Modélisation flexible du risque d'événements iatrogènes radio-induits

Mohamed Amine Benadjaoud

UNIVERSITÉ PARIS SUD

Comprendre le monde, construire l'avenir®

École doctorale 420 Santé publique

Modélisation flexible du risque d'événements iatrogènes radio-induits

THÈSE

présentée et soutenue publiquement le 27 Mars 2015

pour l'obtention du

Doctorat de l'Université Paris-Sud XI

(mention Biostatistique)

par

Mohamed Amine Benadjaoud

Thèse dirigée par Messieurs
Florent de Vathaire et Hervé Cardot

Composition du jury

Président :	Pr. Bruno Falissard	Université Paris XI/INSERM U669
Rapporteurs :	Pr. Sophie Lambert-Lacroix	Université Pierre-Mendès-France de Grenoble
	Pr. Ahmadou Alioum	Université de Bordeaux II - ISPED
Examinateurs :	Dr. Dominique Laurier	Lepid IRSN(Fontenay-aux-Roses)
	Dr. Alejandro Mazal	Institut Curie (Paris)
	Dr. Dietrich Averbeck	IRSN(Fontenay-aux-Roses)-Institut Curie (Paris)
Encadrants :	Dr. Florent de Vathaire	Equipe 3 CESP INSERM 1018
	Pr. Hervé Cardot	Institut de Mathématiques de Bourgogne (Dijon)

Mis en page avec la classe thesul.

Remerciements

Avant tout, je tiens à remercier Monsieur Florent de Vathaire, directeur de l'équipe « épidémiologie des radiations » du CESP, de m'avoir donné l'opportunité de réaliser cette thèse à un moment où je me posais beaucoup de questions sur mon devenir professionnel. Je le remercie de m'avoir accueilli, encadré durant ces trois années et de m'avoir fait découvrir et partager le monde de la recherche en m'offrant des conditions de travail exemplaires.

Je tiens à témoigner ma profonde gratitude à mon co-directeur de thèse, Monsieur Hervé Cardot, pour avoir accepté de se joindre, en cours de route, à l'aventure de cette thèse, qui n'aurait jamais abouti de la sorte sans son aide précieuse et sa constante disponibilité. Ce fut un réel plaisir de travailler avec vous.

Je tiens à remercier tout particulièrement les membres de mon jury de thèse.
Je remercie Madame Sophie Lambert-Lacroix et Monsieur Ahmadou Alioum, qui ont accepté la charge de rapporteurs pour cette thèse. Je suis également très reconnaissant à Messieurs Bruno Falissard, Dominique Laurier, Alejandro Mazal et Dietrich Averbeck d'évaluer mon travail en tant qu'examinateurs.

Mes remerciements s'adressent ensuite à Ibrahima Diallo avec qui j'ai eu grand plaisir à travailler sur la partie physique médicale de cette thèse ainsi qu'à Pierre Blanchard dont la rencontre fut décisive et ô combien enrichissante pour ce qui est des applications de ce travail en radiothérapie clinique.

Je remercie aussi Monsieur Dimitri Lefkopoulos, directeur du service physique médicale de l'institut Gustave Roussy ainsi que Monsieur Jérôme Champoudry de m'avoir permis la collecte des histogrammes dose-volume utilisés dans cette thèse.

Un grand Merci à Madame Laure Sabatier, responsable du Laboratoire de Radiobiologie et Oncologie (LRO) du CEA, dont les conseils et directives ont énormément contribué à l'orientation et l'accomplissement des applications en radiobiologie de ce travail dans le cadre du projet EpiRadBio. Je remercie par la même occasion les autres membres du LRO : Monika Frenzel, Michelle Ricoul, Radhia M'Kacher et Marion Bellamy pour leurs accueil chaleureux à chaque visite et pour avoir profondément révolutionné ma vision de la biologie !

Je tiens également à remercier Monsieur Jean Bouyer, Directeur de l'école doctorale

420, pour son accueil et sa bienveillance vis-à-vis de mon profil un peu atypique ainsi que Madame Audrey Bourgeois pour son aide et sa patience lorsque, trop souvent, ma rigueur administrative était en défaut ! Je suis également très reconnaissant à l'ED 420 pour son aide financière à de nombreux déplacements et séminaires.
Je remercie également la banque publique d'investissement (projet INSPIRA) ainsi que la commission européenne (projet EpiRadBio) pour leurs soutien financier à cette thèse.

L'aboutissement de ces trois années de travaux est l'occasion pour moi de remercier également toutes les personnes que j'ai pu rencontrer , au sein de mon équipe ou ailleurs, dont j'ai sincèrement apprécié la compagnie : Martine pour ses conseils en jardinage et ses récits de voyages en Inde, Isao avec qui j'ai beaucoup partagé sur la famille, la double culture et l'art (si,si ! il n'est pas si réservé que ça !), Karim, Adel et Amar avec qui je retrouve un peu de cette ambiance qui me manque tant, Rodrigue et Boris mes complices de biostat mais pas seulement (Amsterdam, Amsterdam !). Merci à vous pour ces nombreux paris organisés de main de maître (promis je tâcherai de faire aussi bien pour le prochain Rolland Garros que vous pour la Coupe du monde ou l'élection du Pape !), Nadia pour sa bonne humeur et conseils toujours avisés, Angéla incontournable dans tout ce qui est informatique à qui je dois tant de dépannages cruciaux ! Farah pour nos discussions aussi bien sportives, métaphysiques que culinaires (je ne suis pas prêt d'oublier ton incroyable mafé !), Constance pour sa fraîcheur, son écoute et sa patience devant mes blagues pas toujours très drôles (non je ne suis pas réac ! !), Maud pour son coté bon public vis-à-vis de mes blagues justement, Yan pour sa bonne humeur et son adaptation remarquable, Giao pour sa gentillesse et son humeur égale, Stéphanie pour son humour et ses analyses psy dixième degré en toutes circonstances et ma Françoise providentielle sans qui aucun déplacement, séminaire ou démarche administrative ne m'aurait été possible ! Les collègues physiciens de l'équipe bien sûr : Cristina, Safaa, Aymen (je finirai par te rendre visite à Besançon promis !) et Damien (ne t'en fais pas vieux, tu finiras un jour par avoir un revers aussi bon que le mien !) et enfin Carole et Chiraz, mes deux complices de bureau avec qui j'ai tant refait le monde. Une pensée toute particulière pour Jérémi, mon frère d'armes de master 2 physique médicale, mon compagnon de galère de thèse, mon confident sur la reconversion professionnelle et la nature humaine.

Mes derniers remerciements vont à ma famille : à mes parents pour leur soutien sans faille dans ce projet fou de reprise d'études à trente ans passés, mon frère pour son tuyau sur la physique médicale et ses encouragements permanents, Philippe et Elisabeth pour leur relecture attentive et ma compagne, Brunissende, pour sa patience et son inestimable

appui tout au long de cette période. Enfin aux trois lumières de ma vie : mes enfants.

A Ilès, Badis et Léhna

Sommaire

Résumé de la Thèse **xi**

Abstract **xiii**

Valorisation scientifique **xvii**

1 Publications avec comité de lecture xvii

2 Communications orales xviii

3 Communications affichées xviii

Introduction générale **xix**

4 Contexte xix

5 Problématique xx

6 Objectifs xxii

Liste des tableaux **xxxi**

Chapitre 1
Physique médicale **1**

1.1 Introduction 2

1.2 Matériel et Méthode 8

1.2.1 Mesures expérimentales 8

1.2.2 Evaluation semi-empirique de la dose due à la diffusion externe . . 10

1.2.3 Modélisation multi-sources 12

1.3 Résultats 17

1.4 Discussion 18

1.5 Conclusion et perspectives 21

1.6 Vers le prochain chapitre 22

Chapitre 2
Analyse de données fonctionnelles

2.1 L'approche fonctionnelle 26
2.2 Note historique 28
2.3 Statistique fonctionnelle et statistique multivariée 30
2.4 Analyse en composantes principales fonctionnelles 32
2.5 Deux Exemples Importants 35
2.5.1 Analyse en composantes principales fonctionnelles dans un espace engendré par une base [Ramsay and Silverman, 2005] 35
2.5.2 Analyse en composantes principales fonctionnelles d'un ensemble de densités de probabilité 37
2.6 Modèles de régression sur données fonctionnelles à réponses réelles 41
2.6.1 Le modèle de régression linéaire fonctionnel 41
2.6.2 Modèles linéaires généralisés et Modèles de survie fonctionnels . . . 42

Chapitre 3
Modélisation de complications tissus sains après radiothérapie externe

3.1 Introduction 46
3.2 Données de patients et de toxicités 48
3.3 Planification de traitement et HDVs 49
3.4 Analyse statistique 50
3.4.1 L'approche analyse de données fonctionnelles 50
3.4.2 Analyse en composantes principales fonctionnelles 51
3.4.3 Formulation du modèle et interprétation 51
3.4.4 Autres modèles NTCP et comparaison de modèles 53
3.5 Résultats 55
3.5.1 Incidence de la toxicité rectale 55
3.5.2 Comparaisons des distributions de doses entre les patients avec et sans toxicités 56
3.5.3 Résultats de modélisation 56
3.6 Discussion 63

Chapitre 4
Modélisation flexible d'une relation dose-effet en épidémiologie des radiations

4.1 Introduction . . . 68
4.2 Modèle de Cox flexible . . . 71
4.3 Analyse en composantes principales fonctionnelles . . . 72
4.3.1 L'approche analyse de données fonctionelles . . . 72
4.3.2 Analyse en composantes principales fonctionnelles . . . 74
4.4 Estimation de la relation dose–effet . . . 75
4.4.1 Pré-sélection des nœud intérieurs . . . 76
4.4.2 Sélection finale de nœuds et de composantes principales fonctionnelles 77
4.5 Simulations . . . 83
4.5.1 Simulations des données . . . 83
4.5.2 Estimation de la fonction dose–réponse . . . 84
4.5.3 Résultats des simulations . . . 85
4.6 Application : Analyse des données thyroïde . . . 92
4.6.1 Les données thyroïde EURO2K . . . 92
4.6.2 Résultats . . . 93
4.7 Discussion . . . 96

Chapitre 5
Radiobiologie

5.1 Introduction . . . 102
5.1.1 Contexte . . . 102
5.1.2 Problématique . . . 103
5.1.3 Présentation du projet EpiRadBio et Objectifs . . . 105
5.2 La cohorte ANGIO . . . 107
5.2.1 Présentation de la cohorte . . . 107
5.2.2 Design de l'étude . . . 108
5.3 Analyse en composantes principales multiniveau . . . 111
5.3.1 Définition . . . 111
5.3.2 Valeurs propres et vecteurs propres . . . 112
5.3.3 Estimation des scores de l'analyse en composantes principales multiniveau . . . 113

5.4 Résultats 116
5.5 Discussion 120

Conclusion générale **123**

Perspectives **127**

Bibliographie **131**

Annexes

Annexe A
Estimation de densité de probabilité

A.1 Introduction 156
A.1.1 Définitions 156
A.2 Estimation non paramétrique d'une densité de probabilité 157
A.3 Différentes méthodes non paramétriques d'estimation de la densité de probabilité 158
A.3.1 Histogramme de densité 159
A.3.2 Estimateur à noyau d'une densité 161

Annexe B
Régression Spline

B.1 Pourquoi les splines ? 172
B.2 Splines : Définition et premières propriétés 174
B.2.1 La base de puissances tronquée 174
B.2.2 Base B-splines 175
B.3 Splines des moindres carrés 178
B.4 Splines de lissage 181
B.4.1 Définition 181
B.4.2 Une variante : les splines pénalisées 182

Annexe C
Publications

Résumé de la Thèse

La radiothérapie occupe une place majeure dans l'arsenal thérapeutique des cancers. Sur les 355.000 nouveaux patients pris en charge pour un cancer en France, les deux tiers bénéficient d'une radiothérapie à une étape de leur parcours de soin soit environ 175 000 traitements chaque année. Son efficacité est unanimement reconnue puisqu'elle contribue, seule ou associée à d'autres modalités de traitements (chimiothérapie, chirurgie,...etc) à environ 50% des guérisons pour un coût représentant moins de 10% du total des dépenses consacrées au cancer.
Malgré des progrès technologiques importants depuis près de vingt ans, des tissus sains au voisinage ou à distance de la tumeur cible continuent à être inévitablement irradiés à des niveaux de doses très différents. Ces doses sont à l'origine d'effets secondaires précoces (œdème, radionécrose, dysphagie, cystite) ou tardifs (rectorragies, télangiectasie, effets carcinogènes, les pathologie cérébrovasculaires).
Il est donc primordial de quantifier et de prévenir ces effets secondaires afin d'améliorer la qualité de vie des patients pendant et après leur traitement.

La modélisation du risque d'événements iatrogènes radio-induits repose sur la connaissance précise de la distribution de doses au tissu sain d'intérêt ainsi que sur un modèle de risque capable d'intégrer un maximum d'informations sur le profil d'irradiation et des autres facteurs de risques non dosimétriques. Le problème comporte donc deux aspects : un aspect *dosimétrique*, à savoir la détermination quantitative de la dose absorbée par l'organe d'intérêt et un aspect de méthodologie statistique visant à exploiter de manière optimale la richesse des données (dosimétriques, démograpiques, génétiques ...etc) à disposition.

L'objectif de ce travail de thèse a été de développer des méthodes de modélisation capables de répondre à des questions spécifiques aux deux aspects, dosimétriques et statistiques, intervenant dans la modélisation du risque de survenue d'événements iatrogènes radio-induits.

Nous nous sommes intéressé dans un premier temps au développement d'un modèle de calcul permettant de déterminer avec précision la dose à distance due au rayonnements de diffusion et de fuite lors d'un traitement par radiothérapie externe et ce, pour différentes tailles des champs et à différentes distances de l'axe du faisceau. L'écart relatif moyen entre les doses à distance mesurées et calculées est de 10% et ce, pour des niveaux de doses pouvant être inférieurs à 0.01% de la dose maximale à l'axe.

Ensuite, nous avons utilisé des méthodes d'analyse de données fonctionnelles pour développer un modèle de risque de toxicités rectales après irradiation de la loge prostatique. Le modèle proposé a montré des performances supérieures aux modèles de risque existants particulièrement pour décrire le risque de toxicités rectales de grade 3. Ce modèle est également capable d'intégrer d'autres facteurs cliniques non dosimétriques importants en mettant en évidence l'impact de l'âge avancé au moment du traitement.

Dans le contexte d'une régression de Cox flexible sur données réelles, nous avons proposé une application originale des méthodes de statistique fonctionnelle permettant d'améliorer les performances d'une modélisation via fonctions B-splines de la relation dose-effet entre la dose de radiation à la thyroïde. Cette méthode n'a en effet nécessité l'estimation que d'un seul paramètre de régression B-spline avec une réduction conséquente de l'incertitude autour des estimations de risque.

Nous avons également proposé dans le domaine de la radiobiologie une méthodes basée sur l'analyse en composantes principales multiniveau pour quantifier la part de la variabilité expérimentale dans la variabilité des courbes de fluorescence mesurées.
Nous avons ainsi montré que le concept de longueur moyenne surestimait la qualité de la reproductibilité expérimentale en raison de la non prise en compte de la variabilité intra-individuelle. De plus, la méthode proposée permet d'isoler la partie propre à l'individu afin d'en décrire les modes de variabilités les plus dominants. Nous avons ainsi établi que la variabilité inter-individuelle de la longueur des télomères était essentiellement un espace à deux dimensions reposant sur la corrélation contraire entre télomères "courts"/télomères "longs" ainsi que celle des télomères "extrêmes"/télomères "intermédiaires".

Mots-clés : Dose à distance des rayonnements ionisants, analyse de données fonctionnelles, Événements iatrogènes radio-induits, Analyse de survie flexible , Radiobiologie.

Abstract

Radiotherapy plays a major role in the therapeutic arsenal against cancer. Among the 355,000 new patients treated for cancer in France, two thirds receive radiotherapy at some stage of their treatment course (approximately 175,000 treatments each year). Its effectiveness is universally recognized since it helps, alone or in combination with other treatment modalities (chemotherapy, surgery etc ...) about 50 % of healing while its cost is less than 10 % of total expenditure devoted to cancer.

Despite significant advances in technology for nearly twenty years, healthy tissues near or away from the target tumor remain inevitably irradiated at very different levels of doses. These doses are at the origin of early side effects (edema, radiation necrosis, dysphagia, cystitis) or late (rectal bleeding, telangiectasia, carcinogenic, cerebrovascular diseases). It is therefore essential to quantify and prevent these side effects to improve the patient quality of life after their cancer treatment.

The modelling risk of radiation-induced adverse events depends on accurate estimation of the dose distribution to healthy tissue as well statistical risk models able to integrate both radiation profile information and other non dosimetric factors. So the problem consists on two aspects : a dosimetric aspect, ie the quantitative determination of the absorbed dose to the organ of interest and a statistical methodology appearance to optimally exploit the available data (dosimetry, démograpiques , genetic ... etc) .

The objective of this thesis was to propose modelling methods able to answer specific questions asked in both aspects, dosimetry and statistics, involved in the modeling risk of developing radiation-induced iatrogenic pathologies.

Our purpose was firstly to assess the out-of-field dose component related to head scatter radiation in high-energy photon therapy beams and then derive a multisource model for this dose component. For measured doses under out-of-field conditions, the average local difference between the calculated and measured photon dose is 10%, including doses

as low as 0.01% of the maximum dose on the beam axis. This study demonstrates that the multi-plane source approach is suitable for accurate analytical modeling of the out-of-field dose component related to head scatter radiation. These results should be taken into account when evaluating doses to the remaining volume at risk in external beam radiotherapy planning.

We secondly described a novel method to explore radiation dose-volume effects. Functional data analysis is used to investigate the information contained in differential dose-volume histograms. The method is applied to the normal tissue complication probability modeling of rectal bleeding for patients irradiated in the prostatic bed by 3D conformal radiation therapy. Functional principal component analysis was performed to explore the variation modes in the dose distribution. The functional principal components were then tested for association with rectal bleeding using logistic regression adapted to functional covariates. The components which describe the interdependence between the relative volumes exposed at intermediate and high doses were the most correlated to the complication. The regression parameter function leads to a better understanding of the volume-effect by including the treatment specificity in the delivred dose-volum effect information. Patients with advanced age were also significantly at risk and the by inclusion of this clinical factor improved significantly performance of the model.

In the flexible Cox model context, we proposed a new dimension reduction technique based on a functional principal component analysis to estimate a dose-response relationship. A two-stage knots selection scheme was performed : a potential set of knots is chosen based on information from the rotated functional principal components and the final knots selection is then based on statistical model selection. Simulations indicated that the FPCA models with adequate FPCs recovers a variety of clinically plausible shapes of the true dose-response and provides a better fit, in the sense of mean square error, to the data than does the conventional spline regression or penalized spline models.
The method offers a significant dimensional reduction of the problem by replacing the initial spline basis by a smaller number of score functions which summarize the effect of the initial interior knots sequence. The thyroid application illustrates the new insights that the proposed model may offer in real-life prognostic studies by estimating the shape of the estimated dose-response of radio-induced thyroid tumor risk as an unimodal curve using two interior knots and only one estimated parameter. This suggests that the proposed method can be seen as a form of dose-response regularized estimation which could represent an alternative to the classical penalized approach.

Finally, a multilevel functional principal component analysis was applied to radiobiological data in order to quantify the experimental Variability for replicate measurements of fluorescence signals of telomere length.
We have shown that the concept of average telomere length overestimates the quality of experimental reproducibility since the intra-individual variability is not considered. In addition, the proposed method allows to isolate the specific part of the individual signal in order to describe the most dominant modes of variability. Thus, we have established that the inter-individual variability in telomere length was essentially a two-dimensional space based on the opposite correlation between "short" / "long" telomeres on one hand and "extreme" / "intermediate" telomeres on other hand.

Keywords : Out-of-field radiation dosimetry, Functional data analysis, Radio-induced iatrogenic events, flexible survival analysis, Radiobiology

Valorisation scientifique

1 Publications avec comité de lecture

Benadjaoud MA, Bezin J, Veres A, Lefkopoulos D, Chavaudra J, Bridier A, de Vathaire F, Diallo I. A multi-plane source model for out-of-field head scatter dose calculations in external beam photon therapy. *Physics in Medecine and Biology. 2012 Nov 21 ;57(22) :7725-39.*

Benadjaoud MA , Blanchard P, Schwartz B, Champoudry J, Bouaita R, Lefkopoulos D, Deutsch E, Diallo I, Cardot H, de Vathaire F. Functional data analysis in NTCP modeling : a new method to explore the radiation dose-volume effects. *International Journal of Radiation Oncology, Biology, Physics 90 (2014) pp. 654-663.*

Benadjaoud MA , Schwartz B, Diallo I, Cardot H, de Vathaire F. Adaptative knot selection in the flexible B-spline Cox model using functional principal components analysis. *Soumis*

Schwartz B, **Benadjaoud MA**, Cléro E, Haddy N, El-Fayech C, Guibout C, Teinturier C, Oberlin O, Veres C, Pacquement H, Munzer M, N'guyen TD, Bondiau PY, Berchery D, Laprie A, Hawkins M, Winter D, Lefkopoulos D, Chavaudra J, Rubino C, Diallo I, Bénichou J, de Vathaire F. Risk of second bone sarcoma following childhood cancer : role of radiation therapy treatment.*Radiat Environ Biophys. 2014 May ;53(2) :381-90.*

2 Communications orales

Benadjaoud MA , Schwartz B, de Vathaire F. Flexible multi-models applied to thyroid tumor dose-response. *Annual meeting EpiRadBio project. Juin 2013, Alghero, Italie)*

Benadjaoud MA , LLanas D, de Vathaire F. Consequences of the high variability of the radiation dose in the organs such as heart and brain in risk estimates : possibilities from functional data analysis. *Annual meeting Cerebrad/Procardio project. Octobre 2013, Thessalonique, Grèce)*

Benadjaoud MA , Blanchard P, Cardot H, de Vathaire F. Functional data analysis in radiobiology and radiation epidemiology. *35th Annual Conference of the International Society for Clinical Biostatistics, 24-28 Aout 2014, Vienne, Autriche.*

3 Communications affichées

Benadjaoud MA , Cardot H, de Vathaire F. Dimensional reduction in the flexible B-spline Cox model using functional principal components analysis. *35th Annual Conference of the International Society for Clinical Biostatistics, 24-28 Aout 2014, Vienne, Autriche.*

Introduction générale

4 Contexte

La radiothérapie occupe une place majeure dans l'arsenal thérapeutique des cancers. Selon la Société Française de Radiothérapie Oncologique (SFRO), sur les 355.000 nouveaux patients pris en charge pour un cancer en France, les deux tiers bénéficient d'une radiothérapie à une étape de leur parcours de soin soit environ 175 000 traitements chaque année.
Son efficacité est unanimement reconnue puisqu'elle qu'elle contribue, seule ou associée à d'autres modalités de traitements (chimiothérapie, chirurgie,...etc) à environ 50% des guérisons pour un coût représentant moins de 10% du total des dépenses consacrées au cancer.

Un traitement par radiothérapie consiste à délivre une dose de rayonnements ionisants létale aux cellules cancéreuses. Cependant et malgré des progrès technologiques importants depuis près de vingt ans, des tissus sains au voisinage ou à distance de la tumeur continuent à être inévitablement irradiés à des niveaux de doses très différents.
Ces doses sont à l'origine d'effets secondaires *radio-induits* qui peuvent être déterministes ou stochastiques.

Les effets déterministes apparaissent à partir d'une dose seuil et sont d'autant plus sévères que la dose et le débit de dose augmentent. Il s'agit des effets tissulaires, à traduction clinique immédiate ou différée. Ces effets comprennent les effets précoces (transitoires et réversibles survenus au cours du traitement ou dans les 6 premiers mois qui suivent) et les effets tardifs (le plus souvent définitifs) survenant après 6 mois.
Les effets stochastiques apparaissent quant à eux aussi bien à faibles doses qu'à doses élevées et surviennent parfois plusieurs dizaines années après le traitement. Ces effets comprennent notamment les effets carcinogènes (i.e cancer secondaire après radiothérapie) mais aussi les pathologies cardiaques et cérébrovasculaires, le diabète, l'âge à la ménopause,...etc.

La survenue des effets secondaires radio-induits n'est cependant pas liée uniquement aux seuls facteurs physiques de l'irradiation (dose totale, dose par fraction et volume irradié, organe cible et hétérogénéité de la dose). En effet la grande variabilité observée dans l'incidence des complications à court ou long terme chez des patients ayant bénéficié d'un même traitement de radiothérapie standardisé suggère l'existence de différences intrinsèques de radiosensibilité individuelles.
En réalité, la sensibilité aux radiations ionisantes, tant en termes de toxicité tissulaire que de risque de cancer, dépend en effet fortement aussi bien du statut individuel (âge, mode de vie, antécédents médicaux) que du statut génétique et épigénétique.
Ainsi la Commission Internationale de Protection Radiologique (CIPR) a estimé que le pourcentage d'individus radiosensibles dans la population générale se situerait entre 5 et 15%.

Pourtant, la prise en compte de la radiosensibilité individuelle dans le système de radioprotection reste encore limitée notamment parceque les outils d'évaluation du risque individuel sont encore insuffisants et la mise au point de tests prédictifs robustes, rapides, validés et standardisés pour évaluer la radiosensibilité individuelle constituerait une étape-clé en radioprotection aussi bien des patients que du grand public.

5 Problématique

La modélisation du risque d'événements iatrogènes radio-induits repose sur la connaissance aussi précise que possible de la distribution de doses au tissu sain d'intérêt ainsi que sur un modèle de risque adapté capable d'intégrer un maximum d'informations sur le profil d'irradiation en présence ainsi que des autres facteurs de risques non dosimétriques. Le problème comporte donc deux aspects : un aspect *dosimétrique*, c'est-à-dire la détermination quantitative de la dose absorbée par l'organe d'intérêt et un aspect de méthodologie statistique visant à exploiter de manière optimale la richesse des données (dosimétriques, démograpiques, génétiques,...etc) à disposition.

La dosimétrie

Dans les services de radiothérapie, la dosimétrie des organes à risque dans le volume cible est assurée par un logiciel de planification de traitement appelé TPS (treatment

planning system). Les structure critiques sont d'abord contourées sur les coupes scanographiques du patient. La dose de radiation est par la suite calculée à l'aide du choix d'une représentation des appareils de traitement utilisés et de modèles de calcul de dose suffisamment performants pour satisfaire au mieux la prescription médicale avec une précision de 2-3%.
Rien de tout cela en ce qui concerne la dosimétrie des organes situés en dehors du champ d'irradiation : cette partie du corps n'est couverte par aucune procédure d'imagerie préalable au traitement et sa dosimétrie n'est plus assurée par les TPS utilisés en routine clinique.
Or ces gammes de dose intermédiaires voire faibles, sont susceptibles d'être à l'origine d'effets tardifs suffisamment graves pour qu'une attention toute particulière leurs soit portée, d'autant plus que les techniques d'irradiation les plus modernes semblent en augmenter la proportion (bien qu'elles préservent de mieux en mieux les tissus sains des plus fortes doses).

La méthodologie statistique

Dans le cadre de la modélisation du risque d'effets déterministes des organes à risque situés dans le champ d'irradiation, les modèles existants sont capables de prendre en compte l'ensemble de la distribution de dose à l'organe calculée par le TPS.
Cependant, ces modèles sont souvent paramétriques avec une contrainte en terme de forme analytique assez forte. Ce manque de flexibilité a déjà montré ses limites pour différents organes et techniques d'irradiation.

Les données à disposition dans la modélisation du risque à long terme d'événements radio-induits sont longtemps restées en nombre et en qualité inférieures à celles rencontrées dans la problématique de la toxicité des organes à risque et ce, en raison des difficultés spécifiques que pose la question de l'estimation de la dose à distance du champ de radiation.
Du point de vue de la modélisation, on retrouve là encore un manque de flexibilité dans la mesure où seuls quelques modèles (essentiellement de linéaires-exponentiels) sont utilisés pour décrire la relation dose-effet et le plus souvent formulés en terme de dose moyenne.

En plus des données dosimétriques et démographiques, les données issues de la radiobiologie (branche de la biologie étudiant les effets biologiques des rayonnements) sont à présent de plus en plus nombreuses et disponibles en raison des progrès importants des

tests de fonctionnalité de la réponse cellulaire aux dommages radio-induits. Ceci concerne la signalisation de réparation de l'ADN (imunofluorescence de la protéine ATM ou de l'histone H2AX) mais aussi d'autres types de réponses comme par exemple la longueur des télomères.
Au cours de la dernière décennie, plusieurs équipes ont en effet montré le rôle majeur des télomères dans l'intégrité et la stabilité du génome et l'impact des radiations sur les longueurs de télomères constitue à présent un biomarqueur de plus en plus étudié.

Les techniques haut débit pour la quantification de la longueur des télomères permettent aujourd'hui d'en mesurer l'hétérogénéité intercellulaire avec une dizaines de milliers de valeurs d'intensité de fluorescence des brins des régions télomériques. Cependant, les modèles statistiques de risque incluant l'information télomérique le font le plus souvent en terme de longueur moyenne ce qui entraîne une perte importante d'informations en raison de la non prise en compte de la variabilité intra-individuelle.

6 Objectifs

L'objectif de ce travail a été de développer des méthodes de modélisation capables de répondre à des questions spécifiques aux deux aspects, dosimétriques et statistiques, intervenant dans la modélisation du risque de survenue d'événements iatrogènes radio-induits.

Nous nous sommes intéressé dans un premier temps au développement d'un modèle de calcul permettant de déterminer avec précision la dose à distance due au rayonnements de diffusion et de fuite lors d'un traitement par radiothérapie externe et ce, pour différentes tailles des champs et à différentes distances de l'axe du faisceau.

Ensuite, nous avons utilisé des méthodes d'analyse de données fonctionnelles ou statistique fonctionnelle afin de développer un modèle de risque de toxicité tardive déterministe en tenant compte de l'hétérogénéité de la distribution de dose à l'organe d'intérêt. Nous avons appliqué ce modèle pour étudier le risque de toxicités rectales après une irradiation de la loge prostatique et comparé ses performances aux modèles déjà existants.

Dans le contexte d'une régression de Cox flexible sur données réelles, nous avons proposé une application originale des méthodes de statistique fonctionnelle dans un contexte a priori dépourvu de données individuelles fonctionnelles. Cette approche a permis d'améliorer les performances d'une modélisation via fonctions B-splines de la relation dose-effet comparé aux approches de régression splines classiques existantes et celles basées sur les

splines pénalisées. Ce modèle appliqué à la modélisation de la relation dose-effet entre la dose de radiation à la thyroïde et le risque de tumeurs thyroïdiennes radio-induites.

Enfin, nous proposons une méthode d'estimation de la part de la variabilité expérimentale et celle propre à l'individu dans le cas de mesures répétées de signaux de fluorescence en tenant compte de la variabilité intra-individuelle de ces derniers. La part du signal inhérente à individu a par la suite été isolée du signal total mesuré afin d'explorer l'effet des radiations sur la longueur télomérique.

Les travaux réalisés sont présentés et développés selon cinq chapitres.

Tout d'abord, dans le premier chapitre, nous explicitons la problématique de la dose hors champ en radiothérapie en détaillant l'origine, les caractéristiques et les effets de ce dépôt d'énergie. Nous présentons par la suite les travaux expérimentaux menés autour de la mesure de la composante attribuable aux rayonnements diffusés et de fuite ainsi que le modèle théorique que nous proposons pour modéliser ces différentes contributions avant de conclure sur une comparaison calculs/mesures.

Le second chapitre présentera la méthodologie statistique centrale de ce mémoire de thèse, autour de laquelle s'articulent toutes les modélisations présentées par la suite :*l'analyse des données fonctionnelles* ou *statistique fonctionnelle*. Nous détaillerons ses motivations, ses spécificités, ce qui la rapproche et qui la distingue de la statistique multivariée classique.

Les prochains chapitres sont trois applications différentes des outils d'analyse de données fonctionnelles :
Le troisième chapitre utilise une méthode d'analyse en composantes principales fonctionnelles pour caractériser les principaux modes de variabilité des distributions de doses de radiations au rectum lors d'une irradiation de la loge prostatique. L'association avec la toxicité rectale radio-induite est par la suite étudiée via un modèle linéaire généralisé à variables explicatives fonctionnelles et numériques.

Le quatrième chapitre propose une nouvelle méthode de régression de Cox flexible basée sur des B-splines construits à partir de nœuds intérieurs sélectionnés de manière spatio-adaptative. A l'origine de cette sélection de nœuds se trouve l'analyse de données fonctionnelles traduisant la propriété de support minimal au voisinage de chaque valeur

distincte de la variable continue d'intérêt.

La chapitre cinq quant à lui propose de mener une analyse en composantes principales multiniveau dans le but d'une part, quantifier l'effet des normalisations des signaux expérimentaux de fluorescences par ceux de lignées cellulaires de référence et d'autre part d'isoler la partie individuelle du signal afin d'étudier le raccourcissement des télomères en fonction de l'âge et de l'exposition aux radiations.

Enfin, nous conclurons ce travail en essayant de dégager les points importants de cette recherche et les perspectives à envisager.

Table des figures

1.1 Mises en place expérimentales utilisées dans l'article de Kase et al. [Kase et al., 1983]. W, est le diffusé patient, C, le diffusé tête, et L, les fuites. Pour un champ carré de côté A, le point de mesure est à une distance x de l'axe du faisceau, et à une profondeur d. 3

1.2 Dispositif expérimental utilisé dans [Ruben et al., 2014] permettant de ne mesurer que la composante diffusé tête (ou collimateur) de l'accélérateur . 4

1.3 (à gauche) Types de fantômes anthropomorphes utilisés en dosimétrie pour simuler l'anatomie d'un patient.(à droite) Les structures internes au fantôme pour lesquelles les mesures de dose sont effectuées 4

1.4 Les composantes de la dose hors champ. Un repère en bas à gauche indique la convention adoptée en radiothérapie pour définir l'espace. L'origine de ce repère est l'isocentre de l'appareil, défini ici par l'intersection de l'axe central du faisceau et celui de rotation du statif de l'appareil, tous deux représentés sur la figure par des lignes en pointillés. 5

1.5 Modèle multisource décrit dans [Dunscombe and Nieminen, 1992]. *Point source* : une source ponctuelle de photons directs. *Distributed source* : un plan source de photons diffusés par les premières composantes de la tête d'irradiation rencontrées (collimateur primaire et filtre égalisateur en triangle en noir) muni d'une distribution d'intensité Gaussienne. *Jaws* : les mâchoires ou collimations secondaires définissent la portion visible du plan des photons diffusés. 7

1.6 Dispositif expérimental utilisé pour la mesure de la dose due aux rayonnements de diffusions collimateurs et fuites. 10

1.7 Une représentation schématique de la source surfacique $\mathcal{S}$, du blindage linéaire de la tête de l'accélérateur ainsi que des mâchoires inférieures du collimateur secondaire. 12

1.8 Termes géométriques utilisés dans l'expression du modèle final de l'équation 1.9 15

1.9 Estimation des doses du rayonnement de fuite (paramètre $\lambda(r, \varphi, z)$ de l'équation (1.3)) pour l'accélerateur linéaire Varian Clinac 2300 C/D opérantà 6 et 20 MV RX, respectivement. Les doses sont données en pourcentage de la dose à la profondeur de la dose maximale à l'axe du faisceau. 17

1.10 Comparaison des profils de doses à distance mesurés et calculés, à 10cm de profondeur, pour différents appareils et différentes conditions d'irradiation. 19

1.11 Isodoses calculées pour différentes tailles et formes de champs représentés en haut à gauche. Les doses sont données en pourcentage de la dose maximale à l'axe du faisceau. Les calculs sont effectués pour Varian accélérateur linéaire Clinac 2300 C/D opérant à 6 MV. 20

3.1 Coupe transverse au niveau de la loge prostatique représentant la balistique de traitement. Les 4 faisceaux sont centrés sur le barycentre de la loge prostatique (Isocentre). 48

3.2 L'approche statistique fonctionnelle (à droite) comparée à l'approche statistique multivariée matricielle (à gauche). 50

3.3 (a) Exemple d'estimation de la densité de probabilité par la méthode des noyaux. (b) Densités de probabilité de la distribution de dose au rectum chez les cas (rouge) et les non cas (noir) ainsi que leurs moyennes (les moyennes cumulées sont également représentées). 57

3.4 Matrice de corrélation de Spearman des indicateurs dosimétriques standards. 59

3.5 (a) Les trois premiers vecteurs propres issus de l'analyse en composantes principales l'ACP des variables aléatoires $(V_{d_i\ Gy})_{1 \leq i \leq p}$ définies dans la Section 3.4.4 60

3.6 (à gauche) Le paramètre fonctionnel $\beta(\delta)$ décrivant le risque de toxicité rectale de grade ≥ 2 et 3. Les lignes en pointillés représentent les intervalles de confiance de niveau 95%. (à droite) La probabilité de survenue de la toxicité rectale en fonction des scores de l'analyse en composantes principales fonctionnelles. Les carrés sont les taux de complications observés avec leurs barres d'erreurs 68%. 62

4.1 Construction des fonctions à support minimal (MSP).(Gauche) La base B-spline est construite à partir des nœuds désignés par des triangles noirs. En noir (resp en gris) les B-splines prenant des valeurs non nulles (resp nulles) en $z_i = 10$.(Droite) les fonctions MSP $\Phi_{z_i=10}$ comme une moyenne pondérée des quatre B-splines. 73

4.2 Filtrage spectral des quantiles : la rotation d'une fonction propre en noir vue comme un filtre de bande passante mi-hauteur. La séquence initiale de quantiles est représentée avec des triangles et seuls les quantiles en vert sont retenus. 78

4.3 Noeuds *candidats* pré-sélectionnés à partir de cinq scénarios simulés. les fonctions de risques sont représentées dans la ligne du haut. La ligne du milieu donne les histogrammes de répartitions des valeurs des nœuds pré-sélectionnés et la dernière ligne donne la distribution de leurs nombres sur l'ensemble des 1000 cohortes simulées. 86

4.4 Sélection finale de modèle sur les cinq scénarios simulés selon les algorithmes *Backward Search* and *FPCA Forward Search*présentés dans la section 4.4.2). Les lignes représentent les scénarios : la colonne de droite montre la distribution des nœuds, celle du milieu la distribution de leurs nombres, et enfin, celle de gauche la distribution du nombre de paramètres splines estimés dans le modèle. 88

4.5 Estimation de la vraie relation dose-effet (en gris) par différentes méthodes splines : la régression spline classique avec des nœuds intérieurs sélectionnés par la méthode *Backward* (en vert), les deux méthodes proposées *Full* et *forward* FPCA (en lignes solides et pointillés respectivement) et les splines pénalisées (en rouge). Dans chaque cas, les lignes épaisses sont les estimations moyennes sur les 1000 simulations et, en lignes fines, les intervalles de confiance ponctuels de niveau 95%. 89

4.6 Boîtes à moustaches des erreurs de moindres carrés MSE pour chaque scénario simulé et chaque algorithme présenté dans la section 4.4.2 ainsi que les P-splines. 91

4.7 Processus de filtrage spectrale des quantiles de la distribution de dose de radiation de la thyroïde. Pour chaque fonction propre, la variabilité expliquée (Var) et le niveau d'association avec le risque de tumeur thyroïdienne (p-value) sont donnés en haut de chaque cadre ainsi que les nœuds sélectionnés par filtrage. 94

4.8 Risque de tumeurs thyroïdiennes comme fonction de la dose de radiation à la thyroïde en utilisant les algorithmes *FPCA Forward Search* (pointillés), *FPCA Full Search* (tirés), et *Backward Search* (ligne pleine). En gras, l'estimation de la courbe dose–effet et en épaisseur simple les intervalles de confiance bootstrap. 96

5.1 Marquage télomérique de chromosomes de lymphocytes T 110

5.2 Exemple de listing de quantification télomérique d'un microscope 10X. Le listing comporte 10.000 mesures d'intensité de fluorescences cellulaires par lame. 110

5.3 Intensité Moyenne du signal de fluorescence avant et après normalisation par le signal REMB. 117

5.4 Extraction de la partie intrinsèque à l'individu (en rouge) des deux densités de distribution des signaux de fluorescences mesurées lors des deux expériences. En jaune, les parties relatives au niveau 2. 119

5.5 Variation des deux premiers scores du niveau 1 en fonction de l'âge et du statut exposé/non exposé. 120

5.6 (en bas) Les deux premières fonctions propres niveau 1. (en haut) L'effet de ces fonctions propres sur la densité moyenne (ici en bleu) de toutes les courbes de fluorescence. La courbe '-' (resp en '+') représente la densité moyenne à laquelle on retranche (resp on ajoute) un multiple de la fonction propre. 121

A.1 Des exemples de la sensibilité des histogrammes aux modifications des paramètres de fenêtres h avec point d'origine $t_0 = 0$ (gauche) et aux modifications du point d'origine t_0 avec $h = 13.5$ (droite). *Figure tirée de [Scott, 1992]* 160

A.2 Exemple d'estimation à noyau (en pointillées) d'une densité de probabilité (en ligne solide) pour différentes valeurs du paramètre de lissage h. 162

B.1 Exemple de régression polynomiale à plusieurs degrés. Les pointillés représentent le nuage de points généré à partir de la fonction f en tirets. 172

B.2 Exemple de quatre B-splines linéaires (à gauche) de nœuds $\{0.3, 0.6\}$ et sept B-splines cubiques (à droite) de nœuds $\{0.3, 0.6, 0.9\}$ 177

B.3 Effet du nombre et de l'emplacement des nœuds sur la qualité d'une regression spline. Le nuage de points représente les données simulées à partir de la vraie moyenne (en tirets). la courbe en ligne continue est l'estimateur spline de la moyenne 180

Liste des tableaux

1.1 Valeurs des paramètres de régression du modèle (1.9) pour les différentes machines utilisées. 18

3.1 Caractéristiques cliniques des patients et de leurs traitements de radiothérapie externe. 55
3.2 Analyse univariée et multivariée ainsi que la comparaison des différents modèles NTCP 58

4.1 Caractéristiques cliniques de la partie française de la cohorte EURO2K . . 93
4.2 sélection du modèle spline final par chacun des trois algorithmes. Dans chaque cas sont donnés : la liste des nœuds retenue, le nombre de paramètres estimés et la valeur du critère AIC. 95

5.1 Table des caractéristiques du groupe de patients inclus dans l'étude. 116
5.2 Résultats des deux méthodes multivariée /ACP fonctionnelle d'analyse de la variance. Un patient est considéré comme un *groupe* de deux mesures et l'on cherche à quantifier la variabilité inter et intra groupes. 117
5.3 Valeurs propres et pourcentage de variabilité expliquée, pour chaque niveau, obtenus par analyse en composantes principales fonctionnelles multiniveaux des signaux de fluorescences normalisés par REMB. 118

A.1 Tableau des efficacités relatives de plusieurs noyaux 164

Chapitre 1

Physique médicale

Sommaire

1.1 Introduction 2

1.2 Matériel et Méthode 8

1.2.1 Mesures expérimentales 8

1.2.2 Evaluation semi-empirique de la dose due à la diffusion externe 10

1.2.3 Modélisation multi-sources 12

1.3 Résultats 17

1.4 Discussion 18

1.5 Conclusion et perspectives 21

1.6 Vers le prochain chapitre 22

1.1 Introduction

Le principe de base d'un traitement de radiothérapie externe est de délivrer la dose prescrite au volume cible ou PTV (*planned target volume*) tout en épargnant les tissus sains [ICRU, 2010]. Avec les techniques de plus en plus modernes, il est possible, tout en maintenant une dose suffisante pour assurer le contrôle tumoral, de réduire de plus en plus la dose aux tissus sains sans pour autant l'éliminer complètement.
Les recommandations récentes de la commission internationale des unités et mesures de radiations (ICRU) concernant le RVR (*Remaining volume at risk*), invitent à une plus grande prise en compte de ces doses résiduelles dans la pratique clinique en raison de l'accumulation de preuves épidémiologiques quant aux rôles qu'elles pourraient jouer dans l'apparition de pathologies radio-induites à long terme [Xu et al., 2008].

