loss	Whole	-0.564	< 0.000001
	Medial half	-0.596	< 0.000001
	Lateral half	-0.475	$= 0.000001$
	Core	-0.523	< 0.000001
	Peel	-0.561	< 0.000001

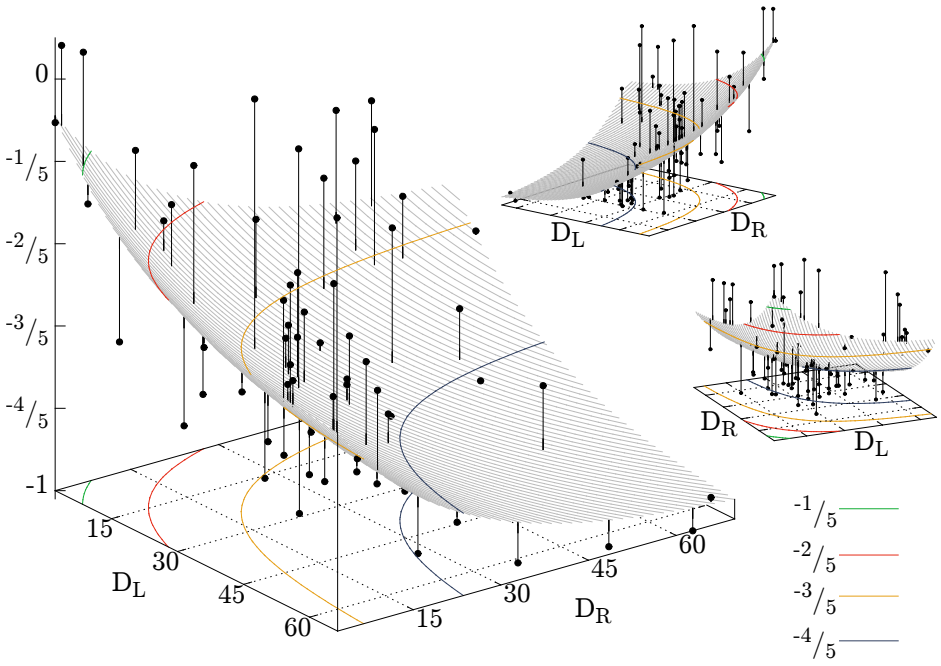

Figure 7.4: Early loss $(S_{3m} - S_b)/S_b$ vs. dose to left (D_L) and right (D_R) whole parotid. Drop lines are given only as a visual guide. The surface is a parametric fit of eq. (7.1). Fit parameters are given in table 7.4.

Table 7.4: Fit parameters describing salivary output loss vs. combined dose for a variety of sub-segments.

		LMS	LSS				
Loss	Sub-segment	$A\left[Gy^{-1}\right]$, 68% c.i.	$A\left[Gy^{-1}\right]$, 68% c.i.	χ^2_{red}	p_a	p_w	p_{eo}
Early	Whole	0.0604(86), [0.0518, 0.0691]	0.102(10) , [0.092, 0.112]	4.12	0.002	—	0.07
	Medial	0.0451(72), [0.0379, 0.0522]	0.0823(95), [0.0728, 0.0917]	4.31	0.004	0.22	0.20
	Lateral	0.083(11) , [0.072, 0.093]	0.157(19) , [0.139, 0.176]	4.58	< 0.001	0.11	0.01
	Core	0.065(11) , [0.054, 0.075]	0.122(17) , [0.105, 0.138]	4.67	< 0.001	0.74	0.03
	Peel	0.065(10) , [0.056, 0.074]	0.104(10) , [0.095, 0.114]	3.94	0.004	0.71	0.06
Overall	Whole	0.0379(73), [0.0306, 0.0451]	0.118(20) , [0.098, 0.138]	5.65	< 0.001	—	
	Medial	0.0304(58), [0.0246, 0.0362]	0.0809(79), [0.0730, 0.0888]	4.91	< 0.001	0.51	
	Lateral	0.043(12) , [0.031, 0.055]	0.220(50) , [0.171, 0.270]	6.56	< 0.001	0.70	
	Core	0.0357(77), [0.0280, 0.0434]	0.115(18) , [0.097, 0.134]	5.40	< 0.001	0.86	
	Peel	0.0415(64), [0.0351, 0.0480]	0.123(20) , [0.103, 0.143]	5.59	< 0.001	0.75	

Notation: A – fit parameter, c.i. – confidence interval of A, χ^2_{red} – reduced chi-square of LSS fit. P-values denote significance of two-tailed t-tests between: LMS and LSS estimates of A (p_a), LMS whole-parotid estimate of A and indicated sub-segment (p_w), and LMS estimate of A for early and overall loss (p_{eo}).

Numbers in parentheses denote uncertainty in the preceding digits.

All fit parameters were estimated via bootstrap using at least 50000 resamples.

Estimates of the pointwise variance used exclusively for χ^2 fit weighting were constructed as follows. Salivary measurement uncertainty was assumed to have a constant base uncertainty of $0.05\,{}^{g}/_{5\,min}$, exhibit a fractional component of uncertainty below an unknown threshold, and was otherwise constant (i.e., a measurement above the threshold was assumed to have the same uncertainty as a measurement at the threshold). Based on reports of variability from both similar and dissimilar saliva collection techniques performed on a wide variety of healthy volunteers (see section 3.1.1), the fractional uncertainty was conservatively taken to be 40%. Ensuring the average pointwise uncertainty did not drop below 25% (from the standard deviation of salivary measurement reported independently by [10] and [14]), the threshold was taken to be $8.0\,{}^{g}/_{5\,min}$. Standard error propagation was performed to construct variance estimates for early loss, recovery, and overall loss.

Parametric regression using the mean dose exponential (eq. (7.1)) resulted in poor fits. Figure 7.4 shows early loss vs. dose to left and right whole parotid. The noise inherent to the salivary measurements was prominent. Approximate normality of pointwise variance estimates was verified after fitting. Performing fits with unit pointwise variance or larger fractional pointwise variance altered the fit parameters slightly, but did not alter findings in any significant way.

Table 7.4 shows the parametric fit parameters. In a comparable study using LMS, Blanco et al. [10] estimates A to be $0.054(2)\,Gy^{-1}$ at both six months and one year for whole parotid. Although a t-test indicates our agreement at three months is not significant (compared with $A = 0.0604(83)\,Gy^{-1}$, due to our large uncertainty), we encompass the result within the 68% confidence interval. One year parameters do not agree; we encounter less salivary output loss (percent difference 35%).

Also shown in table 7.4 are p-values denoting the significance of various two-tailed t-tests. The first, p_a, measures whether the estimated parameter (A) from LSS and LMS are drawn from the same distribution. In all cases, this p-value was significant, indicating a sizable discrepancy between LSS and LMS. This result indicates the presence of strong outliers in the data. Indeed, the χ^2_{red} indicate poor overall LSS fits.

The second, p_w, measures whether the parameter estimated (via LMS) from the indicated sub-segment is distinct from that of the whole parotid. In other words, it measures whether the sub-segment behaves differently from the whole parotid, perhaps due to salivary compensation or heightened dose-susceptibility. No p-values were reasonably significant, though the medial and lateral sub-segments *hint* at such a discrepancy. Regarding the mean dose exponential, the use of sub-segmentation does not improve prediction of salivary output loss compared with mean dose to the whole parotid.

The third, p_{eo}, measures whether the LMS fit parameters for early and overall loss are distinct. In some cases they were not significantly distinct, while in others they were significant, but inconsistently so. This may indicate that, parametrically, our view of recovery (i.e., the difference of overall and late loss) is being masked by noise. Still, the lateral and core sub-segments show a significant difference at three months and one year. In contrast with Blanco et al. [10], our mean dose exponential parametric fits are sensitive to the observed mean recovery.

7.2.3 Output Recovery was Weakly Associated with Mean Dose

Recovery can be constructed from the loss model considered in section 7.2.2 by subtraction of the parametric surfaces

$$\frac{S_{1y} - S_{3m}}{S_b} = \frac{S_{1y} - S_b}{S_b} - \frac{S_{3m} - S_b}{S_b}.$$

Recovery was seen to have little dependence on dose in either whole organ or sub-segmented volumes.

Using NPLLR (fig. 7.5) we are able to see both the significant mean recovery noted in section 7.2.1 (i.e., the vertical offset) and the general arch shape anticipated in section 7.1.6 – though it is minute compared with the variation in the data. The NPLLR trend suggests minimal dependence on dose and qualitatively indicates poor predictive power.

A small distinction exists between the medial and lateral sub-segments as shown in fig. 7.6. The distinction is likely rather benign due to the

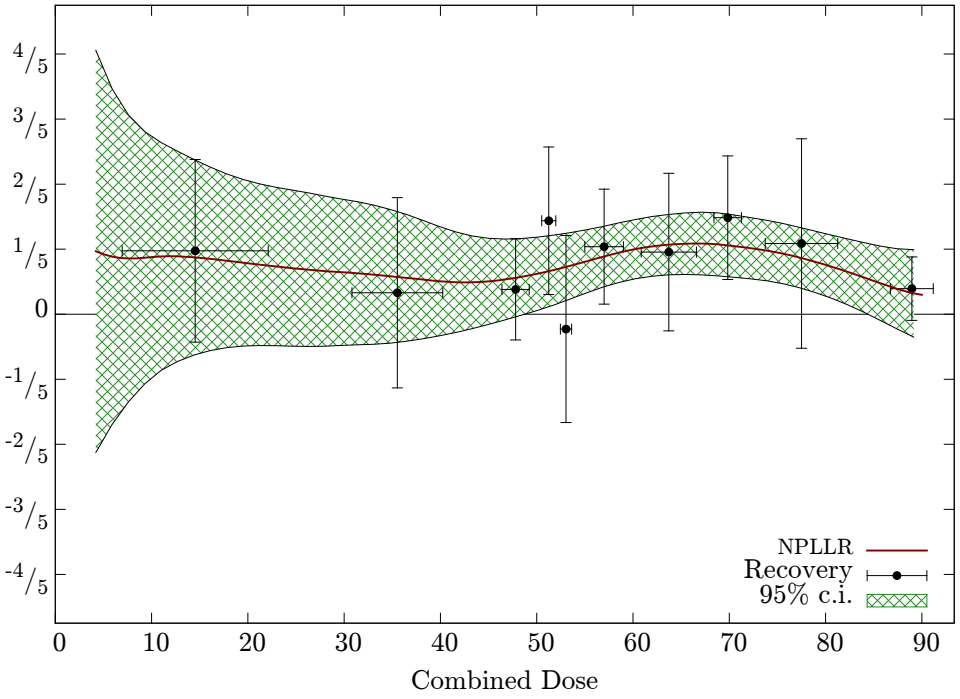

Figure 7.5: Recovery $(S_{1y} - S_{3m})/S_b$ vs. combined dose $\phi = min(D_L, 45) + min(D_R, 45)$ with standard errors for whole parotid. The NPLLR trend and its 95% confidence interval are given. An identical plot with non-binned data is shown in fig. C.3.

nearly-perfect confidence interval overlap. Of notice is that, contrary to early loss (fig. 7.3), the medial and lateral sub-segments produce slightly smaller confidence intervals compared with the whole parotid, particularly at low dose. This is due to the previously mentioned reduction in clustering of ϕ around $40 - 60\,Gy$.

Figure 7.7 shows the fit surface. Fit parameters are given in table 7.5. Due to the extremely low signal-to-noise ratio, the small mean recovery observed in section 7.2.1 was not apparent in the parametric fit. Indeed, the recovery constructed from salivary output loss was no more significant than a flat plane fit to the same data.

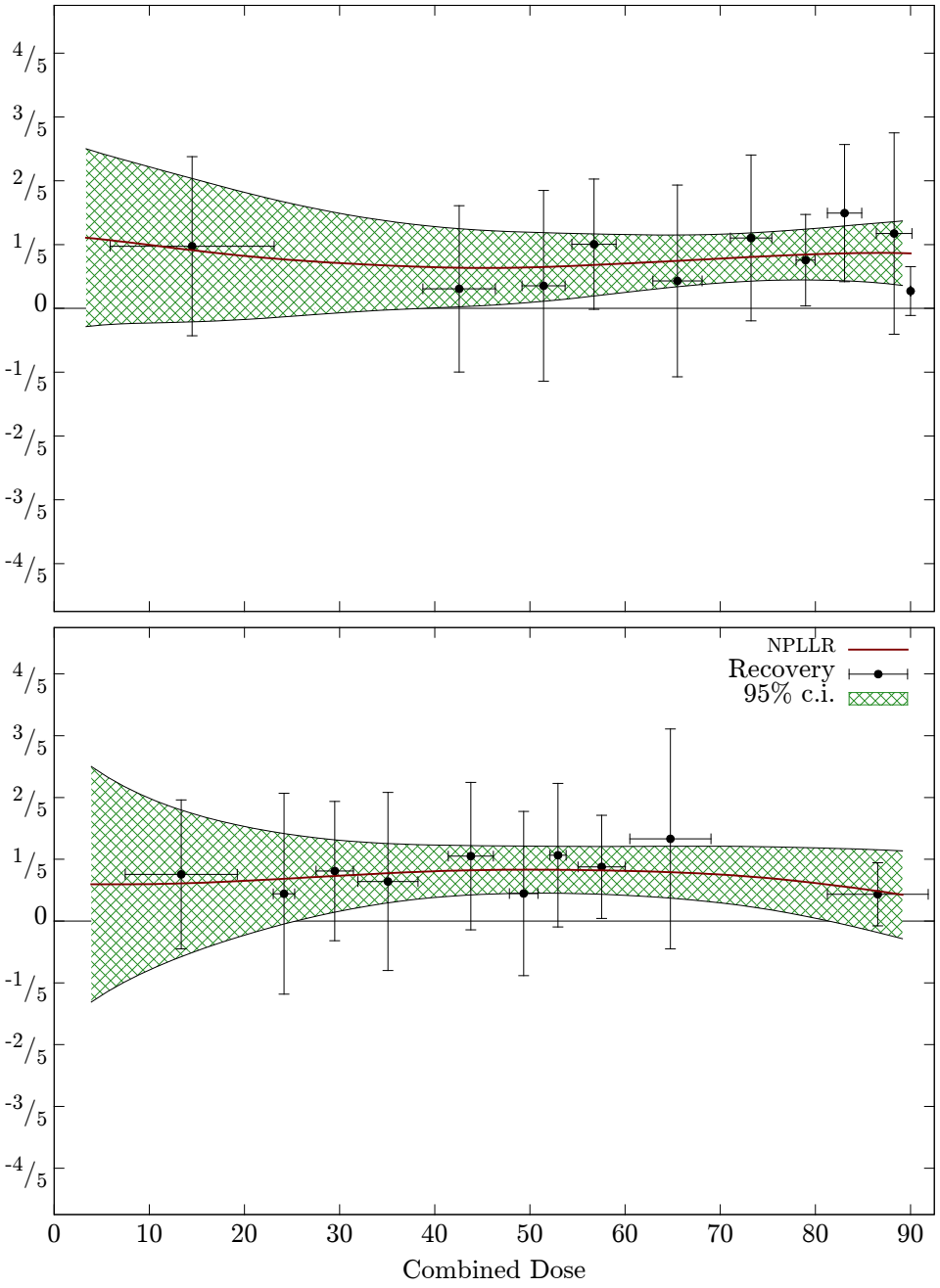

Figure 7.6: Recovery $(S_{1y} - S_{3m})/S_b$ vs. combined dose $\phi = min(D_L, 45) + min(D_R, 45)$ with standard errors for laterally-sub-segmented halves of the parotid. The medial half is on top and the lateral on bottom. As in fig. 7.3, the clustering around $40 - 60\,Gy$ is somewhat dispersed by sub-segmentation. An identical plot with non-binned data is shown in fig. C.4.

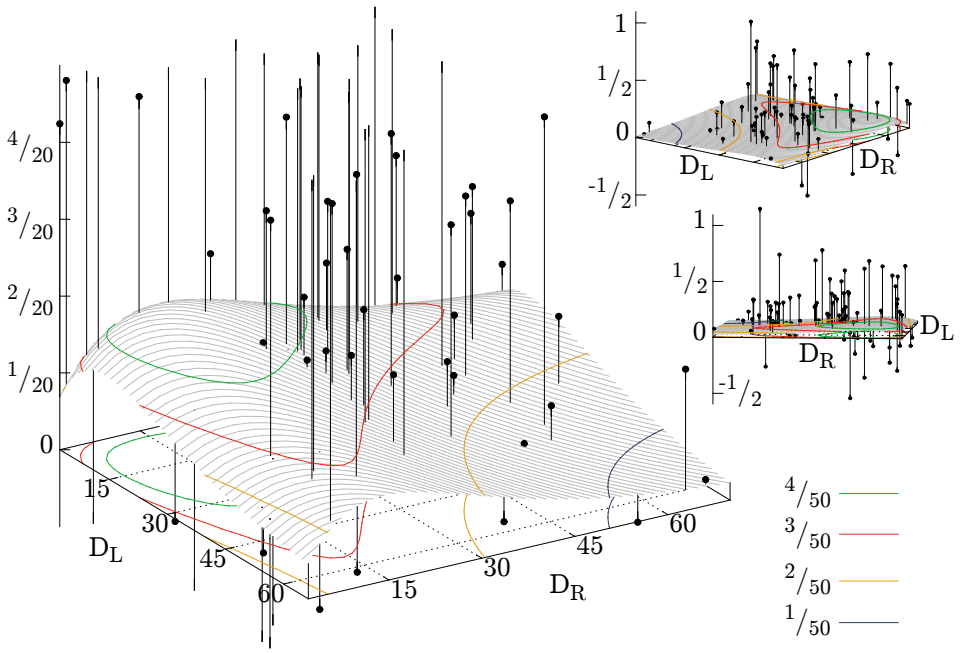

Figure 7.7: Recovery $(S_{1y} - S_{3m})/S_b$ vs. dose to left (D_L) and right (D_R) whole parotid. Drop lines are shown only as a visual guide. The enlarged figure on the left is scaled along the recovery axis to emphasize the surface shape. The figures on the right show the same surface scaled in relation to measurements. The surface is a parametric fit of eq. (7.1).

Table 7.5: Fit parameters describing salivary output recovery vs. combined dose for a variety of sub-segments.

		LMS	LSS				
Sub-segment	Parameter	$A\left[Gy^{-1}\right]$, 68% c.i.	$A\left[Gy^{-1}\right]$, 68% c.i.	χ^2_{red}	p_a	p_w	p_l
Whole	A_{1y}	0.0083(24), [0.0059, 0.0106]	0.061(19), [0.042, 0.080]	1.28	< 0.001	—	0.003
	A_{3m}	0.0111(33), [0.0078, 0.0145]	0.080(33), [0.048, 0.113]		< 0.001	—	< 0.001
Medial	A_{1y}	0.0084(21), [0.0063, 0.0106]	0.063(21), [0.042, 0.084]	1.27	< 0.001	0.98	0.014
	A_{3m}	0.0110(30), [0.0080, 0.0141]	0.094(36), [0.058, 0.131]		< 0.001	0.99	< 0.001
Lateral	A_{1y}	0.0143(32), [0.0110, 0.0175]	0.062(16), [0.045, 0.078]	1.25	< 0.001	0.42	0.02
	A_{3m}	0.0187(49), [0.0138, 0.0236]	0.076(26), [0.050, 0.103]		0.001	0.40	< 0.001
Core	A_{1y}	0.0080(25), [0.0055, 0.0104]	0.048(26), [0.022, 0.075]	1.49	0.02	0.96	0.006
	A_{3m}	0.0109(34), [0.0075, 0.0143]	0.060(40), [0.019, 0.099]		0.02	0.97	< 0.001
Peel	A_{1y}	0.0159(36), [0.0122, 0.0195]	0.062(19), [0.043, 0.080]	1.28	0.003	0.33	0.01
	A_{3m}	0.0199(62), [0.0137, 0.0262]	0.081(32), [0.048, 0.113]		0.002	0.36	< 0.001

Notation: A – fit parameter, c.i. – confidence interval of A, χ^2_{red} – reduced chi-square of LSS fit. P-values denote significance of two-tailed t-tests between: LMS and LSS estimates of A (p_a), LMS whole-parotid estimate of A and indicated sub-segment (p_w), and LMS estimate of A produced by early, overall loss and recovery (p_l).

Numbers in parentheses denote uncertainty in the preceding digits.

All fit parameters were estimated via bootstrap using at least 50000 resamples.

Similar to table 7.4, table 7.5 displays p-values denoting the significance of various two-tailed t-tests. The first, p_a, indicated significant discrepancy between LSS and LMS fit parameters. The χ^2_{red} appear reasonable, but this is most likely happenstance; flat planes fit to the same data (and several other functional forms) produce identical χ^2_{red}.

The second, p_w, which measures whether the sub-segment behaves differently from the whole parotid, was not significant in any case.