La dosimétrie des organes à risque dans le volume cible est assurée par un logiciel de planification de traitement appelé TPS (treatment planning system). Les structures critiques sont d'abord contourées sur les coupes scanographiques du patient. La dose de radiation est par la suite calculée par le TPS à l'aide du choix d'une balistique représentative des appareils de traitement utilisés et de modèles de calcul de dose performants pour satisfaire au mieux la prescription médicale.

Rien de tout cela en ce qui concerne la dosimétrie des organes situés en dehors du champ d'irradiation : Cette partie du corps n'est couverte par aucune procédure d'imagerie préalable au traitement et sa dosimétrie ne peut plus être assurée par les TPS utilisés en routine clinique.

En effet, Les données spécifiques aux accélérateurs intégrées dans les TPS se limitent généralement aux doses dans le champ et à la pénombre et les doses hors champ ne sont pas intégrées et de nombreux auteurs s'intéressant à la dose hors champ ont montré que les TPS n'étaient pas en mesure d'estimer la dose en dehors du faisceau de traitement [Howell et al., 2010, Ruben et al., 2011, Ruben et al., 2014].

Howell *et al.* [Howell et al., 2010] ont notamment montré que la différence moyenne entre les calculs du TPS et des mesures par dosimètres thermoluminescents (TLD) était de 40 % pour des distances supérieures à 3,75 cm du bord du champ, et, augmentant avec la distance, pouvait atteindre 55 % à la plus grande distance observée dans cette étude (11,25 cm du bord du champ).

En fait, l'étude de cette dose à distance doit être menée via une méthodologie spé-

cifique aussi bien du point de vue des mesures métrologiques que du point de vue de la modélisation.

De nombreuses études expérimentales ont fourni une grande quantité d'informations sur les différentes composantes de la dose à distance sur différents types de machines de traitement. Une cuve à eau est souvent utilisée afin de mimer, de façon simple, un milieu semblable à celui du patient (le corps humain est constitué à 80% d'eau). Pour l'étude de la dose hors champ, c'est le corps entier du patient qui est d'intérêt, les dimensions de la cuve doivent donc être adaptées à cette problématique. Greene *et al* [Greene et al., 1983] et surtout Kase *et al* [Kase et al., 1983] ont été les premiers à mener des mesures de doses à distance à grande échelle.
Afin de mesurer séparément les différentes composantes de la dose hors champ, le diffusé patient, qu'ils notent W, le diffusé tête, C, les fuites,L, ils utilisent trois mises en place illustrées dans la Figure 1.1.
La première, (a), permet de mesurer la dose hors champ totale : $W + C + L$. La seconde, (b), en positionnant la cuve en dehors du faisceau de traitement, permet de ne mesurer que la composante venant de l'appareil de traitement : $C + L$. Quant à la troisième, (c), en bloquant l'ouverture du faisceau, elle ne mesure que les fuites. La soustraction de ces différentes séries de mesures permet d'obtenir chaque composante de la dose hors champ séparément.

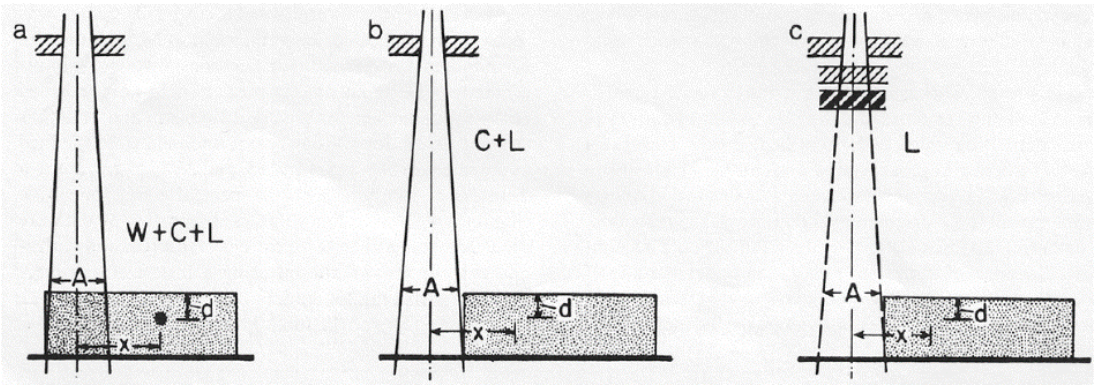

FIGURE 1.1 – Mises en place expérimentales utilisées dans l'article de Kase et al. [Kase et al., 1983]. W, est le diffusé patient, C, le diffusé tête, et L, les fuites. Pour un champ carré de côté A, le point de mesure est à une distance x de l'axe du faisceau, et à une profondeur d.

Ce dispositif a par la suite été repris dans des études plus récentes [Ruben et al., 2011, Ruben et al., 2014] (voir figure 1.2).

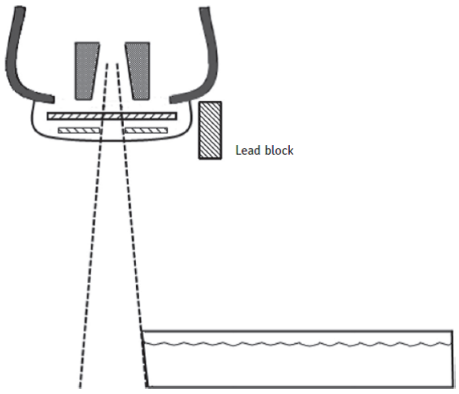

FIGURE 1.2 – Dispositif expérimental utilisé dans [Ruben et al., 2014] permettant de ne mesurer que la composante diffusé tête (ou collimateur) de l'accélérateur

D'autres dispositifs plus sophistiqués comme les fantômes anthropomorphes ont également été utilisés. Citons notamment le travail de l'équipe de Stovall qui a effectué un nombre énorme de mesures de doses à distance dans le cadre de collaborations avec des épidémiologistes des radiations combinant des mesures sur des cuves à eau et des fantômes anthropomorphes afin de déterminer la dose reçue par quatorze organes périphériques (voir figure 1.3) pour différents protocoles de traitement du cancer de l'utérus [Stovall et al., 1989] ainsi que la dose fœtale reçue après traitement par radiothérapie [Stovall et al., 1995]. Ils ont pu reconstituer une cartographie de dose de 20 000 patientes traitées entre 1916 et 1975. Même si ces données ne sont plus d'actualité pour les machines utilisées de nos jours, elles restent d'un intérêt primordial pour le suivi épidémiologique des femmes traitées à cette époque.

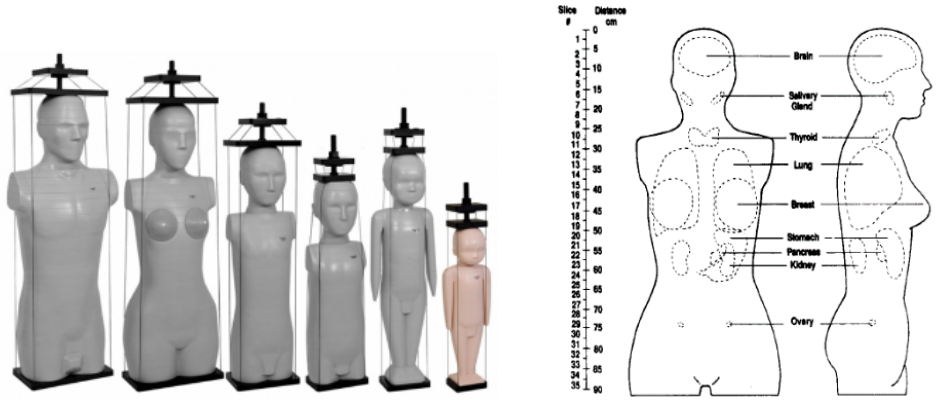

FIGURE 1.3 – (à gauche) Types de fantômes anthropomorphes utilisés en dosimétrie pour simuler l'anatomie d'un patient.(à droite) Les structures internes au fantôme pour lesquelles les mesures de dose sont effectuées

En résumé, la dose à distance du champ d'irradiation comprend essentiellement trois composantes (voir figure 1.4) :

1) La composante de rayonnement de fuite.

2) Le rayonnement diffusé de la tête de l'appareil issu des éléments modificateurs du faisceau tels que collimateurs, blocks...etc

3) Le diffusé interne au patient.

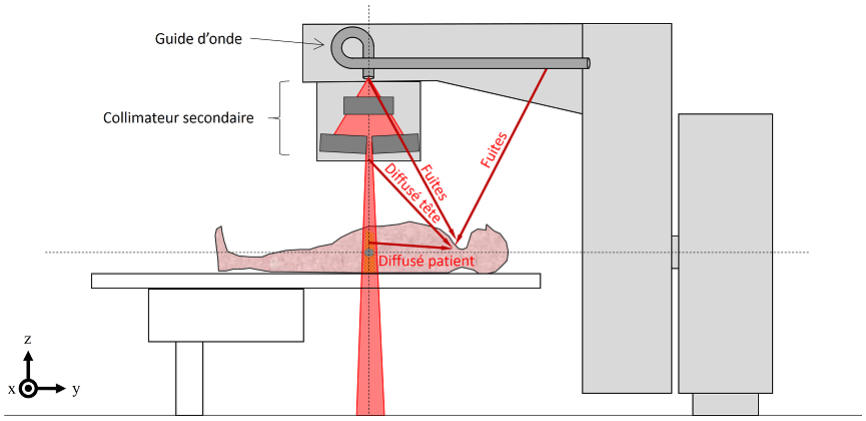

FIGURE 1.4 – Les composantes de la dose hors champ. Un repère en bas à gauche indique la convention adoptée en radiothérapie pour définir l'espace. L'origine de ce repère est l'isocentre de l'appareil, défini ici par l'intersection de l'axe central du faisceau et celui de rotation du statif de l'appareil, tous deux représentés sur la figure par des lignes en pointillés.

Seules les deux premières composantes de la dose à distance sont dépendantes de la géométrie de l'appareil de traitement alors que la troisième ne dépend que de l'énergie du faisceau. Il a été établi que cette composante *diffusé patient* n'est dominante qu'à très courte distance de l'axe du faisceau, les rayonnements de diffusion et fuite constituant par la suite la plus grande contribution.

Par exemple, pour une profondeur de 10 cm et des machines à 6 et 15 MV , il a été récemment établi que plus de la moitié de la dose à distance est constituée de rayonnement de diffusion et de fuite et ce, pour des champs d'irradiation allant jusqu'à $20 \times 20\ cm^2$ [Joosten et al., 2011, Chofor et al., 2012]

Du point de vue de la modélisation, nous pouvons classifier les modèles de calculs de doses à distance en deux grandes catégories : les modèles analytiques et les stochastiques semi-empiriques calibrés à l'aide de mesures.

Les modèles analytiques utilisés jusqu'à présent interpolent les profils de dose d'une grande quantité de données expérimentales sous différentes conditions de distance de champs, de profondeur et de tailles de champs via des formules analytiques simples (type doubles exponentielles pour les profils ou ellipses pour distribution de doses planaires [Francois et al., 1988, McParland and Fair, 1991, Van der Giessen and Hurkmans, 1993, Van der Giessen, 1994, Diallo et al., 1996, Chofor et al., 2012].
Comme on peut le constater, ces modèles parfois désignés par modèles empiriques ne reposent pas sur un formalisme de physique des particules et reposent donc sur un important effort logistique pour collecter toutes les mesures nécessaires à leur mise en œuvre. Ce type de modèles *pragmatique* de calcul des composantes diffusé tête et fuite repose donc uniquement sur la conséquence du phénomène à savoir la dose mesurée et non à la source de celui-ci, notamment la tête de l'appareil d'irradiation.

Avec les modèles stochastiques appelés modèles de Monte Carlo au contraire, la géométrie de la tête de l'appareil ainsi que les lois régissant les interactions rayonnement/matière sont au centre de leur formalisme : reposant sur les dimensions exactes des éléments constituant l'appareil de traitement, la méthode consiste à générer des centaines de millions de trajectoires de photons et d'électrons régis par les probabilités d'occurrence de différentes interactions physiques dans la tête d'irradiation. Ainsi, en plus du temps de calcul conséquent pour estimer la dose hors champ (Kry [Kry et al., 2006, Kry et al., 2007] parle de 900 heures de simulations pour générer 400 millions de scénarios), ces modèles hors champ doivent inclure l'ensemble des éléments constituant l'accélérateur, susceptibles d'intéragir avec le faisceau. Ceci rend le travail de modélisation géométrique de l'accélérateur presque insurmontable (penser également aux secrets industriels qui empêchent l'accès aux plans détaillés de l'appareil) ce qui fait que leurs applications restent limitées par la difficulté à les mettre en place pour plusieurs types de machines.

Objectif

L'objectif du travail présenté dans ce chapitre était de mesurer expérimentalement, pour divers modèles d'appareils de traitement en radiothérapie, la composante *diffusé collimateur et fuites* des rayonnements afin de calibrer un modèle analytique semi-empirique

de calcul de dose à distance.

Ce modèle s'inspire de la modélisation de type multi-source, utilisée avec succès dans les études de l'influence de la taille du champ sur la dose à l'axe du champs d'irradiation, et qui sera présentée dans la section 1.2.3. Brièvement, ces modèles sont basés sur la modélisation d'une source primaire ponctuelle de photons ainsi que d'un plan horizontal considéré comme la source des rayonnements de diffusion issus de la tête de l'appareil. Ce plan est muni d'une d'intensité de rayonnement gouverné par une certaine densité de probabilité plane. Ainsi, la dose en un point d'intérêt P est proportionnelle à l'intégrale de cette intensité sur la partie du plan visible depuis P à travers le système de collimation (voir Figure 1.5.

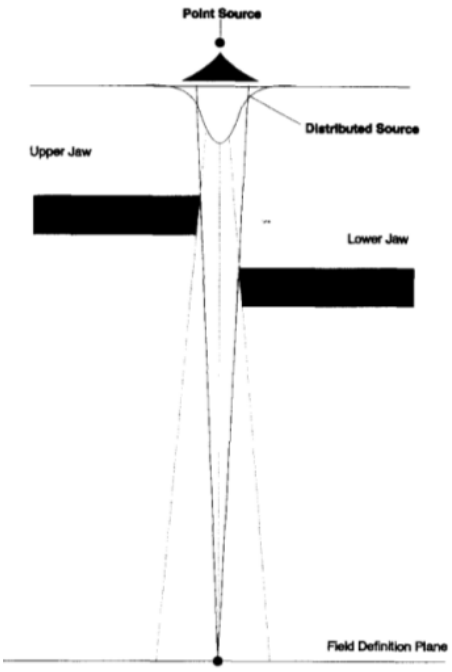

FIGURE 1.5 – Modèle multisource décrit dans [Dunscombe and Nieminen, 1992]. *Point source* : une source ponctuelle de photons directs. *Distributed source* : un plan source de photons diffusés par les premières composantes de la tête d'irradiation rencontrées (collimateur primaire et filtre égalisateur en triangle en noir) muni d'une distribution d'intensité Gaussienne. *Jaws* : les mâchoires ou collimations secondaires définissent la portion visible du plan des photons diffusés.

Ainsi, les collimateurs ne sont utilisés dans ces modèles que pour définir la part visible du plan source du rayonnement diffusé. De plus, l'intensité définie sur ce plan n'est pas motivée par des arguments de physique des rayonnements.

Le but de la partie modélisation du présent travail est de proposer une adaptation de l'approche multisource afin de modéliser la dose hors champs en considérant les mâchoires comme source de rayonnement diffusé munie d'une intensité définie à partir de lois d'énergie et d'interaction rayonnement-matière. Le modèle proposé peut donc être perçu comme un compromis entre la volonté d'une description physique propre aux modèles stochastiques de Monte Carlo et la modélisation géométrique simplifiée de la tête d'irradiation de l'approche multisource.

1.2 Matériel et Méthode

1.2.1 Mesures expérimentales

Modalités d'irradiation

Les mesures expérimentales ont été réalisées sur trois modèles d'appareils d'irradiation de radiothérapie :

1) L'irradiateur au Cobalt60, Alcyon II.

2) Accélérateur linéaire Clinac 2300C/D à 6 et 20 MV.

3) Accélérateur linéaire Novalis Tx à 6 MV.

Une cuve à eau d'une dimension $100cm \times 50cm \times 30cm$ a été placée à l'extérieur du champs du faisceau dans laquelle on a disposé les TLDs à 10cm de profondeur.

Cinq différentes tailles de champs ont été utilisées lors de l'irradiation au Cobalt60 ($5 \times 5, 10 \times 10, 15 \times 15, 20 \times 20$ et 30×30 cm^2) avec des points de mesures disposés de 15 (20 et 25 pour les champs 20×20 et 30×30 cm^2 respectivement) à 70 cm de l'axe du faisceau.
De même, quatre différentes tailles de champs ont été utilisées lors des mesures sur les accélérateurs linéaires ($5 \times 5, 10 \times 10, 15 \times 15,$ et 30×30 cm^2) avec des points de mesures disposés de 15 (25 pour le champ 30×30 cm^2) à 70 cm de l'axe du faisceau.
La distance source surface a été fixée à 80 cm lors de l'irradiation au cobalt et à 100 cm pour les irradiations sur accélérateurs linéaires.

Choix du détecteur

La dose due aux photons a été mesurée à l'aide de dosimètres thermoluminescents (TLD). Le TLD est un dosimètre relatif à lecture différée dont le principe de fonctionnement repose sur la lecture d'une émission de lumière par chauffage.

Le TLD le plus habituellement utilisé en clinique est le TLD-100 mais ce dernier se révèle sensible à la contamination par les neutrons en particulier aux énergies supérieures ou égales à 10 MV. C'est pour ces raisons que nous avons opté pour le TLD-700 composé de 99.99 % de 7Li et de ce fait insensible aux neutrons.

Le TLD-700 (Harshaw Chemical Company, Solon, OH), plus cher et moins utilisé en clinique, reste le dosimètre le plus précis pour la mesure de la dose en dehors du champ irradié. En effet, pour les mesures de doses à distance, les dosimètres utilisés doivent être suffisamment sensibles à des doses qui, nous le verront par la suite, sont de l'ordre du cGy à une distance 20cm jusqu'au mGy à plus de 40cm et ceci pour une dose de 10Gy délivrée à l'axe. La lecture après irradiation a été effectuée par le lecteur automatique PCL-3 (Fimel, Velizy, France).

La préparation et la lecture des TLDs ont été confiées au laboratoire Equal-Estro, basé à l'époque, à l'institut Gustave Roussy pour sa longue expérience en dosimétrie thermoluminescente [Marre et al., 2000, Ferreira et al., 2000, Derreumaux et al., 1995]. La calibration et la dose dépendance en énergie pour chaque appareil d'irradiation a été faite par comparaison avec le signal de référence 2Gy en Cobalt60.

Concernant les incertitudes de mesures, des études menées à l'IGR ont conclu à une incertitude relative de l'ordre de 5% dans l'axe et pouvant atteindre 15% pour les doses les plus à distance (60cm/70cm de l'axe) en raison du rapport de force existant entre le signal perçu et le bruit de fond initial du détecteur.

1.2.2 Evaluation semi-empirique de la dose due à la diffusion externe

La dose totale absorbée $D(f, r, \varphi, z)$ en un point d'intérêt de coordonnées cylindriques (r, φ, z) et situé l'extérieur du champs d'irradiation de taille f est en général décrite à l'aide de trois composantes :

$$D(f, r, \varphi, z) = D_P(f, r, \varphi, z) + D_C(f, r, \varphi, z) + D_F(f, r, \varphi, z) \tag{1.1}$$

avec D_P, D_C et D_F les doses de radiations dues respectivement au diffusé patient, au diffusé collimateur et aux rayonnements de fuite.

Le dispositif expérimental illustré dans la Figure 1.6 proposé par Kase [Kase et al., 1983] et repris par Ruben [Ruben et al., 2011] permet d'exclure la composante du diffusé patient de la mesure expérimentale réduisant ainsi l'équation (1.1) à la somme :

$$D_{Mesure}(f, r, \varphi, z) = D_F(f, r, \varphi, z) + D_C(f, r, \varphi, z) \tag{1.2}$$

En effet, le positionnement de la cuve à eau en dehors du champ d'irradiation élimine de fait la contribution de photons issus de la diffusion directe dans l'eau faute de volume irradié dans le champ.

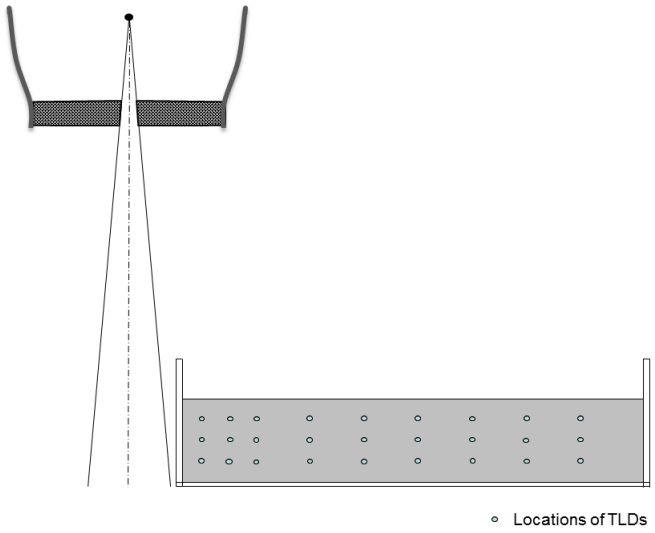

FIGURE 1.6 – Dispositif expérimental utilisé pour la mesure de la dose due aux rayonnements de diffusions collimateurs et fuites.

Extraction de la composante de dose indépendante de la taille du champ

Notre hypothèse de travail à propos de la composante due aux rayonnements de fuite est qu'elle est indépendante de la taille du champ. La représentation graphique semi-logarithmique, à une position fixée (r, φ, z) des doses mesurées en fonction des tailles de champs $5 \times 5, 10 \times 10, 15 \times 15\ cm^2$ suggérant une dépendance linéaire, nous proposons la modélisation suivante :

$$D_{mesure}^{Norm}(f, r, \varphi, z) = \lambda(r, \varphi, z) + \kappa(r, \varphi, z) \times [\exp(\nu(r, \varphi, z) \times f) - 1] \tag{1.3}$$

avec $D_{mesure}^{Norm}(f, r, \varphi, z)$ la dose mesurée par le procédé expérimental illustré dans la Figure 1.6 normalisée par la dose maximale à l'axe et les $\lambda(r, \varphi, z), \kappa(r, \varphi, z)$ et $\nu(r, \varphi, z)$ les trois paramètres de la régression sur la taille du champ f obtenus pour chaque position (r, φ, z) fixée.

En se référant aux résultats expérimentaux de Kase [Kase et al., 1983] dans lesquels le rayonnement de fuite apparait indépendant de la taille du champ, nous avons donc déduit par identification des équations (1.2) et (1.3) que :

$$\begin{cases} D_F^{Norm}(f, r, \varphi, z) &= \lambda(r, \varphi, z) \\ D_C^{Norm}(f, r, \varphi, z) &= D_{mesure}^{Norm} - \lambda(r, \varphi, z) \end{cases} \tag{1.4}$$

avec $D_F^{Norm}(f, r, \varphi, z)$ et $D_C^{Norm}(f, r, \varphi, z)$ les doses dues aux rayonnements de fuite et aux diffusions collimateurs normalisées par la dose maximale à l'axe.

A ce stade de l'étude, il nous est donc possible d'extraire des mesures expérimentales la dose due aux diffusions collimateurs comme différence des doses mesurées et des constantes $\lambda(r, \varphi, z)$ estimées aux positions hors champ correspondantes. Dans la suite, nous proposons une modélisation multi-source de la composante D_C^{Norm}.

1.2.3 Modélisation multi-sources

Concept général

La modélisation du terme diffusé collimateur D_C de l'équation 1.2 repose sur une modélisation de type multisource. Cette modélisation a été proposée par Dunscombe et Nieminen [Dunscombe and Nieminen, 1992] et appliquée avec succès par d'autres auteurs dans l'étude de la dépendance de la dose à l'axe en fonction de l'ouverture du collimateur i.e la taille du champ [Dunscombe and Nieminen, 1992, Yu and Sloboda, 1996, Jian et al., 2001, Yang et al., 2002]
Ces modèles proposent d'une façon générale une représentation de la source de photons diffusée sous forme de disque ou de plan muni d'une distribution gaussienne [Dunscombe and Nieminen, 1992] ou une combinaison de gaussiennes [Yang et al., 2002] ou d'autres distributions multimodales [Yu and Sloboda, 1996, Jian et al., 2001]. Comme illustré dans la Figure 1.7, le modèle proposé dans ce travail modélise, de façon différente et plus complexe, la source de photons diffusés en ensemble de sources planaires qui sont autant d'origines possibles du rayonnement diffusé collimateur.

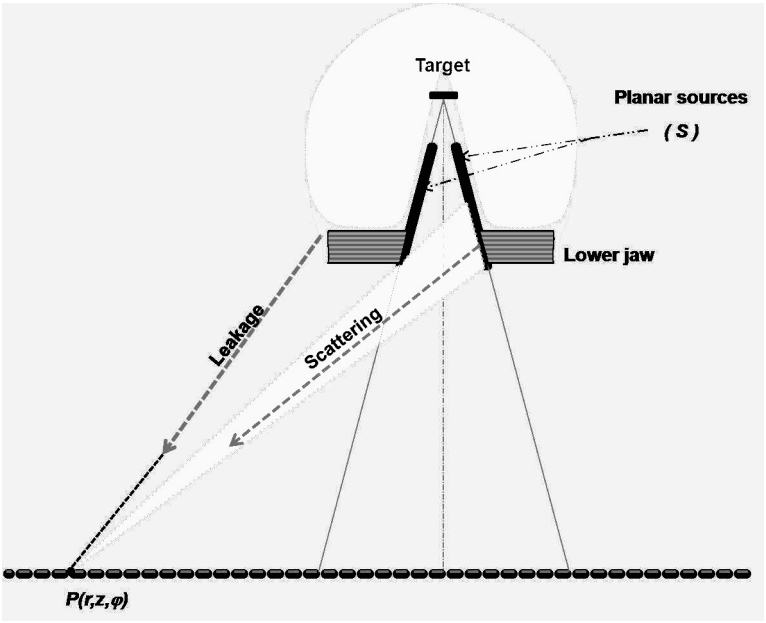

FIGURE 1.7 – Une représentation schématique de la source surfacique $\mathcal{S}$, du blindage linéaire de la tête de l'accélérateur ainsi que des mâchoires inférieures du collimateur secondaire.

Modélisation de la source *diffusé collimateur*

La source de photons diffusés depuis la collimation de l'appareil, notée $\mathcal{S}$, est formée d'un ensemble de plans originaires depuis la base du filtre égalisateur de l'accélérateur jusqu'à la base du système de collimation. $\mathcal{S}$ représente les faces intérieures de l'ensemble mâchoires et/ou collimateurs et/ou multilames et/ou compensateurs,...etc. Le nombre et l'orientation de ces plans constituant la source $\mathcal{S}$ permet donc de rendre compte de la forme et de la taille du champ d'irradiation désiré.

Nous émettons également l'hypothèse que seuls contribuent à la dose du diffusé collimateurs $D_C(f, r, \varphi, z)$ les photons issus des faces visibles de la source $\mathcal{S}$ depuis le point d'intérêt $P(r, \varphi, z)$. Les photons originaires des faces non visibles de $\mathcal{S}$ sont quant à eux considérés comme pris en compte dans le terme fuite $D_F(f, r, \varphi, z)$ via la constante $\lambda(r, \varphi, z)$. La source $\mathcal{S}$ a été modélisée avec une résolution de $5mm \times 5mm$ et un algorithme de type *Z-Buffer* [Catmull, 1984] a été développé afin d'identifier les pixels constituants les phases visibles de $\mathcal{S}$ depuis $P(r, \varphi, z)$.

Distribution d'intensité de la source $\mathcal{S}$

Il a été clairement établi que la majeure partie du rayonnement diffusé dans le champ est originaire du filtre égalisateur dont la portée est limitée par le collimateur primaire de l'appareil [Ahnesjo, 1994, Chaney et al., 1994, Deng et al., 2000]. Par conséquent, nous supposons dans ce travail que les sources planes de S sont alimentées par les photons de diffusion qui provenaient de la partie inférieure du filtre égalisateur.
Partiellement inspirée de précédents travaux de modélisation [Ahnesjo, 1994, Yu and Sloboda, 1996], l'intensité de la source au *ième* pixel est exprimée analytiquement par une distribution quasi-triangulaire :

$$\Omega_D(r_i, \sigma_i) = \begin{cases} (1-\tau_i) \times \left(\frac{DSA}{r_i}\right)^{\alpha} & r_i \leq r_0 \\ (1-\tau_i) & r_i \geq r_0 \end{cases} \tag{1.5}$$

avec σ_i l'angle entre l'axe du faisceau et la droite reliant la cible et le $i^{ème}$ pixel, r_i la distance entre la cible et le $i^{ème}$ et enfin, la DSA est le distance source-axe égale à 100cm pour les accélérateurs linéaires et 80cm pour l'irradiateur au Cobalt60 (voir la Figure 1.8 pour visualiser chacun de ces termes). Les constantes τ, r_0 et α sont quant à elles indépendantes de $P(r, \varphi, z)$.

Distribution énergétique de la source $\mathcal{S}$

L'énergie d'émission du photon diffusé vers le point d'intérêt a été évaluée à l'aide de la formule de l'énergie diffusée de l'interaction inélastique de Compton [Compton, 1923] :

$$E_{\theta_i} = \frac{E_0}{1 + \frac{E_0}{m_e c^2} \times (1 - cos\,(\theta_i))} \tag{1.6}$$

avec E_{θ_i} est l'énergie du photon diffusé selon l'angle θ_i, E_0 est un paramètre à estimer des mesures expérimentales reflétant l'énergie moyenne des photons incidents issus du filtre égalisateur, m_e la masse d'un électron au repos et c la vitesse de la lumière.

Distribution angulaire de la source $\mathcal{S}$

Bien que l'interaction Compton soit prépondérante à basse énergie (Cobalt60 ou 6 MV) il n'en est plus de même à haute énergie (20 MV par exemple) à laquelle un photon qui traverse un matériau à haute densité comme le système de collimation va générer des diffusions supplémentaires qui ne peuvent plus être négligées (rayonnement de freinage, production de paires,...etc) et attribuables aux électrons secondaires mis en mouvement par le faisceau.

Ainsi, la distribution Compton doit être corrigée comme l'ont précédemment proposé des auteurs comme Ahnesjo [Ahnesjo, 1995]. Nous proposons à cette fin l'usage de la fonction de phase de Henyey–Greenstein dont l'avantage réside dans l'existence d'un paramètre libre pouvant tenir compte de l'ensemble des phénomènes physiques en présence. Plus précisement, la fonction de phase d'Henyey–Greenstein [Henyey and Greenstein, 1941] est définie par :

$$p_{HG}\,(\theta) = \frac{1}{4\pi} \times \frac{1 - g^2}{(1 + g^2 - 2 \times g \times cos(\theta))^{3/2}} \tag{1.7}$$

avec $\theta \in [0, \pi]$ est l'angle existant entre la direction du photon avant/après sa diffusion. Le paramètre g est le facteur d'asymétrie d'Henyey–Greenstein.
La probabilité de diffusion dans l'angle solide défini par un cône infinitésimal autour de la direction d'angle θ (voir Figure 1.8) est donnée par [Binzoni et al., 2006] :

$$\mathcal{P}_\theta^g = 2\pi \times p_{HG}(\theta) \times sin(\theta) = \frac{1}{2} \times \frac{(1 - g^2) \times sin(\theta)}{(1 + g^2 - 2 \times g \times cos(\theta))^{3/2}} \tag{1.8}$$

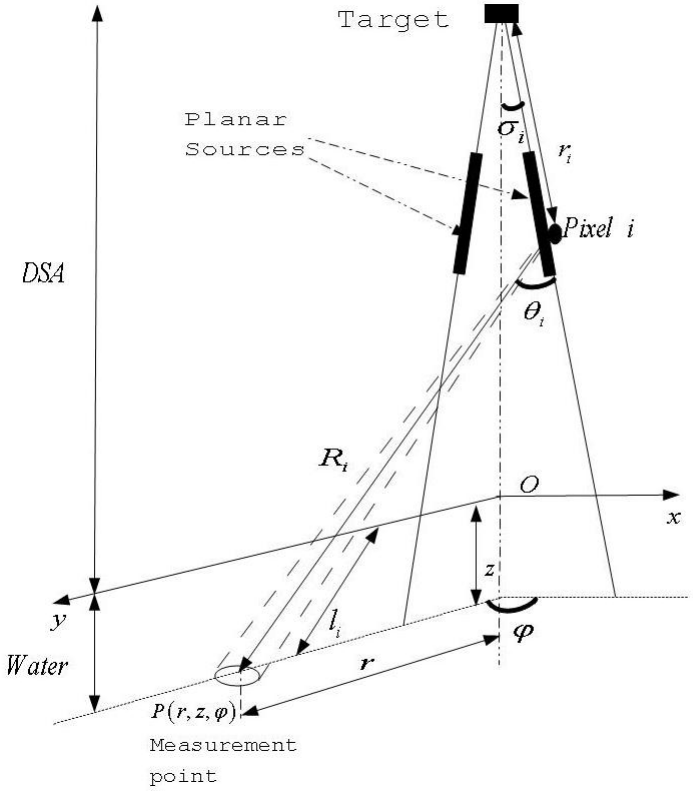

FIGURE 1.8 – Termes géométriques utilisés dans l'expression du modèle final de l'équation 1.9

Modèle Multi-plans final

Le modèle proposé pour relier la dose $D_F^{Norm}(r,\varphi,z)$ de radiation normalisée due à la diffusion du système de collimation pour une taille de champs f aux photons issus de $\mathcal{S}$ est donné par l'expression :

$$D_C^{Norm}(r,\varphi,z) = \omega \times \sum_{i \in \mathcal{S}_{visible}} \Omega_D(r_i,\sigma_i) \times \frac{\mathcal{P}_{\theta_i}^g \times E_{\theta_i} \times \exp(-\mu_{eau,\theta_i} \times l_i)}{R_i^2} \tag{1.9}$$

avec (voir la Figure 1.8) :

- $\Omega_D(r_i,\sigma_i)$ est l'intensité de la source $\mathcal{S}$ au pixel i.
- $\mathcal{P}_{\theta_i}^g$ est la probabilité de diffusion selon l'angle θ_i définie par l'équation 1.8.
- E_{θ_i} est l'énergie Compton à laquelle le photon est diffusé selon l'angle θ_i.
- μ_{eau,θ_i} est le coefficient linéique d'atténuation dans l'eau à l'énergie E_{θ_i}.

- R_i la distance entre et le point $P(r, \varphi, z)$ et le $i^{ème}$ Pixel de $\mathcal{S}$.
- l_i la distance entre et le point $P(r, \varphi, z)$ et le point d'entrée dans l'eau du photo provenant du $i^{ème}$ Pixel de $\mathcal{S}$.
- ω est le coefficient de proportionnalité reliant la fluence énergétique et la dose $D_C^{Norm}(r, \varphi, z)$.

La somme dans l'équation (1.9) est effectuée sur l'ensemble des pixels qui composent la surface $\mathcal{S}_{visible}$ désignant les pixels de $\mathcal{S}$ visibles depuis le point d'intérêt $P(r, \varphi, z)$.

La modélisation de la source $\mathcal{S}$ demande très peu d'informations sur la géométrie de la tête d'irradiation : les distances source-filtre égalisateur, source-mâchoires et source-collimateur multilames ainsi que leurs épaisseurs. Ces informations sont disponibles sans difficulté dans la documentation de base de l'appareil d'irradiation.

Les paramètres $E_0, g, \tau, r_0, \alpha$ et ω du modèle (1.9) ont été estimés à partir des doses mesurées sur chaque appareil d'irradiation et à chaque énergie. En toute rigueur, ces paramètres dépendent de la taille du champ puisque les doses D_C^{Norm} sont normalisées par la dose maximale qui dépend de la taille du champ f. Nous avons cependant décidé, pour des raisons de puissance statistique, d'effectuer la régression sur toutes les doses normalisées mises en commun afin de proposer des paramètres indépendants de f et offrant un bon compromis entre les différentes tailles de champs étudiées.

1.3 Résultats

La Figure 1.9 donne une estimation du rayonnement de fuite obtenue par régression du modèle (1.3). Selon la distance à l'axe du faisceau, les estimations de D_F^{Norm} varient entre 0.013% à 0.055%, de 0.005% à 0.13% et de 0.01% à 0.18% du maximum de dose à l'axe pour le Cobalt60, le 6MV et le 20MV respectivement. L'écart type moyen de ces estimations, toutes énergies confondues est, quant à lui, égal à 15% (min=2%,max=43%).