The third, p_l, measures whether the LMS fit parameters as computed by fitting early and overall loss matched those computed by fitting recovery. In most cases, they were significantly distinct while the medial halves were only poorly distinct. This suggests that they are not being adequately recovered from the data.

7.2.4 Lateral Dimorphism

During our analysis, the lateral and medial sub-segments hinted at possible dimorphism – though in no case were salivary output loss or recovery predictions statistically significant from the whole parotid.

Still, NPLLR trends tended to show noticeable 'shoulders' in lateral sub-segments which were generally less pronounced in the medial or whole sub-segments. In fig. 7.3, the terminus of the lateral sub-segment is shortened by nearly $30\,Gy$ compared to both the medial and whole. Parametric fitting, though not significant for predicting loss, displayed a curious phenomena: the response of the whole parotid was always bounded within the medial and lateral sub-segments. The lateral sub-segment response was consistently steeper, and the medial sub-segment response shallower, than the whole parotid. Indeed, the two-tailed t-test suggests a significant distinction between A predicted by early loss and overall loss for the lateral sub-segment ($p = 0.01$). The medial sub-segment is less distinct ($p = 0.2$), suggesting that the lateral sub-segment may recover more after RT.

One possible explanation for this discrepancy is the inhomogeneous distribution of mean dose amongst the sub-segments. The medial portion of the parotid is generally closer to the Planning Target Volume (PTV) than the

lateral portion during head-and-neck RT. The PTV usually contains cervical lymph node chains which need to be irradiated in case they are harbouring malignant cells, so the medial portion of the parotid is expected to receive a higher dose than the lateral portion, on average. Indeed, this is the case for the data set considered, as can be verified in fig. 7.8.

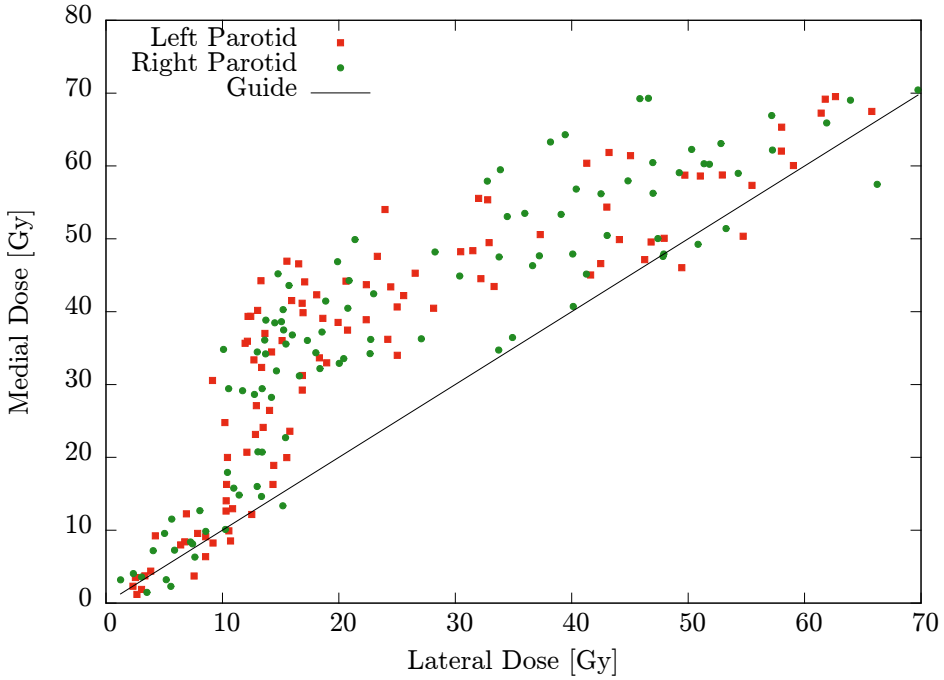

Figure 7.8: Dose to medial sub-segment vs. dose to lateral sub-segment with guide denoting homogeneity. The departure of the dose distribution from homogeneity is apparent.

7.2.5 Monte Carlo Model Sampling

Expressions for $\phi(D_L, D_R)$ were generated independently for whole organ, lateral halves, and core-and-peel sub-segments. Expressions constructed included a maximum of 15 entities and an unlimited number of arithmetical operators $+ - \times /$ (including negation). Entities allowed included variables, constants, and a handful of functions. Dose to left and right sub-segments

comprised the variables while numerical values within $[0, 100]$ (with units of Gy) were selected as constants. Functions included `min`, `max`, and `abs` with logical constructs `if..then..else` and both greater-than and less-than. Monotonic functions are transparent with regard to rank correlation, and exponentials tended to skew the confidence bars during NPLLR. Though monotonic functions were investigated, they were of lesser analytical value and are excluded from this report.

To investigate the suitability of the data for expressing the chosen definitions of early loss, recovery, and overall loss, simultaneous analysis was performed using S_b, S_{3m}, and S_{1y} for modeling the left-hand side of eq. (7.2) (i.e., ΔS).

One million expressions were generated with no specific arithmetical structure, though all expressions were required to be numerically evaluable and have consistent dimensionality. The sampled expressions resulted in a wide range of R_s, which is illustrated in fig. 7.9. No significant differences existed between the various sub-segmentations. The highest R_s occurred predominantly when ΔS described early loss or overall loss and ϕ was of the form $min\,(D_L, N) + max\,(D_R, N)$. The steepest response occurred when N was $40 - 50\,Gy$, though the precise choice of N was of insignificant consequence. The `min` function explicitly limited the amount of dose each parotid contributed to the common dose. In other words, it stopped excessive dose above $N\,Gy$ from spilling over and obscuring the dose contribution signal from the opposing parotid; intuitively, if a parotid were to receive a high dose (e.g., $60\,Gy$) or two times that dose ($120\,Gy$), the parenchymal damage would probably be no higher because it would not be possible to inflict further damage.

Unfortunately, the use of rank correlation meant that salivary output recovery could not be examined in this manner.

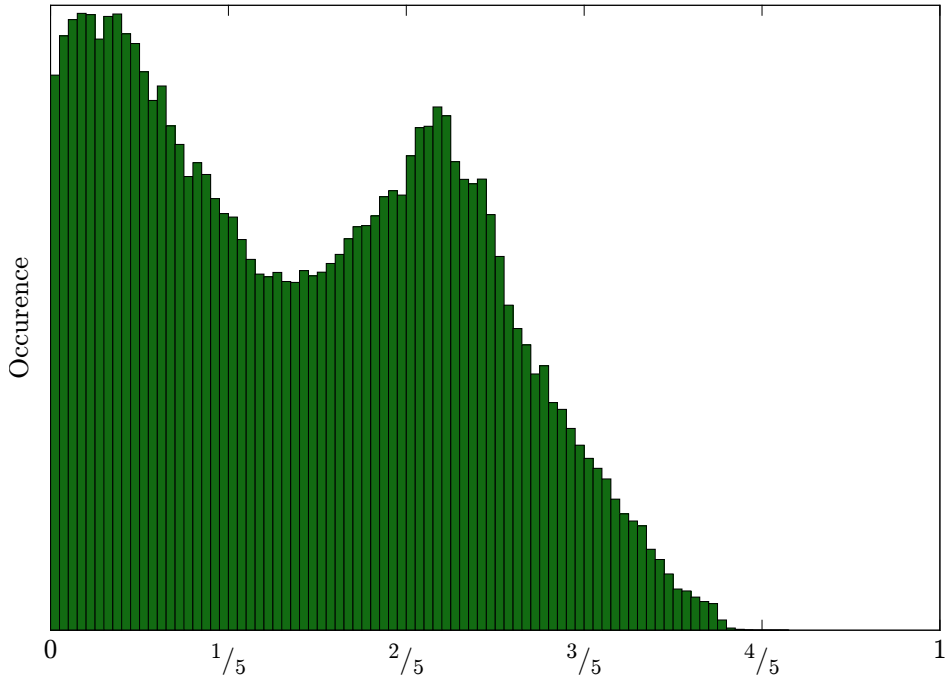

Figure 7.9: Histogram of $|R_s|$ resulting from random sampling of ϕ and ΔS for laterally-sub-segmented halves. The abscissæ shows $|R_s|$ and the ordinate the number of samples within a bin of width 0.02 (i.e., occurence). Duplicate samples were eliminated. The strongest R_s (and thus response) possible occurred when the data described salivary output loss.

7.3 Discussion

> *A good decision is based on knowledge and not on numbers.*
>
> PLATO

Xerostomia is one of the most likely late toxic effects of RT in patients with head-and-neck cancers [37]. It has a strong, broad impact on a patient's QOL, affecting speech, oral comfort, oral health, and taste. Severe cases impose on domains not directly related to xerostomia, such as emotion and pain [46] [66]. Modern RT techniques such as IMRT are able to incorporate

knowledge of complication risks into normal tissue sparing. Therefore, the precise quantification of regional dose-dependence of salivary output loss and recovery could rapidly lead to improvements in patient QOL. To this end, we examined stimulated salivary output in a prospective study.

Previous studies have found that xerostomia can be effectively modeled by mean dose to whole parotid. Additional clinical factors can increase overall predictive power of such models, but only by small amounts. Therefore, simplified models are generally recommended [31]. Following these recommendations we considered a single predictive factor: mean dose to a variety of geometrically defined sub-segments.

Our analysis was constrained by the variability of salivary flow measurements – a result which has been described as unavoidable in the literature (see section 3.1.1). It was found that the salivary measurements were too noisy to accommodate more complex organ sub-segmentation without precise anatomical or functional information as a guide. Such information was absent from our analysis.

The whole mouth salivary output measured (S) is, in reality, the sum of contributions from various organs. Compensation and unequal contributions were not accounted for in parametric models, but were not excluded during non-parametric analysis. Explicit modeling of the disregarded sub-segments and organs was initially attempted. Our findings confirmed those of Blanco et al. [10] – the inclusion of extra factors did not improve the fit nor indicate any additional trend. As Blanco et al., we suspect noise in salivary output measurements has masked the true signal.

The computed Spearman's rank correlation coefficients indicate definitive dose-dependence of salivary loss (vs. combined dose). In a clinical setting it may be possible to spare $5 - 10\,Gy$ to reduce normal tissue complication if one is reasonably confident reduction will occur. The predictive power of our non-parametric models over this dose difference is questionable, considering the large confidence intervals, and forming a clinical opinion based solely on our model or trends is not recommended.

It is noteworthy that early loss in either whole or laterally sub-segmented parotid was strongly predicted by non-parametric combined dose. The NPLLR

trends, despite being unconstrained to any particular functional form, predict regions of no loss ($< 5\,Gy$) and near total loss ($> 80\,Gy$) which are consistent with observation. Non-parametric Spearman's rank correlation coefficients tabulated in table 7.3 for a variety of sub-segments demonstrate strong significance and, therefore, dose dependence. Qualitatively, the NPLLR trends (see fig. 7.2) do not appear to suit the data due to the clustering at $40-60\,Gy$. Lateral sub-segmentation helped to disperse the cluster and qualitatively describes the data more appropriately.

Statistically significant recovery was clearly observed. A median recovery of 12 % of baseline stimulated function occurred over nine months (from three months to one year post-RT). The medians of both early and overall loss of salivary function represent grade III xerostomia on the LENT-SOMA scale, despite the general presence of recovery. No significant dose-dependence was noted in recovery. Possible explanations include systematic biases, such as our timescale being inappropriate, or an ineffective measurement technique resulting in excess noise. Though recovery constructed from measurements was dominated by noise, it seems unlikely that a mean increase of salivary output more than $20-30\%$ of baseline is possible at any dose using current clinical guidelines. Still, the maximum recovery appears to be $\approx 80\%$ of baseline output (fig. 7.1), so we are hopeful that the integration of functional or anatomical information will provide the means to improve prospects of recovery.

Though the parametric analysis was generally not significant, the two-tailed t-test suggests a significant distinction between A predicted by early loss and overall loss for the lateral sub-segment ($p = 0.01$). The medial sub-segment is less distinct ($p = 0.2$), suggesting that the lateral sub-segment underwent more recovery than the medial (i.e., lateral dimorphism).

Compared with whole-organ, sub-segmentation appeared to offer no significant improvement in the prediction of salivary output loss or recovery, though non-significant improvement was seen in the lateral halves sub-segmentation. Stronger rank correlation was noted in the medial sub-segment over a wide range of Monte Carlo sampled expressions. However, steeper combined dose response was seen in the lateral sub-segment in both LMS- and LSS-predicted

salivary output loss. Further, a suspicious 'shelf' in salivary loss occurred preferentially in the lateral sub-segment. It is unclear whether lateral dimorphism was genuinely present, or an artifact of the mean dose distribution of the data (see fig. 7.8) due to the noise. If genuinely present, it would suggest regional dose-dependence.

The possibility of dimorphism notwithstanding, fig. 7.8 and the differences in termination dose of figs. 7.3 and 7.6 show that, in a clinical sense, there are practical differences between the medial and lateral sub-segments. Although we cannot claim higher dose susceptibility in the lateral sub-segment (and, indeed, even if it were slightly *less* susceptible than the medial sub-segment) we recommend that the medial and lateral sub-segments be treated separately during treatment planning. Presently the medial sub-segment is often included in the PTV and is constrained to receive a minimum prescribed dose, skewing the whole parotid mean dose. Little time is spent optimizing the whole parotid mean dose because of the hard limits imposed on the medial edge. The standard PTV dose for treatment of potential microscopic disease is $50\,Gy$, or $70\,Gy$ for macroscopic disease. The medial edge (which may receive more dose) often has little chance of post-RT function or recovery. However, if the lateral sub-segment were separately and aggressively optimized for mean dose reduction, significant post-RT function or recovery may result. Ceteris paribus, based on the non-parametric trend of fig. 7.3, sparing to a combined dose of $\leq 50\,Gy$ is recommended for a patient to keep 40% of baseline function and avoid the severe xerostomia. This recommendation should be treated scrutinously until it can be vindicated with an independent set of data.

To explore and quantify the utility of our chosen definitions of early loss and overall loss, we developed a novel Monte Carlo model sampling technique. To our knowledge, such a technique has not previously been investigated as a means of model evaluation, and warrants further investigation. Currently, the technique allows us to explore the wider functional neighbourhood of suspected models. Such information provides an estimate of the maximum significance attainable using the data.

As discussed in section 7.2.5, the model sampling technique recovered the

structure of early loss and overall loss by random sampling using arithmetical and functional primitives. Our results show that the best possible rank-correlation occurs when the salivary signal ϕ is chosen to express salivary loss. Likewise, the combined dose ϕ which resulted in the strongest response indicated that mean doses larger than $40 - 50\,Gy$ play little part in salivary output loss, in agreement with current clinical recommendations. Variants in definition of loss and recovery (e.g., non-normalized differences in salivary measurement) were of lesser predictive power than early loss, overall loss, and recovery.

Finally, it may seem that our findings differ from those of recent reports which indicate significant regional and morphological dose dependence in parotids [89] [12] [13]. Though we were unable to see any significant regional dose-dependence, our results do not preclude the possibility they exist. In fact, several insignificant hints seem to suggest they exist but are being obscured by noise. Further, subtle discrepancies exist which make comparison of findings difficult. The most important is the form of xerostomia considered: Buettner et al. [13] defined xerostomia as subjective grades II, III, and IV xerostomia on the LENT-SOMA scale, while we have opted to use objective salivary flow. The distinction resulted in completely different approaches; they performed logistic regression while we relied on continuous parametric and non-parametric techniques. Parametric modeling is likely to encounter more snags (e.g., model specification) and is more responsive to measurement error than logistic regression, but is, we believe, capable of more precisely and conclusively demonstrating regional dose dependence. Another key discrepancy is that we have considered different factors. It is possible that the moments considered by Buettner et al. [13] are able to describe dose dependence which is not visible with regional sub-segment mean dose. We feel, however, that the machinery of regional sub-segmentation is an overall superior technique for the conclusive demonstration of regional dose dependence.

7.4 Conclusion

The regional dose dependence of salivary output loss and recovery was investigated using regionally sub-segmented parotid volumes.

Significant salivary output loss and recovery were observed. Significant mean dose dependence was found for loss, but not recovery. The use of sub-segmentation appears to predict neither loss nor recovery with any greater precision than whole-parotid mean dose, though it is not any worse.

Considerable salivary measurement noise was encountered. It is believed, though not significantly demonstrated, that noise has obscured dimorphism between medial and lateral halves of the parotid. Regardless, we recommend considering medial and lateral parotid volumes separately during RT planning.

We stress that our results do not preclude the possibility of inhomogeneous susceptibility in specific anatomical or functionally-important regions of the parotid. However, it seems unlikely that the mean magnitude of recovery would exceed $20 - 30\%$ of baseline output at any dose using the current clinical guidelines.

Chapter 8

Conclusions

Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.

Sir Winston Churchill

The purpose of this thesis was twofold. The first was to demonstrate that the parotid is inhomogeneously responsible for xerostomia. The second was to quantify the regional dose dependence of recovery, if any exists.

Significant salivary output loss and recovery were observed. No significant mean dose dependence was observed in recovery, though we were unable to exclude it as a possibility. The use of sub-segmentation appears to predict neither loss nor recovery with any greater precision than whole-parotid mean dose. Quantification of loss was successful with non-parametric techniques, but attempts to quantify recovery were no more successful at prediction than the mean. Our findings differ from those in recent reports which indicate significant regional and morphological dose dependence, though they do not appear to be inconsistent [89] [12] [13]. Specifically, the extreme noise encountered led to difficulties in all aspects of analysis.

We stress that our results do not preclude the possibility of inhomogeneous dose susceptibility in specific anatomically- or functionally-important regions of the parotid. The use of anatomical or functional information will most likely improve the utility of sub-segmentation. For the purposes of recovery,

it seems unlikely that the amount of mean recovery would exceed $20-30\%$ of baseline output at any dose using the current clinical guidelines. Recovery up to $\approx 80\%$ of baseline output was observed, suggesting there may be an opportunity to improve mean recovery.

8.1 Summary of Contributions

1. Demonstrated both significant salivary output loss and recovery.
2. Demonstrated significant non-parametric dose-dependence in salivary output loss, insignificant dependence in recovery.
3. Showed that, regardless of dose, recovery was not seen to exceed $20-30\%$ of baseline output at any dose using current clinical guidelines.
4. Found evidence that, using existing RT techniques and whole-parotid mean doses, patients are unlikely to recover more than $20-30\%$ of baseline output at any dose.
5. Provided an argument for the clinical treatment of the parotid as a collection of distinct regions (i.e., lateral and medial sub-segments), regardless of the presence of regional dose susceptibility.
6. Developed a set of computational techniques which are able to identify previously unseen contours in a semi-automated fashion. Improvements in recognition over standard techniques was demonstrated.
7. Developed a computational system for performing dose-volume computation on arbitrarily complex sub-segments. Verification against Varian's Eclipse™ showed strong agreement.
8. Developed a computational system which is able to rapidly, precisely, and reversibly sub-segment contours into arbitrarily complex volumes.
9. Introduced a semi-parametric Monte Carlo model sampling technique for exploration and significance-testing of measured data when a model is unknown or partly known.

8.2 Avenues for Future Research

This study could be extended in many ways. Primarily, additional sources of information are required to overcome the high variability of salivary flow measurements.

Anatomically- or functionally-guided sub-segmentation could be achieved with minimal impact on patients using non-ionizing MR techniques. Using such factors, the inclusion of clinical, disease, or non-parotid salivary contribution factors may significantly improve overall predictive power, and should be re-examined when better (or more) data is available.

Recently, researchers have attempted to transplant stem cells into dysfunctional parotid glands following RT [34] [47] [80]. Better anatomical understanding of the nature of xerostomia could lead to identification of more relevant factors pertinent to xerostomia. Alternatively, identification of the regions of the parotid particularly susceptible to radiation damage may provide more accurate targets for implantation.

The use of unstimulated salivary output measurements, patient-reported xerostomia or QOL questionnaires, and non-parotid salivary gland doses are likely to improve prediction of salivary output and xerostomia risk, though the increase in predictive power is thought to be small.

Finally, the use of more powerful statistical techniques which are able to incorporate multiple factors, such as Bayesian multivariate logistic regression, may lead to more confident predictions of xerostomia at the same noise level. In tandem, more powerful continuous methods could help demonstrate regional dose susceptibility of salivary output loss; overall, a more varied barrage of robust statistical techniques would ensure the failings of any particular technique would not spoil the entire analysis. Further development of the Monte Carlo technique of section 7.1.9 may lead to better development of models.