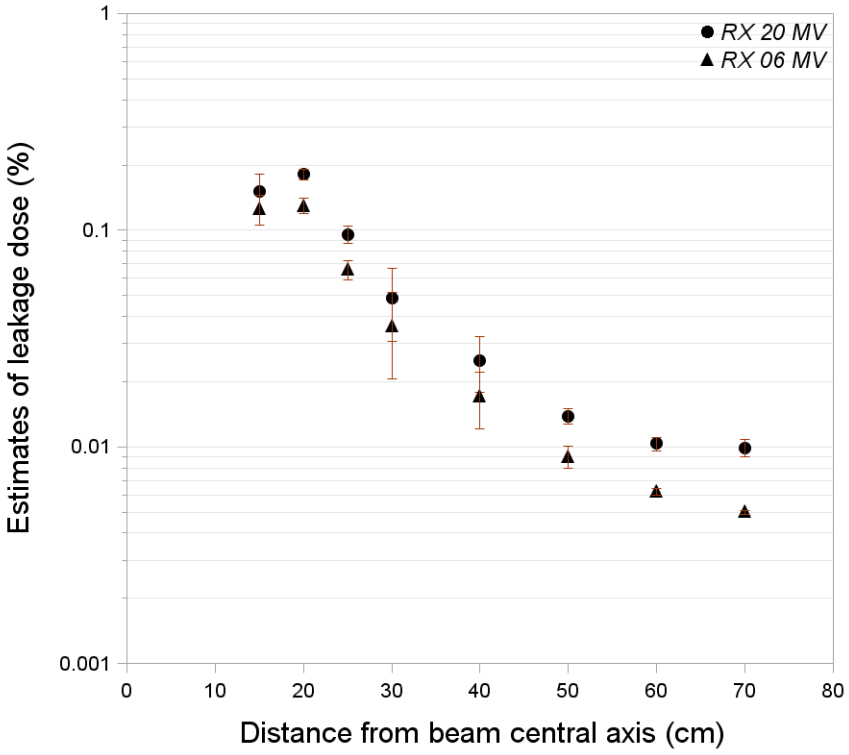

FIGURE 1.9 – Estimation des doses du rayonnement de fuite (paramètre $\lambda(r, \varphi, z)$ de l'équation (1.3)) pour l'accélerateur linéaire Varian Clinac 2300 C/D opérantà 6 et 20 MV RX, respectivement. Les doses sont données en pourcentage de la dose à la profondeur de la dose maximale à l'axe du faisceau.

Le tableau 1.1 donne les valeurs des paramètres de régression obtenues pour le modèle (1.9).

La Figure 1.10 compare les profils de doses du rayonnement diffusé par le système de collimation mesurés et calculés à partir des paramètres du tableau 1.1 pour différentes tailles de champs et énergies de chaque appareil d'irradiation.
Les trois quarts des erreurs relatives entre les doses mesurées et les doses calculées toutes énergies et tailles de champs confondues sont inférieures à 14% avec une médiane égale à

	$E_0\,(MeV)$	g	$\omega\ (cm^2MeV^{-1})$	$\tau\ (rad^-1)$	α	$r_0\ (cm)$
Alcyon II	1	0.537	0.139	0 [a]	1.7	15
6MV Novalis Tx	2	0.223	0.314	3.5	2	20
6 MV Clinac 2300 C/D	2	0.418	0.319	3.8	2	25
20 MV Clinac 2300 C/D	6	0.103	0.303	3.8	2	25

a. Pas de filtre égalisateur pour la machine au Cobalt60 Alcyon II.

TABLE 1.1 – Valeurs des paramètres de régression du modèle (1.9) pour les différentes machines utilisées.

9% et de moins bonnes performances pour les tailles de champs extrêmes : 14% pour le champ $5 \times 5cm^2$, 6% pour le $10 \times 10cm^2$, 8% pour $15 \times 15cm^2$ et 17% pour $30 \times 30cm^2$.

Par exemple, l'accord le plus faible entre les calculs et mesures a été observé pour le Clinac 2300 C/D pour les tailles de champ $5cm \times 5cm$ et $30cm \times 30cm$. Dans le premier cas, le modèle sous-estime la dose normalisée à 15 et 20cm de l'axe de 20% et 13% respectivement. De même, le plus grand écart entre les mesures et les calculs a été observé pour le champ $30cm \times 30cm$ où le modèle surestime la dose normalisée à 30 cm de l'axe de 40%.

Enfin , La figure 1.11 illustre les résultats de modélisation de la dose diffusé-collimateur à distance de champs simples (carrés et rectangles) et complexes représentatifs de ce qui peut être réalisé à l'aide de collimateurs multilames.

1.4 Discussion

Nous avons proposé dans ce travail un nouveau modèle de type multisource pour la modélisation de la dose à distance due aux rayonnements diffusés et de fuite de la tête de l'appareil de radiothérapie. Cette approche, utilisée avec succès dans les modélisations du diffusé collimateur dans le champ, s'est révélée également satisfaisante dans le cadre de la modélisation de la dose diffusée hors champ (erreur relative médiane entre calculs et mesure de l'ordre de 9%).

Le présent travail offre deux améliorations importantes à l'approche multisources classique : tout d'abord une modélisation plus fine (qu'un simple plan parallèle à la surface d'entrée des photons) de la source de diffusion $\mathcal{S}$ basée sur une série de plans inclinés de façon à reproduire la forme et la taille du champ d'irradiation d'intérêt. Ensuite, la

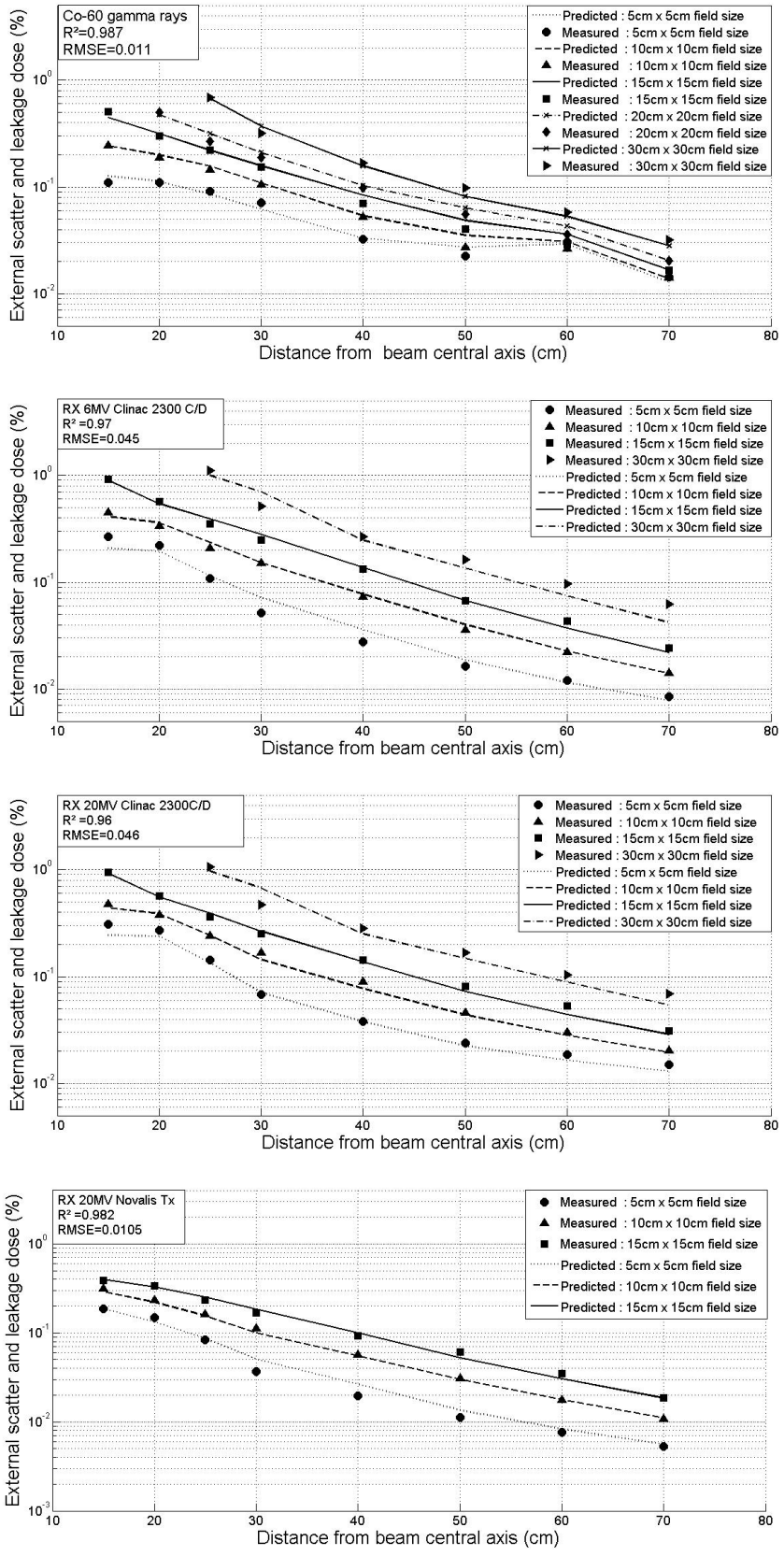

FIGURE 1.10 – Comparaison des profils de doses à distance mesurés et calculés, à 10cm de profondeur, pour différents appareils et différentes conditions d'irradiation.

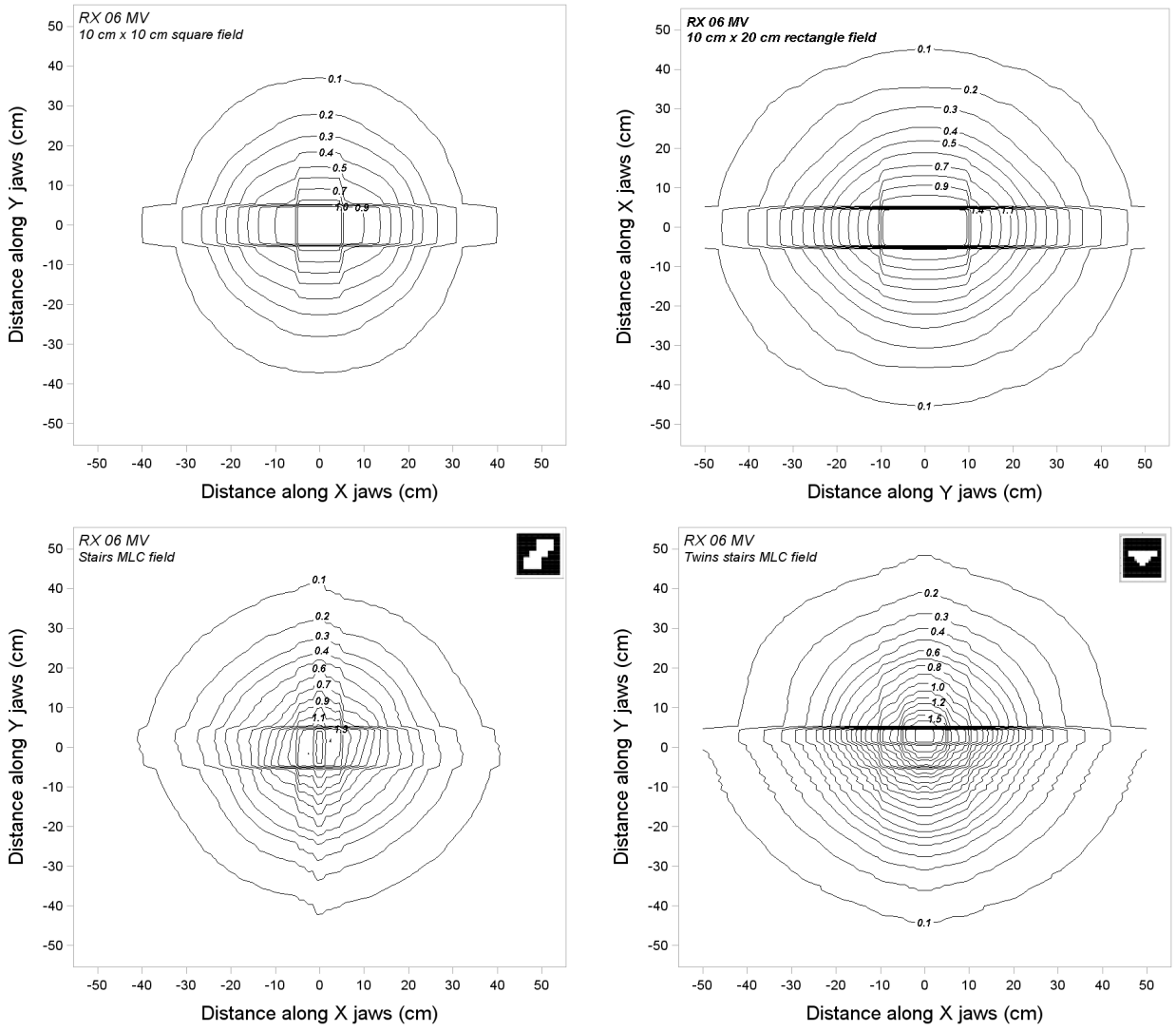

FIGURE 1.11 – Isodoses calculées pour différentes tailles et formes de champs représentés en haut à gauche. Les doses sont données en pourcentage de la dose maximale à l'axe du faisceau. Les calculs sont effectués pour Varian accélérateur linéaire Clinac 2300 C/D opérant à 6 MV.

dose de radiation absorbée au point d'intérêt n'est plus seulement proportionnelle à des quantités purement géométriques relatives aux surfaces visibles de $\mathcal{S}$ mais aussi à la prise en compte d'interactions physiques rayonnement-matière comme l'interaction Compton corrigée.
Cette double amélioration de l'approche multi-source initiale pourrait s'avérer utile dans le développement d'algorithmes de calculs plus précis de la dose au *volume résiduel à risque* (RVR) comme le recommande l'ICRU [ICRU, 2010]. L'intérêt de ce modèle est également de pouvoir tenir compte de la complexité de certains champs complexes utilisés en routine clinique pour des traitements sur accélérateurs équipés de collimateurs multilames.
Autre point important : notre modélisation se base sur une approche *monoénergétique* du photon incident à travers le paramètre E_0 du modèle censé représenter une énergie moyenne du spectre *polyénergétique* accessible en pratique uniquement via des calculs de simulations de type Monte-Carlo. Cependant, si l'information de la composition du spectre énergétique de la machine de radiothérapie est disponible, il est alors aisé d'adapter l'approche proposée tout simplement en considérant la somme pondérée par le spectre de la fluence calculée pour chacune de ses composantes monoénergétiques.

Evidemment, la validité du modèle semi-empirique proposé est étroitement liée à l'incertitude des mesures par TLD. Des études d'incertitude sur les mesures des TLD700 utilisées lors de nos procédés expérimentaux ont estimé des écarts types entre 1.25 à 2.25% [Derreumaux et al., 1995, Marre et al., 2000, Ferreira et al., 2000]. Dans le contexte de la mesure de la dose à distance, l'incertitude est plus grande d'une part en raison des doses mesurées plus faibles et d'autre part en raison du changement dans le spectre énergétique [Scarboro et al., 2011]. L'incertitude globale (en terme de deux fois l'écart type) dans des conditions de mesures de doses à distance avec trois TDLs a été estimée à 10%.

1.5 Conclusion et perspectives

La dose absorbée à distance du champ d'irradiation comprend essentiellement trois composantes : la dose absorbée due à la diffusion interne provenant de l'intérieur du patient, la dose absorbée due à la diffusion externe du système de collimation de la machine de traitement et la dose absorbée due à une fuite externe provenant également de la tête de la machine de traitement.
Ces deux dernières composantes ont été modélisées avec succès avec une modélisation adéquate de la source de diffusion $\mathcal{S}$.

La faisabilité de la méthode pour modéliser les rayonnements diffusés d'irradiation à forme de champs complexes offre un domaine d'application en clinique de toute première importance. En effet, l'usage de plus en plus fréquent de l'IMRT dans des protocoles de traitement par radiothérapie est souvent accompagné d'une augmentation de la dose prescrite comparé aux autres modalités 3D et de fait, d'une augmentation des unités moniteurs [Ruben et al., 2011]. Ces techniques irradient donc les volumes cibles avec des doses importantes via de petits champs complexes fortement collimatés ce qui fait jouer à la dose due au diffusé collimateur et fuite un rôle encore plus important qui doit être précisé et quantifié.

Suite aux recommandations de l'IRCU [ICRU, 2010], une meilleure estimation de la dose hors-champs devrait recevoir un intérêt clinique croissant (choisir entre différentes techniques en fonction de la dose de radiation délivrée en RVR par exemple) et leur intégration dans TPS devrait être bientôt généralisée. Ce travail peut donc répondre à ce nouveau besoin d'une part en raison de sa flexibilité au regard de la taille et de la forme du champ d'irradiation et d'autre part par son faible coût en calculs comparé à des méthodes de simulations de type MonteCarlo par exemple.

Il pourrait ainsi contribuer à une reconstruction d'une dose d'exposition aussi précise que personnalisée notamment pour estimer le risque d'effets déterministes liés par exemple à la dose à l'ovaire ou cristalline et d'effets à long terme comme les maladies cardiovasculaires, pulmonaires, fibrose, seconds cancers, etc...

1.6 Vers le prochain chapitre

Plusieurs publications épidémiologiques [BEIR, 2006, Xu et al., 2008, NCRP, 2012, de Vathaire et al., 1999b] ont été consacrées à l'étude de l'incidence de seconds cancers chez les patients ayant reçu une radiothérapie pour un cancer de l'enfance, de l'adolescence ou à l'âge adulte. Il semble cependant y avoir des incertitudes considérables dans les modèles de risque actuels, à la fois dus aux incertitudes sur les doses mais aussi aux modèles statistiques utilisés.
Pour mieux évaluer les relations dose-effet, plusieurs études ont suggéré que le risque de cancer est lié à la distribution de dose inhomogène à travers un organe plutôt que la dose moyenne [Dasu et al., 2005, Schneider et al., 2005].

Dans le prochain chapitre, nous proposons une nouvelle méthode de modélisation de complications radio-induites tenant compte de l'intégralité de la distribution dose à l'organe.

Chapitre 2

Analyse de données fonctionnelles

Sommaire

2.1 L'approche fonctionnelle 26

2.2 Note historique 28

2.3 Statistique fonctionnelle et statistique multivariée 30

2.4 Analyse en composantes principales fonctionnelles 32

2.5 Deux Exemples Importants 35

2.5.1 Analyse en composantes principales fonctionnelles dans un espace engendré par une base [Ramsay and Silverman, 2005] . . . 35

2.5.2 Analyse en composantes principales fonctionnelles d'un ensemble de densités de probabilité 37

2.6 Modèles de régression sur données fonctionnelles à réponses réelles 41

2.6.1 Le modèle de régression linéaire fonctionnel 41

2.6.2 Modèles linéaires généralisés et Modèles de survie fonctionnels . 42

2.1 L'approche fonctionnelle

Le terme « données fonctionnelles » se réfère aux données dans lesquelles les observations sont des courbes, surfaces ou hypersurfaces par opposition à un point ou un vecteur multidimensionnel. Avec l'émergence des nouvelles technologies, beaucoup de bases de données actuelles contiennent des données mesurées dans le temps et dans l'espace sur des grilles très fines. Bien que ces mesures ne soient pas continues par essence, elles peuvent être néanmoins considérées dans de nombreux cas comme la discrétisation de phénomènes continus. Ainsi, l'analyse des données fonctionnelles considère ce type de données comme générées par des fonctions alétoires définies sur un continuum.

Cette approche a reçu en biostatistique une attention toute particulière comme en témoigne la publication récente d'articles de synthèse dans les revues *Statistics in Medecine* [Sørensen et al., 2013] et *BMC Medical Research Methodology* [Ullah and Finch, 2013]. Ces publications font un état des nombreuses applications déjà faites avec succès de ces techniques dans des domaines aussi nombreux et variés que la médecine [Erbas et al., 2010a, Ullah and Finch, 2010a, West et al., 2007, Stier et al., 2004], l'écologie [Ikeda et al., 2008, Henderson, 2006b, Manté et al., 2005a, Manté et al., 2005b, Bjornstad et al., 1998], l'imagerie médicale [Viviani et al., 2005, Long et al., 2005], l'économie [Grambsch et al., 1995, Bapna et al., 2008, Ramsay and Ramsey, 2002], la psychologie [McAdams, 2004, Vines et al., 2005, Chapados and Levitin, 2008], la neurologie [Harrison et al., 2007, Buckner et al., 2004], la linguistique [Koenig et al., 2008, Lee et al., 2006], l'argriculture [Sauder et al., 2002] ou la spectrométrie [Ferraty and Vieu, 2002, Ferraty et al., 2007]. D'autres applications des statistiques fonctionnelles peuvent être trouvées dans l'ouvrage de référence de Ramsay et Silverman [Ramsay and Silverman, 2002].

Même si les données ne sont pas par essence fonctionnelles, comme c'est le cas dans ce qui va être présenté, on peut être amené à étudier des variables fonctionnelles construites à partir de l'échantillon original. En effet, dans le cadre de cette thèse, les données fonctionnelles se présenteront souvent sous forme de densités de probabilités issues de la distribution de valeurs de comptage en très grand nombre dans des échantillons indépendants (plusieurs milliers de valeurs pour les doses par voxels calculées dans l'organe d'un patient ou de longueurs de télomères mesurées dans des cellules de lymphotyctes d'un prélèvement sanguin). De façon plus abstraite, nous rencontrerons également des données fonctionnelles en régression B-spline adaptatives provenant de la traduction de la propriété de support minimum conditionnellement aux valeurs de la variable continue

en présence.

2.2 Note historique

Le livre de référence [Ramsay and Silverman, 2005] signale les premières traces d'analyses statistiques tenant compte de la nature fonctionnelle des variables en présence dès le $19^{ème}$ siècle dans les travaux des mathématiciens Legendre [Legendre, 1809] et Gauss [Gauss, 1809] dans leurs travaux sur les trajectoires des comètes. Une fois la notion de variables aléatoires fonctionnelles dégagée, de nombreux travaux ont suivi pour développer dans le cadre fonctionnel les méthodes statistiques multivariées existantes.

En particulier, les notions très simples de moyenne et covariance ont dû être étendues au cadre fonctionnel bénéficiant au passage de l'apport conséquent de l'analyse mathématique fonctionnelle et en particulier de la théorie des opérateurs linéaires d'espaces fonctionnels dont certains résultats importants dans l'étude de l'opérateur de covariance par exemple ont été établis dès le premier quart du vingtième siècle [Riesz, 1918, Hilbert, 1912].

L'étude spectrale des opérateurs de covariance, vue sous l'angle de l'opérateur linéaire fonctionnel ou de celui de l'opérateur intégral à noyau, a mené à une rapide généralisation au cadre fonctionnel de *l'analyse en composantes principales* décrite pour la première fois dans le cas multivarié par Pearson [Pearson, 1901] et Hotelling [Hotelling, 1933]. L'analyse du spectre des matrices symétriques a alors été étendue aux opérateurs de covariance, vus comme un cas particulier d'opérateur à noyau symétrique, via la très importante décomposition de Karhunen-Loève [Karhunen, 1947, Loève, 1945].
Une des toutes premières applications de l'analyse en composantes principales remonte à Rao [Rao, 1958] et Tucker [Tucker, 1958] et ont concerné les modèles de croissance ainsi que [Deville, 1974] pour les courbes de naissances durant les vingt premières années de mariage. Peu après, des résultats théoriques asymptotiques très importants de convergence des éléments spectraux de l'opérateur de covariance empirique ont été établis par Dauxois [Dauxois et al., 1982] :

La régression fonctionnelle est une autre importante généralisation au cadre fonctionnel de la statistique multivariée. Comme son nom l'indique, elle concerne deux variables dont l'une d'elles au moins est fonctionnelle.

Pour estimer le paramètre fonctionnel inconnu de ce type de modèles de régression, différentes méthodes ont été proposées dont la régression sur composantes principales,

une technique qui sera abondamment utilisée dans cette thèse.

La régression sur composantes principales permet d'obtenir une estimation du paramètre fonctionnel inconnu en se ramenant à un modèle paramétrique multivarié équivalent avec la troncature comme outil de réduction dimensionnelle.

La régression sur composantes principales a été appliquée dans le cadre des modèles linéaires fonctionnels avec réponse réelle en absence de bruit [Cardot et al., 1999] ou avec bruit [Crambes, 2007], ainsi que dans le cas où les données sont collectées de façon dense ou éparse [Yao et al., 2005b]. Elle a même été appliquée avec succès dans des modèles de régression linéaire généralisée comme le modèle logistique [Escabias et al., 2004, Escabias et al., 2005, HG Müller and Stadtmüller, 2005] ainsi que pour des modèles à variable de réponse fonctionnelle [Müller et al., 2008].

Un autre type de méthode d'estimation où les paramètres fonctionnels des modèles de régression fonctionnelle sont estimés par des techniques de pénalisation dans des bases appropriées telles que les splines. Ce type de méthode ne sera pas abordé dans cette thèse mais on pourra trouver des exemples d'application pour le modèle linéaire ([Cardot et al., 2003, Ramsay and Silverman, 2005, Reiss and Ogden, 2009] et pour le modèle linéaire généralisé [Cardot and Sarda, 2005, James, 2002, Reiss and Ogden, 2010] ainsi que pour le cas où les variables réponses sont fonctionnelles [Reiss et al., 2010, Ivanescu et al., 2012].

Il serait inutile, en quelques pages, de tenter de présenter tous les outils méthodologiques à disposition en statistique fonctionnelle. Cette section présente les deux méthodes d'analyse les plus utilisées dans cette thèse : l'analyse en composantes principales fonctionnelles et les modèles de régression à variables fonctionnelles. Une multitude d'autres méthodes d'analyse peuvent être consultées dans des livres de référence tels que [Ramsay and Silverman, 2005, Ferraty and Vieu, 2006].

2.3 Statistique fonctionnelle et statistique multivariée

D'une façon générale, l'approche fonctionnelle apparaît dans un contexte de données de grande dimension avec des observations consécutives hautement corrélées, deux caractéristiques qui rendent leur étude par les méthodes d'analyses multivariées classiques difficile voire impossible. En effet, un cadre multivarié de l'étude aboutirait à un paradoxe bien connu sous le nom de « malédiction de la dimension » : loin d'être un avantage pour la connaissance des phénomènes, l'abondance de données aboutirait à une détérioration des résultats statistiques ! !

En effet, les travaux de Stone [Stone, 1980b, Stone, 1982, Stone, 1983] ont montré que la vitesse optimale de convergence des estimateurs non-paramétriques (de la fonction de régression, densité, quantiles,...) se détériore très rapidement lorsque la dimension de la variable explicative augmente et ce, en raison de la raréfaction des données la dimension augmente.

Des solutions à ces problème ont été proposées dans le cadre spécifique de la statistique multivariée.

Tout d'abord pour la prise en compte des colinéarités avec la *ridge regression* (introduite initialement par [Hoerl and Kennard, 1981], la régression sur composantes principales [Massy, 1965], ou la régression des moindres carrés partielle [Wold, 1975, H Martens and Naes, 1989, Helland, 1990].
Ces différentes méthodes ont été comparées dans [Frank and Friedman, 1993] et la non prise en compte du caractère fonctionnel des objets mathématiques en présence a été discutée par Hastie et Mallows [Hastie and Mallows, 1993].

Des réponses aux difficultés liées aux grandes dimensions ont été proposées via le modèle additif ([Stone, 1980a], le modèle additif généralisé[Hastie and Tibshirani, 1980, Hastie and Tibshirani, 1990], le modèle à transformations optimales [Breiman and Friedman, 1985], le modèle à projections révélatrices [Diaconis and Shashahani, 1984].

Avec l'approche fonctionnelle, les données observées sont donc perçues comme la réalisation de variables aléatoires fonctionnelles. Ainsi, les vecteurs de grande dimension sont remplacés par une unique fonction régulière ou lisse (dans le sens fonctionnel de continuité, dérivabilité, etc...) dont il constitue un ensemble de valeurs ponctuelles. Notons au passage que la notion de régularité fonctionnelle, *smoothness* en anglais, centrale en statistique

fonctionnelle, n'a pas son équivalent en analyse multivariée [Green and Silverman, 1994]. Cette approche reposant sur une entité fonctionnelle unique associée à chaque individu évacue les problèmes dus aux corrélations de type données répétées ce qui représente un changement notable dans la façon d'appréhender les problèmes de séries temporelles et données corrélées [Ramsay, 2007].

En effet, il est donc plus naturel dans le cadre fonctionnel d'envisager l'étude de données distribuées de façon irrégulière d'un individu à un autre tantôt éparse tantôt dense et souvent entachées d'erreurs de mesure. C'est une situation rencontrée souvent en milieu médical où les visites de suivi par exemple diffèrent d'un patient à l'autre tant au niveau des dates qu'au niveau du nombre. En plus d'une discrétisation différente d'une donnée fonctionnelle à l'autre, elles contiennent une certaine proportion d'erreurs de sources multiples : erreurs de saisie, données manquantes, valeurs aberrantes, etc...

Ces situations requièrent une attention particulière dans la modélisation des processus fonctionnels sensés générer les données observées [James and Sugar, 2003, Rice and Wu, 2001, Yao et al., 2005a]. Ceci en général au moyen de fonctions non-paramétriques telles que les splines [Crane et al., 2010, Ullah and Finch, 2010b, Erbas et al., 2010b, Parker and Wen, 2009, Ramsay, 2000, Ramsay et al., 1996], les séries de Fourier [Gao, 2007, Henderson, 2006a, Laukaitis, 2005, Ratcliffe et al., 2002b, Ratcliffe et al., 2002a], les ondelettes [Ogden and Greene, 201 Laukaitis, 2008, Lucero, 2005], les noyaux [Maslova et al., 2010, Baladandayuthapani et al., 2008, Hall et al., 2001] ou les polynômes locaux [Wu and Muller, 2010, Jiang et al., 2009].

2.4 Analyse en composantes principales fonctionnelles

Bien que les variables aléatoires fonctionnelles soient par essence de dimension infinie, il est possible de les approcher par leur projection dans des espaces fonctionnels de dimension finie munis d'une base.
L'intérêt de cette approche étant que la fonction considérée peut alors être vue comme un vecteur fini dimensionnel de coefficients relatifs à la base de l'espace de projection.
En pratique, nous disposons de nombreuses bases de projection telles que les fonctions trigonométriques, les polynômes orthogonaux, les splines, les ondelettes, etc. Toutefois, le recours à ces bases induit un certain *a priori* sur la nature des données fonctionnelles que l'on étudie.

Dans ce chapitre, nous présentons une méthode de construction d'une base alternative appelée **analyse en composantes principales fonctionnelles** et qui a pour but de construire une base uniquement à partir des données observées à disposition : à partir des éléments spectraux (valeurs propres et fonctions propres) de l'opérateur de covariance, une base orthogonale est construite en maximisant (pour toute dimension finie) la proportion de variance totale expliquée.

Il s'agit d'une généralisation au cadre fonctionnel de l'analyse en composantes principales multivariée. Ainsi, elle permet de réduire la dimension du problème tout en gardant le caractère fonctionnel des données et le maximum d'informations.
L'analyse en composantes principales fonctionnelles est l'une des méthodes les plus utilisées en statistique fonctionnelle notamment parce qu'elle permet dans certaines situations de se ramener à un problème de statistique multivariée via les coordonnées ou *scores* issues de la décomposition dans la base de fonctions propres.

Après une définition générale, nous nous intéresserons plus particulièrement à l'analyse en composantes principales de deux cas particuliers importants de données fonctionnelles que l'on rencontrera par la suite : les fonctions B-splines et les fonctions de densité de probabilité.

Etant donné que les données fonctionnelles sont par essence de dimension infinie, la réduction dimensionnelle a été étudiée très tôt de façon à représenter les données dans un espace de dimension finie tout en conservant leurs plus importantes caractéristiques.

Désignons par X une variable aléatoire fonctionnelle de carré intégrable définie sur un intervalle I i.e $\int_I E\left(X^2\right) < \infty$. Notons également μ sa moyenne (Il s'agit donc d'une fonction définie sur I).

Par analogie avec les statistiques multivariées, la représentation de X en terme de composante principale peut être construite à partir de l'opérateur de covariance définie sur $I \times I$ par :

$$K\left(u,v\right) = E\left[\left(X(u) - \mu(u)\right)\left(\left(X(v) - \mu(v)\right)\right)\right] \tag{2.1}$$

Supposons à présent que nous disposons d'un ensemble $\{X_1, ..., X_n\}$ de variables aléatoires fonctionnelles indépendantes identiquement distribuées selon la loi de X.
L'approximation empirique de l'opérateur de covariance $K(u,v)$ défini dans (2.1) est donnée par :

$$\widehat{K_n}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i(u) - \overline{X}(u)\right)\left(X_i(v) - \overline{X}(v)\right) \tag{2.2}$$

où $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Ainsi, il est possible de définir l'application empirique $\widehat{K}$ définie de $L^2\left(I\right)$ dans lui-même par :

$$\left(\widehat{K}_n\varphi\right)(u) = \int_I \widehat{K}_n(u,v)\varphi(v)\,dv \tag{2.3}$$

Comme l'opérateur $\widehat{K}_n$ est un opérateur linéaire auto-adjoint, compact et semi-défini positif, le théorème de Mercer assure l'existence d'une base orthonormée $\left(\widehat{\xi}_j\right)_{j\in\mathbb{N}}$ et d'une suite croissante de réels positifs $\left(\widehat{\lambda}_j\right)_{j\in\mathbb{N}}$ tel que :

$$\left(\widehat{K}_n\widehat{\xi}_j\right)(u) = \widehat{\lambda}_j\widehat{\xi}_j(u),\ u \in I \text{ and } j \in \mathbb{N} \tag{2.4}$$

Beaucoup de résultats sur la convergence de la suite des opérateurs aléatoires $\left(\widehat{K}_n\right)_{n\in\mathbb{N}}$ vers l'opérateur de covariance K ainsi que celle de la suite $\left(\widehat{\lambda}_n\right)_{n\in\mathbb{N}}$ and $\left(\widehat{\xi}_n\right)_{n\in\mathbb{N}}$ vers les valeurs propres et les fonctions propres de K peuvent être consultés dans [Dauxois et al., 1982].

Théorème de Karhunen-Loève [Karhunen, 1947, Loève, 1945]
Avec les notations précédentes, on montre que la variable aléatoire X peut se décomposer dans la base de fonctions propres (ξ) de la façon suivantes :

$$X \overset{L^2}{=} \mu + \sum_{j \geq 1} \theta_j \times \xi_j \tag{2.5}$$

avec $\theta_j = \int (X_u - \mu_u) \times \xi_j(u)\, du$ des variables aléatoires ,appelées *scores*, centrées et décorrélées i.e pour tout $j, k \in \mathbb{N}$: $E(\theta_j \theta_k) = \sqrt{\lambda_j \lambda_k}\ \mathbb{1}_{j=k}$.

Ainsi, chaque donnée fonctionnelle individuelle est identifiée de façon unique à l'aide de ses scores qui serviront à leur tour à traduire les modèles initialement formulés en termes de données fonctionnelles en modèles statistiques multivariés classiques.

2.5 Deux Exemples Importants

2.5.1 Analyse en composantes principales fonctionnelles dans un espace engendré par une base [Ramsay and Silverman, 2005]

Notons $\{X_1, ..., X_n\}$ n variables aléatoires fonctionnelles *centrées* après soustraction de leur moyenne commune.
Notons également $(B_k)_{1\leq k\leq K}$ une base choisie pour approximer les trajectoires $\{X_1, ..., X_n\}$:

$$X = \Gamma B \tag{2.6}$$

où $X = (X_1, ..., X_n)$, $B = (B_1, ..., B_K)^T$ et Γ une matrice $n \times K$ ayant pour ligne les coefficients de chaque fonction dans la base $(B_k)_{1\leq k\leq K}$.
Alors, l'opérateur de covariance empirique $\widehat{K}_n$ défini dans l'équation (2.2) devient :

$$\widehat{K}_n(u, v) = \frac{1}{n} B(u)^T \Gamma^T \Gamma B(v) \tag{2.7}$$

Notons à présent $\widehat{\xi}$ la fonction propre associée à la valeur propre $\widehat{\lambda}$ de l'opérateur $\widehat{K}_n$:

$$\left(\widehat{K}_n \widehat{\xi}\right)(u) = \widehat{\lambda} \times \widehat{\xi} u) \tag{2.8}$$

En écrivant $\widehat{\xi} = \rho^T B$ la décomposition de $\widehat{\xi}$ dans la base $B = (B_1, ..., B_K)^T$, le membre de gauche de l'équation (2.8) devient :

$$\begin{aligned}\left(\widehat{K}_n \widehat{\xi}\right)(u) &= \int_I \widehat{K}_n(u, v)\widehat{\xi}(v)\, dv \\ &= \int_I \frac{1}{n} B(u)^T \Gamma^T \Gamma B(v) B(v)^T \rho\, dv \\ &= B(u)^T \frac{1}{n} \Gamma^T \Gamma \Omega \rho\end{aligned}$$

où Ω est la matrice symétrique $K \times K$ de Gram de la base B i.e $\Omega_{ij} = \int_I B_i(v) B_j(v)\, dv$ for $1 \leq i, j \leq K$.
Alors, l'équation spectrale (2.8) devient :

$$B(u)^T \frac{1}{n} \Gamma^T \Gamma \Omega \rho = \widehat{\lambda} B(u)^T \rho \text{ pour tout} u \in I \tag{2.9}$$

Cette égalité étant vraie pour tout $B(u)$, $u \in I$, il en découle l'égalité matricielle

suivante de vecteur inconnu ρ :

$$\frac{1}{n}\Gamma^T\Gamma\Omega\rho = \widehat{\lambda}\rho \tag{2.10}$$

Pour déterminer ρ, nous effectuons le changement de variables $\kappa = \Omega^{1/2}\rho$ et nous résolvons le problème spectral matriciel suivant de vecteur inconnu κ :

$$\frac{1}{n}\Omega^{1/2}\Gamma^T\Gamma\Omega^{1/2}\kappa = \widehat{\lambda}\kappa \tag{2.11}$$

Ainsi, pour chaque vecteur propre κ_k dans l'équation (2.11), il est possible de définir la fonction propre $\widehat{\xi}_k = \rho_k^T B = \Omega^{1/2}\kappa_k B$ pour tout $1 \leq k \leq K$. Cette construction assure ainsi l'orthonormalité de la base propre $\left(\widehat{\xi}_k\right)_{1\leq k\leq K}$.

Notons enfin qu'avec cette méthode, il n'est possible d'estimer qu'un nombre de fonctions propres égal au nombre d'éléments de la base B à savoir K fonctions propres.

2.5.2 Analyse en composantes principales fonctionnelles d'un ensemble de densités de probabilité

Notons $\{f_1, ..., f_n\}$ une famille de densités de probabilité indépendantes définies sur l'intervalle I.

L'objectif de cette section est de présenter, dans le cadre spécifique des densités de probabilité, une méthode d'estimation des paramètres (λ_k),(θ_k) des fonctions propres (ξ_k) dans la décomposition de Karhunen-Loève :

$$f_i(\delta) = \overline{f}\,(\delta) + \sum_{k=1}^{n} \theta_{ik} \times \xi_k\,(\delta) \tag{2.12}$$

où $\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$.

Nous allons tout d'abord établir une correspondance importante entre les éléments spectraux (valeurs propres et fonctions) de l'opérateur de covariance empirique défini dans l'équation (2.3) avec les éléments spectraux (valeurs propres et vecteurs propres) de la matrice de Gram des densités centrées $(f_i)_{1\leq i\leq n}$.

Correspondance avec le spectre de la matrice de Gram des observations fonctionnelles centrées

Soit $\mathbf{M}$ la matrice de Gram des observations fonctionnelles centrées dont les éléments sont définis par :

$$\mathbf{M}_{ij} = \left\langle f_i - \overline{f} | f_j - \overline{f} \right\rangle \tag{2.13}$$

avec $\langle a|b\rangle = \int_I a(u)b(u)\,du$ le produit scalaire fonctionnel usuel dans $L^2\,(I)$.

Rappelons tout d'abord que l'opérateur de covariance empirique de la famille de densités $\{f_1, ..., f_n\}$ défini pour toute fonction φ de carré intégrale sur I par :

$$\left(\widehat{K}_n\varphi\right)(u) = \int_I \widehat{K}_n(u,v)\varphi(v)\,dv \tag{2.14}$$

avec $\widehat{K_n}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\left(f_i(u) - \overline{f}(u)\right)\left(f_i(v) - \overline{f}(v)\right)$.

Pour alléger le texte, nous adopterons l'abus de notation consistant à désigner par f_i pour tout $1 \leq i \leq n$ la densité centrée $f_i - \overline{f}$.