Bibliography

[1] *The Practitioner*. Number v. 42. The Practitioner Limited, 1889. URL http://books.google.ca/books?id=52ECAAAAYAAJ. → pages 9

[2] K. Al-Qahtani, M. P. Hier, K. Sultanum, and M. J. Black. The role of submandibular salivary gland transfer in preventing xerostomia in the chemoradiotherapy patient. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 101(6):753–756, 2006. → pages 8

[3] C. M. Amosson, B. S. Teh, T. J. Van, N. Uy, E. Huang, W.-Y. Mai, A. Frolov, S. Y. Woo, J. K. Chiu, L. S. Carpenter, et al. Dosimetric predictors of xerostomia for head-and-neck cancer patients treated with the smart (simultaneous modulated accelerated radiation therapy) boost technique. *International Journal of Radiation Oncology Biology Physics*, 56(1):136–144, 2003. → pages 9, 21

[4] E. Astreinidou, J. M. Roesink, C. P. Raaijmakers, L. W. Bartels, T. D. Witkamp, J. J. Lagendijk, and C. H. Terhaard. 3d mr sialography as a tool to investigate radiation-induced xerostomia: feasibility study. *International Journal of Radiation Oncology Biology Physics*, 68(5): 1310–1319, 2007. → pages 10

[5] C. L. Bartels. Xerostomia information for dentists. *Oral Cancer Foundation Cite*, 7, 2005. → pages 8, 12

[6] A. Bartley. Suppression of the saliva. *Medical Times and Gazette*, 54: 603, 1868. (Published as a letter to the editor). → pages 9

[7] S. M. Bentzen, L. S. Constine, J. O. Deasy, A. Eisbruch, A. Jackson, L. B. Marks, R. K. Ten Haken, and E. D. Yorke. Quantitative analyses of normal tissue effects in the clinic (quantec): an introduction to the scientific issues. *International Journal of Radiation Oncology Biology Physics*, 76(3):S3–S9, 2010. → pages 10, 22

[8] M. Bergdahl and J. Bergdahl. Low unstimulated salivary flow and subjective oral dryness: association with medication, anxiety, depression, and stress. *Journal of Dental Research*, 79(9):1652–1658, 2000. → pages 14

[9] W. D. Bidgood and S. C. Horii. Introduction to the acr-nema dicom standard. *Radiographics*, 12(2):345–355, 1992. → pages xv

[10] A. I. Blanco, K. Chao, I. El Naqa, G. E. Franklin, K. Zakarian, M. Vicic, and J. O. Deasy. Dose–volume modeling of salivary function in patients with head-and-neck cancer receiving radiotherapy. *International Journal of Radiation Oncology Biology Physics*, 62(4):1055–1069, 2005. → pages 7, 9, 11, 14, 15, 16, 17, 18, 19, 21, 22, 39, 43, 45, 57, 58, 67

[11] S. T. Buckland, K. P. Burnham, and N. H. Augustin. Model selection: an integral part of inference. *Biometrics*, pages 603–618, 1997. → pages 2

[12] F. Buettner, S. L. Gulliford, S. Webb, M. Partridge, A. B. Miah, K. J. Harrington, and C. M. Nutting. Using a bayesian feature-selection algorithm to identify dose-response models based on the shape of the 3d dose-distribution: an example from a head-and-neck cancer trial. In *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, pages 740–745. IEEE, 2010. → pages 70, 72

[13] F. Buettner, A. Miah, S. Gulliford, E. Hall, K. Harrington, S. Webb, M. Partridge, and C. Nutting. Novel approaches to improve the therapeutic index of head and neck radiotherapy: an analysis of data from the parsport randomised phase iii trial. *Radiotherapy and Oncology*, 103(1):82, 2012. → pages 8, 17, 21, 22, 24, 33, 39, 70, 72

[14] F. R. Burlage, J. Pijpe, R. P. Coppes, M. E. Hemels, H. Meertens, A. Canrinus, and A. Vissink. Variability of flow rate when collecting stimulated human parotid saliva. *European Journal of Oral Sciences*, 113(5):386–390, 2005. → pages 14, 15, 22, 57

[15] S. Buus, C. Grau, O. L. Munk, A. Rodell, K. Jensen, K. Mouridsen, and S. Keiding. Individual radiation response of parotid glands investigated by dynamic ^{11}c-methionine pet. *Radiotherapy and Oncology*, 78(3):262–269, 2006. → pages 10

[16] K. Chao, J. O. Deasy, J. Markman, J. Haynie, C. A. Perez, J. A. Purdy, and D. A. Low. A prospective study of salivary function sparing in

patients with head-and-neck cancers receiving intensity-modulated or three-dimensional radiation therapy: initial results. *International Journal of Radiation Oncology Biology Physics*, 49(4):907–916, 2001. → pages 8, 12, 16, 17, 18, 19, 89

[17] S. Cheng, V. Wu, D. Kwong, and M. Ying. Assessment of post-radiotherapy salivary glands. *British Journal of Radiology*, 84 (1001):393–402, 2011. → pages 3, 5, 6, 7, 9

[18] H. Clark, J. Wu, V. Moiseenko, R. Lee, B. Gill, C. Duzenli, and S. Thomas. Semi-automated contour recognition using DICOMautomaton. *Journal of Physics: Conference Series*, 2013 – In press. → pages 25

[19] H. Clark, S. Thomas, V. Moiseenko, R. Lee, B. Gill, C. Duzenli, and J. Wu. Automated segmentation and dose-volume analysis with DICOMautomaton. *Journal of Physics: Conference Series*, 2013 – In press. → pages 33

[20] L. Clifton Stephens, K. Kian Ang, T. E. Schultheiss, G. K. King, W. A. Brock, and L. J. Peters. Target cell and mode of radiation injury in rhesus salivary glands. *Radiotherapy and Oncology*, 7(2):165–174, 1986. → pages 8

[21] J. D. Cox, J. Stetz, and T. F. Pajak. Toxicity criteria of the radiation therapy oncology group (rtog) and the european organization for research and treatment of cancer (eortc). *International Journal of Radiation Oncology Biology Physics*, 31(5):1341–1346, 1995. → pages 10

[22] C. Dawes. Circadian rhythms in human salivary flow rate and composition. *The Journal of physiology*, 220(3):529–545, 1972. → pages 5

[23] J. O. Deasy, V. Moiseenko, L. Marks, K. Chao, J. Nam, and A. Eisbruch. Radiotherapy dose–volume effects on salivary gland function. *International Journal of Radiation Oncology Biology Physics*, 76(3):S58–S63, 2010. → pages 21

[24] J. Del Regato. Dental lesions observed after roentgen therapy in cancer of the buccal cavity, pharynx and larynx. *Am J Roentgenol*, 42:404–410, 1939. → pages 11

[25] F. Denis, P. Garaud, E. Bardet, M. Alfonsi, C. Sire, T. Germain, P. Bergerot, B. Rhein, J. Tortochaux, P. Oudinot, et al. Late toxicity results of the gortec 94-01 randomized trial comparing radiotherapy with concomitant radiochemotherapy for advanced-stage oropharynx carcinoma: comparison of lent-soma, rtog-eortc, and nci-ctc scoring systems. *International Journal of Radiation Oncology Biology Physics*, 55(1):93–98, 2003. → pages 10

[26] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. → pages 28

[27] T. Dijkema, C. H. Terhaard, J. M. Roesink, P. M. Braam, C. H. van Gils, M. A. Moerland, and C. P. Raaijmakers. Large cohort dose–volume response analysis of parotid gland function after radiotherapy: intensity-modulated versus conventional radiotherapy. *International Journal of Radiation Oncology Biology Physics*, 72(4): 1101–1109, 2008. → pages 20

[28] B. Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26, 1979. → pages 46

[29] A. Eisbruch, R. K. Ten Haken, H. M. Kim, L. H. Marsh, and J. A. Ship. Dose, volume, and function relationships in parotid salivary glands following conformal and intensity-modulated irradiation of head and neck cancer. *International Journal of Radiation Oncology* Biology* Physics*, 45(3):577–587, 1999. → pages 21

[30] A. Eisbruch, H. M. Kim, J. E. Terrell, L. H. Marsh, L. A. Dawson, and J. A. Ship. Xerostomia and its predictors following parotid-sparing irradiation of head-and-neck cancer. *International Journal of Radiation Oncology Biology Physics*, 50(3):695–704, 2001. → pages 8, 9, 17, 18, 39

[31] I. El Naqa, J. Bradley, A. I. Blanco, P. E. Lindsay, M. Vicic, A. Hope, and J. O. Deasy. Multivariable modeling of radiotherapy outcomes, including dose–volume and clinical factors. *International Journal of Radiation Oncology Biology Physics*, 64(4):1275–1286, 2006. → pages 2, 15, 16, 19, 20, 45, 67

[32] C.-M. Eneroth, C. Henrikson, and P. Jakobsson. The effect of irradiation in high doses on parotid glands. *Acta otolaryngologica*, 71 (1-6):349–356, 1971. → pages 12

[33] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004. → pages 38

[34] J. Feng and R. P. Coppes. Can we rescue salivary gland function after irradiation. *Sci World J*, 8:959–962, 2008. → pages 8, 20, 74

[35] D. Fischer and J. A. Ship. Effect of age on variability of parotid salivary gland flow rates over time. *Age and Ageing*, 28(6):557–561, 1999. → pages 15

[36] P. Fox, K. Busch, and B. Baum. Subjective reports of xerostomia and objective measures of salivary gland performance. *The Journal of the American Dental Association*, 115(4):581–584, 1987. → pages 9

[37] L. Franzén, U. Funegård, T. Ericson, and R. Henriksson. Parotid gland function during and following radiotherapy of malignancies in the head and neck: a consecutive study of salivary flow and patient discomfort. *European Journal of Cancer*, 28(2):457–462, 1992. → pages 8, 9, 39, 66

[38] U. Hoeller, S. Tribius, A. Kuhlmey, K. Grader, F. Fehlauer, and W. Alberti. Increasing the rate of late toxicity by changing the score? a comparison of rtog/eortc and lent/soma scores. *International Journal of Radiation Oncology Biology Physics*, 55(4):1013–1018, 2003. → pages 10

[39] A. C. Houweling, M. E. Philippens, T. Dijkema, J. M. Roesink, C. H. Terhaard, C. Schilstra, R. K. Ten Haken, A. Eisbruch, and C. P. Raaijmakers. A comparison of dose–response models for the parotid gland in a large group of head-and-neck cancer patients. *International Journal of Radiation Oncology Biology Physics*, 76(4):1259–1265, 2010. → pages 16, 20

[40] P. Jaccard. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901. → pages 28

[41] K. Jensen, A. Bonde Jensen, and C. Grau. The relationship between observer-based toxicity scoring and patient assessed symptom severity after treatment for head and neck cancer. a correlative cross sectional study of the dahanca toxicity scoring system and the eortc quality of life questionnaires. *Radiotherapy and Oncology*, 78(3):298–305, 2006. → pages 10

[42] M. Kam, S. Leung, B. Zee, P. Choi, M. Chau, K. Cheung, J. Suen, P. Teo, W. Kwan, and A. Chan. Impact of intensity-modulated

radiotherapy (imrt) on salivary gland function in early-stage nasopharyngeal carcinoma (npc) patients: A prospective randomized study. *J Clin Oncol*, 23(Suppl 1):5501, 2005. → pages 18

[43] D. E. Knuth. *The art of computer programming, Vol. 3*, volume 109. Addison-Wesley, Reading, MA, 1973. → pages 27

[44] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady*, volume 10, page 707, 1966. → pages 27

[45] H. Liem, R. A. V. Olmos, A. J. Balm, R. B. Keus, H. van Tinteren, R. P. Takes, S. H. Muller, A. M. Bruce, C. A. Hoefnagel, and F. J. Hilgers. Evidence for early and persistent impairment of salivary gland excretion after irradiation of head and neck tumours. *European Journal of Nuclear Medicine*, 23(11):1485–1490, 1996. → pages 9

[46] A. Lin, H. M. Kim, J. E. Terrell, L. A. Dawson, J. A. Ship, and A. Eisbruch. Quality of life after parotid-sparing imrt for head-and-neck cancer: a prospective longitudinal study. *International Journal of Radiation Oncology Biology Physics*, 57(1):61–70, 2003. → pages 8, 9, 10, 18, 22, 39, 66

[47] I. M. Lombaert, J. F. Brunsting, P. K. Wierenga, H. Faber, M. A. Stokman, T. Kok, W. H. Visser, H. H. Kampinga, G. de Haan, and R. P. Coppes. Rescue of salivary gland function after stem cell transplantation in irradiated glands. *PloS one*, 3(4):e2063, 2008. → pages 20, 74

[48] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *ACM Siggraph Computer Graphics*, volume 21, pages 163–169. ACM, 1987. → pages 38

[49] T. Makkonen and E. Nordman. Estimation of long-term salivary gland damage induced by radiotherapy. *Acta Oncologica*, 26(4):307–312, 1987. → pages 8

[50] L. B. Marks, S. M. Bentzen, J. O. Deasy, F.-M. S. Kong, J. D. Bradley, I. S. Vogelius, I. El Naqa, J. L. Hubbs, J. V. Lebesque, R. D. Timmerman, et al. Radiation dose–volume effects in the lung. *IJROBP*, 76(3):S70–S76, 2010. → pages 33

[51] L. Marmiroli, G. Salvi, A. Caiazza, L. Di Rienzo, M. Massaccesi, P. Murino, and G. Macchia. Dose and volume impact on radiation-induced xerostomia. *Rays*, 30(2):145–148, 2004. → pages 4

[52] H. E. Martin. Treatment of pharyngeal cancer. *Archives of Otolaryngology – Head and Neck Surgery*, 27(6):661, 1938. → pages 9, 10

[53] A. Meirovitz, C. A. Murdoch-Kinch, M. Schipper, C. Pan, and A. Eisbruch. Grading xerostomia by physicians or by patients after intensity-modulated radiotherapy of head-and-neck cancer. *International Journal of Radiation Oncology* Biology* Physics*, 66(2): 445–453, 2006. → pages 18

[54] V. Moiseenko, J. Wu, A. Hovan, Z. Saleh, A. Apte, J. O. Deasy, S. Harrow, C. Rabuka, A. Muggli, and A. Thompson. Treatment planning constraints to avoid xerostomia in head-and-neck radiotherapy: an independent test of quantec criteria using a prospectively collected dataset. *International Journal of Radiation Oncology Biology Physics*, 82(3):1108–1114, 2012. → pages 10, 21

[55] C. Montani, R. Scateni, and R. Scopigno. Discretized marching cubes. In *Proceedings of the Conference on Visualization '94*, pages 281–287. IEEE Computer Society Press, 1994. → pages 38

[56] G. B. Moore, J. L. Kuhns, J. L. Trefftzs, and C. A. Montgomery. *Accessing individual records from personal data files using non-unique identifiers*. US Department of Commerce, National Bureau of Standards, 1977. → pages 27

[57] H. R. Mortensen, K. Jensen, K. Aksglæde, M. Behrens, and C. Grau. Late dysphagia after imrt for head and neck cancer and correlation with dose–volume parameters. *Radiotherapy and Oncology*, 2013. → pages 22

[58] K. Mossman, A. Shatzman, and J. Chencharick. Long-term effects of radiotherapy on taste and salivary function in man. *International Journal of Radiation Oncology Biology Physics*, 8(6):991–997, 1982. → pages 12

[59] M. Navazesh and C. Christensen. A comparison of whole mouth resting and stimulated salivary measurement procedures. *Journal of Dental Research*, 61(10):1158–1162, 1982. → pages 19

[60] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965. → pages 44

[61] T. S. Newman and H. Yi. A survey of the marching cubes algorithm. *Computers & Graphics*, 30(5):854–879, 2006. → pages 38

[62] Y. Nishimura, K. Nakamatsu, T. Shibata, S. Kanamori, R. Koike, M. Okumura, and M. Suzuki. Importance of the initial volume of parotid glands in xerostomia for patients with head and neck cancers treated with imrt. *Japanese Journal of Clinical Oncology*, 35(7): 375–379, 2005. → pages 8, 17

[63] M. Nixon and A. S. Aguado. *Feature extraction & image processing.* Academic Press, 2008. → pages 28

[64] L. Philips. The double metaphone search algorithm. *CC Plus Plus Users Journal*, 18(6):38–43, 2000. → pages 27

[65] E. H. Pow, D. L. Kwong, A. S. McMillan, M. Wong, J. S. Sham, L. H. Leung, and W. K. Leung. Xerostomia and quality of life after intensity-modulated radiotherapy vs. conventional radiotherapy for early-stage nasopharyngeal carcinoma: initial report on a randomized controlled clinical trial. *International Journal of Radiation Oncology* Biology* Physics*, 66(4):981–991, 2006. → pages 18

[66] S. Rathod, T. Gupta, S. Ghosh-Laskar, V. Murthy, A. Budrukkar, and J. Agarwal. Quality-of-life (qol) outcomes in patients with head and neck squamous cell carcinoma (hnscc) treated with intensity-modulated radiation therapy (imrt) compared to three-dimensional conformal radiotherapy (3d-crt): Evidence from a prospective randomized study. *Oral Oncology*, 2013. → pages 10, 22, 39, 66

[67] S. Rauch. *Die Speicheldrüsen des Menschen: Anatomie, Physiologie und klinische Pathologie.* Georg Thieme, 1959. → pages 3

[68] J. M. Roesink, M. Schipper, W. Busschers, C. P. Raaijmakers, and C. H. Terhaard. A comparison of mean parotid gland dose with measures of parotid gland function after radiotherapy for head-and-neck cancer: Implications for future trials. *International Journal of Radiation Oncology Biology Physics*, 63(4):1006–1009, 2005. → pages 18

[69] F. S. Rosen and B. J. Bailey. Anatomy and physiology of the salivary glands. *Grand rounds presentation. UTMB, Otolaryngology*, 2001. → pages 5, 7

[70] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984. → pages 46

[71] M. Schiødt and N. Hermund. Management of oral disease prior to radiation therapy. *Supportive Care in Cancer*, 10(1):40–43, 2002. → pages 11

[72] G. Seifert. Spezielle pathologische anatomie. *Mundhöhle, Mundspeicheldrüsen, Tonsillen und Rachen*, pages 257–259, 1966. → pages 3

[73] H. Seikaly, N. Jha, T. McGaw, L. Coulter, R. Liu, and D. Oldring. Submandibular gland transfer: a new method of preventing radiation-induced xerostomia. *The Laryngoscope*, 111(2):347–352, 2001. → pages 8

[74] K. W. Shertzer and M. H. Prager. Least median of squares: a suitable objective function for stock assessment models? *Canadian Journal of Fisheries and Aquatic Sciences*, 59(9):1474–1481, 2002. → pages 46

[75] D. J. Sheskin. *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2003. → pages 44

[76] G. C. Smith. Salivary gland dysfunction and xerostomia. Master's thesis, University of Sydney, 1980. → pages 12

[77] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. → pages 44

[78] L. M. Sreebny and A. Vissink. *Dry mouth, the malevolent symptom: a clinical guide*. Wiley. com, 2010. → pages 9, 12

[79] L. C. Stephens, T. E. Schultheiss, R. E. Price, K. K. Ang, and L. J. Peters. Radiation apoptosis of serous acinar cells of salivary and lacrimal glands. *Cancer*, 67(6):1539–1543, 1991. → pages 8

[80] Y. Sumita, Y. Liu, S. Khalili, O. M. Maria, D. Xia, S. Key, A. P. Cotrim, E. Mezey, and S. D. Tran. Bone marrow-derived cells rescue salivary gland function in mice with head and neck irradiation. *The International Journal of Biochemistry & Cell Biology*, 43(1):80–87, 2011. → pages 20, 74

[81] L. A. Tabak. In defense of the oral cavity: structure, biosynthesis, and function of salivary mucins. *Annual review of physiology*, 57(1):547–564, 1995. → pages 4, 7

[82] J. M. Taylor. Kendall's and spearman's correlation coefficients in the presence of a blocking variable. *Biometrics*, pages 409–416, 1987. → pages 44

[83] K. Teshima, R. Murakami, E. Tomitaka, T. Nomura, R. Toya, A. Hiraki, H. Nakayama, T. Hirai, M. Shinohara, N. Oya, et al. Radiation-induced parotid gland changes in oral cancer patients: correlation between parotid volume and saliva production. *Japanese Journal of Clinical Oncology*, 40(1):42–46, 2010. → pages 17

[84] J. J. Thorn, H. S. Hansen, L. Specht, and L. Bastholt. Osteoradionecrosis of the jaws: clinical characteristics and relation to the field of irradiation. *Journal of Oral and Maxillofacial Surgery*, 58 (10):1088–1093, 2000. → pages 11

[85] C. Toldt and A. Dalla Rosa. *An atlas of human anatomy for students and physicians*, volume 3-4. Rebman Company, New York, 1919. URL http://archive.org/details/atlasofhumananat34told. Alternate location: http://www.biodiversitylibrary.org/item/95995 [Accessed 30th July 2013]. → pages x, 4, 5

[86] A. Trotti, A. D. Colevas, A. Setser, V. Rusch, D. Jaques, V. Budach, C. Langer, B. Murphy, R. Cumberlin, C. N. Coleman, et al. Ctcae v3. 0: development of a comprehensive grading system for the adverse effects of cancer treatment. In *Seminars in Radiation Oncology*, volume 13, pages 176–181. Elsevier, 2003. → pages 10