Alors l'équation l'opérateur empirique de covariance dans l'équation (2.14) peut s'écrire :

$$\widehat{K_n}\varphi = \frac{1}{n}\sum_{i=1}^{n} \langle \varphi | f_i \rangle \times f_i \tag{2.15}$$

Dans le cas où ξ est une fonction propre de l'opérateur $\widehat{K_n}$ associée à la valeur propre λ, ceci peut se traduire par :

$$\lambda \times \xi = \frac{1}{n}\sum_{i=1}^{n} \langle \xi | f_i \rangle \times f_i \tag{2.16}$$

Nous pouvons facilement remarquer que cette dernière équation implique que toute fonction propre de l'opérateur de covariance empirique appartient au sous-espace fonctionnel engendré par les $\{f_1, ..., f_n\}$. En d'autres termes, la fonction propre ξ peut s'écrire comme une combinaison linéaire de ces densités :

$$\xi = \sum_{j=1}^{n} \alpha_j \ f_j \tag{2.17}$$

En prenant le produit scalaire des deux membres de l'équation (2.16) avec $f_k,\ 1 \leq k \leq n$ et en la combinant avec l'équation (2.17), nous en déduisons que la détermination des éléments spectraux de l'opérateur empirique $\widehat{K_n}$ revient à déterminer les réels $(\alpha_i)_{1 \leq i \leq n}$ tels que pour tout $1 \leq k \leq n$:

$$\begin{aligned} \lambda \times \sum_{j=1}^{n} \underbrace{\langle f_k | f_j \rangle}_{\mathbf{M_{kj}}} \ \alpha_j &= \frac{1}{n}\sum_{i=1}^{n} \langle f_k | f_i \rangle \times \left\langle \sum_{j=1}^{n} \alpha_j \ f_j | f_i \right\rangle \\ &= \frac{1}{n}\sum_{j=1}^{n} \alpha_j \ \underbrace{\sum_{i=1}^{n} \langle f_k | f_i \rangle \langle f_i | f_j \rangle}_{\mathbf{M^2_{kj}}} \end{aligned} \tag{2.18}$$

L'équation (2.18) peut alors se réécrire sous forme matricielle simple :

$$n\lambda \ \mathbf{M}\alpha = \mathbf{M^2}\alpha \tag{2.19}$$

avec $\mathbf{M}$ la matrice définie par l'équation (2.13) et $\alpha = (\alpha_1, ..., \alpha_n)^T$.
Comme la matrice $\mathbf{M}$ est symétrique et semi définie positive, il existe une base de vecteurs

propres telle que :

$$n\lambda\ \alpha = \mathbf{M}\alpha \tag{2.20}$$

Nous fournissant ainsi toutes les solutions de α l'équation (2.19).

Si on note $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ les valeurs propres de la matrice $\mathbf{M}$ et $\alpha^1, ..., \alpha^n$ les vecteurs propres orthonormés correspondants, il reste alors à déterminer les constantes $(c_k)_{1\leq k\leq n}$ telles que les fonctions propres $\xi^k = \sum_{j=1}^n c_k \alpha_j^k \times f_j$ forment une base fonctionnelle orthonormée. Pour tout $1 \leq k \leq n$:

$$\begin{aligned}
1 &= \left\langle \xi^k | \xi^k \right\rangle \\
&= c_k^2 \sum_{i,j=1}^{n} \alpha_i^k \alpha_j^k \left\langle f_i | f_j \right\rangle \\
&= c_k^2 \sum_{i,j=1}^{n} \alpha_i^k \alpha_j^k \mathbf{M}_{ij} \\
&= c_k^2 \left\langle \alpha^k | \mathbf{M}\alpha^k \right\rangle \\
&= c_k^2 \lambda_k \left\langle \alpha^k | \alpha^k \right\rangle \\
&= c_k^2 \lambda_k
\end{aligned} \tag{2.21}$$

Par conséquent, les fonctions propres $(\xi_k)_{1\leq k\leq n}$ s'écrivent :

$$\xi^k = \sum_{j=1}^{n} \frac{\alpha_j^k}{\sqrt{\lambda_k}} \times f_j \tag{2.22}$$

et les scores $(\theta_{ik})_{ik}$ dans l'équation (2.12) s'obtiennent par :

$$\begin{aligned}
\theta_{ik} &= \left\langle f_i, \xi^k \right\rangle \\
&= \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{n} \langle f_i, f_j \rangle \alpha_j^k \\
&= \frac{1}{\sqrt{\lambda_k}} \lambda_k \alpha_i^k \\
&= \sqrt{\lambda_k} \times \alpha_i^k
\end{aligned} \tag{2.23}$$

Formulation en terme d'estimateurs à noyaux de densités

Nous allons donner, dans le cas des estimateurs à noyaux de densités, l'expression d'un estimateur de la matrice de Gramm $\mathbf{M}$ définie dans l'équation (2.13). Nous avons utilisé l'estimateur proposé dans [Kneip and Utikal, 2001] et construit à partir de l'estimateur naturel $\widetilde{M}_{ts} = \left\langle \widehat{f}_{t,h} - \widehat{\overline{f}} | \widehat{f}_{s,h} - \widehat{\overline{f}} \right\rangle$ avec $\widehat{f}_{t,h}(x) = \frac{1}{n_t} \sum_{i=1}^{n_t} K\left(\frac{x-x_{it}}{k}\right)$ l'estimateur à noyau des densités $(f_t)_{1 \leq t \leq n}$ avec comme paramètres de lissage h par exemple la médiane des paramètres de lissage optimaux $(h_t)_t$ de chaque densité comme suggéré dans [Kneip and Utikal, 2001].

La procédure est la suivante :

- Déterminer la matrice $\widehat{M}^{[1]}$ définie par :

$$\begin{cases} \widehat{M}_{ts}^{[1]} = \dfrac{1}{n_t n_s h^2} \displaystyle\sum_i^{n_t} \sum_j^{n_s} \int K\left(\frac{x-x_{it}}{h}\right) K\left(\frac{x-x_{js}}{h}\right) dx & \text{si } t \neq s \\ \widehat{M}_{ts}^{[1]} = \dfrac{1}{n_t^2 h^2} \displaystyle\sum_i^{n_t} \sum_{j \neq i}^{n_t} \int K\left(\frac{x-x_{it}}{h}\right) K\left(\frac{x-x_{js}}{h}\right) dx & \text{sinon} \end{cases}$$

- L'estimateur $\widehat{M}$ est défini par :

$$\widehat{M}_{ts} = \widehat{M}_{ts}^{[1]} - \widehat{M}_{ts}^{[2]} + \widehat{M}_{ts}^{[3]} \tag{2.24}$$

$$\text{avec : } \widehat{M}_{ts}^{[2]} = \frac{1}{n} \sum_l^n \widehat{M}_{tl}^{[1]} + \widehat{M}_{ls}^{[1]} \text{ et } \widehat{M}_{ts}^{[3]} = \frac{1}{n^2} \sum_l^n \sum_k^n \widehat{M}_{lk}^{[1]}$$

L'équation(2.24) résultant du passage de la matrice à sa version centrée correspondante.

Ainsi, les valeurs propres (λ_k) et vecteurs propres $\left(\alpha^k\right)$ dans les équations (2.23) et (2.22) peuvent être estimés par les valeurs propres $\left(\widehat{\lambda}_k\right)$ et les vecteurs propres $\left(\widehat{\alpha}^k\right)$ de la matrice $\widehat{\mathbf{M}}$.

2.6 Modèles de régression sur données fonctionnelles à réponses réelles

Dans cette section, nous étendons aux données fonctionnelles quelques modèles classiques de la statistique multivariée.
Nous expliquerons également comment l'analyse en composantes principales fonctionnelles sert de lien entre les modèles de régression fonctionnelle et leurs homologues multivariés.

De manière générale, une approche fonctionnelle du problème de régression consiste à modéliser le lien entre la réponse Y (qui peut être une variable binaire indiquant l'apparition ou non d'une pathologie radio-induite, continue comme par exemple le temps de survie) et une variable explicative réelle ou fonctionnelle (par exemple la densité de distribution de la dose de radiation dans un organe).

Nous commençons par la généralisation du modèle linéaire aux données fonctionnelles avant de passer aux autres modèles linéaires généralisés ou de survie.

2.6.1 Le modèle de régression linéaire fonctionnel

Le modèle de régression linéaire fonctionnel consiste à supposer que la variable réponse Y est reliée à la variable explicative fonctionnelle X par un opérateur linéaire L :

$$E\left(Y|X\right) = L\left(X\right) \tag{2.25}$$

Le théorème de représentation des opérateurs linéaires continus permet de ramener le problème de l'estimation de L à celui d'une fonction β de carré intégrable :

$$\begin{aligned} E\left(Y|X\right) &= \alpha + \langle \beta | X \rangle \\ &= \alpha + \int \beta(t) X(t)\ dt \end{aligned} \tag{2.26}$$

Le problème revient donc à estimer la fonction inconnue β à partir des données. Pour ce faire, l'analyse en composante principale apporte une réponse aussi simple que puissante.
Considérons en effet un ensemble de variables réponses $(Y_i)_{1\leq i\leq n}$ et de variables explicatives fonctionnelles $(X_i)_{1\leq i\leq n}$.

La section 2.4 a montré que l'analyse en composantes principales fonctionnelles permettait d'obtenir une base orthogonale $(\xi_k)_{1\leq k\leq n}$ dans laquelle il est possible de décomposer la fonction inconnue β :

$$\beta = \sum_{k=1}^{n} \beta_k \times \xi_k \tag{2.27}$$

Par conséquent, le modèle linéaire ci-dessus peut s'écrire pour tout $1 \leq i \leq n$:

$$\begin{aligned} E\left(Y_i|X\right) &= \alpha + \int \beta(t) \times X_i(t)\ dt \\ &= \underbrace{\alpha + \int \beta(t) \times \mu(t)\ dt}_{\widetilde{\alpha}} + \int \beta(t) \times \left(X_i(t) - \mu(t)\right)\ dt \\ &= \widetilde{\alpha} + \int \left(\sum_{k=1}^{n} \beta_k \times \xi_k(t)\right) \times \left(\sum_{l=1}^{n} \theta_{il} \times \xi_l(t)\right)\ dt \\ &= \widetilde{\alpha} + \sum_{k=1}^{n}\sum_{l=1}^{n} \beta_k \times \theta_{il} \int \xi_k(t)\xi_l(t)\ dt \\ &= \widetilde{\alpha} + \sum_{k=1}^{n} \beta_k \times \theta_{ik} \end{aligned} \tag{2.28}$$

La dernière égalité s'obtenant par orthogonalité des fonctions propres (ξ).

Ce résultat est d'une importance capitale en régression sur données fonctionnelles : L'analyse en composantes principales fonctionnelles est un puissant outil pour non seulement réduire la dimension du problème en restreignant le développement précédant à une somme sur les K ($<$ n) premières fonctions propres mais aussi pour ramener le problème de régression sur données fonctionnelles à un modèle multivarié classique portant sur les scores (θ_{ik}) et ayant pour objet l'estimation des coefficients (β_k) d'un développement dans la base des vecteurs propres (ainsi que l'intercept $\widetilde{\alpha}$).

2.6.2 Modèles linéaires généralisés et Modèles de survie fonctionnels

L'idée du modèle linéaire pour données fonctionnelles peut être étendue avec les mêmes arguments aux modèles linéaires généralisés et aux modèles à risques proportionnels.

Modèle de Poisson pour données fonctionnelles

$$\begin{aligned}\log(\lambda_i) &= \alpha + \int \beta(t) \times X_i(t)\, dt \\ &= \widetilde{\alpha} + \sum_{k=1}^{n} \beta_k \times \theta_{ik}\end{aligned} \tag{2.29}$$

Modèle logistique pour données fonctionnelles

$$\begin{aligned}\text{logit}(p_i) &= \alpha + \int \beta(t) \times X_i(t)\, dt \\ &= \widetilde{\alpha} + \sum_{k=1}^{n} \beta_k \times \theta_{ik}\end{aligned} \tag{2.30}$$

Modèle de Cox pour données fonctionnelles

$$\begin{aligned}\lambda(t|X) &= \lambda_0(t) \times \exp\left(\int \beta(t) \times X_i(t)\, dt\right) \\ &= \widetilde{\lambda}_0(t) \times \exp\left(\sum_{k=1}^{n} \beta_k \times \theta_{ik}\right)\end{aligned} \tag{2.31}$$

Chapitre 3

Modélisation de complications tissus sains après radiothérapie externe

Sommaire

3.1 Introduction 46

3.2 Données de patients et de toxicités 48

3.3 Planification de traitement et HDVs 49

3.4 Analyse statistique 50

3.4.1 L'approche analyse de données fonctionnelles 50

3.4.2 Analyse en composantes principales fonctionnelles 51

3.4.3 Formulation du modèle et interprétation 51

3.4.4 Autres modèles NTCP et comparaison de modèles 53

3.5 Résultats 55

3.5.1 Incidence de la toxicité rectale 55

3.5.2 Comparaisons des distributions de doses entre les patients avec et sans toxicités 56

3.5.3 Résultats de modélisation 56

3.6 Discussion 63

3.1 Introduction

Le risque de complications des tissus sains a toujours été une préoccupation dans l'optimisation d'un plan de traitement en radiothérapie et ce, par l'intermédiaire de la probabilité de complication de tissus sains (NTCP). La variabilité des histogrammes dose-volumes (HDVs) des tissus sains étant plus importante que celles des volumes tumoraux, de nombreux modèles NTCP ont été proposés afin d'étudier l'impact de l'hétérogénéité de cette distribution de dose.

Le modèle NTCP le plus étudié est sans doute le modèle de Lyman-Kutcher-Burman [Lyman, 1985, Burman et al., 1991, Kutcher et al., 1991] basé sur la réduction d'une distribution de dose non uniforme à une dose uniforme équivalente (EUD) dans laquelle l'effet dose-volume est exprimé sous forme d'une puissance [Niemierko, 1997].

Une des limitations majeures de ce modèle est l'absence de prise en compte des paramètres cliniques non dosimétriques dans sa formulation. Au prix d'une perte de puissance statistique, certains auteurs ont proposé des méthodes pour remédier à ce problème basées soit sur la stratification [Fiorino et al., 2008] soit sur l'ajout de facteurs de correction supplémentaires [Peeters et al., 2006, Defraene et al., 2012].

Les modèles multi-métriques sont employés afin de faciliter la prise en compte de variables cliniques non dosimétriques, par exemple via une régression logistique. Dans leur expression la plus simple, ces modèles n'incluent qu'un nombre restreint d'indicateurs de distribution de dose à l'organe comme par exemple les V_{xGy} i.e volumes exposés à plus d'une certaine dose x Gy, ce qui potentiellement peut amener à perdre des informations sur la dynamique d'irradiation susceptible de discriminer des plans de traitements selon leurs risques d'engendrer des complications.

Des modèles plus sophistiqués proposent d'intégrer dans les modèles multi-métriques l'ensemble de l'information relative à la distribution de la dose à l'organe via une analyse en composantes principales (ACP) sur les vecteurs définissant les histogrammes dose-volume cumulés [Dawson et al., 2005, Söhn et al., 2007]. Malgré leurs propriétés intéressantes, l'interprétation des résultats de ces ACP n'est pas toujours évidente et l'information supplémentaire qu'elles fournissent en comparaison aux indicateurs dosimétriques standards n'est pas claire [Vesprini et al., 2011].

.

L'approche adoptée ici est issue du domaine de *l'analyse de données fonctionnelles* présentée dans le chapitre 2.
L'idée est de considérer, pour chaque patient, les doses calculées dans chaque voxel de la structure d'intérêt par les logiciels de planification de traitements (TPS) comme un échantillon de réalisations d'une variable aléatoire d'une certaine loi de probabilité définie par sa fonction densité de probabilité (*pdf*). Bien que les histogrammes soient des outils intéressants pour donner une idée de la distribution de dose à l'organe (voir section A.3.1), il existe d'autres estimateurs des *pdf* avec des propriétés mathématiques supérieures tant en terme de régularité qu'en terme de propriétés asymptotiques de vitesse de convergence. Parmi les plus populaires d'entre eux se trouvent les estimateurs à noyaux présentés en détails dans la section A.3.2 de ce mémoire.

Par conséquent, toutes les données dosimétriques pour chaque patient sont résumées en un unique objet fonctionnel : la *pdf*. Dans ce contexte, étudier l'effet dose-volume d'un tissu sain après irradiation revient à décrire la corrélation entre un prédicteur fonctionnel (la *pdf*) et une réponse binaire qui est la toxicité (oui/non). Étant donné que les objets en présence sont des fonctions, il est dès lors naturel de privilégier l'usage d'outils statistiques spécialement conçus pour généraliser aux espaces fonctionnels les différentes méthodes de la statistique classique multivariée telles que la régression, l'analyse en composantes principales, etc. comme présenté dans le chapitre 2.4.

Dans cette étude, nous allons appliquer les outils de l'analyse de données fonctionnelles à 141 HDVs du rectum après irradiation par 4 faisceaux en boîte de la loge prostatique. La corrélation entre la distribution de dose au rectum et le risque de toxicité rectale de grade ≥ 2 est étudiée via un modèle linéaire généralisé pour données fonctionnelles. Les résultats obtenus seront comparés à ceux obtenus par trois autres modèles NTCP : le modèle LKB, le modèle multi-métrique logistique et le modèle d'analyse en composantes principales multivarié sur les HDVs cumulés.

3.2 Données de patients et de toxicités

Entre septembre 2005 et décembre 2010, 141 patients de l'institut Gustave Roussy (IGR) ont bénéficié d'une radiothérapie externe de la loge prostatique pour une augmentation de PSA après prostatectomie. Le traitement a consisté en une irradiation 3D conformationnelle à 4 faisceaux en boîte avec 18 ou 20 MV (Voir Figure 4.1). Une dose totale égale à 65 ou 66 Gy a été prescrite avec une dose par fraction égale à 2.5 Gy et 2 Gy respectivement.
Après la radiothérapie de rattrapage, les patients étaient suivis, tous les trois à six mois les cinq premières années puis tous les ans au delà, par un urologue et un oncologue radiothérapeute avec un examen clinique et un dosage de PSA pour suivre l'évolution éventuelle de la maladie et les effets secondaires qui en découlent. La présente étude a été menée sur les données enregistrées jusqu'au 31 décembre 2013.

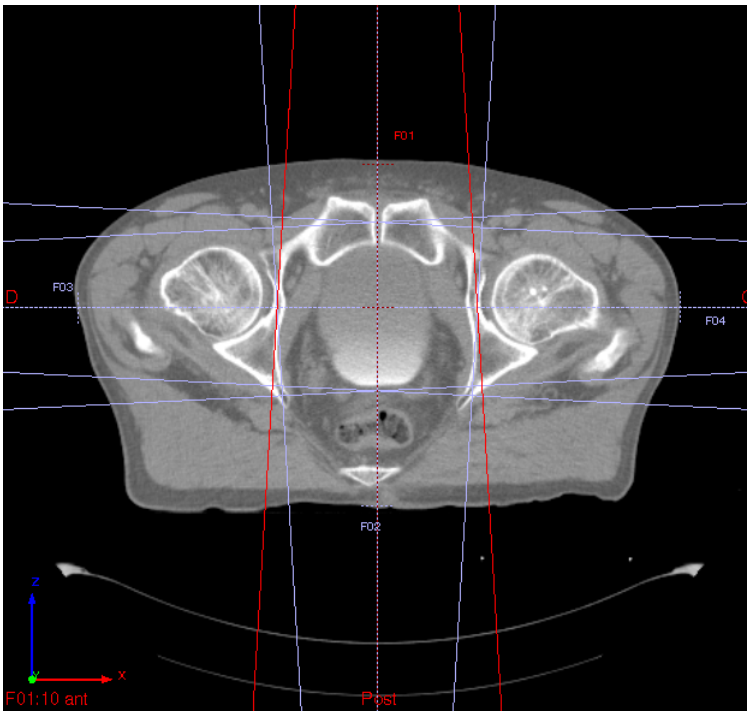

FIGURE 3.1 – Coupe transverse au niveau de la loge prostatique représentant la balistique de traitement. Les 4 faisceaux sont centrés sur le barycentre de la loge prostatique (Isocentre).

Les toxicités rectales ont été recueillies rétrospectivement à partir des bases de données de l'IGR. Les variables cliniques enregistrées furent : l'âge à l'irradiation, le diabète, les maladies cardiaques, les antécédents chirurgicaux, l'hormonothérapie, la prise d'anticoa-

gulants ou d'antiaggrégants et hypertension. Les toxicités ont été gradées selon l'échelle RTOG : Le Grade 2 concerne les cas de saignements intermittents, le Grade 3 celui des saignements nécessitant la prise de médicaments, le recours à une coagulation par laser ou à une transfusion sanguine et le Grade 4 pour perforation ou nécrose.

3.3 Planification de traitement et HDVs

Tous les patients ont été scannés en décubitus dorsal avec une épaisseur de coupe de 0.3 cm. Le rectum a été contouré de l'anus à la jonction rectosigmoïdienne et considéré comme un organe solide. Chaque patient a été traité en décubitus dorsal avec 4 faisceaux en boîte : Antérieur (0^o), Postérieur (180^o), Latéral droite (270^o) et Latéral gauche (90^o) et protégé par des caches personnalisés et/ou collimateur multi-lames en utilisant la fonction Beam Eye View (BEV) du TPS. Le bon positionnement du patient et de l'isocentre ont été vérifiés grâce à la réalisation d'une imagerie de contrôle au cours des 2 à 3 premières séances de la première semaine, de façon hebdomadaire par la suite, et à chaque modification du traitement. Ce mode de contrôle de repositionnement fait partie de la radiothérapie guidée par l'image (IGRT). Ces contrôles s'accompagnent du repérage laser et du renouvellement des marques de repérage.
Les contraintes de dose suivantes ont été utilisées pour valider le plan de traitement tout en veillant à la bonne couverture du volume cible (95% de la dose prescrite dans 100% du volume cible / homogénéité de dose dans la tolérance $(-5\%, +7\%)$ par rapport à la dose prescrite) :

- **Paroi rectale** : La dose de 60 Gy ne doit pas être délivrée dans plus de 50% du volume rectal i.e $V_{60Gy} \leq 50\%$ et la dose maximale de 60 Gy au niveau de la totalité de la circonférence rectale.
- **Paroi vésicale** : La dose de 60 Gy ne doit pas être délivrée dans plus de 50% du volume vésical i.e $V_{60Gy} \leq 50\%$
- **Têtes fémorales et grands trochanters** : La dose de 50 Gy ne doit pas être délivrée dans plus de 10% d'un volume osseux contouré par convention du sommet des têtes fémorales au petit trochanter exclu : $V_{50Gy} \leq 10\%$.

Enfin, les HDVs différentiels de la somme des doses reçues pendant les séances étaient disponibles avec un pas de 0.33 Gy. Ces HDVs physiques ont été convertis en terme de dose

biologique effective (BED) pour une dose par fraction de 2 Gy à l'aide du modèle linéaire quadratique [Thames and Hendry, 1987, Fowler, 1992] avec le rapport $\alpha/\beta = 3Gy$.

3.4 Analyse statistique

3.4.1 L'approche analyse de données fonctionnelles

L'analyse de HDVs amène naturellement à considérer des données de nature fonctionnelles. La Figure 3.2 illustre l'approche des statistiques fonctionnelles en comparaison avec l'approche statistique multivariée classique :

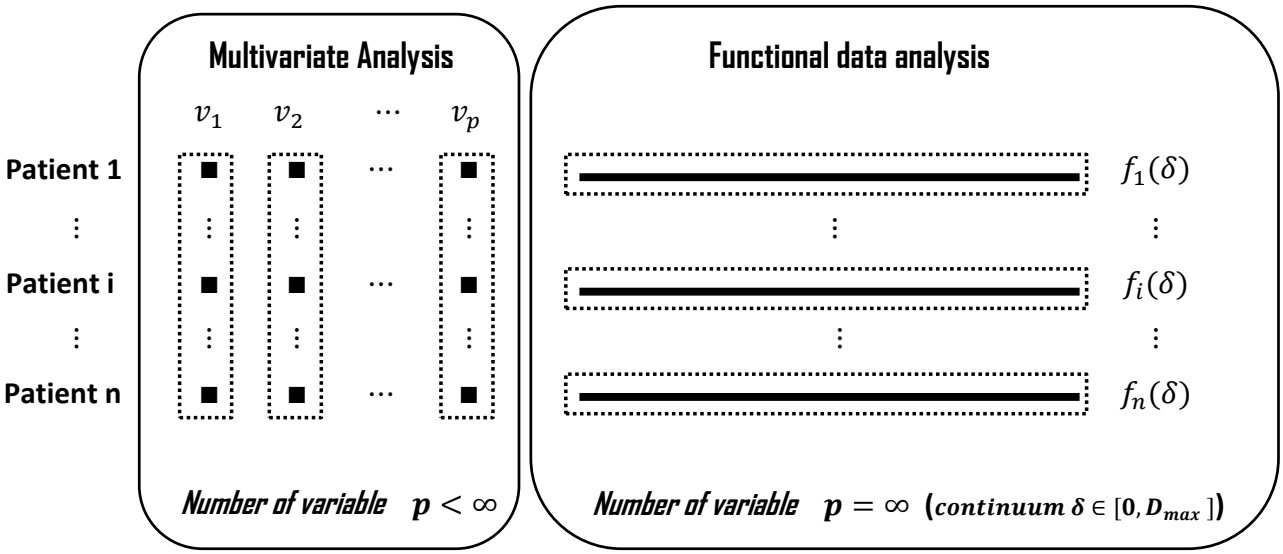

FIGURE 3.2 – L'approche statistique fonctionnelle (à droite) comparée à l'approche statistique multivariée matricielle (à gauche).

A gauche, nous avons l'approche multivariée classique et sa vision *verticale* : le $i^{ème}$ patient fournit les $i^{èmes}$ observations ou réalisations $\nu_{i1},...,\nu_{ip}$ des variables aléatoires $\nu_1,...,\nu_p$ (pourcentage du volume rectal exposé aux doses $d_1,...,d_p$ de la grille de calcul de dose du TPS).

Si, à présent, comme les TPS modernes permettent de le faire, nous pouvons évaluer la distribution de la dose sur une grille de dose $d_1,...,d_p$ de plus en plus fine, cela reviendrait à définir un intervalle continu de doses. Comme représenté dans la Figure 3.2, l'approche *horizontale* de la statistique fonctionnelle associe à chaque patient une fonction de densité de probabilité désignée par $f_i(\delta)$ définie en tout point δ du continuum de doses.

3.4.2 Analyse en composantes principales fonctionnelles

Une fois avoir estimé les fonctions de densité de probabilité de chaque patient par la méthode des noyaux présentée dans la section A.3.2 , nous avons mené une analyse spectrale afin de dégager leurs principaux modes de variabilité en appliquant la méthode présentée en section 2.5.2.

Il s'agit de représenter chaque *pdf* $f_i(\delta)$ en terme de décomposition de Karhunen-Loève :

$$f_i(\delta) = \overline{f}(\delta) + \sum_{k=1}^{n} \theta_{ik} \times g_k(\delta) \tag{3.1}$$

où $\overline{f}(\delta) = \frac{1}{N}\sum_{i=1}^{N} f_i(\delta)$ est la moyenne commune, $(g_k)_{k\geq 1}$ les fonctions propres, une base orthonormée dans le sens L^2 qui détermine de façon optimale dans le sens des moindres carrés les modes de variabilités les plus importants relativement à la moyenne commune. Les scores $(\theta_{ik})_{k\geq 1}$ décrivent les coordonnées de la $i^{ème}$ fonction f_i dans la base $(g_k)_{k\geq 1}$. Les variances des scores appelées *valeurs propres* notées $\lambda_1 \geq \lambda_2 \geq \lambda_3 ...$ sont donc la contribution du $k^{ième}$ terme de (3.1) dans la description de la variance totale.

3.4.3 Formulation du modèle et interprétation

Le modèle NTCP que nous proposons est basé sur une généralisation aux données fonctionnelles de la régression logistique multivariée comme présenté dans la section 2.6.2. L'objectif est d'expliquer une variable réponse binaire (toxicité oui/non) par une covariable fonctionnelle i.e la fonction de densité de probabilité représentant la distribution de la dose de radiations dans le volume rectal.

Le modèle s'écrit de la façon suivante :

$$logit(p_i) = \alpha_0 + \sum_{j\geq 1} \alpha_j \times \text{covariables\_cliniques}_{ij} + \int_0^{D_{max}} \beta(\delta) \times f_i(\delta)\ d\delta \tag{3.2}$$

où $logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$ est la transformation *logit* de la probabilité p_i de survenue d'une toxicité rectale, α_0 (l'intercept) et $(\alpha_j)_{j\geq 1}$ des paramètres de régression réels et $\beta(\delta)$ le paramètre fonctionnel.

Pour estimer le paramètre $\beta(\delta)$, nous transformons à l'aide du résultat de l'analyse en composantes principales fonctionnelles le modèle décrit dans l'équation (3.2) en un

modèle de régression logistique classique basé sur les scores :

$$logit\left(p_i\right) = \alpha_0 + \sum_{j \geq 1} \alpha_j \times \text{covariables\_cliniques}_{ij} + \sum_{k=1}^{K} \beta_k \times \theta_{ik} \tag{3.3}$$

Le choix du nombre K de scores à intégrer dans le modèle (3.3) est en général fixé à partir de la part de variabilité expliquée cumulée. Dans la présente étude, nous avons fixé ce nombre à 10 (plus de 80% de variabilité expliquée).

A présent que nous sommes en présence d'un modèle de régression logistique classique, nous pouvons utiliser toutes les méthodes de sélection de modèles disponibles afin de construire le modèle le plus parcimonieux à partir du sous-ensemble I_{FPCS} de scores.

Enfin, les coefficients estimés $(\beta_k)_{k \in I_{FPCS}}$ permettent d'estimer la fonction $\beta\left(\delta\right)$ par la formule :

$$\hat{\beta}\left(\delta\right) = \sum_{k \in I_{FPCS}} \beta_k \times g_k\left(\delta\right) \tag{3.4}$$

Ce modèle NTCP bien que construit sur des covariables fonctionnelles, permet cependant de quantifier les risques exactement comme un modèle logistique multivarié en fournissant notamment un calcul d'*odds ratio* :

$$\text{OR}_i^{\text{Radiation}} = \exp\left(\int_0^{D_{max}} \beta\left(\delta\right) \times f_i\left(\delta\right)\ d\delta\right) \tag{3.5}$$

3.4.4 Autres modèles NTCP et comparaison de modèles

Modèle NTCP de Lyman-Kutcher-Burman (LKB)

Le modèle LKB a été proposé pour la première fois par Lyman [Lyman, 1985] dans le cas des irradiations uniformes pour modéliser, via une fonction *probit*, la relation entre la dose et la probabilité de survenue d'une complication radio-induite. La prise en compte d'irradiations non uniformes a été possible par l'introduction du concept de dose équivalente uniforme (EUD) [Kutcher et al., 1991, Niemierko, 1997] visant à trouver un équivalent en terme d'irradiation homogène d'un profil d'irradiation initialement hétérogène.

Ce modèle NTCP utilisant ce schéma de réduction de HDVs s'écrit :

$$NTCP = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-x^2/2\right)\ dx \tag{3.6}$$

où $t = \frac{EUD - TD_{50}}{m \times TD_{50}}$, $EUD = \left(\sum_{i=1}^{p} \nu_i \times d_i^{\frac{1}{n}}\right)^n$ et $(d_i, \nu_i)_{1 \leq i \leq N}$ sont N couples formés par le niveau de dose et le pourcentage de volume de l'organe exposé à ce niveau de dose de la maille de calcul du TPS.

Le paramètre TD_{50} peut s'interpréter comme étant la dose uniforme réalisant un risque de complication de 50%. Le paramètre m est la mesure de la pente de la sigmoïde et enfin le paramètre n, compris entre 0 et 1, est le paramètre mesurant la nature de l'*effet-volume* : n proche de 1 signifie que l'organe est de nature parallèle (l'organe peut voir une partie de son volume endommagé sans conséquence significative sur son fonctionnement comme le foie par exemple). Par opposition, n proche de 0 signifie que l'organe est plutôt en série (l'organe peut connaître de sérieux dysfonctionnements à la suite de l'altération d'une portion limitée de son volume, par exemple le rectum). Remarquons enfin que l'EUD est réduite à la dose moyenne lorsque $n = 1$.

Les valeurs des paramètres TD_{50}, m et n sont obtenues en maximisant la fonction de vraisemblance suivante :

$$\sum_j A_j \times \ln\left[NTCP_j\left(n, m, TD50\right)\right] + \left(1 - A_j\right) \times \ln\left[1 - NTCP_j\left(n, m, TD50\right)\right] \tag{3.7}$$

où $A_j = 1$ en cas de complication survenue chez le$j^{ième}$ patient et $A_j = 0$ sinon.

Modèles multi-métriques simples

Comme il a été évoqué ci-dessus, les modèles multi-métriques simples sont des modèles linéaires généralisés logistiques basés sur les covariables non dosimétriques décrites dans la Section 3.2 et sur les indicateurs dosimétriques standards à savoir les $V_{xGy}, x = 10, 15, 20, ..., 60$ et $65Gy$ i.e pourcentages de volume rectal exposés à des doses $\geq xGy$.

Modèles multi-métriques basés sur une analyse en composantes principales multivariée(ACP)

Les modèles multi-métriques ACP sont des modèles linéaires généralisés logistiques de même nature que les précédents basés sur les mêmes covariables non dosimétriques décrits dans la Section 3.2 à l'exception de l'usage d'indicateurs différents de la distribution de la dose au rectum. En effet, en lieu et place des indicateurs dosimétriques V_{xGy}, ces modèles utilisent les scores issus de l'ACP des variables aléatoires $(V_{d_i\ Gy})_{1\leq i\leq p}$ avec $(d_i)_{1\leq i\leq p}$ les valeurs de doses utilisées dans la maille de calcul de la Section 3.4.1.

Comparaison des Modèles

Pour comparer les modèles présentés ci-dessus, différents tests statistiques ont été menés : le test du rapport de vraisemblance a été utilisé pour tester l'existence d'une différence significative entre deux modèles emboîtés. Le critère d'Akaike (AIC) a été utilisé dans la sélection de modèles non emboîtés. Enfin, l'aire sous la courbe (AUC) de chaque modèle a permis d'évaluer leurs performances en terme de prédiction.

3.5 Résultats

3.5.1 Incidence de la toxicité rectale

Un total de 141 patients traités selon le protocole décrit dans la Section 5.2.2 entre septembre 2005 et décembre 2010 ont été recrutés dans l'étude. Les caractéristiques des patients sont résumées dans la Table 3.1.
L'âge médian au début du traitement est de 65 ans (valeurs comprises entre 51 et 80 ans) et le suivi médian est de 4 ans (valeurs comprises entre 6 mois et 8 ans). Vingt patients (soit 14% du total) sur les 141 ont développé une toxicité rectale de Grade ≥ 2 dont neuf de Grade 3 et aucun de Grade 4.
Parmi les cas de Grade ≥ 2, cinq (25% du total) ont développé une toxicité rectale entre six mois et un an, onze (55% du total) entre un et deux ans et enfin quatre (50% du total) après plus de deux ans.

Caractéristiques	Nb de patients (% du total)	p-value* Grade≥ 2	p-value* Grade 3
Suivi moyen	4 ans (6 mois-8 ans)		
Toxicités rectales de Grade ≥ 2	20/141 (14%)		
Temps de latence Toxicités ≥ 2	Moyenne 17 mois (5-37 mois)		
Toxicités rectales de Grade 3	9/141 (6.4%)		
Temps de latence Grade 3	Moyenne 22 mois (5-37 mois)		
Age	Moyenne 65 ans (51-80 ans)	0.01	0.18
Chirurgie Abdominale	35/141 (25%)	0.28	>0.5
Hypertension	39/141 (28%)	>0.5	> 0.5
Hypercholesterolemie	18/141 (13%)	0.28	> 0.5
Diabète	11/141 (8%)	>0.5	>0.5
Usage anticoagulants/antiaggrégants	14/141 (10%)	0.5	0.5
Hormonothérapie	34/141 (24%)	>0.5	>0.5
Dose par fraction		0.44	0.37
2Gy	68/141 (48%)		
Grade≥ 2	8/68		
Grade 3	3/68		
2.5Gy	73/141 (52%)		
Grade≥ 2	12/73		
Grade 3	6/73		

* p-value de l'analyse univariée d'associations avec une toxicité Grade ≥ 2 ou Grade 3.

TABLE 3.1 – Caractéristiques cliniques des patients et de leurs traitements de radiothérapie externe.

3.5.2 Comparaisons des distributions de doses entre les patients avec et sans toxicités

Le Figure 3.3 donne deux exemples d'estimation de la densité de probabilité de la distribution de dose au rectum à partir des histogrammes dose volumes différentielles.

Dans la même figure se trouvent les estimations des densités de probabilités des patients ayant connu ou pas de toxicités rectales (à gauche) ainsi que leurs moyennes (à droite).
Les patients ayant connu une toxicité rectale de grade ≥ 2 ont en moyenne une distribution de doses plus importantes autour de l'équivalent biologique de la dose prescrite (66Gy et 71.5Gy pour le Grade ≥ 2, et seulement autour de 71.5Gy pour le Grade 3) alors que le gradient de dose cumulée est moins important entre 20 et 40Gy.

3.5.3 Résultats de modélisation

Le modèle LKB

Les estimations des paramètres du modèle de LKB pour les toxicités rectales de Grade ≥ 2 et Grade 3 sont données dans la Table 3.2 : $(TD_{50} = 72.6Gy, m = 0.17, n = 0.12)$ et $(TD_{50} = 73.8Gy, m = 0.16, n = 0.24)$ respectivement. Les intervalles de confiance de ces paramètres ont été construits par la méthode de la vraisemblance partielle [Pawitan, 2001].

Modèle Multimétrique Simple

Les résultats de l'analyse univariée sont résumés dans la Table 3.2. Les meilleures p-value sont obtenues pour les volumes exposés à des doses supérieures à 35-65Gy. L'âge du patient au moment de sa radiothérapie de rattrapage est la seule covariable clinique non dosimétrique significativement associée à une complication de Grade ≥ 2 (p=0.01). La Figure 3.4 montre la forte corrélation des variables V_{60Gy} et V_{60Gy} avecV_{35Gy} à V_{55Gy} avec les variables. Ces dernières n'ont donc pas été intégrées à l'analyse multivariée afin d'éviter les problèmes d'estimation liés à la colinéarité.

L'ajout de la variable V_{60Gy} n'améliore pas significativement la qualité d'ajustement du modèle logistique construit à partir de V_{65Gy} (p=0.59 du test du rapport de vraisemblance).

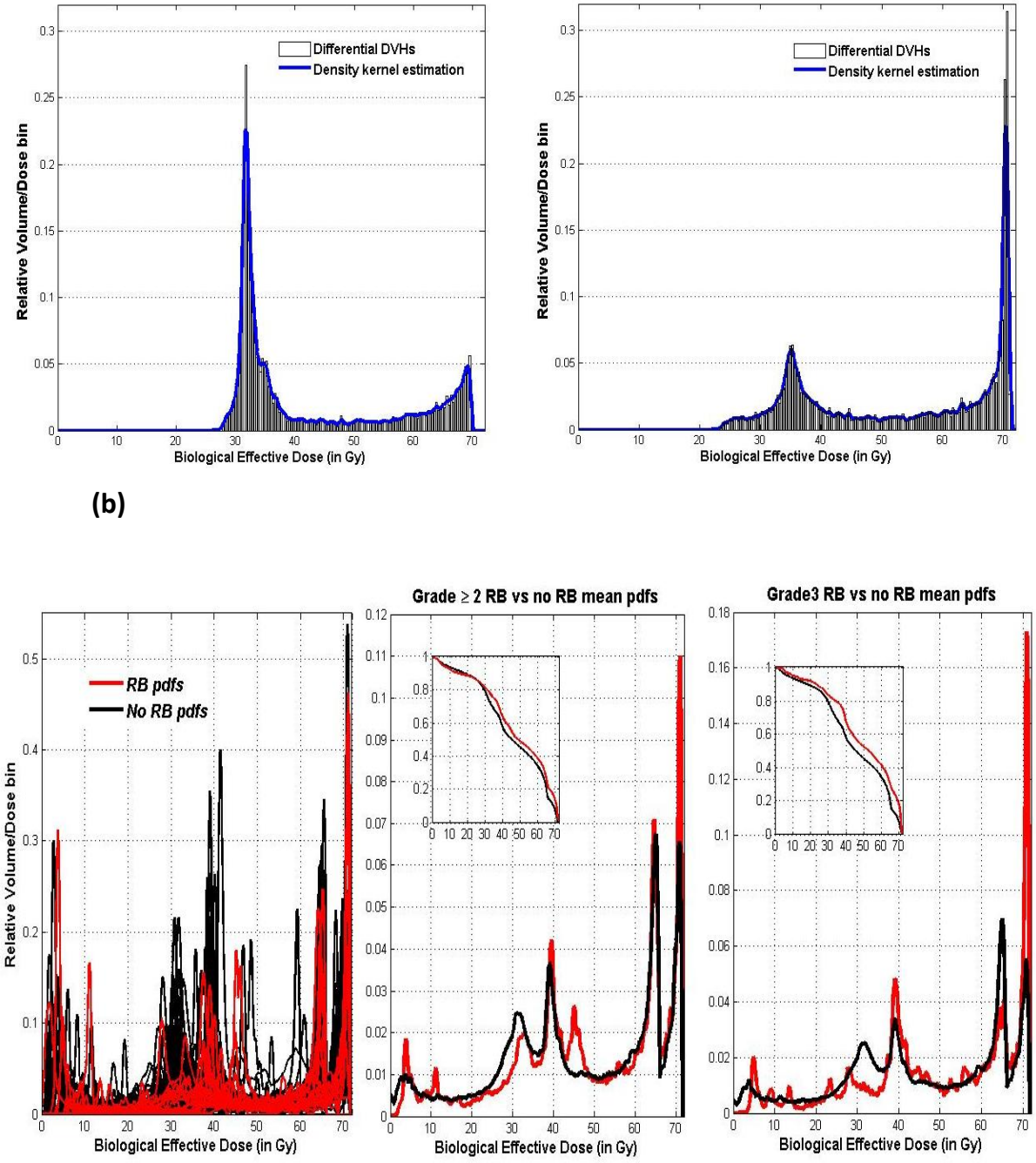

FIGURE 3.3 – (a) Exemple d'estimation de la densité de probabilité par la méthode des noyaux. (b) Densités de probabilité de la distribution de dose au rectum chez les cas (rouge) et les non cas (noir) ainsi que leurs moyennes (les moyennes cumulées sont également représentées).

Analyse univariée	#p-value Grade ≥ 2 (p-value Grade 3)				
Multimétrique Simple		Multimétrique ACP		Modèle Proposé	
Covariables	$p-$value#	Covariables	$p-$value#	Covariables	$p-$value#
Mean dose	0.23(0.08)	CP1	0.23(0.09)	FP1	0.11(0.02)
V_{10Gy}	>0.5(>0.5)	CP2	0.055(0.14)	PF2	0.06(0.04)
V_{15Gy}	>0.5(>0.5)	CP3	0.37(0.36)	FP3	0.78(0.32)
V_{20Gy}	>0.5(>0.5)	CP4	>0.5(>0.5)	FP4	0.35(>0.5)
V_{25Gy}	>0.5(>0.5)			FP5	> 0.5
V_{30Gy}	>0.5(0.4)			FP6	0.12(0.22)
V_{35Gy}	0.13(0.09)			FP7	0.39(>0.5)
V_{40Gy}	0.12(0.07)			FP8	>0.5(>0.5)
V_{45Gy}	0.15(0.13)			FP9	0.14(>0.5)
V_{50Gy}	0.23(0.10)			FP10	>0.5(0.34)
V_{55Gy}	0.12(0.08)				
V_{60Gy}	0.09(0.07)				
V_{65Gy}	0.06(0.03)				

Analyse Multivariée	*Intercept inclus **Modèle à 2 vs 3 paramètres ***68% CI				
Modèle NTCP	Nb paramètres*	Modèle Retenu	AIC	p-value LRT**	AUC(I.C 95%)
Grade≥2					
Multimétrique	2	V_{65Gy}	111.95		0.62(0.46-0.75)
Simple	3	V_{65Gy} + V_{60Gy}	116.66	0.59	
	3	V_{65Gy} + Age	108.79	0.023	0.70(0.56-0.82)
Multimétrique	2	CP2	111.83		0.62(0.47-0.75)
ACP	3	CP2 + CP1	112.29	0.21	
	3	CP2 + Age	108.2	0.017	0.72(0.59-0.82)
Modèle	2	FP2	112.04		0.62(0.48-0.74)
Proposé	3	FP2 + FP1	112.21	0.17	
	3	FP2 + Age	107.87	0.01	0.73(0.59-0.84)
Modèle LKB	3	n = 0.12 (0.03-0.36)*** m =0.17 (0.11- 0.27) TD_{50} =72.6Gy (67.2-82.3Gy)	113.01		0.63(0.47- 0.77)
Grade 3					
Multimétrique	2	V_{65Gy}	66.22		0.67(0.40-0.88)
Simple	3	V_{65Gy} + V_{60Gy}	68.17	0.82	
	3	V_{65Gy} + Age	67.06	0.28	
Multimétrique	2	CP1	67.76		0.62(0.39-0.83)
ACP	3	CP1 + CP2	68.14	0.21	
	3	CP2 + Age	68.18	0.21	
Modèle	2	FP1	64.67		0.75(0.52-0.91)
Proposé	3	FP1 + FP2	64.71	0.16	
	3	FP1 + Age	65.64	0.31	
Modèle LKB	3	n = 0.24 (0.07-0.92)*** m =0.16(0.09-0.26) *** TD_{50} =73.8Gy (66.2-84.1Gy)	113.01		0.63(0.47- 0.77)

TABLE 3.2 – Analyse univariée et multivariée ainsi que la comparaison des différents modèles NTCP

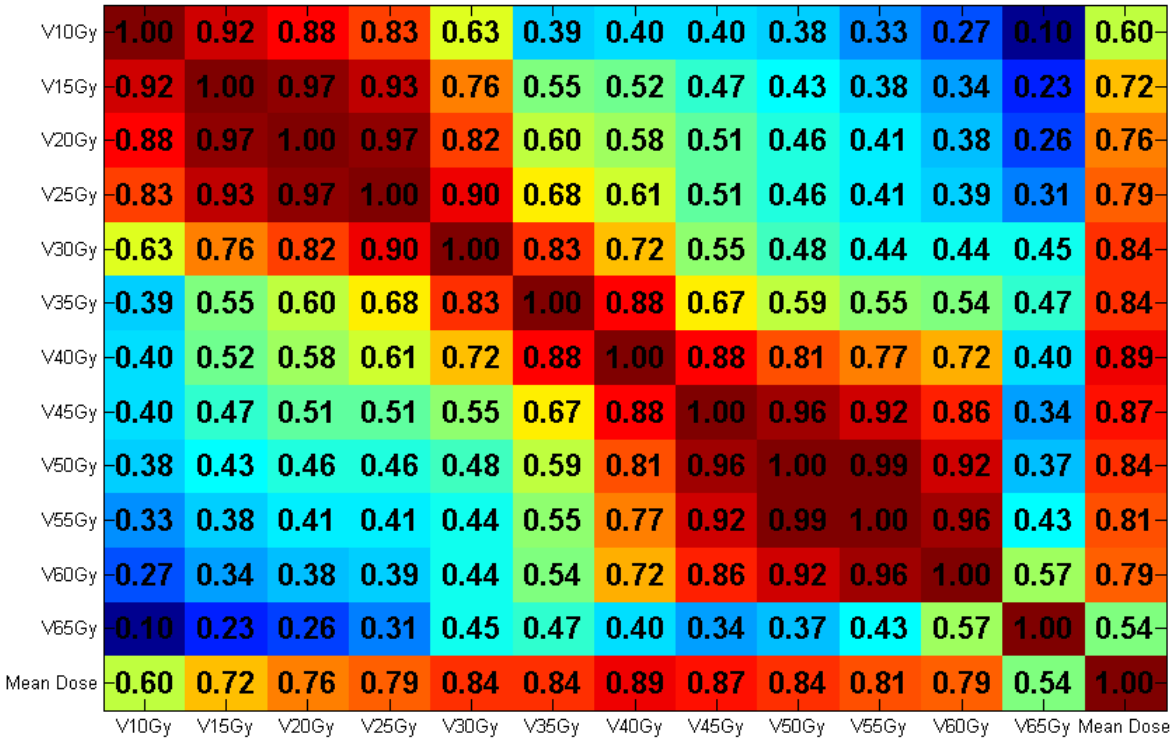

FIGURE 3.4 – Matrice de corrélation de Spearman des indicateurs dosimétriques standards.

Modèle multimétrique ACP

Les trois premiers vecteurs propres issus de l'ACP des vecteurs $(V_{d_i\ Gy})_{1\leq i\leq N}$ sont représentés en haut de la Figure 3.5. Les cinq premières composantes principales (CPs) expliquent plus de 96% de la variabilité observée dans les HDVs cumulés (61% pour la première, 16% pour la seconde).

Comme on peut le voir dans la Table 3.2, le modèle Modèle ACP le plus parcimonieux pour les Grades ≥ 2 et 3 est construit à partir des CP2 et CP1 respectivement.

Modèle NTCP issu de l'analyse de données fonctionnelles

Sont représentées dans le bas de la Figure 3.5 les deux première fonctions propres issues de l'analyse en composantes principales fonctionnelles. La première fonction propre caractérise *la dose par fraction* dans la mesure où son signe discrimine les deux groupes de patients ayant reçu une dose par fraction de 2Gy ou 2.5Gy (voir le scatterplot à droite). La seconde fonction propre est liée à *la technique de traitement de quatre faisceaux en boîte* dans la mesure où elle capte l'anti-corrélation entre le volume rectal exposé aux

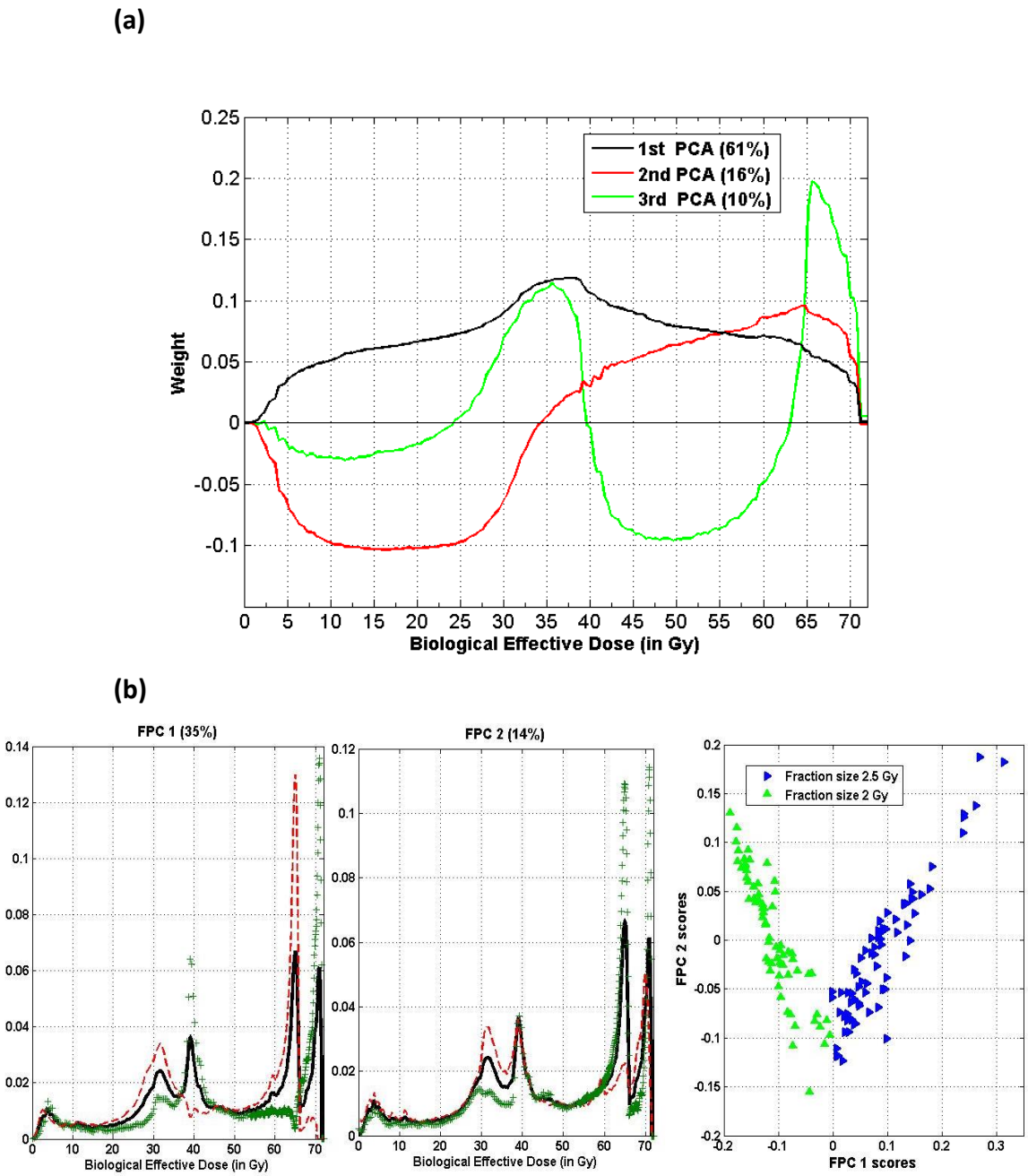

FIGURE 3.5 – (a) Les trois premiers vecteurs propres issus de l'analyse en composantes principales l'ACP des variables aléatoires $(V_{d_i\ Gy})_{1\leq i\leq p}$ définies dans la Section 3.4.4 .(b) L'effet sur la moyenne globale des distributions de doses au rectum des deux premières fonctions propres g_1 et g_2 issues de l'analyse en composantes principales fonctionnelles définie par l'équation (3.1). La courbe en ligne continue est la distribution moyenne, et les courbes $+$ (resp $-$) sont obtenues en retranchant à la courbe moyenne un multiple de la fonction propre (la racine carrée de la valeur propre correspondante). A droite, le scatterplot des scores 1 vs 2. (Entre parenthèses les pourcentages de la variabilité expliquée).

doses intermédiaires (zone d'intersection de deux faisceaux) et le volume rectal exposé aux plus fortes doses (zone d'intersection des quatre faisceaux du traitement).

L'analyse en composantes principales fonctionnelles explique 90% de la variabilité observée par les seize premières fonctions propres avec 50% par les deux premières (FP). L'analyse univariée et multivariée a été menée avec les dix premières fonctions propres (80% de la variabilité cumulée).
Le résultat de l'analyse univariée est donné dans la table 3.2. Les deux meilleures p-values sont obtenues dans le modèle (3.3) par les scores des deux premières fonctions propres. Les tests du rapport de vraisemblance et du critère AIC sélectionnent les modèles basés sur *FP2* et *FP1* pour modéliser le risque de toxicité rectale de Grade ≥ 2 et 3 respectivement.

Le paramètre fonctionnel de la régression 3.2 est représenté dans la Figure3.6 pour les toxicités de Grade ≥ 2 et Grade 3 respectivement.

Ce paramètre permet donc d'évaluer quantitativement le risque de toxicité rectale à partir de la densité de probabilité décrivant la distribution de dose de radiation dans le rectum : si le signe de la fonction $\beta(\delta)$ est positif sur une certaine gamme de doses, alors son intégrale également ce qui s'interprète comme une augmentation du risque alors qu'un signe négatif signifie une diminution.
Par exemple, comparons le risque de développer une toxicité rectale de Grade ≥ 2 chez deux patients P_1 et P_2 à partir de leurs densités de probabilité de distribution de dose au rectum notées f_{P_1} (à gauche) et f_{P_2} (à droite) représentées dans la Figure 3.3.
Alors selon la relation (3.5) :

$$\frac{OR_{P_1}^{Radiation}}{OR_{P_2}^{Radiation}} = \exp\left(\int_0^{71.5Gy} \beta_{Grade2}(\delta) \times \left(f_{P_1}(\delta) - f_{P_2}(\delta)\right)\ d\delta = 0.57\right) \tag{3.8}$$

Ainsi, le risque de toxicité rectale de grade ≥ 2 est réduit de 43% chez le patient P_1 comparé à P_2.

Comparaison des modèles

Selon la Table 3.2, le modèle proposé présente un meilleur AIC et AUC particulièrement pour les toxicités rectales de Grade 3 (AIC=64.61 and AUC=0.75). Excepté pour le modèle LKB, l'inclusion de la covariable *âge* améliore significativement la qualité des modèles dans le cas des complications de Grade ≥ 2.

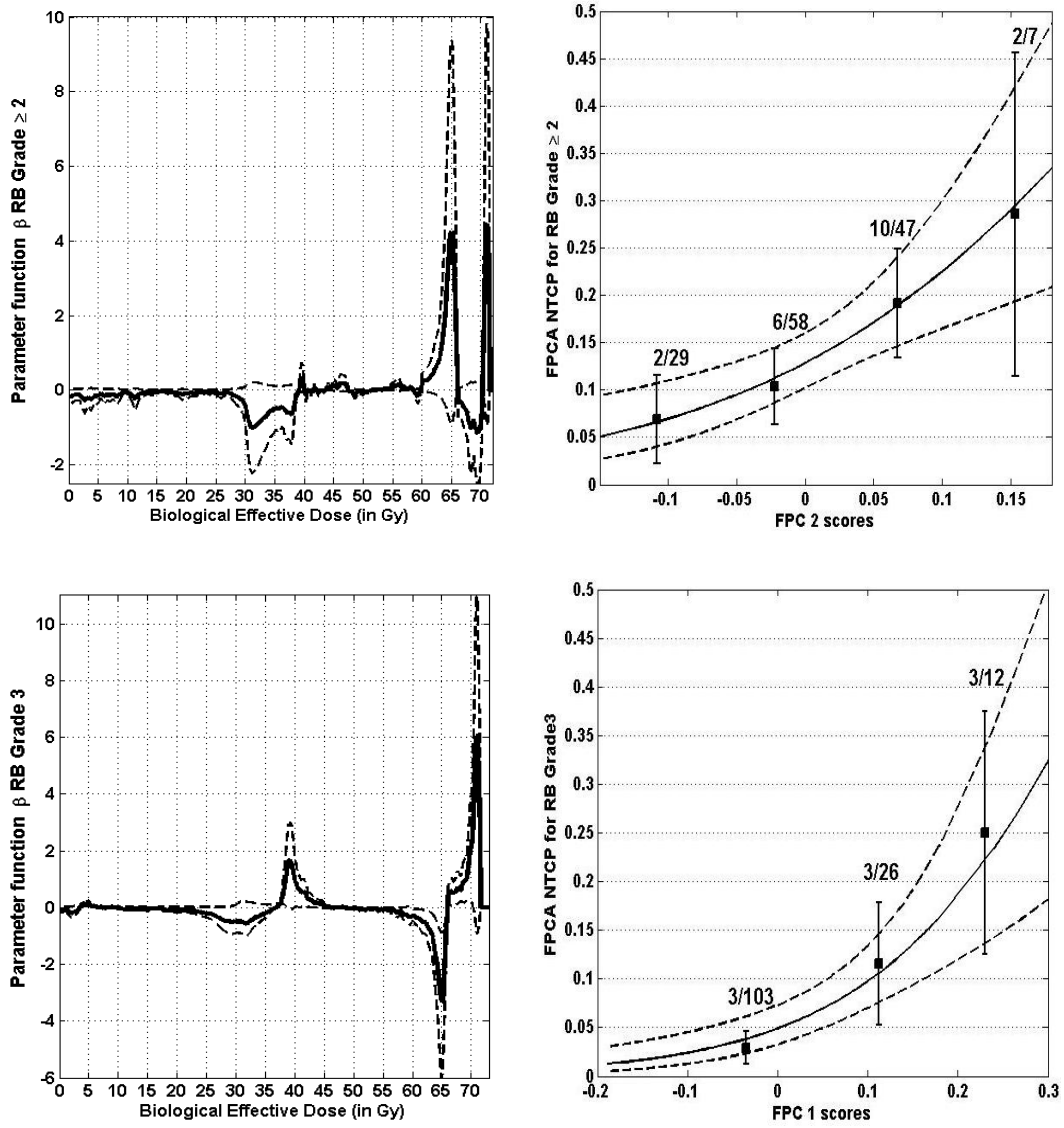

FIGURE 3.6 – (à gauche) Le paramètre fonctionnel $\beta(\delta)$ décrivant le risque de toxicité rectale de grade ≥ 2 et 3. Les lignes en pointillés représentent les intervalles de confiance de niveau 95%. (à droite) La probabilité de survenue de la toxicité rectale en fonction des scores de l'analyse en composantes principales fonctionnelles. Les carrés sont les taux de complications observés avec leurs barres d'erreurs 68%.

3.6 Discussion

L'objectif de ce travail a été de proposer une application de l'analyse des données fonctionnelles à l'analyse des HDVs et à la prédiction des toxicités radio-induites.
Dans ce but, nous avons choisi pour cette première application une technique d'irradiation très simple de quatre faisceaux en boîte utilisée dans la radiothérapie de rattrapage de la loge prostatique et ceci, avant d'explorer des techniques d'irradiation plus complexes comme la radiothérapie à modulation d'intensité (IMRT) par exemple.

Après avoir estimé par la méthode des noyaux les densités de probabilité représentant la distribution de dose au rectum, une analyse en composantes principales fonctionnelles a été par la suite menée afin d'analyser leurs structure spectrales et les principaux modes de variabilité. Ces derniers, incarnés par les fonctions propres, ont enfin été intégrés dans un modèle logistique fonctionnelle afin de tester leurs associations avec la complication d'intérêt.

Le modèle proposé peut être perçu comme un modèle de dose équivalente uniforme (EUD) non paramétrique où le terme $\int \beta(\delta) \times f_i(\delta)\ d\delta$ apparaît comme une dose moyenne pondérée par le paramètre fonctionnel $\beta(\delta)$.
Contrairement au modèle NTCP de LKB qui impose une forme analytique avec des coefficients de pondération de type puissance du paramètre n de l'effet volume i.e $\left(\sum_{i=1}^{N} d_i^{\frac{1}{n}} \times \nu_i\right)^n \approx \left(\int \delta^{\frac{1}{n}} \times f_i(\delta)\ d\delta\right)^n$, notre approche flexible a l'avantage d'éviter toute hypothèse a priori sur la forme des coefficients de pondération de la dose uniforme équivalente pouvant ne pas convenir à toutes les situations étudiées. L'approche proposée est, quant à, elle capable par construction de s'adapter aux spécificités du profil d'irradiation en présence dans l'organe d'intérêt.

Le paramètre de l'effet volume estimé à partir du modèle NTCP de LKB a été estimé à 0.12 (IC68% : 0.03-0.36) pour le Grade ≥ 2 ce qui est en accord avec les données publiées par le QUANTEC [Michalski et al., 2010] : 0.09 (0.04-0.14 CI95%) confirmant ainsi la nature d'organe *en série* du rectum et donc, sa sensibilité aux fortes doses. Signalons à ce propos que les paramètres estimés dans le modèle LKB le sont à l'issue d'une optimisation non linéaire contrairement au modèle proposé dans lequel tous les paramètres sont estimés par une procédure linéaire, ce qui assure l'existence et l'unicité de maximum de la fonction de vraisemblance et évite ainsi le piège lié aux minima locaux.

L'effet volume est donc décrit par notre modèle par le biais d'une fonction $\beta(\delta)$ et non plus un nombre, ce qui fournit une information plus riche : l'effet volume d'une part par la pondération qu'apporte la fonction $\beta(\delta)$ à la gamme de doses en présence en affectant des poids positifs très importants aux plus fortes doses (intersection des quatre faisceaux) confirmant ainsi la nature sérielle du rectum. D'autre part, cette pondération est conditionnelle à la spécificité du plan de traitement étudié puisque la gamme de dose intermédiaire (25-40Gy i.e intersection de deux faisceaux) est affectée d'un poids négatif. Cette fonction effet-volume quantifie donc la manière avec laquelle le risque de toxicité rectale à mesure que le volume du rectum exposé aux plus fortes doses augmente (et donc celui exposé aux doses intermédiaires diminue). Cette corrélation inverse est également signalée par la seconde composante principale du Modèle multi-métrique ACP comme précédemment décrit par Sohn *et al* [Söhn et al., 2007]. Cependant, l'analyse en composante principale menée sur les histogrammes dose-volume cumulés n'a permis de détecter que la transition autour de 35Gy (voir Figure 3.5) sans être capable de spécifier exactement le rôle des gammes de doses intermédiaires/fortes, ce qui en limite donc l'interprétation.

La nature même des statistiques fonctionnelles qui intègrent *l'ordre* dans les gammes de doses comme une dimension à part entière de l'analyse a permis de dégager en une seule composante le rôle joué simultanément par les fortes doses et la dose par fraction 2.5 Gy dans la description de la toxicité de Grade 3 surpassant ainsi significativement les autres modèles NTCP.

Le modèle proposé permet également d'intégrer d'autres covariables cliniques non dosimétriques importantes car même si les contraintes dose-volumes sont respectées de façon optimale, des complications peuvent encore survenir pour d'autres raisons. Les résultats de notre étude suggèrent qu'un âge avancé au moment de la radiothérapie est un facteur de risque de complication rectale de Grade ≥ 2 avec un odds ratio OR=1.123 (1.03-1.22 IC95%). Un historique de chirurgie abdominale ainsi qu'une hypercholestérolémie sont aussi de possibles facteurs de risque. Ces résultats sont en accord avec ceux déjà établis dans de précédentes études [Peeters et al., 2006, Skwarchuk et al., 2000, Tomita et al., 2013].

La méthode proposée dans cette étude appliquée à une technique d'irradiation 3D conformationnelle est tout à fait applicable à toute autre technique d'irradiation dans la mesure où elle exploite des HDVs. Dans le cadre spécifique des toxicités rectales, elle peut également être appliquée sur les HDVs des parois rectales si l'organe est considéré comme

creux. Un des grands potentiels de cette méthode réside dans la possibilité de l'étendre aux données fonctionnelles spatiales [Ramsay et al., 2011, Hörmann and Kokoszka, 2013]. Ceci permettrait d'explorer l'effet-volume sans perdre l'information spatiale fournie par le TPS et ainsi d'adapter les contraintes de dose aux structures particulièrement radiosensibles mises en évidence au sein de l'organe d'intérêt.

Chapitre 4

Modélisation flexible d'une relation dose-effet en épidémiologie des radiations

Sommaire

4.1 Introduction 68

4.2 Modèle de Cox flexible 71

4.3 Analyse en composantes principales fonctionnelles 72

4.3.1 L'approche analyse de données fonctionelles 72

4.3.2 Analyse en composantes principales fonctionnelles 74

4.4 Estimation de la relation dose–effet 75

4.4.1 Pré-sélection des nœud intérieurs 76

4.4.2 Sélection finale de nœuds et de composantes principales fonctionnelles 77

4.5 Simulations 83

4.5.1 Simulations des données 83

4.5.2 Estimation de la fonction dose–réponse 84

4.5.3 Résultats des simulations 85

4.6 Application : Analyse des données thyroïde 92

4.6.1 Les données thyroïde EURO2K 92

4.6.2 Résultats 93

4.7 Discussion 96

4.1 Introduction

Dans les pays développés, plus de la moitié des cancers comprennent la radiothérapie dans leur protocole de traitement [Altekruse et al., 2007, Rosenblatt et al., 2013]. Cependant, l'allongement de la durée de vie des patients traités pour un cancer pose inévitablement la question de la qualité de vie, particulièrement en terme de cancers secondaires radio-induits dont la connaissance de la relation dose-effet devient dès lors un enjeu de santé publique majeur [BEIR, 2006, UNSCEAR, 2006, NCRP, 2012].

Le modèle de Cox [Cox, 1972] est l'un des modèles d'analyse de survie le plus utilisé en épidémiologie des radiations pour étudier l'association entre dose de radiation et événement radio-induit d'intérêt. Les méthodes les plus utilisées pour obtenir de telles relations dose–effet reposent soit sur la catégorisation de la covariable continue qu'est la dose de radiation et ce, dans le but d'obtenir des risques par classes relativement à une classe référence, soit sur des modèles paramétriques reposant sur une expression analytique *a priori* de la nature de l'association dose–effet [Guibout et al., 2005b, Menu-Branthomme et al., 2004b, Lundell et al., 1999, Allard et al., 2010a, Richardson and Wing, 2007, Gilbert et al., 2003, Neglia et al., 2006].

Cette dernière approche assume donc un excès de risque relatif (ERR) issu des différentes simplifications de la formule générale $ERR = (b_1 D + b_2 D^2)\exp(b_3 D + b_4 D^2)$, avec D la dose de radiation et $(b_i)_{1\leq i\leq 4}$ les coefficients de régression [Preston et al., 1993]. le modèle $ERR = b_1 D$ correspond , quant à lui, à une relation dose–effet linéaire avec comme unité pour le paramètre $b1$ l'ERR/Gy. Le terme exponentiel désigne pour sa part le phénomène de mort cellulaire du fait de l'exposition aux doses les plus fortes. Ce type d' a priori sur la forme analytique de la relation dose–effet est très restrictif et peut ne pas être adapté à toutes les situations considérées. D'autre part, l'approche reposant sur la catégorisation de la dose de radiation (covariable par nature continue donc) est connue pour être très sensible aux valeurs de doses définissant chaque catégorie et entraîne , qui plus est, une perte de puissance statistique de l'analyse [Zhao and Kolonel, 1992, Taylor and Yu, 2002, Royston et al., 2005].

L'approche flexible pour modéliser la relation dose–effet permet de s'affranchir de toute hypothèse a priori d'une quelconque expression analytique. Une classe de fonctions flexibles particulièrement utilisées pour explorer la non-linéarité des relations dose-effet ne sont autres que les fonctions splines présentées en annexe B.

Comme exposé dans la section B.4 des annexes, les fonctions splines apparaissent naturellement dans des problèmes d'optimisation pénalisés [Silverman, 1985, Wahba, 1990]. Dans ce contexte, les nœuds sont situés à chaque valeur distincte de la variable continue. Ce résultat théorique est cependant difficile à mettre en oeuvre en pratique sur de grandes cohortes comme l'on rencontre en épidémiologie des radiations , où l'ensemble des nœuds peut se compter par milliers, entraînant ainsi des instabilités numériques des algorithmes d'optimisation. En pratique, deux alternatives sont souvent utilisées : Les splines pénalisées (P-splines) et les splines de régression.

Rappelons que les P-splines sont des splines (construites sur un grille dense de 40-50 nœuds) auxquelles on applique une pénalité de lissage portant en général sur la courbure de l'estimateur fonctionnel de façon à réduire la dimension du modèle à un nombre "effectif" de paramètres. Les splines de régression, de leur coté, utilisent un nombre sensiblement plus réduit de nœuds en omettant le terme de pénalisation de façon à obtenir un modèle totalement paramétrique.

Les splines de régression sont particulièrement sensibles au nombre et à l'emplacement des nœuds intérieurs (voir section B.3 des annexes). De ce fait, beaucoup d'algorithmes ont été proposés pour choisir de façon adaptative le nombre et l'emplacement des nœuds. Nombre de ces modèles reposent sur le principe de sélection de modèle en adoptant une représentation spline sous forme de puissances tronquées, où chaque nœud est associé à un unique terme de la base. Citons parmi les plus connus les algorithmes MARS [Friedman, 1991] and POLYMARS [Stone et al., 1997] pour la régression linéaire et HARE [Kooperberg et al., 1995] pour les données de survie où le critère des moindres carrés est remplacé par la vraisemblance partielle.

L'objectif de ce chapitre est de proposer une nouvelle méthode de régression spline basée sur les techniques d'analyse de données fonctionnelles. Elle est constituée de deux étapes : un ensemble de nœuds *potentiels* est dérivé de l'analyse en composantes principales fonctionnelles, conditionnellement à la distribution de la covariable continue, d'un ensemble de fonctions résumant la propriété de support minimum d'une base B-spline construite à partir d'une grille dense de quantiles équidistants. L'estimation spline finale de la relation dose–effet résulte, dans un second temps, d'un processus de sélection de variables ayant pour arguments d'optimisation non seulement le nombre et l'emplacement des nœuds, mais aussi le nombre de composantes principales incorporées dans le modèle

final. C'est précisément ce dernier point qui est à l'origine de la réduction dimensionnelle significative qu'offre la méthode proposée en comparaison des méthodes de régression spline classiques.

Nous évaluerons les performances de la méthode proposée à travers plusieurs scénarios simulés de relations dose–effet. Les performances seront comparées aux méthodes de régression splines classiques ainsi que P-splines en tant que concurrent naturel en terme réduction dimensionnelle. Enfin, la méthode proposée sera appliquée pour étudier la relation dose–effet entre la dose de radiation reçue par la thyroïde et le risque de tumeurs secondaires radio-induites de la Thyroïde dans une population de 3289 survivants d'un cancer de l'enfance dans cinq centres hospitaliers en France.

4.2 Modèle de Cox flexible

Le modèle de Cox à risques proportionnels assume dans sa version initiale la proportionnalités des risques ainsi que l'effet log-linéaire des covariables d'intérêt.
Plus précisément, le modèle de Cox pour données de survie avec le vecteur de covariables z est défini par la fonction de risque instantané $h(t,z)$ par :

$$h(t,z) = h_0(t) \exp\left(\sum_{l=1}^{L} b_l z_l\right) \tag{4.1}$$

où $h_0(t) = h(t,0)$ est la fonction de risque de base.

Afin de s'affranchir de l'hypothèse de log-linéarité à laquelle est soumise l'action de certaines covariables, ce modèle se généralise en remplaçant le terme $b_l z_l$ par une classe de fonctions plus générales :

$$h(t,z) = h_0(t) \exp\left(\sum_{l=1}^{L} r_l(z_l)\right) \tag{4.2}$$

où $(r_l)_{1\leq l\leq L}$ sont des fonctions régulières (dans le sens continuité et/ou dérivabilité, etc.) en $(z_l)_{1\leq l\leq L}$. Dans cette partie, une attention particulière sera portée à la classe de fonctions B-splines cubiques.

Ainsi, chaque fonction r_l, $1 \leq l \leq L$, peut être exprimée comme une fonction cubique de z_l :

$$r_l(z_l) = b_l^T B_l(z_l) = \sum_{k=1}^{K_l} b_{lk} B_{kl}(z_l) \tag{4.3}$$

où $b_l^T = (b_{l1}, ..., b_{lK_l})$, $B_l(z_l)^T = (B_{l1}, ..., B_{lK_l})$ et $(B_{lk})_{1\leq k\leq K_l}$ sont des bases B-splines construites à partir de m_l interiors knots.

Ainsi, les fonctions $(r_l)_{1\leq l\leq L}$ de (4.2) peuvent être estimées par $\left(\widehat{r_l} = \widehat{b_l}^T B_l\right)_{1\leq l\leq L}$ où les coefficients splines $\left(\widehat{b_l}\right)_{1\leq l\leq L}$ sont obtenus par maximisation de la vraisemblance partielle. Notons au passage que dans cette approche usuelle, le nombre de paramètres à estimer est étroitement lié aux nombres de nœuds intérieurs servant à la construction des bases splines :

$$K_l = 4 + m_l \tag{4.4}$$

4.3 Analyse en composantes principales fonctionnelles

4.3.1 L'approche analyse de données fonctionelles

Sans perte de généralité et dans le but de simplifier l'exposé, considérons les modèles de Cox flexibles (4.2) et (4.3) avec une seule variable continue à effet non log-linéaire :

$$\begin{aligned} h(t,z) &= h_0(t)\exp\left(r(z)+\sum_{m=1}^{M} c_m\nu_m\right) \\ &= h_0(t)\exp\left(b^T B(z)+\sum_{m=1}^{M} c_m\nu_m\right) \\ &= h_0(t)\exp\left(\sum_{k=1}^{K} b_k B_k(z)+\sum_{m=1}^{M} c_m\nu_m\right) \end{aligned} \tag{4.5}$$

où, $b^T = (b_1, ..., b_K)$, $B(z)^T = (B_1(z), ..., B_K(z))$ et $(B_k)_{1\leq k\leq K}$ est une base B-spline cubique. Des covariables additionnelles à effet log-linéaire sont désignées par $(\nu_m)_{1\leq m\leq M}$ et leurs coefficients de régression par $(c_m)_{1\leq m\leq M}$.

Dans cette section nous établissons un lien entre le modèle (4.5) et le modèle de risques proportionnels de Cox pour *données fonctionnelles*.
La phrase *données fonctionnelles* désigne des données où chaque observation consiste en une courbe, surface ou hypersurface, par opposition à un point ou vecteur.

Notons N la taille de l'échantillon et $1 \leq i \leq N$. Dans notre cas, la $i^{ème}$ observation de la variable prédictive continue z est une valeur réelle z_i.

Selon la propriété de support minimal des bases B-splines cubiques établie dans la Section B.2.2, il existe au plus 4 fonctions B-splines non nulles en z_i comme illustré dans la Figure 4.1

De plus, la positivité et le fait que les bases B-splines soient des partitions de l'unité en toute valeur de la covariable implique que les nombres $(B_k(z_i))_{1\leq k\leq K}$ peuvent être considérés comme des *poids* de la propriété locale de la base $(B_1, ..., B_K)$ au voisinage d'une valeur donnée z_i de z.

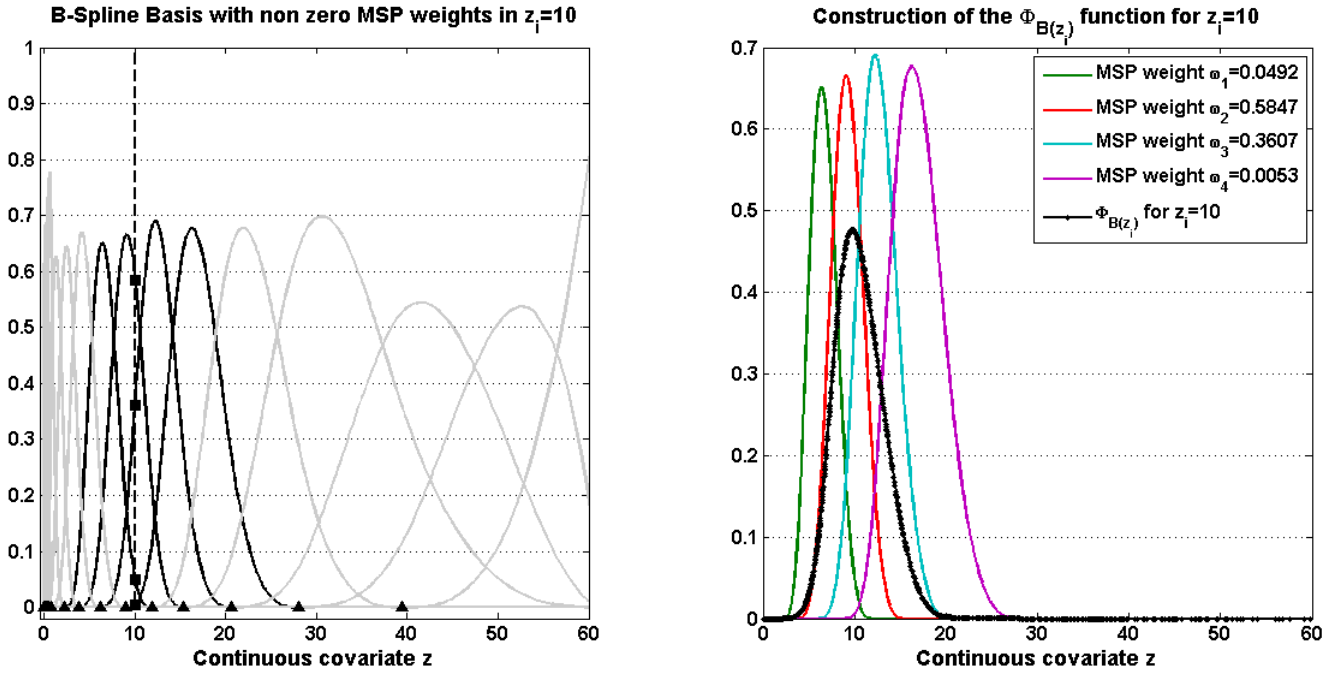

FIGURE 4.1 – Construction des fonctions à support minimal (MSP).(Gauche) La base B-spline est construite à partir des nœuds désignés par des triangles noirs. En noir (resp en gris) les B-splines prenant des valeurs non nulles (resp nulles) en $z_i = 10$.(Droite) les fonctions MSP $\Phi_{z_i=10}$ comme une moyenne pondérée des quatre B-splines.