[87] H. Tsuni. Quantitative dose-response analysis of salivary function following radiotherapy using sequential ri-sialography. *International Journal of Radiation Oncology Biology Physics*, 11(9):1603–1612, 1985. → pages 8

[88] P. van Luijk, H. P. Bijl, A. W. Konings, A. J. Kogel, and J. M. Schippers. Data on dose–volume effects in the rat spinal cord do not support existing ntcp models. *International Journal of Radiation Oncology Biology Physics*, 61(3):892–900, 2005. → pages 20

[89] P. van Luijk, H. Faber, J. M. Schippers, S. Brandenburg, J. A. Langendijk, H. Meertens, and R. P. Coppes. Bath and shower effects in the rat parotid gland explain increased relative risk of parotid gland dysfunction after intensity-modulated radiotherapy. *International Journal of Radiation Oncology Biology Physics*, 74(4):1002–1005, 2009. → pages 17, 20, 22, 23, 70, 72

[90] A. Vissink, F. Burlage, F. Spijkervet, J. Jansma, and R. Coppes. Prevention and treatment of the consequences of head and neck radiotherapy. *Critical Reviews in Oral Biology & Medicine*, 14(3): 213–225, 2003. → pages 10, 11

[91] A. Wada, N. Uchida, M. Yokokawa, T. Yoshizako, and H. Kitagaki. Radiation-induced xerostomia: objective evaluation of salivary gland injury using mr sialography. *American Journal of Neuroradiology*, 30(1): 53–58, 2009. → pages 10

[92] L. Wasserman. *All of nonparametric statistics*, volume 4. Springer New York, 2006. → pages 44

[93] W. E. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Survey Research Methods*, pages 354–359. American Statistical Association, 1990. → pages 27

[94] W. E. Yancey. Evaluating string comparator performance for record linkage. *Statistics*, page 05, 2005. → pages 28

[95] K. Yu and M. Jones. Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association*, 99(465):139–144, 2004. → pages 44

[96] S. Yue, P. Pilon, and G. Cavadias. Power of the mann–kendall and spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology*, 259(1):254–271, 2002. → pages 44

Appendix A

Extended Comments: Recognition System

The lexicographical recognition system described in chapter 5 (hereafter referred to as simply `DICOMautomaton` for brevity) does not, in any way, require DICOM contour labels to function. It has been found to be practically useful in many situations. We discuss here how it differs from related techniques.

A.1 Comparison with Spell-Checkers

The use of indirection via a lexicon results in maximum flexibility in recognition. For instance, in the toy lexicon of section 5.1.1 the raw labels 'l parotid', 'lt_par', and 'Left Parotid' are made to refer to the unique label 'Left Parotid', but the fact that the unique label appears among the raw labels is coincidental; instead, we could have the raw labels 'l parotid', 'lt_par', and 'Left Parotid' refer to the unique label 'Left-side Salivary Organs.' This distinction would be impossible with spell-checkers. For example, a (good) spell checker would never suggest that 'lfet paoritd' should be replaced with 'Left-side Salivary Organs.' But `DICOMautomaton` can be easily configured to do so using the above lexicon.

This indirection – the ability to gauge similarity between *elements* of a

group, instead of the group itself – is the primary logical distinction between `DICOMautomaton` and a spell-checker.

A.2 Comparison with Macros

Indirection allows `DICOMautomaton` to be used as a powerful fuzzy-macro replacement system. A hierarchy of instances of `DICOMautomaton` could be used for compounded logical grouping. To illustrate this point, we give an extended example. Consider two toy lexicons. The first, Φ_1,

$$\begin{aligned} d_1 &= \{\text{Vancouver}, \text{Victoria}, \text{Kamloops}\} & c_1 &= \text{a\_British\_Columbian\_city} \\ d_2 &= \{\text{Edmonton}, \text{Fort McMurray}, \text{Calgary}\} & c_2 &= \text{an\_Albertan\_city} \\ d_3 &= \{\text{Melbourne}, \text{Sydney}, \text{Canberra}, \text{Perth}\} & c_3 &= \text{an\_Australian\_city} \end{aligned}$$

and the second, Φ_2,

$$\begin{aligned} d_1 &= \{\text{British Columbia}, \text{Alberta}, \text{Toronto}\} & c_1 &= \text{Canada} \\ d_2 &= \{\text{Melbourne}, \text{New Zealand}, \text{Australia}\} & c_2 &= \text{Australasia} \\ d_3 &= \{\text{Munich}, \text{Black Forest}\} & c_3 &= \text{Central Europe.} \end{aligned}$$

Now, consider the following sample text.

> Your author was born in Edmonton, grew up in Ft. MacMurray, and currently lives in Vancooover. He recently attended a conference in Melbin. Some day, he hopes to visit Munick.

Note the intentional spelling mistakes. Barring possible uncertainties of similarity, application of Φ_1 would result in the following.

> Your author was born in **an_Albertan_city**, grew up in **an_Albertan_city**, and currently lives in **a_British_Columbian_city**. He recently attended a conference in **an_Australian_city**. Some day, he hopes to visit Munick.

Application of Φ_2 to this text would result in the following.

> Your author was born in **Canada**, grew up in **Canada**, and currently lives in **Canada**. He recently attended a conference in **Australasia**. Some day, he hopes to visit **Central Europe**.

The composition of the lexicon can be used to hold single layers of a limited knowledge graph, the order of application of which is controlled by the user. However, more than a simple macro replacement system, `DICOM-automaton` is able to perform recognition. Thus, *fuzzy* macro replacement is possible, where the particular notion of fuzziness is specified and tunable by the user.

Appendix B

Saliva Collection Procedure

The saliva collection procedure used for this study and reproduced here has been detailed elsewhere. It is modeled after that described by Chao et al. [16] and is a standard technique which has seen extensive use in the literature.

Patients were asked to avoid consuming food one hour prior to measurement. After resting for a few minutes prior to collection and clearing the mouth of excess saliva, the patient was asked to lean forward and expectorate saliva into a collection cup. Over a five minute span, whole mouth unstimulated saliva was collected *without* the aid of suction cups. After collection the cup was weighed.

For stimulated measurement, patient's were given a small block of paraffin wax to chew on. The collection procedure was identical, save for the paraffin. Both procedures were supervised by qualified personnel to ensure conformity.

Measurements were taken prior to RT (baseline), and both three months and one year after the RT completion date. Measurements have also been taken two years after the RT completion date, but were not considered herein due to sparsity of the data. QOL questionnaires were issued at the same appointments, but likewise have not been considered.

No specific attempts were made to ensure patient collection occurred at a consistent time of day, though collection predominantly occured between 9:00 and 15:00.

Appendix C

Supplementary Plots

The purpose of this appendix is to provide the reader the opportunity to view the *raw* combined dose data of the plots in chapter 7; previously the data was displayed using histograms, and an estimate of the standard error (of each bin) was derived directly from the raw data. In contrast, these plots show the combined dose with an estimate of the standard error *of each point* which has been constructed from variances reported in the literature.

The pointwise standard errors shown were used only for χ^2 fitting, and are merely estimates which may not capture all sources of uncertainty. Their motivation is detailed in section 7.1.8 and their construction in section 7.2.2.

Conversely, the standard errors derived from the histogram bins do capture all observed uncertainty. However, they introduce many subtle effects: the choice of bin size and location affect the resulting errors, reduction of statistical power occurs due to the act of discarding the abscissæ coordinates and averaging, de-emphasis of clustering, suppression of divergent trends, and possible systematic bias which may be difficult to detect. The latter would occur if, for example, the ordinate changes too rapidly compared with the width of the bin, resulting in a consistently over-reported standard error.

Despite the nuances, the pointwise standard errors are in general agreement with the standard errors derived from binning data. The findings and conclusions described in chapter 7 agree in either case.

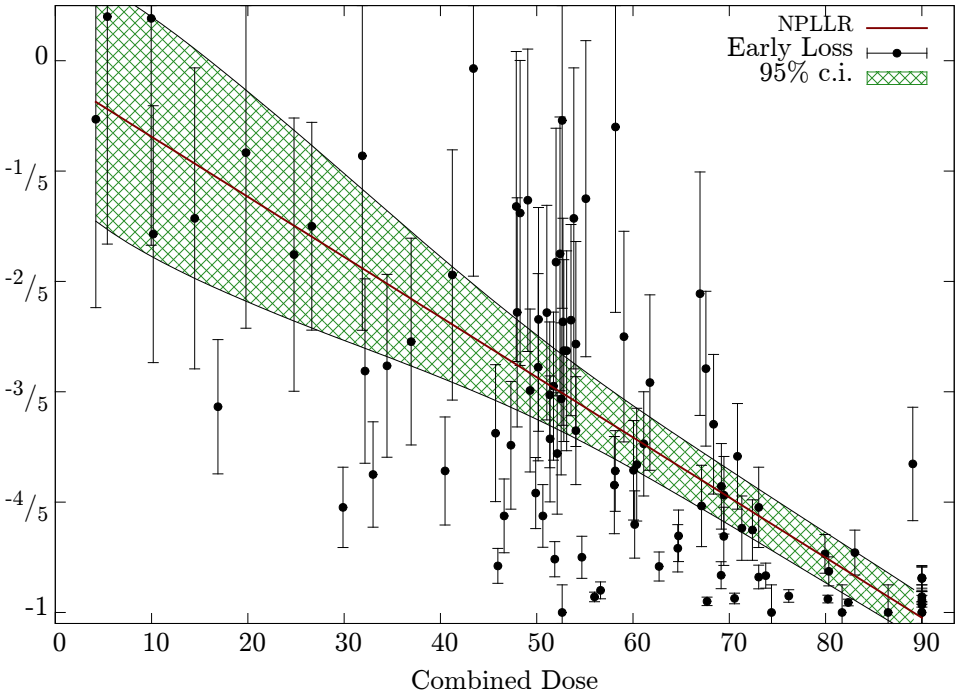

Figure C.1: Early loss $(S_{3m} - S_b)/S_b$ vs. combined dose $\phi = min(D_L, 45) + min(D_R, 45)$ with χ^2 pointwise variance for whole parotid (cf. fig. 7.2). The NPLLR trend and its 95% confidence interval are given.