Ainsi, et par analogie avec la physique du solide, nous sommes donc en présence au voisinage de z_i d'un système mécanique $\langle \{B_1, B_1(z_i)\}, ..., \{B_K, B_K(z_i)\} \rangle$ avec comme représentant son **centre de masse** :

$$\Phi_{B(z_i)} = \sum_{k=1}^{K} B_k(z_i) B_k \tag{4.6}$$

Par suite, si nous remplaçons chaque vecteur $B(z_i)^T = (B_1(z_i), ..., B_K(z_i))$ par la fonction B-spline $\Phi_{B(z_i)}$ définie par (4.6), nous pouvons écrire le modèle à risque proportionnel pour données fonctionnelles suivant :

$$h(t, z) = \widetilde{h}_0(t) \exp\left(\int_{\Omega(z)} \beta(u) \times \Phi_{B(z)}(u)\, du + \sum_{m=1}^{M} c_m \nu_m \right) \tag{4.7}$$

où $\Omega(z)$ désigne l'ensemble des valeurs prises par la variable aléatoire z et β est une fonction à carré intégrable sur $\Omega(z)$.

La reformulation du problème en terme de données fonctionnelles nous permettra de faire usage une nouvelle fois de *l'analyse en composante principale fonctionnelle* dans le cas particulier où toutes les fonctions sont exprimées dans la même base B-spline. Ce cas de figure a été présenté en détail dans la section 2.5.1.
Le problème d'estimation du paramètre fonctionnel β est, quant à lui, traité dans la section 4.4.

4.3.2 Analyse en composantes principales fonctionnelles

Rappelons (voir section 2.4) que nous pouvons écrire chaque $\Phi_{B(z_i)}$ selon la représentation de Karhunen-Loève :

$$\Phi_{B(z_i)} = \overline{\Phi}_B + \sum_{l\geq 1} \theta_{il} \times \psi_l \tag{4.8}$$

où $\overline{\Phi}_B = \frac{1}{N}\sum_{i=1}^{N} \Phi_{B(z_i)}$ est la moyenne commune, $(\psi_l)_{l\geq 1}$ les fonctions propres, une base orthonormée dans le sens $L^2(\Omega(z))$ qui détermine de façon optimale dans le sens des moindres carrés les modes de variabilités les plus importants relativement à la moyenne commune. Les scores $(\theta_{il})_{l\geq 1}$ décrivent les coordonnées de la $i^{ème}$ fonction $\Phi_{B(z_i)}$ dans la base $(\psi_l)_{l\geq 1}$. Les variances des scores appelées *valeurs propres*, notées $\lambda_1 \geq \lambda_2 \geq \lambda_3...$, sont donc la contribution du $l^{ième}$ terme de (4.8) dans la description de la variance totale.

Une telle analyse spectrale dans un espace fonctionnel peut être grandement facilitée par l'usage d'une base fonctionnelle permettant de se ramener à un problème spectral matriciel équivalent (voir section 2.5.1.)

Dans ce cas, le nombre maximal de fonctions propres est de K, i.e la dimension du sous-espace engendré par la base B-spline $(B_k)_{1\leq k\leq K}$. La décomposition (4.8) devient alors :

$$\Phi_{B(z_i)} = \overline{\Phi}_B + \sum_{l=1}^{K} \theta_{il} \times \psi_l \tag{4.9}$$

4.4 Estimation de la relation dose–effet

Les notations sont similaires à la section 4.3.2. Considérons à présent le modèle de Cox pour données fonctionnelles défini par(4.7).
Nous reprenons ici le même raisonnement que dans la section 2.6.2 à propos de la régression sur composantes principales fonctionnelles et développons la fonction de risque β selon $\sum_{j=1}^{K} \beta_j \times \psi_j$, il suit :

$$
\begin{aligned}
h(t, z_i) &= \widetilde{h}_0(t) \times \exp\left(\int_{\Omega(z)} \beta(u)\ \Phi_{B(z_i)}(u)\, du + \sum_{m=1}^{M} c_m \nu_m \right) \\
&= \widetilde{h}_0(t) \times \exp\left(\int_{\Omega(z)} \beta(u) \left(\overline{\Phi}_B(u) + \sum_{l=1}^{K} \theta_{il} \times \psi_l(u) \right) du + \sum_{m=1}^{M} c_m \nu_m \right) \\
&= \widetilde{h}_0(t) \times \exp\left(\int_{\Omega(z)} \beta(u) \times \overline{\Phi}_B(u)\, du + \sum_{l=1}^{K} \theta_{il} \int_{\Omega(z)} \beta(u) \times \psi_l(u)\, du + \sum_{m=1}^{M} c_m \nu_m \right) \\
&= \widetilde{h}_0(t) \times \exp\left(\int_{\Omega(z)} \beta(u) \times \overline{\Phi}_B(u)\, du \right) \times \exp\left(\sum_{j=1}^{K} \sum_{l=1}^{K} \beta_j \times \theta_{il} \times \int_{\Omega(z)} \psi_j(u)\ \times \psi_l(u)\, du + \sum_{m=1}^{M} c_m \nu_m \right) \\
&= \underbrace{\left[\widetilde{h}_0(t) \times \exp\left(\int_{\Omega(z)} \beta(u) \times \overline{\Phi}_B(u)\, du \right) \right]}_{h_0(t)} \times \exp\left(\sum_{l=1}^{K} \beta_l \times \theta_{il} + \sum_{m=1}^{M} c_m \nu_m \right) \\
&= h_0(t) \times \exp\left(\sum_{l=1}^{K} \beta_l \times \theta_{il} + \sum_{m=1}^{M} c_m \nu_m \right)
\end{aligned}
\tag{4.10}
$$

L'avant-dernière égalité est due à l'orthonormalité des fonctions propres $(\psi_l)_{1 \le l \le K}$ et la dernière résulte de l'absorption de l'intercept exponentiel $\exp\left(\int_{\Omega(z)} \beta(u) \times \overline{\Phi}_B(u)\, du \right)$ par la fonction risque de base.
L'estimateur $\left(\widehat{\beta}_1, ..., \widehat{\beta}_K \right)^T$ du vecteur coefficients $(\beta_1, ..., \beta_K)^T$ est alors obtenu par maximisation de la vraisemblance partielle.

Dans ce contexte, l'estimateur fonctionnel $\widehat{r}$ dans (4.5) de la relation dose–effet est égal en chaque valeur z_i à :

$$
\widehat{r}(z_i) = \sum_{l=1}^{K} \widehat{\beta}_l \times \theta_{il} \tag{4.11}
$$

Le lien avec la régression spline classique sur la base $(B_k)_{1 \le k \le K}$ est alors établi puisque, pour tout i, les scores $(\theta_{il})_{1 \le l \le K}$ sont des combinaisons linéaires des $(B_k(z_i))_{1 \le k \le K}$.

Pour des problème d'unicité de la solution, nous devons imposer une contrainte initiale à la fonction de risque r en lui attribuant une valeur fixe en une valeur de référence de la covariable continue. Comme il est souvent d'usage en épidémiologie des radiations, nous avons adopté la restriction $r(0) = 0$. Cette condition a une conséquence simple en terme de développement dans la base B-spline $(B_k(z_i))_{1\leq k\leq K}$ puisqu'il suffit simplement de supprimer la première fonction de base $B_1(z)$. Dans ce cas, le nombre de paramètres splines à estimer vaut $4 + K - 1 = K + 3$.

Etant donné que l'usage de toutes les fonctions propres dans (4.11)(et donc de tous les scores par individu) donne, par construction, le même résultat qu'une régression spline classique sur la même base B-spline, avec le même nombre de coefficients splines estimés, l'analyse en composantes principales fonctionnelles permet une réduction dimensionnelle en terme de nombre de coefficients estimés, dès lors que le modèle intègre un nombre plus réduit de scores dans l'équation (4.11) selon la part de variabilité expliquée ou d'association statistique significative avec l'événement d'intérêt.

Avec cette approche, l'estimation de la relation dose–effet consiste à déterminer le couple **ensemble de nœud optimal et le nombre de composantes principales** que compte le modèle (et donc de paramètres estimés). Remarquons la différence avec l'approche régression spline classique dans laquelle le nombre de paramètres estimés est directement lié aux nombre de nœuds intérieurs.

Ainsi, le nombre et l'emplacement des nœuds intérieurs vise à capter les changements de courbature de la relation dose–effet alors que le choix des composantes principales, quant à lui, peut être vu comme un paramètre de lissage.
Nous proposons dans la suite un processus d'estimation de la relation dose-effet reposant sur deux étapes principales : déterminer par une méthode spatialement adaptative un ensemble potentiel de nœuds intérieurs. Achever le processus par une sélection finale de la combinaison *nœud intérieurs/composantes principales fonctionnelles.*

4.4.1 Pré-sélection des nœud intérieurs

L'objectif de cette première section est donc d'identifier un premier ensemble de nœuds intérieurs *candidats* offrant une distribution spatiale décrivant les intervalles de valeurs de la covariable continue z pour lesquels la propriété de support minimal présente les corrélations les plus fortes avec l'événement d'intérêt.

L'idée est donc d'étudier la propriété de support minimal par l'intermédiaire des fonctions $\left(\Phi_{B(z_i)}\right)_{1\leq i\leq N}$ et, plus précisément, à partir de leurs décompositions en composantes en analyse composantes principales. Comme dans le contexte de l'analyse statistique multivariée classique, une rotation des composantes principales originales peut amener à une meilleure interprétation des parts de variabilité en présence. En général, cette rotation est appliquée seulement aux premières composantes principales, à savoir les plus informatives. Dans notre cas, nous avons appliqué une rotation Varimax fonctionnelle sur les premières fonctions propres avec comme critère un pourcentage de variabilité expliqué de 95%. L'effet de cette rotation Varimax fonctionnelle est donc de produire de nouvelles fonctions propres davantage *concentrées* sur certaines régions distinctes de la gamme de valeurs couvertes par la covariable continue d'intérêt.

Ainsi, comme chaque intervalle de valeurs de z correspond à une rotation d'une fonction propre, le modèle de Cox (4.12) basé sur les scores correspondant à ces rotations $\left(\theta_{il}^{Rot}\right)$ permet de détecter, par le biais d'inférence sur les coefficients (par exemple la p-value) les intervalles dans lesquels la propriété de support minimum est la plus corrélée avec la toxicité étudiée.

$$h(t, z_i) = h_0(t) \times \exp\left(\sum_{l=1}^{K} \beta_l \times \theta_{il}^{Rot} + \sum_{m=1}^{M} c_m \nu_m\right) \tag{4.12}$$

Ainsi, une séquence initiale de quantiles équidistants est passée par un filtre construit à partir des fonctions propres, et ne sont retenus que les quantiles présents dans les zones de support minimum les plus significatives comme illustré dans la figure 4.2.

4.4.2 Sélection finale de nœuds et de composantes principales fonctionnelles

Guidés par l'analyse du profil de la propriété de support minimal au voisinage d'un certain nombre de quantiles, nous obtenons, à l'issue de l'étape précédente, un ensemble potentiel de nœuds qu'il faut à présent affiner afin d'obtenir une estimation optimale de la relation dose–effet.

Nous présentons trois algorithmes de sélection finale de nœuds.
Le premier est l'algorithme que l'on désignera par *Backward Search* : un algorithme de sélection classique de nœuds basé sur le développement spline en terme de puissance tronquée.

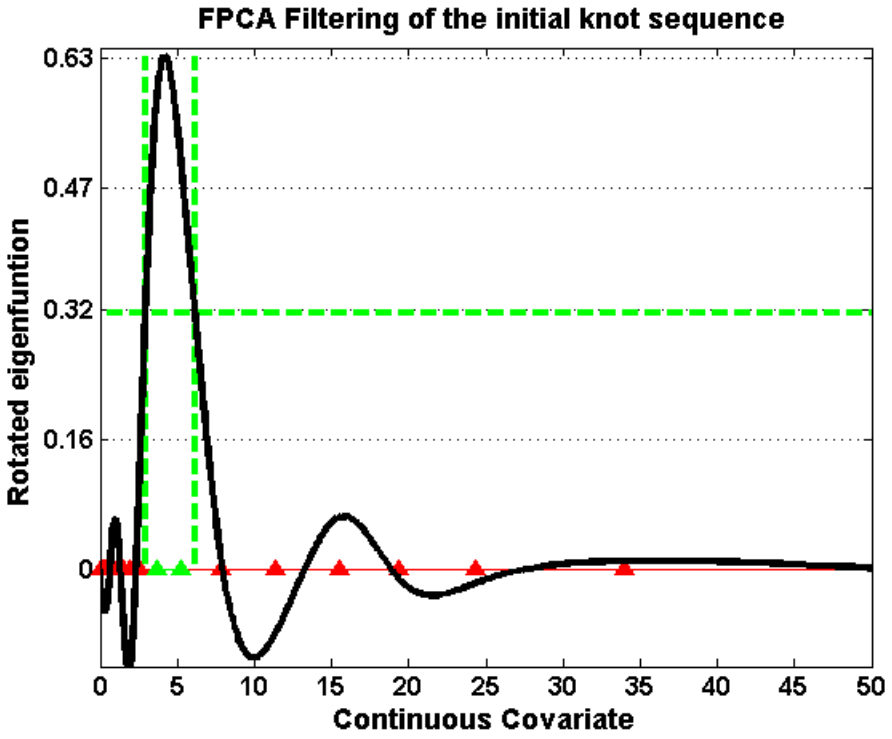

FIGURE 4.2 – Filtrage spectral des quantiles : la rotation d'une fonction propre en noir vue comme un filtre de bande passante mi-hauteur. La séquence initiale de quantiles est représentée avec des triangles et seuls les quantiles en vert sont retenus.

Le deuxième, nommé *Full FPCA Search*, est nouveau et vise à s'affranchir de l'interdépendance nombre de nœuds et nombre de paramètres splines à estimer en réduisant la dimension du modèle via une optimisant jointe nœuds intérieurs/ nombre de fonctions propres incorporées.
Le troisième algorithme, nommé *Forward FPCA Search* est une version accélérée du précédent algorithme au prix d'une optimisation marginale plutôt que jointe.

- **Algorithme 1 : Backward Search**
  Dans cette méthode, les nœuds jugés les moins importants dans la représentation de la relation dose–effet sont éliminés à chaque étape par un processus de sélection de variable. Ce processus est mené via la base *puissance tronquée* présentant l'avantage d'associer l'effet d'un nœud intérieur à une unique fonction de base.

  Pour les splines cubiques, la base associée à l'ensemble de nœuds intérieurs $(z_1, ..., z_m)$ est $\left\{z, z^2, z^3, (z-z_1)_+^3, (z-z_2)_+^3, ..., (z-z_m)_+^3\right\}$ où $x_+^3 = max\,(0, x^3)$.

L'estimateur de la relation dose-effet est donc une combinaison linéaire des fonctions de base ci-dessus :

$$\hat{r}(z) = a_1 \times z + a_2 \times z^2 + a_3 \times z^3 + \sum_{i=1}^{m} \alpha_i \times (z - z_i)_+^3 \tag{4.13}$$

Basé sur le critère de Schwarz :

$$BIC(S_q) = -2 \times \log(l_{S_q}) + \log(N_{Events}) \times q \tag{4.14}$$

où S_q la taille d'une séquence de nœuds intérieurs, l_{S_q} la vraissemblance partielle du modèle spline construit à partir de la séquence S_q et N_{Events} le nombre d'événements, l'algorithme *BackwardSearch* procède comme suit :

- **Etape 0** : Commencer par l'ensemble de nœuds potentiels *candidats* obtenus lors de la phase de pré-sélection présenté en Section 4.4.1.

- **Etape 1 :** Déterminer le meilleur modèle par élimination descendante :

  1. Identifier le $i^{ème}$ nœud dont la covariable puissance tronquée correspondante présente la statistique de Wald la plus large (la moins bonne p-value).

  2. Évaluer la BIC du modèle réduit privé du $i^{ème}$ nœud de l'étape précédente.

  3. Supprimer le $i^{ème}$ nœud de la séquence de nœuds potentielle.

  4. Revenir à 1. tant que la séquence de nœuds potentiels est non vide.

- **Etape 2** : Choisir l'ensemble de nœuds S_q^* correspondant au modèle choisi à l'**Etape 1** possédant le $BIC\left(S_q^*\right)$ le plus bas.

- **Algorithme 2 : FPCA Full Search Algorithm**
  Dans la régression spline cubique usuelle, le nombre de coefficients estimés est égal,

sous la restriction $r(0) = 0$, à K avec $K - 3$ le nombre de nœuds intérieurs.

La décomposition (4.11) permet une réduction dimensionnelle du modèle en ne gardant dans le modèle que les scores associés aux premières composantes principales (dans l'ordre décroissant de part de variabilité expliquée). Dans le présent travail, le choix du nombre de composantes principales est effectué par le critère AIC.

Ainsi, le problème d'estimation de la relation dose–effet comporte à présent deux dimensions : la séquence de nœuds intérieurs et le nombre de fonctions propres de l'ensemble de fonctions $\left(\Phi_{B(z_i)}\right)_{1\leq i\leq N}$ construites à partir des ces nœuds intérieurs via la base B-splines $(B_k)_{1\leq k\leq K}$.

La procédure d'optimisation peut alors s'écrire :

$$\min_{\substack{S\subset S_0\\ L\leq \#S}} AIC\left(M_{S,L}\right) = \min_{\substack{S\subset S_0\\ L\leq \#S}} \left\{-2\times \log\left(l_{M_{S,L}}\right) + 2\times dim(M_{S,L})\right\} \tag{4.15}$$

où S_0 représente l'ensemble de nœuds *candidats* obtenu à l'issue de l'étape de pré-sélection de la section 4.4.1, $\#S$ le nombre de nœuds que comporte la sous-séquence S de S_0, $M_{S,L}$ le modèle de Cox flexible construit à partir de S et basé sur les L premiers scores :

$$h(t, z_i) = h_0\left(t\right) \times \exp\left(\sum_{l=1}^{L} \beta_l \times \theta_{il} + \sum_{m=1}^{M} c_m \nu_m\right) \tag{4.16}$$

- **Algorithme 3 : Forward Search Algorithm**
  L'algorithme précédent repose sur une optimisation jointe de la séquence de nœuds intérieurs et du nombre de fonctions propres. Une telle procédure consiste donc à tester tous les modèles possibles et à comparer leurs AICs, ce qui demande des ressources en calculs très conséquentes.
  Nous proposons dans ce premier algorithme une version plus rapide qui consiste à mener l'optimisation (4.15) de façon ascendante en ajoutant à la séquence de nœuds en présence un nouveau nœud de façon à ce que la réduction dimensionnelle qui s'ensuit soit plus efficace que la précédente en terme de AIC.
  Cette méthode a l'avantage de s'affranchir de tester toutes les combinaisons possibles dans (4.15) en construisant le modèle à l'aide d'optimisations marginales successives.

En utilisant les mêmes noations que dans (4.15), l'algorithme peut être décrit comme suit :

- **Etape 1**

$$\begin{cases} S_{Addition} \leftarrow Vide \\ L_{Addition} \leftarrow 3 \\ S_{Restants} \leftarrow Séquence\ de\ nœud\ candidats\ S_0 \\ AIC_{Test} \leftarrow +\infty \end{cases}$$

- **Etape 2**

  Trouver le nœud u^* dans $S_{Restants}$ qui fournit la meilleure reduction dimensionnelle du modèle en terme de AIC :

$$(u^*, L^*) = \underset{\substack{u \in S_{Restants} \\ L \leq \#(S_{Addition} \cup \{u\})+3}}{argmin} AIC\left(M_{S_{Addition} \cup \{u\}, L}\right) \tag{4.17}$$

- **Step 3**

  Tester si $AIC\left(M_{S_{Addition} \cup \{u^*\}, L^*}\right) > AIC_{Test}$.

  SI OUI : $\begin{cases} S_{Finale} \leftarrow S_{Addition} \\ L_{Finale} \leftarrow L_{Addition} \end{cases}$ et STOP.

  SINON

$$\begin{cases} S_{Addition} \leftarrow S_{Addition} \cup \{u^*\} \\ L_{Addition} \leftarrow L^* \\ S_{Restants} \leftarrow S_{Restants} \text{ privé de } \{u^*\} \\ AIC_{Test} \leftarrow AIC\left(M_{S_{Addition} \cup \{u^*\}, L^*}\right) \end{cases}$$

  et revenir à **Etape 1**

Le modèle final selectionné par cet algorithme est alors construit à partir de la séquence de nœuds intérieurs finale S_{Final} en utilisant les L_{Final} premières fonctions propres dans la décomposition (4.9).

4.5 Simulations

4.5.1 Simulations des données

Dans le but d'évaluer la performance de la méthode proposée, nous avons simulé des cohortes présentant les mêmes caractéristiques que celles observées et étudiées en épidémiologie des radiations. L'objectif de cette section est d'évaluer les capacités de la méthode spatialement adaptative issue de l'analyse spectrale fonctionnelle à reconstituer les relations dose–effet utilisées pour générer les événements d'intérêt.
La variable aléatoire d'exposition, nommée `Dose`, est générée à partir d'une loi exponentielle par morceaux $e_1 I_{\left(\frac{1}{3} \leq u < \frac{2}{3}\right)} + e_2 I_{\left(u \geq \frac{2}{3}\right)}$ où u est généré à partir d'une loi uniforme de support [0,1], e_1 et e_2 des variables de loi exponentielles de moyennes 0.17 et 17, respectivement. Ainsi, l'exposition de la cohorte simulée se divise en trois groupes : un groupe non exposé, un groupe exposé à moins d'un Gy et un troisième exposé à des doses de radiation supérieures à 1Gy. Ce profil de distribution de dose étant assez représentatif de l'exposition observée en épidémiologie des radiations.

Deux autres variables catégorielles sont considérées : `Sexe` et `NbDrug`, représentant respectivement le sexe et le nombre de drogues que comprend un traitement par chimiothérapie, sont générées à partir d'une *loi Bernouilli* de paramètre 0.7 et d'une *loi multinomiale* [n=3, valeurs=(0,1,2), probabilités=(0.33,0.22,0.45)].

Ainsi, le risque instantané au temps t s'écrit :

$$h\left(t|\texttt{Dose},\texttt{Gender},\texttt{Nbdrug}\right) = h_0\left(t\right)\exp\left(r\left(\texttt{Dose}\right) - 0.22 \times \ \texttt{Sexe} + 0.4 \times \ \texttt{Nbdrug}\right) \tag{4.18}$$

où r est la fonction dose–effet décrivant l'impact de la variable `Dose` sur le risque.

Un millier de cohortes, chacune de taille 3000, ont été simulées selon l'algorithme permutationnel décrit en [Sylvestre and Abrahamowicz, 2008] avec une répartition des instants d'événements simulés selon une loi de Weibull (paramètre d'échelle = 26, paramètre de forme = 2), de façon à obtenir 95% de censures. Ce taux de censure élevé reflétant la rareté des pathologies radio-induites observées en pratique.

4.5.2 Estimation de la fonction dose–réponse

Nous apportons dans cette section quelques précisions sur le processus d'estimation de la relation dose–effet. Commençons tout d'abord par rappeler que, pour des problèmes d'identifiabilité et d'unicité, la fonction r dans l'équation (4.18) est sujette à la restriction $r(0) = 0$.

Comme décrit dans la section 4.4, le processus d'estimation débute avec la pré-sélection, de façon spatialement adaptative, d'un ensemble de nœuds potentiels *candidats*.

Nous avons considéré un ensemble initial de 30 quantiles équidistants à partir desquels nous avons construit une base B-spline $(B_k)_{1\leq k\leq K}$. Dans le but d'explorer l'association entre *la propriété de support minimal* de $(B_k)_{1\leq k\leq K}$ et l'événement d'intérêt, une analyse en composantes principales fonctionnelles a été menée sur la famille de fonctions $\{\Phi_{B(z_i)}\}_{1\leq i\leq N}$ définies dans l'équation (4.6).

Cette étape de pré-sélection est par la suite suivie d'une seconde phase pour déterminer l'ensemble final de nœuds intérieurs, et ceci, via les trois algorithmes présentés dans la section 4.4.2. Ainsi, il sera possible d'étudier le mode de sélection des algorithmes *FPCA Full Search* et *FPCA Forward Search* en comparaison avec la sélection descendante de nœuds en régression spline classique.

En terme de réduction dimensionnelle, une classe de splines constituant un concurrent naturel à la méthode proposée sont bien évidement les splines pénalisées (P-splines) qui, partant également d'un ensemble de nœuds situés sur les quantiles de la distribution de la covariable continue, réduisent à un nombre "effectif" de paramètres la complexité du modèle par le biais d'une pénalisation de la courbure. Nous avons utilisé, à cet effet, dans les simulations le package R `smoothHR` avec comme critère de sélection du paramètre de lissage optimal le critère de AIC conseillé dans de précédents travaux de simulations, voir [Malloy et al., 2009] par exemple.

Enfin, le critère des moindres carrés empirique normalisé a été utilisé pour comparer la qualité de l'estimation de la relation dose–effet fournie par chaque méthode :

$$MSE = \frac{\sum_i (r(z_i) - \hat{r}(z_i))^2}{\sum_i r^2(z_i)} \tag{4.19}$$

Ce critère est usuellement utilisé dans ce genre de simulation dans la mesure où il permet de quantifier le compromis biais/variance.

4.5.3 Résultats des simulations

Cette section présente les résultats des simulations menées sous les cinq scénarios ci-dessous :

$r(\texttt{Dose}) = \log(1 + 0.4 \times \texttt{Dose})$	HR linéaire
$r(\texttt{Dose}) = \log(1 + 2.8 \times \texttt{Dose} \times \exp(-0.05 \times \texttt{Dose}))$	HR linéaire exponentiel
$r(\texttt{Dose}) = \log\left(1 + 0.01 \times \texttt{Dose} + 0.008 \times \texttt{Dose}^2\right)$	HR linéaire quadratique
$r(\texttt{Dose}) = 3/(1 + \exp(-0.5 \times (\texttt{Dose} - 10)))$	HR log-logistique
$r(\texttt{Dose}) = 0.6 \times \left(2.4 \times \texttt{Dose} \times \exp(-0.2 \times \texttt{Dose}) + 0.0011 \times \texttt{Dose}^2\right)$	HR bimodal

Pré-sélection des nœuds

La Figure 4.3 illustre la capacité spatio-dépendante de la méthode de filtrage spectral présentée dans la section 4.4.1. Elle permet en effet de pré-sélectionner les nœuds en tenant compte des changements de courbure de la fonction dose–effet *vraie*.
Par exemple, les distibutions des nœuds intérieurs pré-sélectionnés lors du scénario HR bimodal et linéaire exponentiel présentent une concentration de nœuds dans la région des maxima, particulièrement pour le scénario bimodal où le pic est plus prononcé sur un intervalle de valeurs de `Dose` plus réduit.

Le cas du HR log-logistique est également intéressant dans la mesure où la distribution des nœuds *candidats* est concentrée autour du point d'inflexion de la sigmoïde.

L'effet de la réduction dimensionnelle dans cette phase de pré-sélection peut être intuitivement évalué à partir du nombre de nœuds *candidats* retenus par le filtrage spectral. Rappelons que ce filtrage spectral est construit à partir de la rotation des premières fonctions propres de variabilité expliquée cumulée de 95%. Dans nos simulations, cela concernait en moyenne 7 à 8 fonctions propres seulement ; un chiffre à rapporter aux 3000 fonctions $\left\{\Phi_{B(z_i)}\right\}$ faisant l'objet de l'étude spectrale.
A partir des 30 quantiles à l'origine de la construction de la base $(B_k)_{1 \leq k \leq K}$, le nombre median de nœuds *candidats* sur l'ensemble des simulations et des scénarios est compris entre 5 et 8 le maximum étant 11. Le plus grand (resp le plus petit) nombre de nœuds pré-sélectionnés est obtenu à l'issue du scénario linéaire exponentiel (resp quadratique).

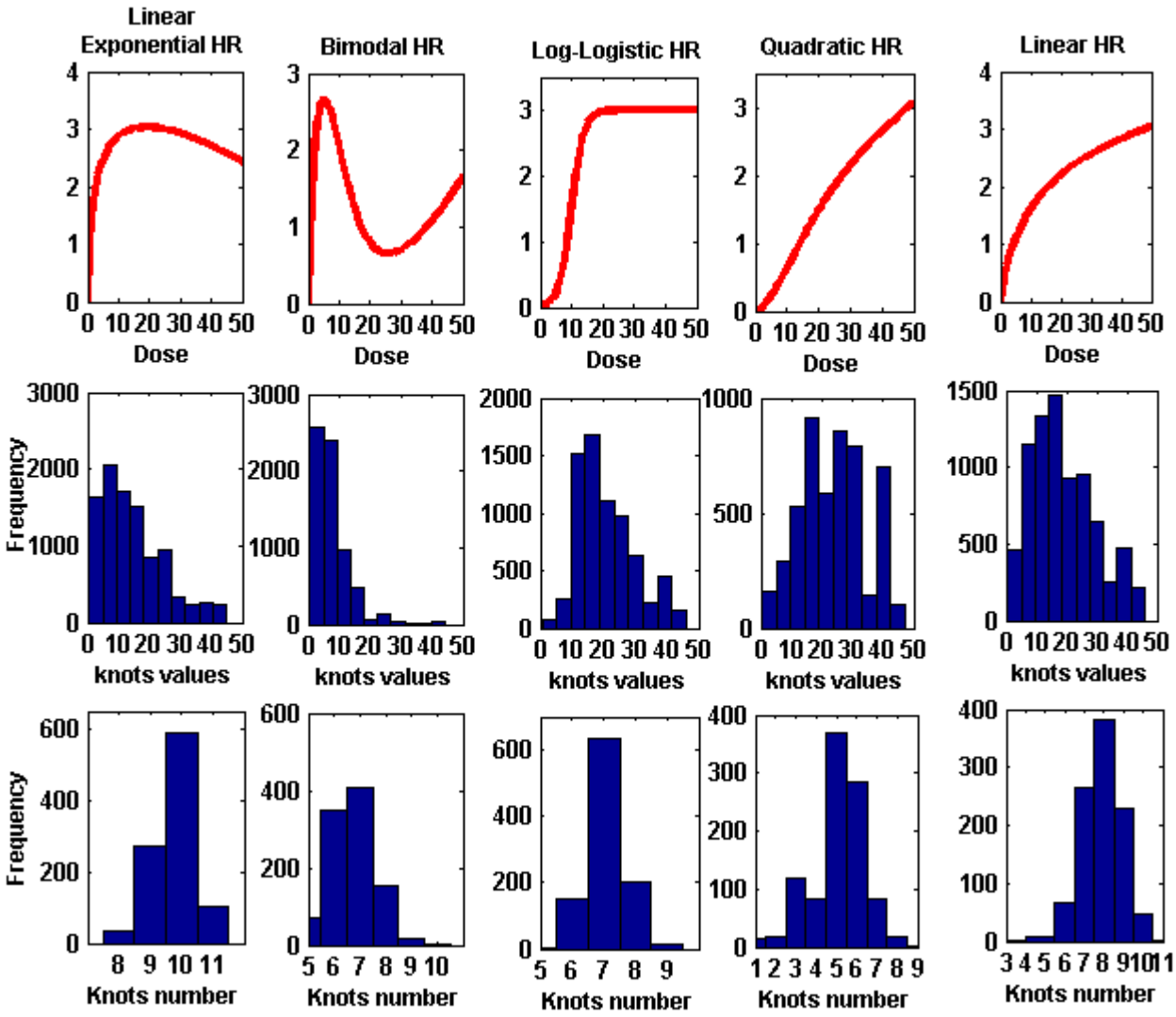

FIGURE 4.3 – Noeuds *candidats* pré-sélectionnés à partir de cinq scénarios simulés. les fonctions de risques sont représentées dans la ligne du haut. La ligne du milieu donne les histogrammes de répartitions des valeurs des nœuds pré-sélectionnés et la dernière ligne donne la distribution de leurs nombres sur l'ensemble des 1000 cohortes simulées.

Sélection finale du modèle

Dans la Figure 4.4 sont présentés les résultats de la sélection finale de nœuds obtenue par les algorithmes 1 et 3 présentés dans la section 4.4.2.

La différence majeure entre l'approche régression spline classique et l'approche proposée apparaît dès lors très clairement : l'algorithme *PFCA Forward Search* sélectionne *davantage de nœuds intérieurs* et bien *moins de paramètres à estimer* que la régression spline classique *Backward Search*.

Par exemple, le nombre de nœuds intérieurs utilisés par la méthode que nous proposons est *strictement* plus grand dans 70% des cohortes simulées sous le scénario Log-Logistique. C'est également le cas dans 50% des cohortes simulées sous le scénario bimodal, et dans un tiers des cas pour les autres scénarios simulés.

D'autre part, l'effet de la réduction dimensionnelle proposée dans la section 4.4.2 est manifeste lorsqu'on analyse le nombre de paramètres estimés par chaque modèle : sur l'ensemble des scénarios et l'ensemble des cohortes simulées, les modèles construits à partir de l'approche proposée comptent un nombre inférieur (resp strictement) de paramètres dans 98% (resp 80%) des cas.

La Figure 4.5 présente les courbes des estimateurs des fonctions dose–effet des cinq scénario simulés. La moyenne des 1000 cohortes simulées est à chaque fois représentée en gras, et en épaisseur simple les intervalles de confiance construits ponctuellement à partir du $5^{ème}$ et du $95^{ème}$ centile. Visuellement, nous pouvons remarquer que les fonctions dose–effet estimées à partir des méthodes de régression spline présentées en 4.4.1 paraissent moins biaisées que celles estimées par la méthode P-splines. En revanche, cette dernière semble fournir des estimateurs à variabilité plus réduite si l'on se réfère aux intervalles de confiance (à l'exception peut être du scénario bimodal).

Dans le but de préciser davantage le compris biais/variance de chaque méthode, les boites à moustache des MSE de chaque scénario sont représentées dans la Figure 4.6. Aucune des méthodes ne semble l'emporter sur l'ensemble des scénarios simulés : la méthode proposée possède un meilleur MSE dans trois scénarios (log-logistique, lineaire exponentiel et bimodal) alors que la régression spline usuelle et les P-splines sont meilleures dans un seul chacune (le HR linéaire et quadratique respectivement).

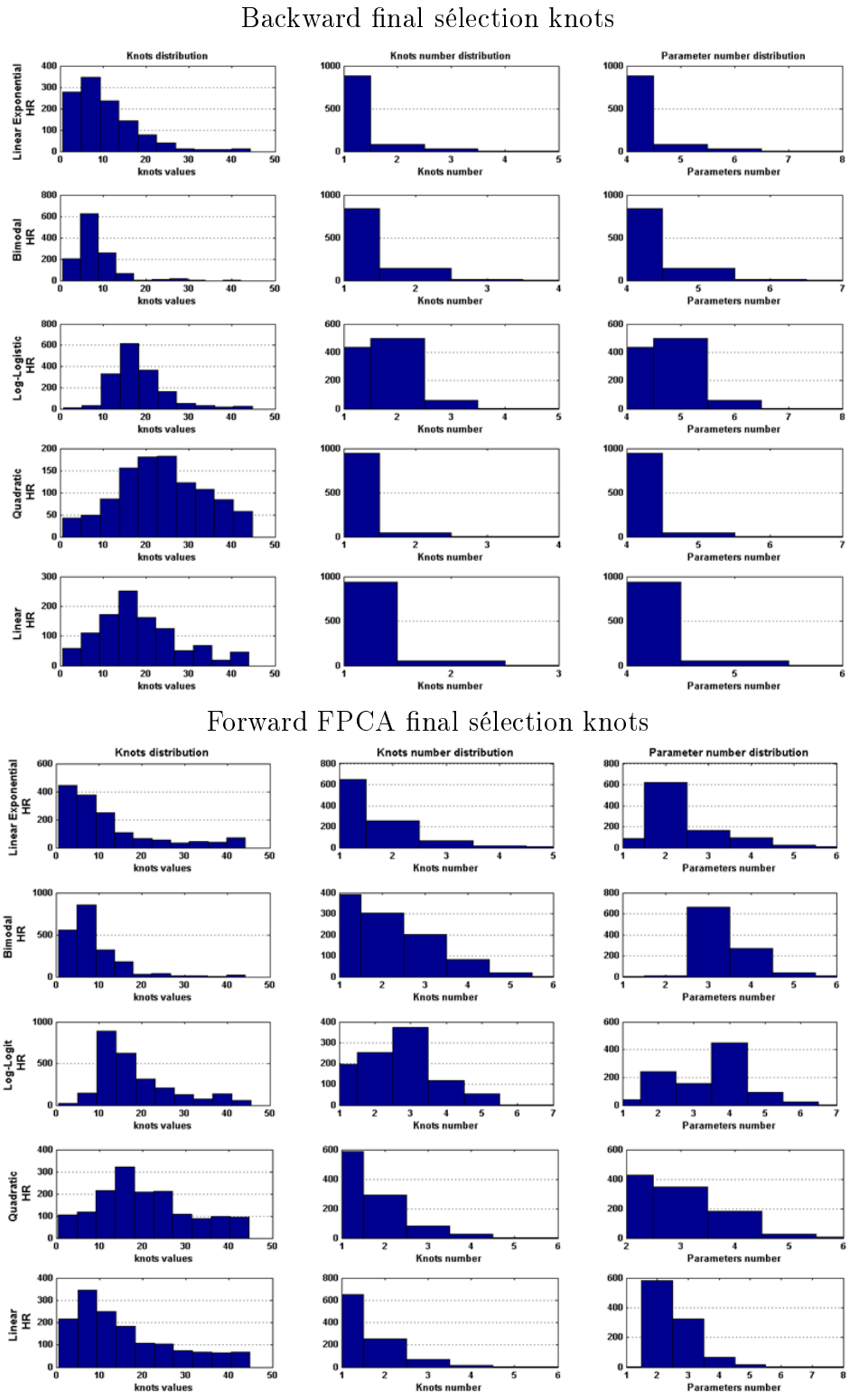

FIGURE 4.4 – Sélection finale de modèle sur les cinq scénarios simulés selon les algorithmes *Backward Search* and *FPCA Forward Search*présentés dans la section 4.4.2). Les lignes représentent les scénarios : la colonne de droite montre la distribution des nœuds, celle du milieu la distribution de leurs nombres, et enfin, celle de gauche la distribution du nombre de paramètres splines estimés dans le modèle.

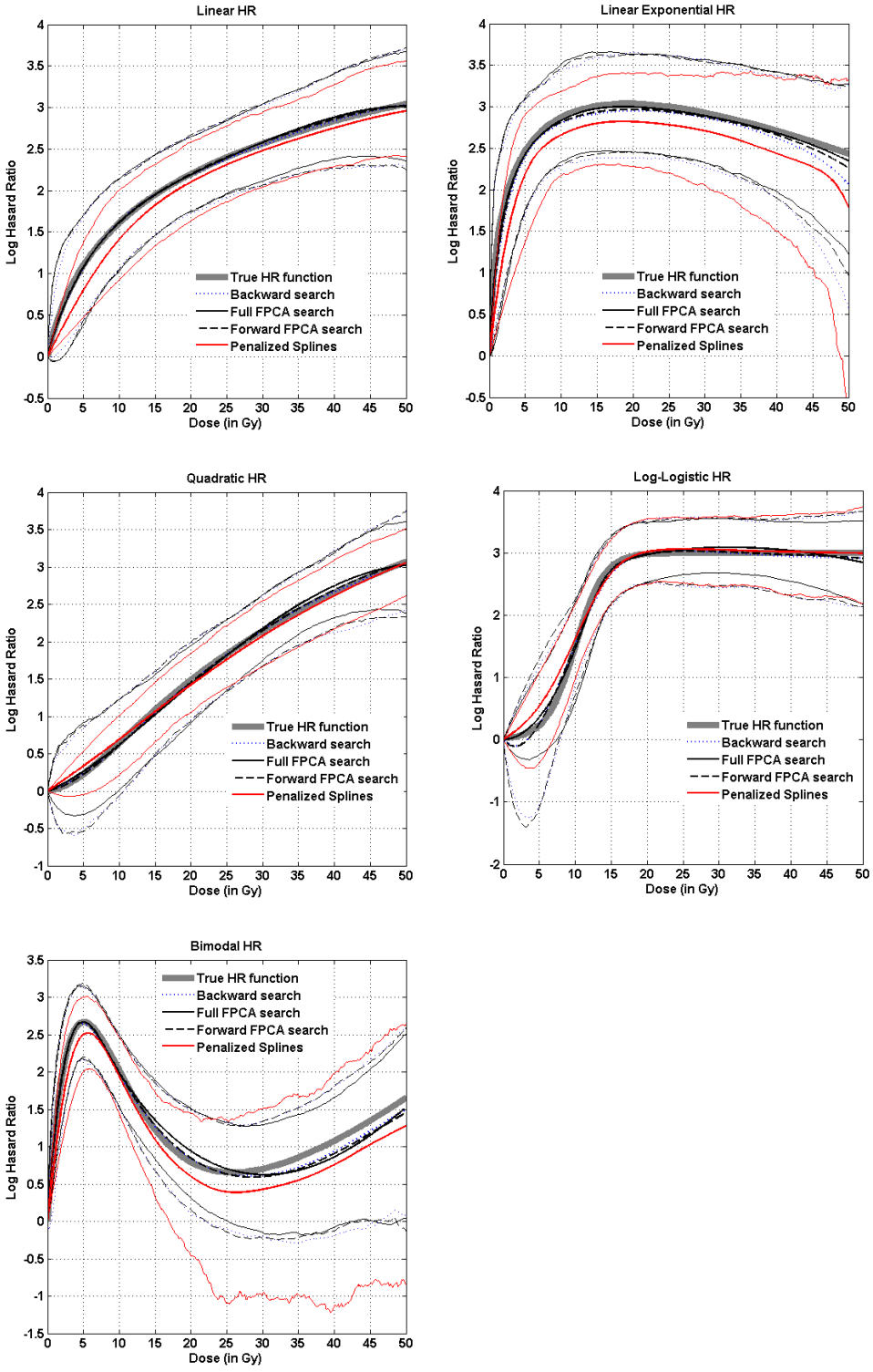

FIGURE 4.5 – Estimation de la vraie relation dose-effet (en gris) par différentes méthodes splines : la régression spline classique avec des nœuds intérieurs sélectionnés par la méthode *Backward* (en vert), les deux méthodes proposées *Full* et *forward* FPCA (en lignes solides et pointillés respectivement) et les splines pénalisées (en rouge). Dans chaque cas, les lignes épaisses sont les estimations moyennes sur les 1000 simulations et, en lignes fines, les intervalles de confiance ponctuels de niveau 95%.