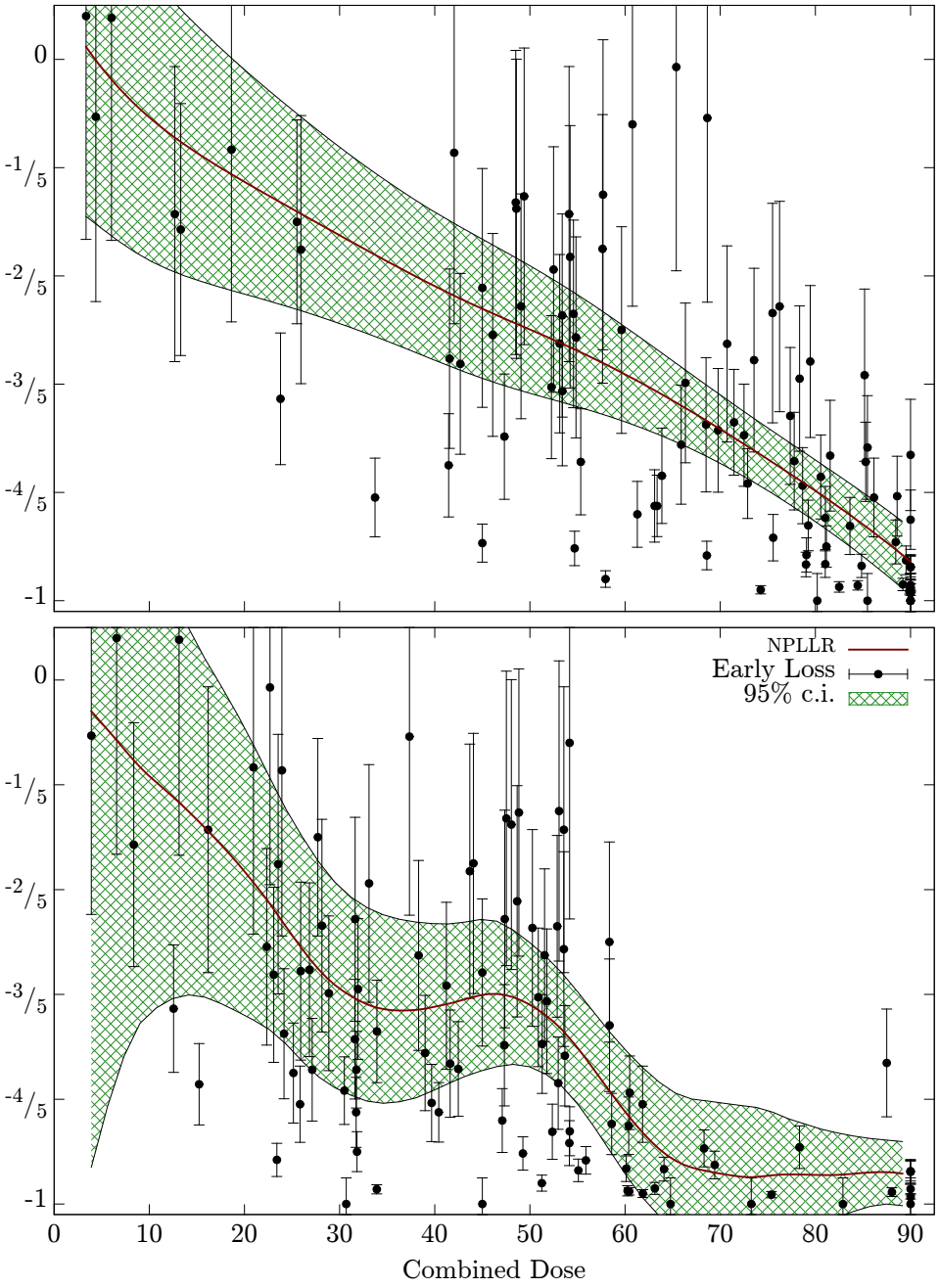

Figure C.2: Early loss $(S_{3m} - S_b)/S_b$ vs. combined dose $\phi = min(D_L, 45) + min(D_R, 45)$ with χ^2 pointwise variance for laterally-sub-segmented halves (of equal volume) of the parotid (cf. fig. 7.3). The medial half is on top and the lateral on bottom. The NPLLR trends appear, qualitatively, to better suit the data compared to the whole parotid of fig. C.1.

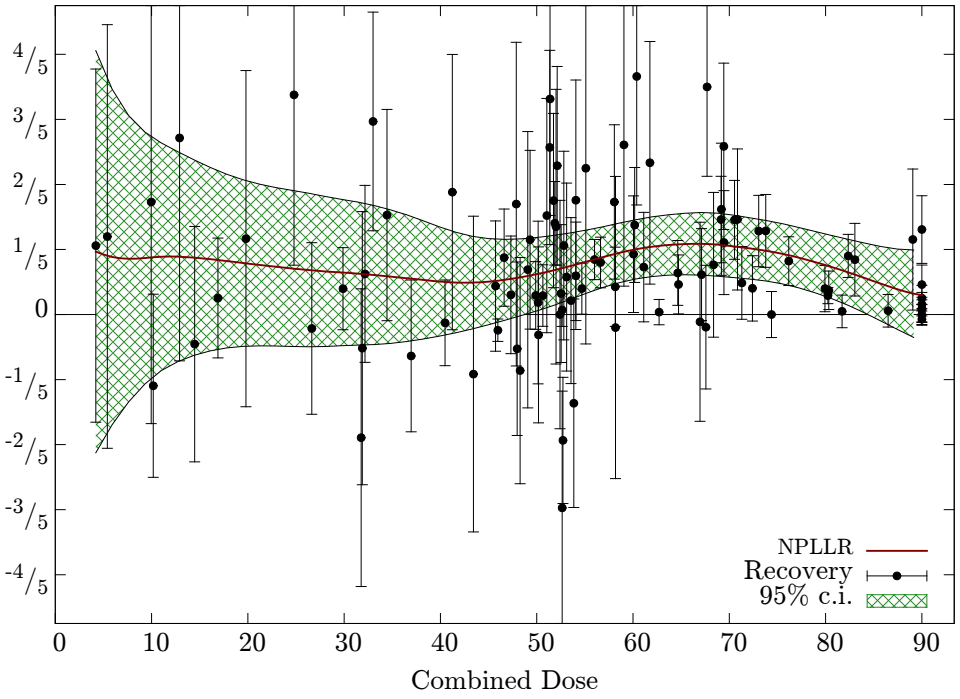

Figure C.3: Recovery $(S_{1y} - S_{3m})/S_b$ vs. combined dose $\phi = min(D_L, 45) + min(D_R, 45)$ with χ^2 pointwise variance for whole parotid (cf. fig. 7.5). The NPLLR trend and its 95% confidence interval are given.

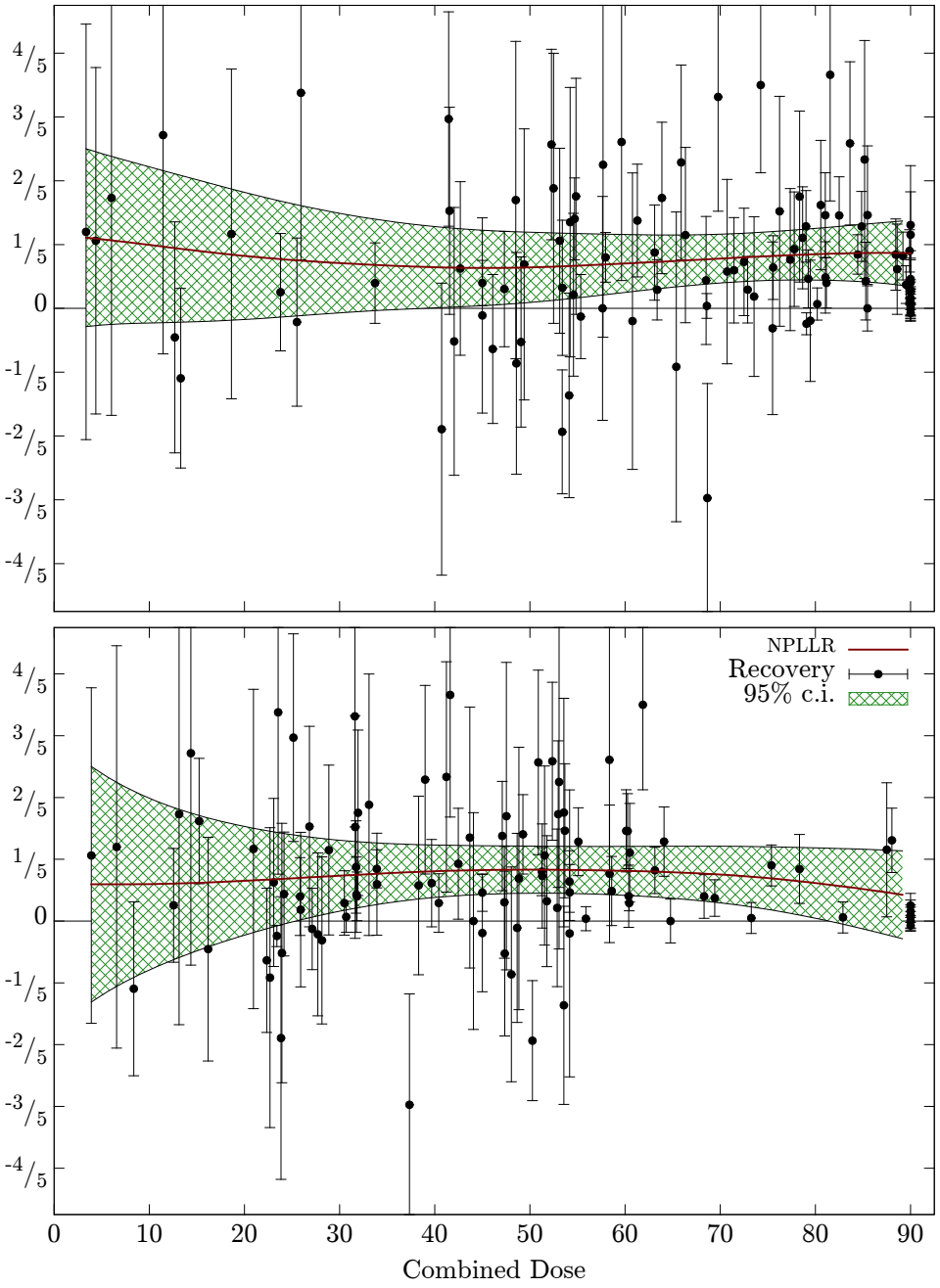

Figure C.4: Recovery $(S_{1y} - S_{3m})/S_b$ vs. combined dose $\phi = min(D_L, 45) + min(D_R, 45)$ with χ^2 pointwise variance for laterally-sub-segmented halves of the parotid (cf. fig. 7.6). The medial half is on top and the lateral on bottom. As in fig. C.2, the clustering around $40 - 60\,Gy$ is somewhat dispersed by sub-segmentation.