Les algorithmes *FPCA Full Search* et *FPCA Forward Search* présentent des performances similaires excepté pour le scénario log-logistique dans lequel la recherche jointe présente de meilleurs résultats notamment en sélectionnant un nombre de nœuds plus important situés au niveau du point d'inflexion de la sigmoïde avec un nombre de paramètres à estimer plus réduit. C'est donc un exemple dans lequel la recherche marginale du *FPCA Forward Search* bien que bien plus rapide semble arrêter le processus d'optimisation prématurément.

La pré-sélection spatio-adaptative proposée dans la section 4.4.1 a permis aux méthodes basées sur les splines de régression d'être plus performantes que les P-splines. C'est particulièrement le cas pour le scénario linéaire exponentiel pour lequel les méthodes reposant sur la réduction dimensionnelle via analyse en composantes principales fonctionnelles offrent un meilleur MSE dans 80% des cohortes simulées. Dans le scénario quadratique en revanche, la méthode proposée n'est meilleure que dans 22% des cohortes simulées.

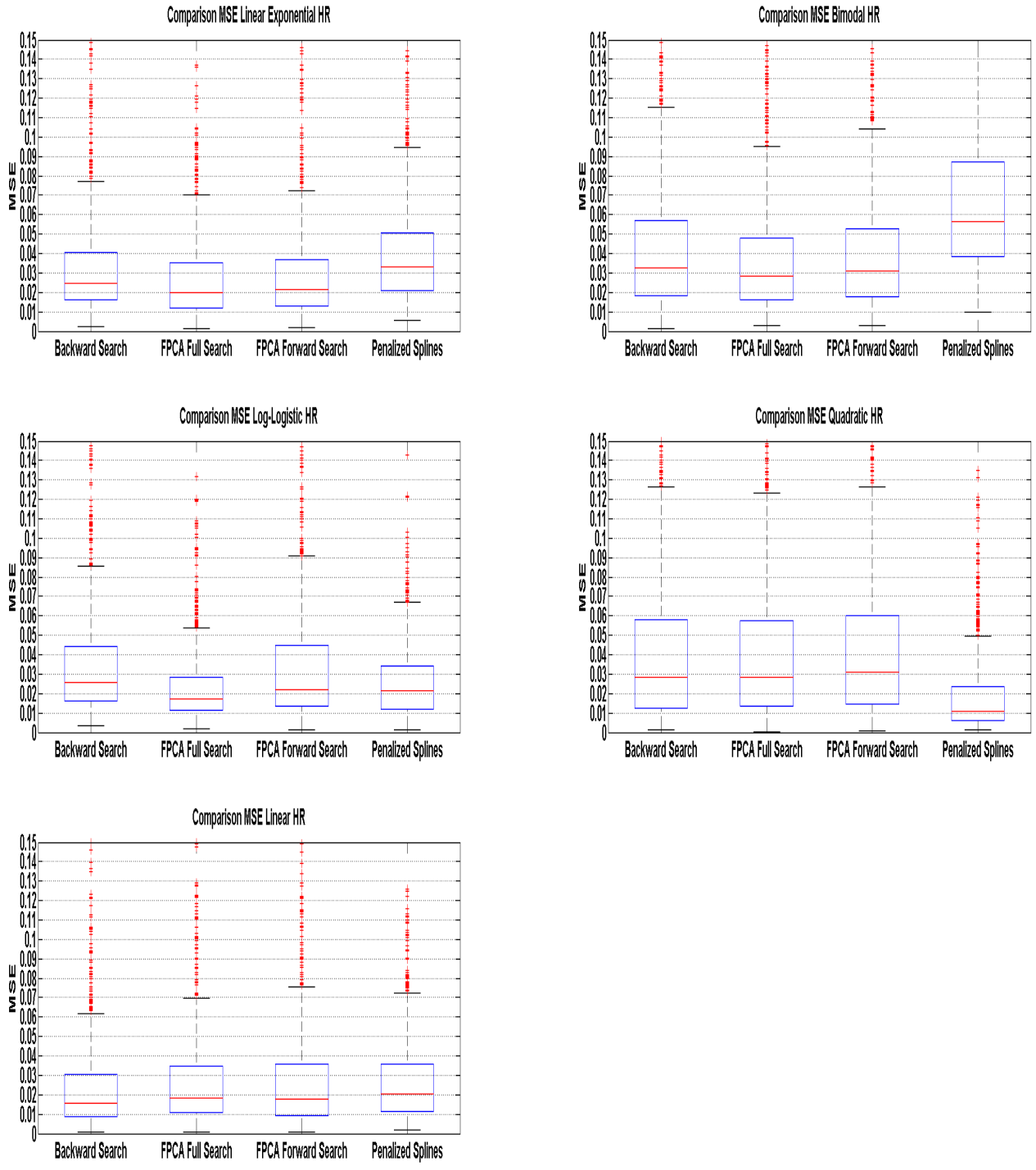

FIGURE 4.6 – Boîtes à moustaches des erreurs de moindres carrés MSE pour chaque scénario simulé et chaque algorithme présenté dans la section 4.4.2 ainsi que les P-splines.

4.6 Application : Analyse des données thyroïde

4.6.1 Les données thyroïde EURO2K

L'équipe d'épidémiologie des radiations CESP 1018 INSERM dispose des données d'une cohorte hospitalière, la cohorte Euro2K, constituée de 4 590 sujets traités pour une tumeur solide ou un lymphome entre 1942 et 1986 avant l'âge de 18 ans, dans 5 centres anticancéreux en France et 3 centres en Grande Bretagne et ayant survécu au moins 2 ans (3 ans dans les centres britanniques) à leur cancer initial.
Cette cohorte a été mise en place afin quantifier les effets iatrogènes à long terme des traitements des cancers chez les enfants.

La constitution de la cohorte eu lieu de 1985 à 1995 à partir des dossiers médicaux des centres. Une estimation des doses de rayonnements ionisants reçues à tous les organes situés dans ou hors des champs de radiothérapie a été réalisée entre 1992 et 1997. L'étude des cancers secondaires, a commencé en 1995, à partir du seul suivi recueilli dans les dossiers médicaux. En 2005, les données ont été mises à jour sur la base de près de 2000 questionnaires retournés.

Euro2K a donné un grand nombre de résultats sur le risque de cancers après traitement par radiothérapie pour les cancers secondaires radio-induits du sein [Guibout et al., 2005a], du cerveau [Little et al., 1998], des sarcomes de l'os [Le Vu et al., 1998, Schwartz et al., 2014], des sarcomes des tissus mous [Menu-Branthomme et al., 2004a], des mélanomes[Guérin et al., 2003] et des leucémies [Allard et al., 2010b, Haddy et al., 2006] ainsi que de la thyroïde [de Vathaire et al., 1999a, Kovalchik et al., 2013, Haddy et al., 2012b, Haddy et al., 2009a, Rubino et al., 2002].

La méthode de régression spline que nous proposons a été appliquée sur une cohorte de 3289 patients d'EURO2K traités pour un cancer de l'enfant avant l'âge de 16 ans et antérieur à 1986 uniquement dans cinq centres en France.
Trois variables catégorielles ont été incluses dans l'analyse : sexe, chimiothérapie (0,1 ou plus de deux drogues) et le type du premier cancer (maladie d'Hodgkin, tumeurs cérébrale, autres) et deux variables continues : l'âge au premier cancer et la dose de radiations à la thyroïde.
La dose de radiation à la thyroïde est définie par la moyenne des doses de radiations reçues à l'isthme et aux deux lobes.

Nous avons donc appliqué l'analyse en composantes principales fonctionnelles pour analyser la relation dose–effet entre la dose de radiations à la thyroïde et le risque de tumeurs thyroïdiennes.

4.6.2 Résultats

Caractéristiques des patients

Les caractéristiques des patients EURO2K sont présentées dans la table 4.1. Au sein de la cohorte, 116 (3.5% du total) patients ont développé une tumeur de la thyroïde parmi lesquels 106 exposés aux radiations.

Caractéristiques	Nb de patients (% du total)
Sexe	
Hommes	1854(56%)
Femmes	1435(44%)
Suivi moyen (min-max)	23ans(2-61ans)
Age moyen premier cancer (min-max)	5ans(1-16ans)
Type de premier cancer	
Maladie d'Hodgkin	223(16%)
Tumeur cérébrale	514(7%)
Autre	2552(77%)
Chimiothérapie	2325(71%)
Nombre de drogues	
0	964(29%)
1	305(9%)
≥2	2020(62%)
Dose moyenne à la thyroïde (écart-type)	4Gy(9Gy)
Tumeurs secondaires de la thyroïde	118 chez 116 patients(3.5%)
Adénomes thyroïdiens	75
Carcinomes thyroïdiens	53

TABLE 4.1 – Caractéristiques cliniques de la partie française de la cohorte EURO2K

Pré-sélection des nœuds dans les données thyroïde

Le processus de pré-sélection des nœuds *candidats* est initié à partir d'une base B-spline construite à partir de 30 équi-quantiles de la distribution des doses de radiations chez la population exposée de la cohorte. Une analyse en composantes principales fonctionnelles a ensuite été menée sur les fonctions $\left\{\Phi_{B(z_i)}\right\}$ définies par l'équation (4.6). Plus de 95%

de la variabilité expliquée cumulée des $\left\{\Phi_{B(z_i)}\right\}$ est atteinte en seulement huit fonctions propres.

La Figure 4.7 montre le processus de pré-sélection des nœuds intérieurs *candidats* impliquant les huit premières composantes principales perçues comme un filtre à bande passante mi-hauteur par lequel sont passés les trente quantiles. Seuls sont retenus les nœuds issus du filtrage par des rotations des fonctions propres présentant une association significative avec le risque de tumeur thyroïdienne. Ainsi, les nœuds filtrés par la rotation de la seconde et quatrième composante principale ne sont pas retenus. A l'issue de cette étape, onze nœuds intérieurs potentiels sont retenus pour l'étape de sélection finale.

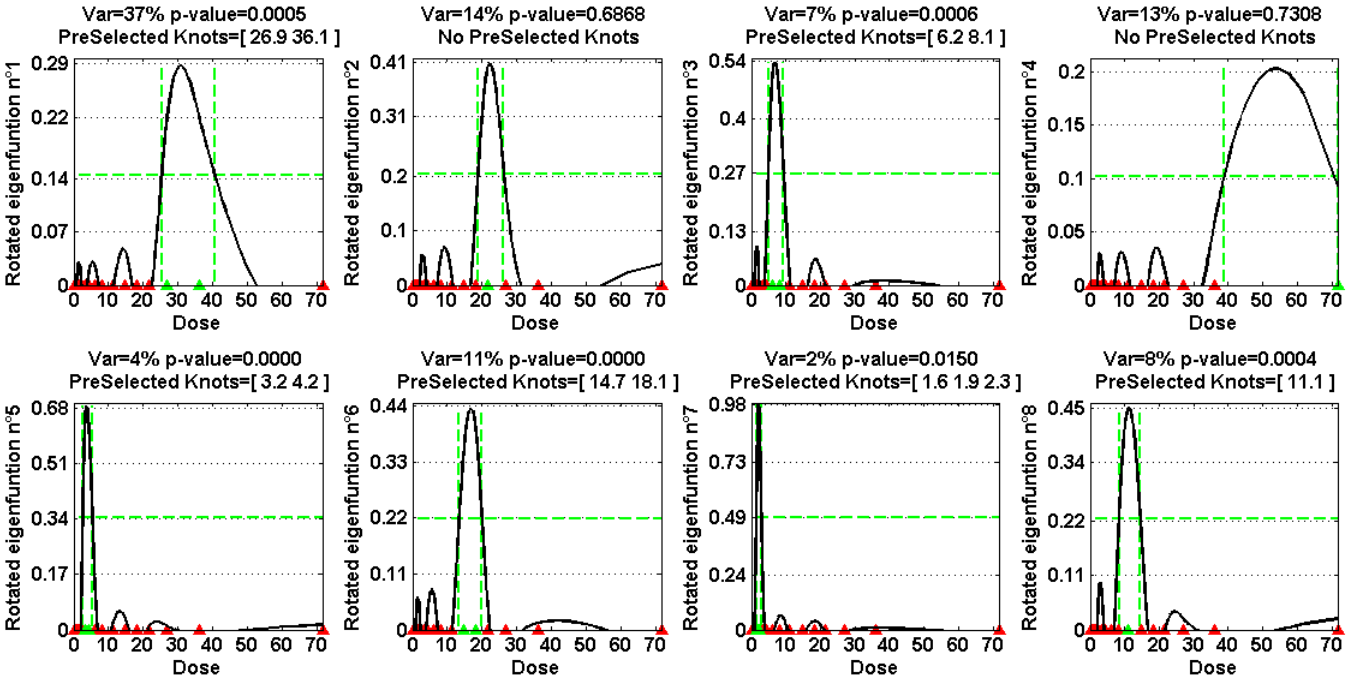

FIGURE 4.7 – Processus de filtrage spectrale des quantiles de la distribution de dose de radiation de la thyroïde. Pour chaque fonction propre, la variabilité expliquée (Var) et le niveau d'association avec le risque de tumeur thyroïdienne (p-value) sont donnés en haut de chaque cadre ainsi que les nœuds sélectionnés par filtrage.

Sélection finale des nœuds dans les donnés thyroïde

Les résultats de la sélection finale des nœuds intérieurs sont donnés dans le tableau 4.2.

Les algorithmes *FPCA Full Seach* et *FPCA forward Seach* sélectionnent des modèles splines à deux nœuds intérieurs ([4.2 36.1] and [6.2 36.1] respectivement) et un paramètre

spline estimé. C'est une illustration supplémentaire de la réduction dimensionnelle dont est capable l'approche proposée dans la mesure où, dans une régression spline classique, un modèle spline à deux nœuds intérieurs nécessiterait l'estimation de cinq paramètres (sous la restriction $r(0) = 0$). L'approche spline classique via l'algorithme *Backward Search* fournit, quant à elle, un modèle spline construit à partir d'un unique nœud intérieur et quatre paramètres estimés.

FPCA Full search			FPCA Forward search			Conventional Backward search		
Final knots	nb coeff	AIC	Final knots	nb coeff	AIC	Final knots	nb coeff	BIC
4.2 36.1	1	1586.3	6.2	2	1587.0	1.9 3.2 4.2 6.2 14.7 18.1	9	1626.3
6.2 36.1	1	1586.8	6.2 36.1	1	1586.8	3.2 4.2 6.2 14.7 18.1	8	1621.6
3.2 36.1	2	1586.9				4.2 6.2 14.7 18.1	7	1617.6
6.2	2	1587.0				6.2 14.7 18.1	6	1612.9
4.2 36.1	2	1587.2				6.2 18.1	5	1612.1
4.2	3	1587.4				6.2	4	1608.1

TABLE 4.2 – sélection du modèle spline final par chacun des trois algorithmes. Dans chaque cas sont donnés : la liste des nœuds retenue, le nombre de paramètres estimés et la valeur du critère AIC.

Dans ce type de régression spline spatio-adaptatif, il est important de noter que la sélection des nœuds et du nombre de paramètres est *données-dépendante*. En terme d'inférence, il est impératif de tenir compte de l'incertitude en terme de sélection de modèle et de ne pas seulement se contenter d'intervalles de confiance construits à partir de la normalité asymptotique des coefficients.
C'est pour cette raison que nous avons décidé de générer les intervalles de confiance par bootstrap (ré-échantillonnage avec remise) non paramétrique de la base thyroïde originale : sur chacune des 1000 copies bootstrap sont appliqués les trois algorithmes présentés en section 4.4.2 ; les intervalles de confiance sont par la suite dérivés en prenant ponctuellement les 2.5 % et 97.5% centiles.

Les relations dose–effet du risque de tumeurs thyroïdiennes radio-induites avec leurs intervalles de confiance bootstrap de niveau 95% sont représentées dans la Figure 4.8. La relation dose–effet obtenue en fonction de la dose de radiation à la thyroïde est de nature unimodale : le risque augmente avec la dose jusqu'à un maximum puis décroît pour les plus fortes doses. Les estimateurs issus de l'analyse en composantes principales fonctionnelles donnent des courbes très similaires situant le maximum du risque autour de 20 Gy. La régression spline classique situe quant à elle ce maximum autour de 30 Gy. Le critère AIC des régressions issues des algorithmes *FPCA Full Search*, *FPCA Forward Search*

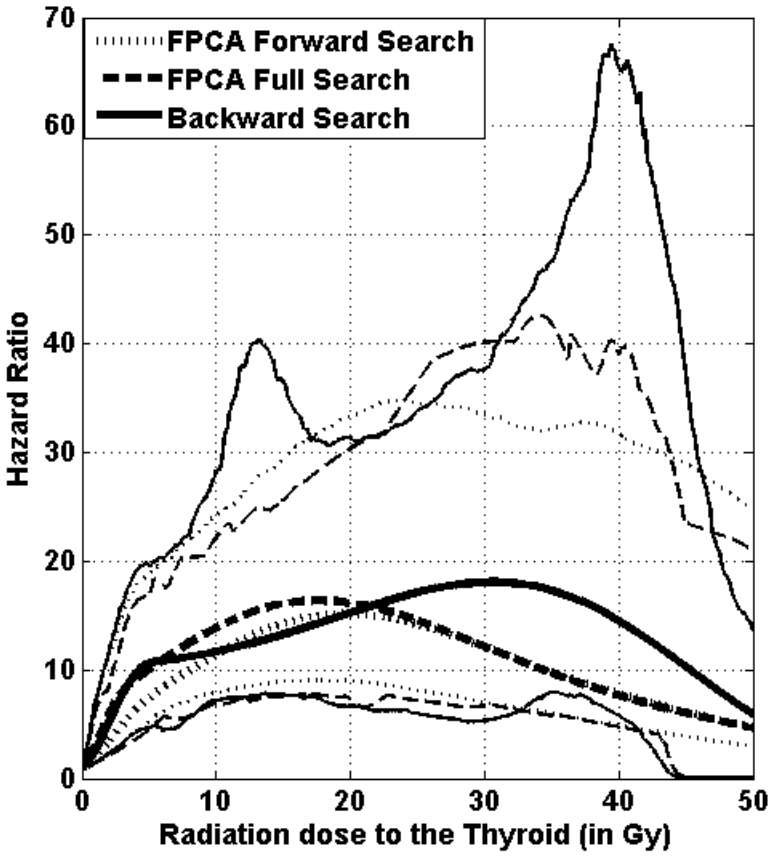

FIGURE 4.8 – Risque de tumeurs thyroïdiennes comme fonction de la dose de radiation à la thyroïde en utilisant les algorithmes *FPCA Forward Search* (pointillés), *FPCA Full Search* (tirés), et *Backward Search* (ligne pleine). En gras, l'estimation de la courbe dose–effet et en épaisseur simple les intervalles de confiance bootstrap.

et *Backward Search* vaut respectivement 1586.3, 1586.8 et 1588.8. Notons enfin qu'au regard des intervalles de confiance bootstrap, la méthode proposée offre une variabilité plus réduite que la régression spline classique, ce qui est sans doute la conséquence de la réduction dimensionnelle synonyme d'un nombre de coefficients à estimer sensiblement plus faible.

4.7 Discussion

Nous avons donc proposé dans ce chapitre un nouveau modèle flexible d'analyse de survie afin de décrire l'effet d'une covariable continue sur le risque d'incidence d'un événement d'intérêt. Contrairement à la plupart des méthodes de régression splines utilisant

la représentation en polynômes tronqués, notre méthode consiste à pleinement exploiter les avantages bien connus de la forme B-spline et particulièrement, *la propriété de support minimal*.

Pour une base fixée, le point clé de notre approche réside dans la construction de $\Phi_{B(z_i)}$ de l'équation (4.6), une fonction qui, intuitivement, peut être perçue comme un centre de masse des éléments de la base pondérés par leurs poids en terme de *support minimal* au voisinage de chaque valeur z_i de la covariable d'intérêt.

Une analyse spectrale de l'opérateur de covariance des fonctions $\left\{\Phi_{B(z_i)}\right\}$ est, par la suite, menée afin de déterminer les principales directions fonctionnelles de variabilité incarnées par les *fonctions propres* $\{\psi_l(u)\}_l$ dans l'équation (4.8). La réduction dimensionnelle du problème en découle directement dans la mesure où elle permet de décrire l'ensemble des fonctions $\left\{\Phi_{B(z_i)}\right\}$ par un nombre réduit de $\{\psi_l(u)\}_l$.

La première application de cette approche issue de l'analyse de données fonctionnelles est la sélection spatio-adaptative d'un ensemble potentiel de *nœuds candidats* au voisinage desquels sont sensés être capturées les principales variations de la fonction dose–effet. Partant d'une base B-spline construite à partir d'un ensemble dense de quantiles, l'analyse en composantes principales permet de sélectionner un nombre réduit de fonctions propres $\{\psi_l(u)\}_{1\leq l\leq L}$ sur un critère de variabilité expliquée cumulée par exemple. Ces fonctions $\{\psi_l(u)\}_{1\leq l\leq L}$ subissent par la suite une rotation Varimax afin d'identifier sur quels intervalles elles expriment le mieux la variabilité de la *propriété de support minimale*.
Le modèle (4.12) construit à partir des scores $\left(\theta_{il}^{Rot}\right)$ des fonctions propres résultant de cette rotation notée $\left\{\psi_l^{Rot}(u)\right\}_{1\leq l\leq L}$, permet d'identifier les intervalles sur lesquels le support minimum de la base B-spline est le plus corrélé avec l'événement d'intérêt. Les nœuds potentiels sont alors sélectionnés par un processus de filtrage dans lequel les fonctions $\left\{\psi_l^{Rot}(u)\right\}_{1\leq l\leq L}$ sont perçues comme des filtre à bande mi-hauteur comme illustré dans la Figure 4.3.

Pour obtenir des estimations stables en données de survie, i.e en présence de censures, il est recommandé de restreindre le rapport entre nombre d'événements et nombre de paramètres estimés dans le modèle. Ainsi une régression spline classique se doit de réduire son nombre de nœuds intérieurs dans la mesure où le nombre de coefficients splines en dépend directement via la relation (4.4). Cette contrainte peut donc être à l'origine d'un déficit de flexibilité qui pourrait être préjudiciable dans la description de certaines relations dose–effet complexes.

La méthode présentée propose donc une solution à cette problématique. Une fois une séquence de nœuds intérieurs fixée, les fonctions $\left(\Phi_{B(z_i)}\right)_{1\leq i\leq N}$ peuvent être construites afin d'explorer la variabilité de la propriété de support minimum de la base B-spline construite. L'analyse en composantes principales offre alors la possibilité de remplacer la base spline initiale par les fonctions scores qui résument une grande partie de la propriété de support minimal induite par la séquence de nœuds intérieurs.

Les résultats de simulations indiquent que les modèles construits à partir de ces fonctions scores permettent de reconstruire une variété de profils de relation dose–effets cliniquement plausibles et offrent une qualité de régression en terme de MSE (équation (4.19) meilleure que les méthodes splines classiques. A l'exception du scénario quadratique, cette conclusion est également valable en comparaison aux P-splines.

Nous avons proposé deux algorithmes *FPCA Full Search* and *FPCA Forward search* qui optimisent de façon jointe et marginale respectivement. L'algorithme *FPCA Forward search* est bien plus rapide et donne des résultats très proches du *FPCA Forward search* à l'exception du scénario log-logistique. Ainsi, la différence entre les deux algorithmes semble plus importante pour des relations dose–effet qui nécessitent davantage de nœuds intérieurs dans certaines zones de modifications de courbure, comme le point d'inflexion d'une sigmoïde. Dans ce cas, l'algorithme *FPCA Forward Search* semble stopper sa recherche avant de sélectionner le nombre de nœuds intérieurs adéquat.

Le fait de travailler avec des splines cubiques rend plus compliquée la reconstruction de relations dose–effet présentant un profil avec des lignes droites, ce qui expliquerait les performances mitigées des méthodes régressions splines pour le scénario quadratique comparé aux P-splines.
Une solution possible serait d'abord de mener une analyse avec des splines linéaires afin de mettre en évidence une relation de type linéaire entre le logarithme du HR et la covariable d'intérêt dans le modèle de Cox ; Les splines cubiques pouvant par la suite être introduits pour davantage de flexibilité.

L'application aux données thyroïde illustre les nouvelles possibilités offertes par les modèles issus de l'analyse en composantes principales fonctionnelles. La forme unimodale de relation dose-effet obtenue confirme les résultats rapportés dans la littérature : dans les années 90, les premières relations dose–effet que l'on y trouve décrivaient une

relation linéaire avec la dose de radiation avec un risque plus élevé pour ceux exposés durant l'enfance comme dans [Ron et al., 1995]. Cependant, l'analyse de Ron *et al* comportait en majorité des gammes de doses relativement faibles et les études les plus récentes incluant des patients en milieu médical exposés à des doses fortes au delà des 15-20 Gy a permis de mettre en évidence une croissance du risque de tumeurs thyroïdiennes jusqu'à 10Gy puis un plateau, voire un déclin, à partir des 20-25Gy, en général attribué à un phénomène de mort cellulaire [de Vathaire et al., 1999b, Ronckers et al., 2006, Veiga et al., 2012, Haddy et al., 2012a].

L'analyse de notre base de données thyroïde a révélé l'efficacité de la méthode proposée en terme de réduction dimensionnelle où un seul paramètre a permis l'estimation de la relation dose–effet contre quatre pour la régression spline classique avec sélection descendante des nœuds. Cependant, l'inférence de ces modèles doit tenir compte non seulement de l'incertitude sur les coefficients estimés, mais aussi celle relative à la section des modèles. C'est pourquoi, la construction des intervalles de confiance par Bootstrap semble dans ce contexte plus indiquée. Il ressort enfin de cette analyse que la réduction du nombre de paramètres estimés induit un intervalle de confiance plus réduit que ceux de la régression spline usuelle, réduisant de fait les incertitudes autour des indicateurs de risque obtenus.

Notre expérience dans l'analyse de données d'épidémiologie des radiations en milieu clinique a montré que la variance d'un modèle spline construit à partir de deux à six nœuds intérieurs est à 95% captée par au plus trois fonctions propres. Ainsi, les fonctions propres à faible variabilité explicative tendent à augmenter et apparaissent dès lors comme *un bruit* dans le processus d'estimation. Le fait de réduire le nombre de fonctions propres, et, par conséquence, le nombre de scores dans les modèles 4.10 peut donc être perçu comme un paramètre de régularisation contre le risque d'*overfitting*.

La régression spline issue de ce travail diffère ainsi de l'approche classique dans le sens où elle propose de dissocier le nombre de nœuds intérieurs du nombre de paramètres à estimer dans le modèle, en se proposant de garder l'information fournie par les nœuds tout en limitant le nombre de paramètres à estimer à l'essentiel. Ce lien entre la régression flexible de données de survie et l'analyse de données fonctionnelles devrait être approfondi dans des investigations futures.

Chapitre 5

Radiobiologie

Sommaire

5.1 Introduction 102

5.1.1 Contexte 102

5.1.2 Problématique 103

5.1.3 Présentation du projet EpiRadBio et Objectifs 105

5.2 La cohorte ANGIO 107

5.2.1 Présentation de la cohorte 107

5.2.2 Design de l'étude 108

5.3 Analyse en composantes principales multiniveau 111

5.3.1 Définition 111

5.3.2 Valeurs propres et vecteurs propres 112

5.3.3 Estimation des scores de l'analyse en composantes principales multiniveau 113

5.4 Résultats 116

5.5 Discussion 120

5.1 Introduction

5.1.1 Contexte

Plusieurs facteurs influencent la probabilité qu'un patient développe des complications à court et long terme après exposition aux rayonnements ionisants.
Ces facteurs sont largement liés à la physique (par exemple, la dose de rayonnement et le volume irradié), aux traitements concomitants (chirurgie, chimiothérapie), aux antécédents et mode de vie du patient (âge, tabagisme, taux d'hémoglobine, facteur de comorbidité tels que le diabète, l'hypertension, les maladies vasculaires et des tissus conjonctifs). En plus de ces facteurs, il est également important de considérer les différences inhérentes à la sensibilité génétique aux rayonnements ionisants.
Le problème qui se pose est alors de quantifier avec précision les facteurs de risque non génétiques de telle sorte que l'influence de la génétique puisse être détectée de façon fiable.

Dans la pratique clinique de la radiothérapie, le concept de radiosensibilité individuelle prend une place centrale.
Par exemple, des patients hautement radiosensibles sont susceptibles de développer des effets radio-induits précoces ou tardifs tandis que le contrôle tumoral chez des patients radio-résistants peut nécessiter des doses supérieures à celles prévues par le protocole adopté.

Ainsi, une méthode clinique rapide et fiable pour une radiothérapie personnalisée déterminant à la fois le niveau de radiorésistance (respectivement radiosensibilité) des tumeurs (respectivement des tissus sains) reste encore à établir [Andreassen et al., 2012, Fernet and Hall, 2004].

Un des biomarqueurs de radiosensibilité les plus étudiés est la longueur de télomères (LT) [Pernot et al., 2012, Genesca et al., 2006]. Une revue de la littérature très complète à ce propos peut être trouvée dans [Shim et al., 2014] dont plusieurs éléments ont été puisés pour cette mise en contexte.
Les télomères sont des structures nucléoprotéiques, coiffant les chromosomes eucaryotes, qui sont essentielles dans le contrôle de la prolifération cellulaire et la stabilité du génome. L'ADN télomérique se raccourcit à chaque division cellulaire, car la réplication des extrémités d'une molécule d'ADN linéaire est incomplète. Cette dégradation peut être compensée grâce à une enzyme de type transcriptase inverse, la télomérase [Cech, 2004].

La variabilité inter-individuelle de la longueur de télomère est fortement liée à des

facteurs génétiques et environnementaux [Gilson and Londono-Vallejo, 2007], et des observations ont permis d'établir que les personnes à télomères courts présentaient davantage de signatures de cassures radio-induites (par exemple en fréquence de micro-noyaux) de l'ADN que les personnes à longs télomères [Castella et al., 2007]. La corrélation entre LT et radiosensibilité a été en effet observée dans de nombreuses études in vivo et in vitro sur des cellules animales et humaines [Bouffler et al., 2001]. Ces études ont établi plus précisément une corrélation inverse entre la sensibilité aux rayonnements et la LT moyenne, à savoir que le raccourcissement des télomères améliorait la radiosensibilité sans pour autant conclure à la radiorésistance en cas de télomères longs [Rubio et al., 2002].

D'autres études suggèrent quant à elles que, à niveau de télomérase radio-induit égal, la radiosensibilité est davantage liée à la proportion de perte de télomères et non à la longueur moyenne [Berardinelli et al., 2012]. L'inhibition de l'activité de la télomérase a également été, dans les cellules humaines, fortement corrélée à la radiosensibilité, avec, comme conséquence, la présence de télomères courts à l'origine de fusions et réarrangements chromosomiques [Latre et al., 2003, Soler et al., 2009]. D'autre part, des cellules cancéreuses radio-résistantes ont montré une surexpression de l'activité de la télomérase [Zhou et al., 2010, Kurvinen et al., 2006] leur permettant de rallonger préférentiellement les télomères courts quasi-dysfonctionnels [Rubio et al., 2004], contrecarrant ainsi les processus de mort cellulaire radio-induits[Gorbunova et al., 2002, Pirzio et al., 2004].

Ainsi, l'usage d'inhibiteurs de télomérase durant la radiothérapie pourrait radio-sensibiliser les cellules tumorales assurant ainsi une plus grande efficacité au traitement [Nakamura et al., 2005, Hauguel and Bunz, 2003, Wesbuer et al., 2010, Kovalenko et al., 2010].

5.1.2 Problématique

Nous l'avons vu, le raccourcissement ou l'allongement des télomères dans les cellules humaines normales est important à considérer car il peut induire des dysfonctionnements et instabilités chromosomiques, qui peuvent induire un processus de cancérogenèse. L'impact des rayonnements ionisants sur la longueur des télomères devient, dès lors, une question importante en épidémiologie des radiations.

L'effet des rayonnements ionisants sur la longueur des télomères humains semble être à la fois dépendant de la dose de rayonnement et du transfert d'énergie linéique (TEL) des particules en présence (TEL=énergie transférée au milieu cible par unité de longueur).

Par exemple, Nieri et al [Nieri et al., 2013] n'ont observé aucun changement dans la taille des télomères de fibroblastes humains AG1522 normaux 24h et 48h après irradiation à des doses comprises entre 0.1-1Gy, que ce soit avec des particules légères

comme des photons (bas TEL) ou lourdes (haut TEL). En revanche, il a été observé que l'irradiation à 4 Gy de fibroblastes humains HFFF2 normaux provoquait une élongation des télomères en cas d'irradiation protonique (haut TEL) tandis qu'aucun effet substantiel n'a été rapporté après une irradiation photonique à dose égale (4Gy) [Berardinelli et al., 2013, Berardinelli et al., 2011]. En revanche, une irradiation photonique in vitro de lymphocytes et fibroblates à haute dose (18-50Gy) a montré un raccourcissement des télomères à 24h post-irradiation [Li et al., 2012]. Une récente étude clinique, quant à elle, a montré qu'une irradiation photonique à 52Gy de lymphocytes patients avait entraîné une baisse de la proportion de cellules à télomères courts, tout en conservant une longueur moyenne de télomères stables avant/après irradiation [Maeda et al., 2013], ce qui pourrait suggérer une élongation par activation d'une télomérase radio-induite. Au-delà de 24h d'irradiation, les effets sur les longueurs de télomères semblent parfois s'inverser : après n'avoir observé aucun changement dans la longueur de télomères 24h après une irradiation photonique à 4 Gy, Berardinelli et al [Berardinelli et al., 2013, Berardinelli et al., 2011] ont signalé un raccourcissement après 4 jours puis un rallongement au bout de 15 jours post irradiation. Alors que la même dose protonique induisait un allongement qui semblait persister sur toute la période d'observation. Ce qui suggère donc que l'effet des radiations sur les télomères dépend de la complexité des lésions ADN infligées aux cellules qui sont de natures différentes entre les particules à bas et haut TEL. L'induction d'un allongement des télomères après une irradiation de 4Gy à haut EL dans des cellules déficientes en télomérase suggère également l'activation d'un mécanisme alternatif d'élongation de l'ADN télomérique, l'ALT, fondé sur des échanges non réciproques entre chromosomes.

En plus de ces résultats in vitro, il existe de nombreux résultats in vivo concernant la longueur des télomères. Citons, par exemple, l'effet de la chimiothérapie : le raccourcissement des télomères accompagné d'une réduction de l'activité de la télomérase a été observé chez des patients ayant reçu une chimiothérapie, par rapport à des échantillons prélevés avant exposition [Schroder et al., 2001]. En outre, par rapport aux témoins sains d'âge équivalent, des patients ayant reçu une chimiothérapie présentaient des télomères significativement plus courts, ce qui suggère que certaines drogues de chimiothérapie peuvent accélérer le raccourcissement naturel qui se produit avec le processus de vieillissement [Beeharry and Broccoli, 2005]. Dans une étude prospective de patients ayant survécu à une maladie de Hodgkin, ceux qui possédaient des télomères plus courts ou présentaient des réarrangements chromosomiques plus complexes accompagnés d'une radiosensibilité in vitro étaient les plus à risque de développer un second cancer [M'Kacher et al., 2007] ou une pathologie cardiaque [M'kacher et al., 2014]. Enfin, dans une récente étude, une association a également été décrite entre le raccourcissement de télomères et le cancer de

la thyroïde après cancer de l'enfant [Gramatges et al., 2014].

L'exposition aux faibles doses de rayonnements ionisants perturbe également le maintien des télomères in vivo, comme en témoignent des résultats sur les travailleurs de Tchernobyl exposés à de faibles doses de rayonnements ionisants où le raccourcissement télomérique a été constaté même 20 ans après l'exposition, suggérant ainsi un effet prolongé de la perte radio-induite des télomères [Ilyenko et al., 2011].

Cependant, les effets à long terme des faibles doses sur les longueurs télomériques sont largement inconnus et davantage d'études sont nécessaires pour mieux comprendre les mécanismes impliqués et l'impact des rayonnements ionisants sur leur hétérogénéité.

5.1.3 Présentation du projet EpiRadBio et Objectifs

EpiRadBio est un projet du septième programme-cadre de la Communauté européenne de l'énergie atomique ayant pour thème de recherche la fission nucléaire et la radioprotection (7ème PC - Euratom - Fission).

EpiRadBio associe l'épidémiologie à la radiobiologie dans le but de quantifier les risques de cancer après exposition aux faibles doses et aux faibles débits de doses de radiations ionisantes.
EpiRadBio étudie la carcinogénèse radio-induite en mesurant les instabilités génomiques de différents prélèvements biologiques (tissu et sang) de patients de cohortes déjà existantes : cohorte française de sujets traités pour un hémangiome durant l'enfance, cohorte des travailleurs de Mayak, cohorte de patients ayant développé un cancer de la thyroïde après l'accident de Tchernobyl.

Les communications intercellulaires après exposition à de faibles doses et l'influence sur l'apoptose, la prolifération cellulaire, la différenciation et l'instabilité génomique seront explorées grâce à des modèles de cultures cellulaires 2D et des modèles tissulaires 3D. Des travaux sur cellules souches isolées de tissus sains seront réalisés et les résultats permettront le développement de modèles de carcinogenèse dans des cohortes existantes : survivants des bombes nucléaires, cohorte française, suédoise et italienne du cancer de la thyroïde, cohorte des travailleurs de Mayak, cohorte suédoise des hémangiomes, cohorte sur le cancer de la thyroïde après l'accident de Tchernobyl, registre anglais des travailleurs exposés.

Le risque de cancer sera déterminé après exposition à de faibles débits de dose de rayonnement à bas transfert d'énergie linéique (rayonnement gamma externe et interne de 131I) et à haut TLE (particules alpha du plutonium). La prise en compte des facteurs de risques individuels permettra de définir de nouvelles limites de doses et de mieux estimer les risques liés à l'exposition médicale.

Au sein du projet EpiRadBio, notre équipe est impliquée dans deux Work packages : le Work Package 1 étudiant l'instabilité génomique et la susceptibilité individuelle aux rayonnements et le Work Package 3 dédié aux études de l'incidence de cancers radio-induits de la thyroïde, des poumons et du sein et sur l'évaluation des risques pour la radioprotection.

Cette partie du mémoire de thèse portera plus précisément sur la méthodologie employée dans le cadre du Workpackage 1.1 intitulé *Individual susceptibility to genomic instability : Epidemiology and radiation biology studies*. Ce workpackage est coordonné par le Commissariat à l'Energie Atomique et aux énergies alternatives (CEA), et dans lequel l'équipe 3 du CESP INSERM1018 participe à une étude fondée sur l'analyse d'échantillons de sang de patients traités pour un hémangiome (voir la section pour davantage de détails sur cette cohorte 5.2.

L'objectif est d'investiguer l'effet d'une exposition durant l'enfance aux faibles doses sur la longueur des télomères et leurs hétérogénéités permettant ainsi de détecter des signes d'instabilités chromosomiques radio-induites chez ces patients (perte des télomères, réarrangements chromosomiques, etc.).

La primeur des résultats sur l'ensemble des participants recrutés (au nombre de cent cinquante environ) étant réservée au congrès de clôture du projet en Mars 2015, les résultats de la méthodologie présentée ici seront ceux d'une étude de faisabilité menée sur un sous-groupe de vingt et une personnes.

5.2 La cohorte ANGIO

5.2.1 Présentation de la cohorte

La cohorte ANGIO est constituée de 8320 enfants traités, principalement par radiothérapie, pour un hémangiome cutané à l'Institut Gustave Roussy (IGR) de 1941 à 1973. Ces enfants ont été traités dans leur grande majorité avant l'âge d'un an, c'est-à-dire à un âge auquel la sensibilité aux rayonnements ionisants est très importante. La période 1941-1973 a été retenue car elle correspond à la période de traitement par radiothérapie des hémangiomes à l'IGR.

Cette cohorte a été mise en place par l'équipe 3 d'épidémiologie des radiations du CESP 1018 INSERM dans le but de quantifier les effets iatrogènes à long terme des traitements de radiothérapie administrés pour une pathologie bénigne.
L'objectif principal de l'étude était d'informer ces patients des traitements qu'ils avaient reçus afin qu'ils puissent en informer leur médecin traitant et être mieux suivis. En effet, s'agissant de personnes traitées pour une affection bénigne à l'enfance, ils n'étaient pas toujours au courant de leur exposition aux radiations et ce, souvent à moins d'un an d'âge, c'est-à-dire à un âge auquel la sensibilité aux rayonnements ionisants est très importante.

L'espérance de vie au sein de la cohorte ANGIO étant celle de la population générale, elle constitue donc une opportunité unique pour étudier les risques de pathologie radio-induites à long terme.

Les reconstructions dosimétriques corps-entier ont été estimées par notre équipe en collaboration avec le service de physique de l'IGR pour 5357 patients entre 1989 et 1999 grâce à l'obtention des dossiers techniques d'irradiation du département de dermatologie et du service de physique. Cette estimation des doses de rayonnements ionisants a demandé la mise au point de deux logiciels spécifiques (ICTA et DosEG).

L'étude de la mortalité par cancer a été réalisée à partir de l'obtention du statut vital et des causes des éventuels décès.

L'étude d'incidence des effets cancérogènes des rayonnements ionisants reçus dans l'enfance a été réalisée entre 2000 et 2008 à partir d'un auto-questionnaire portant sur les tumeurs malignes ou bénignes et les éventuels traitements, ainsi que les facteurs pouvant influencer le risque de cancer.
Quatre mille sept cent soixante-neuf réponses ont été obtenues. Les tumeurs déclarées ont été vérifiées à partir des comptes-rendus anatomo-pathologiques et par contact avec les

médecins des patients.

Après un suivi moyen de 35 ans de la cohorte ANGIO, nous avons confirmé que les traitements par radiothérapie d'un hémangiome cutané au cours de l'enfance augmentaient le risque de cancer de la thyroïde [Haddy et al., 2009b] , du sein [Haddy et al., 2010] et de mélanome [Haddy et al., 2012c].

5.2.2 Design de l'étude

Cette étude a été menée en partenariat avec le Laboratoire de radiobiologie et oncologie (LRO) du CEA Fontenay aux Roses dirigé par le docteur Laure Sabatier.

Vingt et un patients ont été sélectionnés parmi ceux qui ont accepté de participer à l'étude EpiRadBio (environ cent cinquante personnes).

Ce groupe de patients est composé de onze personnes dont l'angiome n'a pas été traité par radiothérapie et de dix autres parmi celles qui ont été exposées aux doses les plus fortes à la moelle active calculée selon la méthode de Cristy ($\geq 50mGy$) [Cristy, 1981].

Chaque échantillon de sang fourni par les participants a fait l'objet de deux analyses de quantification de fluorescence au microscope 10x dont les détails techniques sont donnés plus bas.

Protocole de marquage

Le marquage des télomères se fait sur des noyaux interphasiques étalés sur lame par cytocentrifugation ainsi qu'après culture. A la suite de la cytocentrifugation, les lames sont passées dans du PBS (*Buffers et General Solutions*) (5 min) puis au moins 15 minutes dans le fixateur Ethanol/Acide acétique (3v/1v) avant de sécher à l'air pendant la nuit. Toutes les lames sont ensuite rincées au PBS (5 min) et sont fixées pendant 2 minutes dans du formaldéhyde à 4%. Après une série de rinçage au PBS, les lames sont incubées à 37^oC avec de la pepsine (7 min) à 0.3mg/mL. La pepsine permet de couper les ponts disulfures et facilite l'accès de la sonde aux extrémités chromosomiques. Après rinçage, on effectue une deuxième fixation identique à la première suivie de rinçages. Les lames sont déshydratées par des bains successifs d'éthanol (50%, 70%, 100%) puis laissées à sécher à l'air. Elles sont incubées avec la sonde PNA (Peptid Nucleic Acid) télomère marquée à la Cyanine 3 (dilution 1/100è) après 3 minutes de dénaturation à 80^oC, pendant 1h30 en chambre noire. Les lames sont ensuite lavées au formamide (70%)/Tris pH 7.2 (10mM) puis au Tris pH

7.2 (50mM)/NaCl (150mM)/Tween20 (0.05%). Elles sont contre-colorées au DAPI (4',6'-Diamidino-2-Phénylindole, 1μL/mL) et on termine par un montage entre lame et lamelle avec du PPD. Les noyaux sont ensuite capturés à l'aide du logiciel Metacyte (version 3.9.1, MetaSystems, Newton, MA, USA).

Méthode de quantification

Le Teloquant-FISH est une technique de quantification de la longueur des télomères automatisée permettant également d'explorer l'hétérogénéité intercellulaire en rapport avec la morphologie cellulaire dans un grand nombre d'échantillons. Les conditions d'acquisition sont dictées par le nombre important de noyaux dont on souhaite mesurer la quantité de fluorescence.

Optiquement, ceci peut se réaliser en utilisant un objectif 10X pour un grand champ d'observation et une profondeur de champ au moins équivalente à l'épaisseur de l'échantillon. Ceci permet l'analyse d'un très grand nombre de cellules avec une sensibilité importante.

Avec cette technique, 10 000 cellules pourront être analysées en moins de 2 minutes. Elle permet ainsi d'accéder aux données sur la morphologie cellulaire (irrégularités, rondeurs, concavité...) ainsi qu'à une estimation très précise de la distribution des intensités de fluorescence.
Pour accéder à une distribution de la longueur de télomères relative du patient, il est nécessaire de normaliser cette distribution de fluorescences des cellules du patient par le signal de fluorescence, mesuré lors de la même manip, d'une lignée cellulaire dont on connaît la longueur de télomère moyenne. Dans cette étude, nous avons normalisé par le signal des lignées cellulaires REMB.

Ce procédé permet donc d'avoir accès à des informations précises, que ne reflète pas le concept de longueur moyenne, relative aux longueurs télomériques extrêmes (courtes et longues) dont le rôle dans l'apparition de pathologies radio-induites a été souvent mis en évidence (Section 5.1).

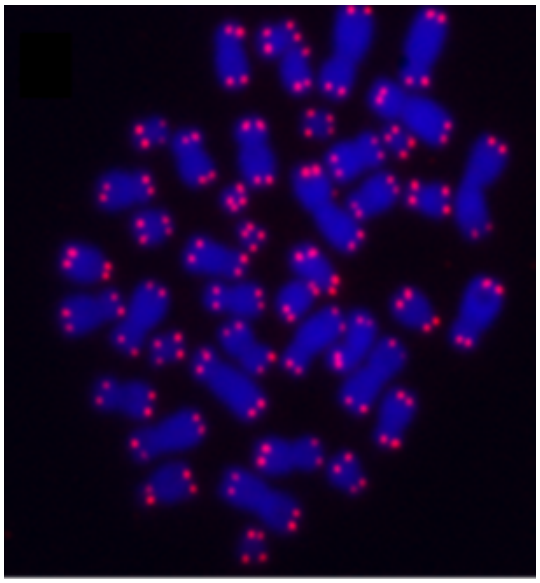

FIGURE 5.1 – Marquage télomérique de chromosomes de lymphocytes T

MetaCyte Exported Cell Data V3.9.6 07/03/14 18:37:54

File EPI 5.TXT N Cells : 8262 Scanned : 20/02/14 12:58:10

ColNo	ChanNo	FeatNo	Param1	Param2	Param3	FeatDesc
1	1	1				Total Area within Contour in µm²
2	1	2				Circumference of Contour in µm
3	1	3				Irregularity of Contour (0..1)
4	1	4				Roundness = 4 Pi Area / Circumference² (0..1)
5	1	5				Maximum Relative Concavity Depth (0..1)
6	1	6				Relative Area of Deepest Concavity (0..1)
7	1	7				Total Relative Concavity Area (0..1)
8	1	8				Integrated Intensity
9	1	9				Mean Intensity
10	1	10				Standard Deviation of Intensity
11	1	111				Horizontal Object Position on the Slide (µm)
12	1	112				Vertical Object Position on the Slide (µm)
13	2	8				Integrated Intensity
14	2	13	20.000			Maximum Intensity, Absolute Spot Area X/100 µm²
15	2	14	20.000			Minimum Intensity, Absolute Spot Area X/100 µm²
16	2	10				Standard Deviation of Intensity
17	2	61	0.000	300.000	0.000	Number of Objects at X% Intensity (Maximum Gain Y%, Upper Thr. with Z Sat. Pixels)
18	2	66				Mean of Object Intensity
19	2	29				Capture Integration Time (seconds)

B:Contour Area	B:Circumference	B:Irregularity	B:Roundness	B:Max. Conc. Depth	B:Max. Conc. Area	B:Tot. Conc. Area	B:Total Int
9,402165000E+0001	3,501335495E+0001	5,780433897E-0001	9,637630672E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	3,457365807
1,327119750E+0002	4,215419369E+0001	6,585872583E-0001	9,385089164E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	5,336757092
9,610177500E+0001	3,501335495E+0001	6,218522339E-0001	9,850852589E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	3,271474632
8,570115000E+0001	3,410118720E+0001	8,141674416E-0001	9,260992148E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	3,567902790
7,030822500E+0001	2,916251621E+0001	6,065000707E-0001	1,038881497E+0000	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	3,689992333
8,278897500E+0001	3,296769044E+0001	6,979532603E-0001	9,572058020E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	3,162250438
8,445307500E+0001	3,303251621E+0001	5,488127565E-0001	9,726173615E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	3,788714200
9,110947500E+0001	3,410118720E+0001	5,266806819E-0001	9,845423692E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	6,236018778
1,065024000E+0002	3,888335495E+0001	8,298734139E-0001	8,852008024E-0001	8,862269255E-0002	3,125000000E-0002	4,531250000E-0002	3,100549272
7,946077500E+0001	3,212034846E+0001	6,727402121E-0001	9,678368734E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	2,854532279
8,944537500E+0001	3,356685171E+0001	5,684490347E-0001	9,975772430E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	3,480278923
1,418645250E+0002	4,275335495E+0001	5,275374425E-0001	9,753113346E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	3,078675096
1,331280000E+0002	4,344419369E+0001	7,085550191E-0001	8,863715115E-0001	3,467911385E-0002	6,562500000E-0003	6,562500000E-0003	3,995979648
9,734985000E+0001	3,576901945E+0001	7,473716808E-0001	9,561611217E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	5,947489545
7,904475000E+0001	3,136468396E+0001	5,721430351E-0001	1,009720242E+0000	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	2,490720191
1,102466250E+0002	3,721552270E+0001	5,179472048E-0001	1,000292289E+0000	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	2,614091506
8,570115000E+0001	3,318901945E+0001	6,370892088E-0001	9,777046265E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	3,722923555
8,986140000E+0001	3,463552270E+0001	6,951821493E-0001	9,413249116E-0001	6,633011500E-0002	2,037037037E-0002	2,037037037E-0002	3,850177185
7,696462500E+0001	3,227685171E+0001	7,499190197E-0001	9,283648482E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	4,397152107
1,780587000E+0002	4,806985819E+0001	3,976258232E-0001	9,683389040E-0001	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	7,721491413
9,984600000E+0001	3,539118720E+0001	5,174838819E-0001	1,001729059E+0000	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	3,965783872
1,065024000E+0002	3,630335495E+0001	4,841352597E-0001	1,015490196E+0000	0,000000000E+0000	0,000000000E+0000	0,000000000E+0000	3,501415966
8,611717500E+0001	3,410118720E+0001	7,543726472E-0001	9,305948421E-0001	5,543732642E-0002	1,739130435E-0002	1,739130435E-0002	3,804236467

FIGURE 5.2 – Exemple de listing de quantification télomérique d'un microscope 10X. Le listing comporte 10.000 mesures d'intensité de fluorescences cellulaires par lame.

5.3 Analyse en composantes principales multiniveau

A l'issue de la phase expérimentale, les données en présence sont donc de nature fonctionnelles et répétées puisqu'il s'agit de densités de probabilité estimées à chaque expérience (et donc à différents instants) pour un même individu.
Comme dans le cas multivarié, la difficulté majeure dans le traitement statistique de ces données provient de ce qu'il n'est en général pas réaliste de supposer que les observations réalisées sur un même individu, au cours du temps, sont indépendantes.

Il est donc nécessaire d'introduire une structure de covariance pour ces variables aléatoires fonctionnelles associées à chaque individu, afin de tenir compte de cette situation particulière.

Nous présentons, dans la suite, la méthode de statistique fonctionnelle que nous avons utilisée pour tenir compte des contraintes de ce problème.
Il s'agit d'une généralisation de l'analyse en composantes principales fonctionnelles au contexte multiniveau qui fut proposée par Di *et al* [Di et al., 2009].

5.3.1 Définition

Notons pour $1 \leq i \leq I$ et $1 \leq j \leq J$ la fonction $X_{ij}(t)$ définie pour tout réel $t \in \mathcal{I}$ représentant la $j^{ième}$ mesure fonctionnelle du $i^{ème}$ individu.

Dans notre application, $\mathcal{I}$ représentera le continuum des différentes valeurs de fluorescence mesurées à chaque cellule et $X_{ij}(t)$ la $j^{ième}$ densité de probabilité estimée à partir des dix mille valeurs de fluorescence cellulaire du $i^{ème}$ patient.
En particulier $I = 21$ et $J = 2$.

La décomposition en composantes principales fonctionnelles multiniveau de l'ensemble des fonctions $X_{ij}(t)$ est définie par :

$$X_{ij}(t) = \mu(t) + \sum_{k \geq 1} \xi_{ik} \phi_k^{(1)}(t) + \sum_{l \geq 1} \zeta_{ijl} \phi_l^{(2)}(t) \tag{5.1}$$

avec $\mu, \phi_k^{(1)}$ et $\phi_l^{(2)}$ des fonctions déterministes et les (ξ_{ik}) et (ζ_{ijl}) des variables aléatoires de moyenne nulle.

Nous reconnaissons dans la formulation (5.1) la spécification usuelle en contexte multiniveaux des résidus qui sont estimés simultanément aux niveaux des patients et des

mesures. Plus précisément, voici les hypothèses que nous formulons concernant la décomposition (5.1) :

1. $E(\xi_{ik}) = 0$,$Var(\xi_{ik}) = \lambda_k^{(1)}$ et pour tout $i, k_1 \neq k_2, E(\xi_{ik_1}\xi_{ik_2}) = 0$.
2. $\left(\phi_k^{(1)}\right)$ est une base orthonormale de $L^2(\mathcal{I})$.
3. $E(\zeta_{ijl}) = 0$,$Var(\zeta_{ijl}) = \lambda_l^{(2)}$ et pour tout $i, j, l_1 \neq l_2, E(\zeta_{ijl_1}\zeta_{ijl_2}) = 0$.
4. $\left(\phi_l^{(2)}\right)$ est une base orthonormale de $L^2(\mathcal{I})$.
5. Les variables aléatoires (ξ_{ik}) et (ζ_{ijl}) sont décorrélées.

Ces hypothèses sont assez habituelles en analyse en composantes principales et analyse multiniveaux. Notons enfin que bien que les fonctions $\left(\phi_k^{(1)}\right)$ et $\left(\phi_l^{(2)}\right)$ soient deux bases orthogonales de $L^2(\mathcal{I})$, il n'est pas demandé qu'elles soient mutuellement orthogonales.

5.3.2 Valeurs propres et vecteurs propres

Nous nous intéressons dans cette section à l'estimation des valeurs propres $\left(\lambda_k^{(1)}\right), \left(\lambda_l^{(2)}\right)$ ainsi que des fonctions propres $\left(\phi_k^{(1)}\right)$ et $\left(\phi_l^{(2)}\right)$.

Définissons pour tout $s, t \in I$ les noyaux de covariance suivants :

$$\begin{aligned} K_T(s,t) &= cov(X_{ij}(s), X_{ij}(t)) \\ K_B(s,t) &= cov(X_{ij}(s), X_{ik}(t)) \\ K_W(s,t) &= K_T(s,t) - K_B(s,t) \end{aligned} \tag{5.2}$$

K_T, K_B et K_W désignent le noyau de covariance totale, inter et intra individus respectivement.

Ainsi, l'estimation des valeurs propres et des fonctions propres de chaque niveau, individu et mesure, est obtenu en estimant les éléments spectraux des noyaux K_B et K_W :

$$\begin{aligned} K_B(s,t) &= \sum_{k\geq 1} \lambda_k^{(1)} \phi_k^{(1)}(t) \phi_k^{(1)}(s) \\ K_W(s,t) &= \sum_{l\geq 1} \lambda_l^{(2)} \phi_l^{(2)}(t) \phi_l^{(2)}(s) \end{aligned} \tag{5.3}$$

En pratique, l'estimation de ces noyaux est relativement simple lorsque les fonctions $(X_{ij})_{1\leq i\leq n, 1\leq j\leq J}$ sont mesurées sur une grille commune et régulière $\{t_s, s = 1...T\}$, comme

c'est le cas pour les densités de probabilité des distributions de télomères :

$$\begin{array}{lcl} \widehat{K}_T(t_s,t_r) & = & \frac{1}{I\times J}\sum_{ij}\left(X_{ij}(t_s)-\widehat{\mu}(t_s)\right)\left(X_{ij}(t_r)-\widehat{\mu}(t_r)\right) \\ \widehat{K}_B(t_s,t_r) & = & \frac{2}{I\times J\times(J-1)}\sum_{i}\sum_{j_1<j_2}\left(X_{ij_1}(t_s)-\widehat{\mu}(t_s)\right)\left(X_{ij_2}(t_r)-\widehat{\mu}(t_r)\right) \\ \widehat{K}_W(t_s,t_r) & = & \widehat{K}_T(t_s,t_r)-\widehat{K}_B(t_s,t_r) \end{array} \tag{5.4}$$

avec $\widehat{\mu}=\frac{1}{n\times J}\sum_{ij}X_{ij}$ l'estimateur empirique de la moyenne fonctionnelle des (X_{ij}).

Un important paramètre qui découle de la connaissance des valeurs propres $\left(\lambda_k^{(1)}\right)$ et $\left(\lambda_l^{(2)}\right)$ est la mesure de la part de la variance totale expliquée par le niveau de l'individu ou ce qui revient au même, de la corrélation entre les mesures au sein d'un même individu :

$$\rho_W=\frac{\sum_{k\geq 1}\lambda_k^{(1)}}{\sum_{k\geq 1}\lambda_k^{(1)}+\sum_{l\geq 1}\lambda_l^{(2)}} \tag{5.5}$$

5.3.3 Estimation des scores de l'analyse en composantes principales multiniveau

Considérons le cas général où les mesures des fonctions (X_{ij}) sont entachées d'erreurs de mesures.
Alors, le modèle (5.1) devient :

$$X_{ij}(t)=\mu(t)+\sum_{k=1}^{N_1}\xi_{ik}\phi_k^{(1)}(t)+\sum_{l=1}^{N_2}\zeta_{ijl}\phi_l^{(2)}(t)+\epsilon_{ij}(t) \tag{5.6}$$

avec $\xi_{ik}\sim\mathcal{N}\left(0,\lambda_k^{(1)}\right)$. $\zeta_{ijl}\sim\mathcal{N}\left(0,\lambda_l^{(2)}\right)$ et $\epsilon_{ij}\sim\mathcal{N}\left(0,\sigma^2\right)$, On remarquera que la sommation sur les éléments propres de niveau 1 et 2 a été tronquée aux N_1 et N_2 premiers termes respectivement par un critère d'AIC ou de variabilité expliquée suffisamment grande (dans notre cas plus de 90%). En cas de fonctions mesurées sur une grille suffisamment dense, il est possible d'utiliser un modèle projectif pour l'estimation des scores (ξ_{ik}) et (ζ_{ijl}).
L'idée consiste alors à projecter les mesures centrées dans les espaces engendrés par les fonctions propres $\left(\phi_k^{(1)}\right)_{1\leq k\leq N_1}$ et $\left(\phi_k^{(2)}\right)_{1\leq l\leq N_2}$:

$$\begin{aligned} A_{ijk} &= \int_I \left(X_{ij}(t) - \mu(t)\right) \phi_k^{(1)}(t)\ dt \\ &= \xi_{ik} + \sum_{l=1}^{N_2} \zeta_{ijl} \times c_{kl} + \epsilon_{ijk}^{(1)} \end{aligned} \tag{5.7}$$

$$\begin{aligned} B_{ijl} &= \int_I \left(X_{ij}(t) - \mu(t)\right) \phi_l^{(2)}(t)\ dt \\ &= \zeta_{ijl} + \sum_{l=1}^{N_1} \xi_{ik} \times c_{kl} + \epsilon_{ijl}^{(2)} \end{aligned} \tag{5.8}$$

avec $c_{kl} = \int_I \phi_k^{(1)}(t)\phi_l^{(2)}(t)\ dt$ le produit scalaire des fonctions propres des deux niveaux individu/mesure et $\epsilon_{ijk}^{(1)}$ et $\epsilon_{ijl}^{(2)}$ les résidus rendant compte aussi bien de l'erreur de mesure que la réduction dimensionnelle en N_1 et N_2. Remarquons que les intégrales dans les équations (5.7) et (5.8) peuvent toutes deux être estimées par quadrature numérique via les estimateurs des fonctions propres obtenus comme expliqué dans la section précédente.

La réécriture matricielle des modèles (5.7) et (5.8) donne :

$$\begin{cases} \mathbf{A}_{ij} &= \xi_i + C\zeta_{ij} + \epsilon_{ij}^{(1)} \\ \mathbf{B}_{ij} &= \zeta_{ij} + C^T\xi_i + \epsilon_{ij}^{(2)} \\ \xi_i &\sim \mathcal{N}\left(0, \Lambda^{(1)}\right) \\ \zeta_{ij} &\sim \mathcal{N}\left(0, \Lambda^{(2)}\right) \\ \epsilon_{ij}^{(1)} &\sim \mathcal{N}\left(0, \sigma_1^2\mathbf{I}_{N_1}\right) \\ \epsilon_{ij}^{(2)} &\sim \mathcal{N}\left(0, \sigma_2^2\mathbf{I}_{N_2}\right) \end{cases} \tag{5.9}$$

avec

$$\begin{array}{ll} \mathbf{A}_{ij} = (A_{ij1}, ..., A_{ijN_1})^T & \mathbf{B}_{ij} = (B_{ij1}, ..., B_{ijN_2})^T \\ \xi_i = (\xi_{i1}, ..., \xi_{iN_1})^T & \zeta_{ij} = (\zeta_{ij1}, ..., \zeta_{ijN_2})^T \\ \epsilon_{ij}^{(1)} = \left(\epsilon_{ij1}^{(1)}, ..., \epsilon_{ijN_1}^{(1)}\right)^T & \epsilon_{ij}^{(2)} = \left(\epsilon_{ij1}^{(2)}, ..., \epsilon_{ijN_2}^{(2)}\right)^T \\ \Lambda^{(1)} = diag(\lambda_1^{(1)}, ..., \lambda_{N_1}^{(1)}) & \Lambda^{(2)} = diag(\lambda_2^{(2)}, ..., \lambda_{N_2}^{(2)}) \end{array}$$

Le modèle (5.9) apparaît alors comme un modèle linéaire mixte avec comme variances résiduelles $\sigma_1^2 = var\left(\epsilon_{ijk}^{(1)}\right)$ et $\sigma_2^2 = var\left(\epsilon_{ijk}^{(2)}\right)$. Il est alors possible d'estimer les scores avec les outils des modèles linéaires mixtes, et, particulièrement, les estimateurs BLUPs (*Best Linear Unbiased Predictor*) [Harville, 1977] qui possèdent d'excellentes propriétés statistiques particulièrement robustes vis à vis de l'écart à l'hypothèse de normalité.

Ceci a été implémenté dans un contexte Bayésien via le logiciel MCMC Winbug avec des distributions a priori pour $1/\sigma_1^2$ et $1/\sigma_2^2$ de loi gamma de moyenne 1 et de variance très large.

5.4 Résultats

Vingt et un patients ont été inclus dans cette étude dont les caractéristiques générales sont données dans le tableau 5.1 :

	Groupe de patients (n=21)	
	Radiothérapie	Pas de radiothérapie
Nombre de patients	10	11
Age actuel median (min-max)	51 ans (44-65)	53 ans (49-65)
Age median à l'irradiation (min-max)	3 mois (1-8)	-
Dose moyenne à la moelle active (min-max)	0.122 Gy (0.05-0.29)	-
Type de Radiothérapie		
Rayons X	5	-
Radium (^{226}Rad)	3	-
Strontium (^{90}Sr)	2	-

TABLE 5.1 – Table des caractéristiques du groupe de patients inclus dans l'étude.

On remarque que les patients exposés aux radiations sont plutôt moins âgés que les non exposés. Toutes les personnes exposées l'ont été à des doses moyennes à la moelle active supérieures à 50 mGy.

La Figure 5.3 montre les intensités moyennes du signal de fluorescence de la première et la seconde série de mesures avant et après normalisation par le signal de fluorescence de la lignée REMB. Bien que les secondes mesures soient en moyenne d'intensités moindres, nous constatons que la normalisation par un signal de référence permet de réduire sensiblement le biais expérimental.

Plus précisément, le tableau 5.2 quantifie la part de variabilité des mesures de fluorescence due à la différence intrinsèque entre individus et celle attribuable au protocole expérimental. Les parts de variabilités inter-patients obtenues par analyse de la variance multivariée des signaux moyens sont supérieures à celles obtenues par analyse en composantes principales multiniveau via l'équation 5.5, et ce, particulièrement dans le cas non normalisé (12% d'écart contre 5% dans le cas normalisé).

Le résultat de l'analyse spectrale multiniveau des courbes de fluorescences normalisées par REMB est donné dans la table 5.3 :

La majeure partie de l'information au niveau patient est contenue dans un espace à 2 dimensions seulement. Par exemple, la première composante principale explique 93%, la

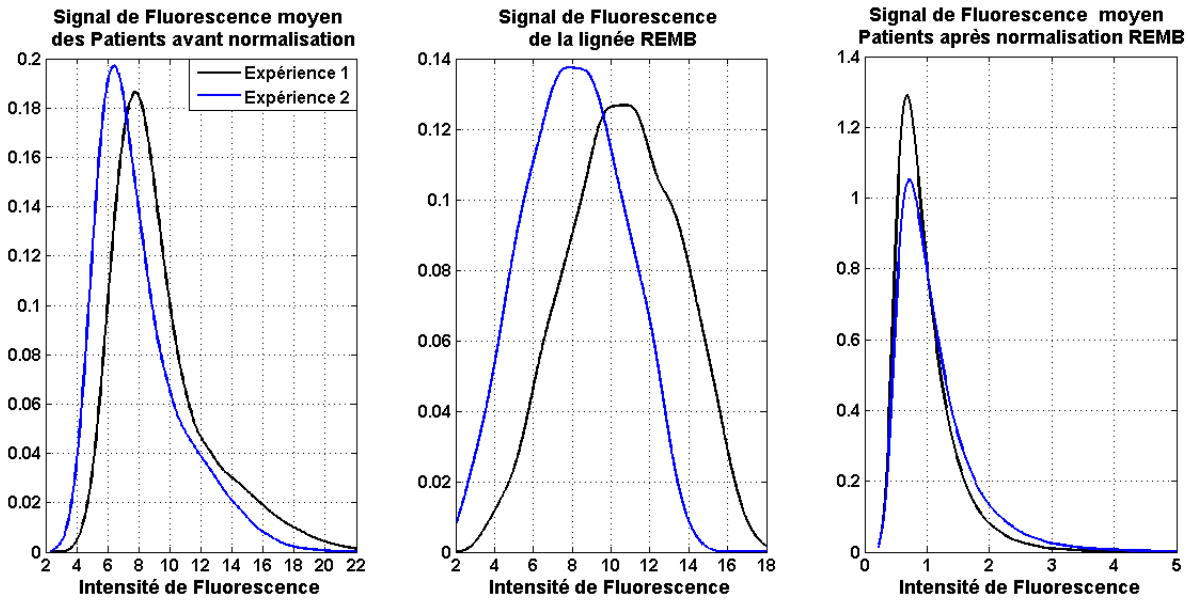

FIGURE 5.3 – Intensité Moyenne du signal de fluorescence avant et après normalisation par le signal REMB.

	Anova multivariée [a]		ACPFM [b]	
	Sans REMB	Avec REMB	Sans REMB	Avec REMB
Variabilité inter-groupe (%)	56%	67%	42%	62%
Variabilité intra-groupe (%)	44%	33%	58%	38%

a. Calculs menés sur les valeurs moyennes des signaux de fluorescence
b. Analyse en composantes principales fonctionnelles multiniveau sur les signaux de fluorescence

TABLE 5.2 – Résultats des deux méthodes multivariée /ACP fonctionnelle d'analyse de la variance. Un patient est considéré comme un *groupe* de deux mesures et l'on cherche à quantifier la variabilité inter et intra groupes.

seconde 6 %. Ensemble, elles expliquent plus de 99 % de la dispersion du niveau patient. De même, il suffit de deux composantes principales pour expliquer la majeure partie de la dispersion due au niveau 2 (niveau mesure).

Un des objectifs principaux de l'analyse en composante principale est, on le sait, d'obtenir une réduction dimensionnelle : d'après les résultats ci-dessus, le niveau patient peut être correctement représenté par les scores dans la base formée par les deux premières fonctions propres de niveau 1 dans la figure 5.6.

L'interprétation de ces tendances est assez immédiate : la première fonction propre

	Valeurs propres niveau 1 (niveau patient)		
Numéro	1	2	3
Valeur propre ($\times 10^{-3}$)	28.82	1.9	0.04
Pourcentage de variabilité expliquée	93.55%	6.18%	0.14%
Pourcentage cumulé de variabilité expliquée	93.55%	99.73%	99.87%
	Valeurs propres niveau 2 (niveau mesure)		
Numéro	1	2	3
Valeur propre ($\times 10^{-3}$)	15.25	1.77	0.429
Pourcentage de variabilité expliquée	85.91%	9.98%	2.42%
Pourcentage cumulé de variabilité expliquée	85.91%	95.89%	98.31%

TABLE 5.3 – Valeurs propres et pourcentage de variabilité expliquée, pour chaque niveau, obtenus par analyse en composantes principales fonctionnelles multiniveaux des signaux de fluorescences normalisés par REMB.

exprime la corrélation inverse entre télomères courts et télomères longs avec un seuil approximativement à 0.8. Les individus qui présenteront des scores négatifs auront tendance à avoir plus de télomères courts que la moyenne (au-dessus de la courbe moyenne en bleu) et moins de télomères longs que la moyenne (en dessous de la courbe moyenne en bleu) et vice versa avec les patients aux scores positifs. La seconde fonction propre exprime, quant à elle, la corrélation contraire entre télomères extrêmes et télomères de taille intermédiaire avec deux seuils de basculement situés à 0.5 et 1.3. Les patients avec des scores 2 fortement positifs auront donc tendance à avoir plus de télomères intermédiaires que la moyenne et moins de télomères de longueur très petite ou très grande.
Ces observations fournissent donc des métriques simples pour évaluer le raccourcissement ou le rallongement des télomères en considérant, par exemple, le % de télomères de longueur inférieure ou supérieure aux seuils fournis par l'analyse en composantes principales fonctionnelles.

La capacité de l'analyse en composantes principales multivineau à séparer le signal propre du niveau patient de celui du niveau mesure est illustré à travers les trois exemples de la Figure 5.4.
En bleu, les deux densités de probabilité estimées à partir des deux séries d'analyse du signal de fluorescence au microscope 10x. Ce sont donc deux mesures des distributions

de tailles de télomères du même patient et en noir, leurs approximations par projection dans les quatre premières bases orthogonales de niveau 1 et 2 de la décomposition 5.1. En rouge, la partie de la décomposition n'utilisant que les scores et fonctions propres du niveau 1 (niveau patient). Enfin en jaune, la partie de la décomposition utilisant les scores et fonctions propres du niveau 2 (niveau mesure). On remarque qu'en progressant du panneau de gauche à celui de droite, la contribution « mesure » au signal total est de plus en plus importante.

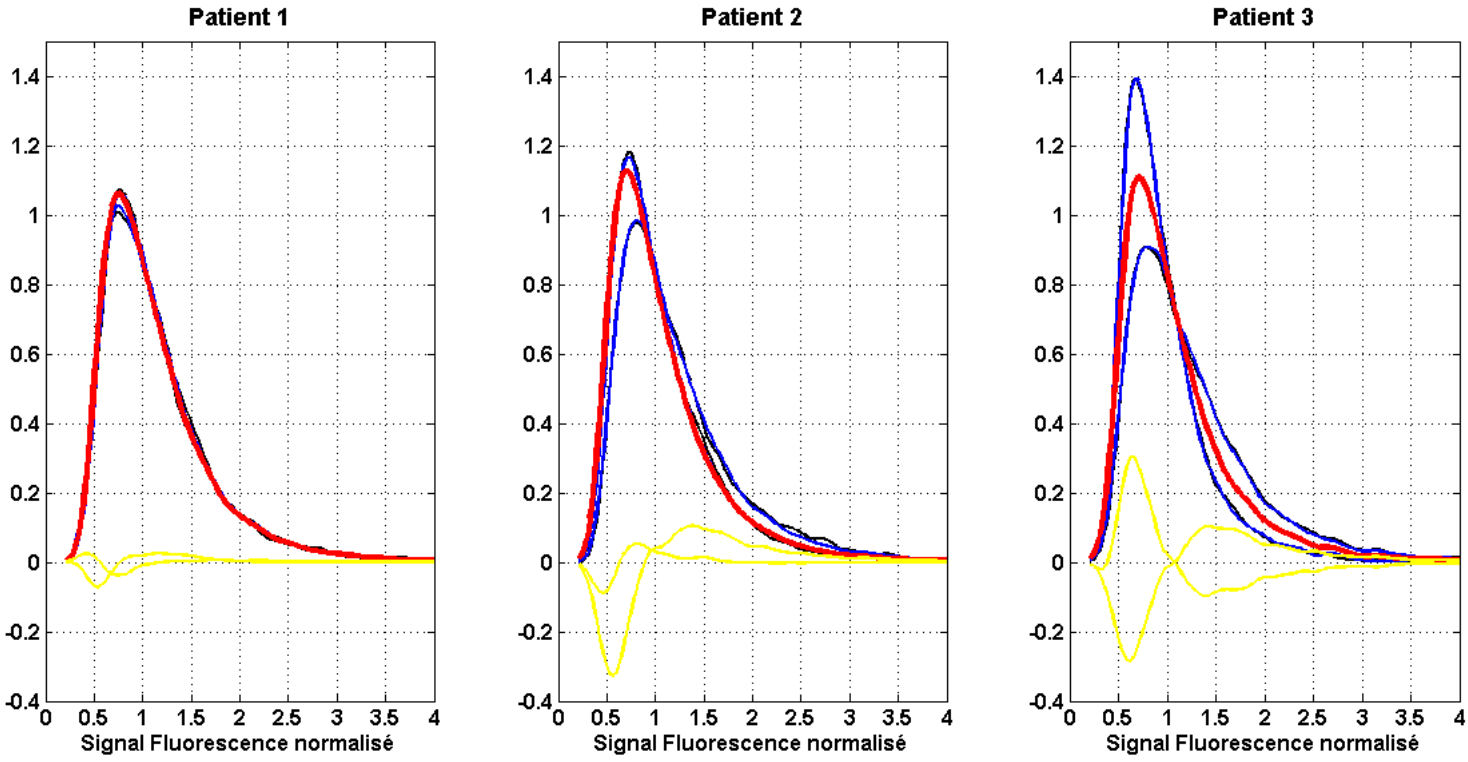

FIGURE 5.4 – Extraction de la partie intrinsèque à l'individu (en rouge) des deux densités de distribution des signaux de fluorescences mesurées lors des deux expériences. En jaune, les parties relatives au niveau 2.

L'information du niveau 1 du patient étant ainsi débarrassée du *bruit expérimental* des signaux mesurés grâce à une normalisation adéquate par le REMB puis par l'analyse en composantes principales fonctionnelles, nous sommes en mesure à présent d'apprécier la tendance, en fonction de l'âge, des deux premiers scores de niveau 1 chez les patients exposés et non exposés représentés dans la figure 5.5.

Nous remarquons une décroissance de chacun des scores en fonction de l'âge avec un changement de signe au début de la cinquantaine chez les deux groupes exposés/non exposés. D'après la dynamique décrite par la Figure 5.6, des scores négatifs s'interprètent par une augmentation avec l'âge de la proportion des télomères courts (fonction propre

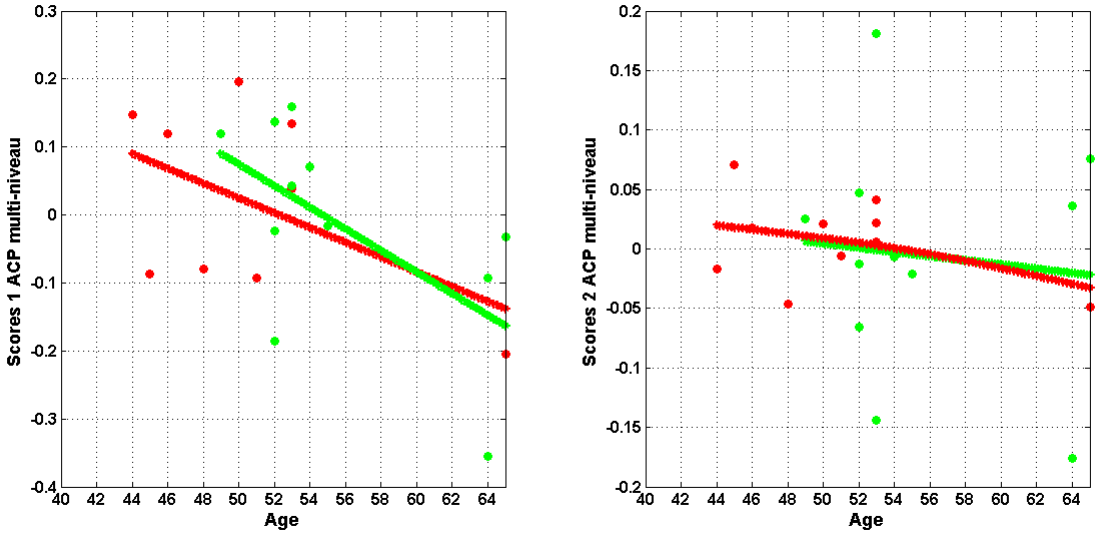

FIGURE 5.5 – Variation des deux premiers scores du niveau 1 en fonction de l'âge et du statut exposé/non exposé.

1) et très courts (fonction propre 2). Même si les tendances des scores 1 suggèrent un raccourcissement de télomères plus précoce chez les exposés, il n'a pas été possible en raison du nombre limité de données de mettre en évidence une différence significative.

5.5 Discussion

Notre objectif dans cette étude était de proposer une nouvelle méthode d'analyse des longueurs de télomères. Il s'agit d'une extension de l'analyse en composantes principales fonctionnelles aux situations où différents niveaux apparaissent naturellement dans la structure des données.

Les données fonctionnelles sur lesquelles nous avons appliqué cette méthode sont les densités de probabilités estimées à partir de milliers de valeurs d'intensités de fluorescences obtenues en analysant les lymphocytes T des échantillons sanguins au microscope 10x. Le fait de travailler directement sur les densités de probabilité et non sur les densités moyennes permet ainsi de prendre en compte la variabilité intra-individuelle des longueurs de télomères.

Ces expériences ayant été effectuées sur deux échantillons sanguins prélevés sur chacun des vingt-et-un patients, nous sommes bien en présence de données multiniveaux avec comme niveau 1 celui du patient et comme niveau 2 celui de l'expérience.

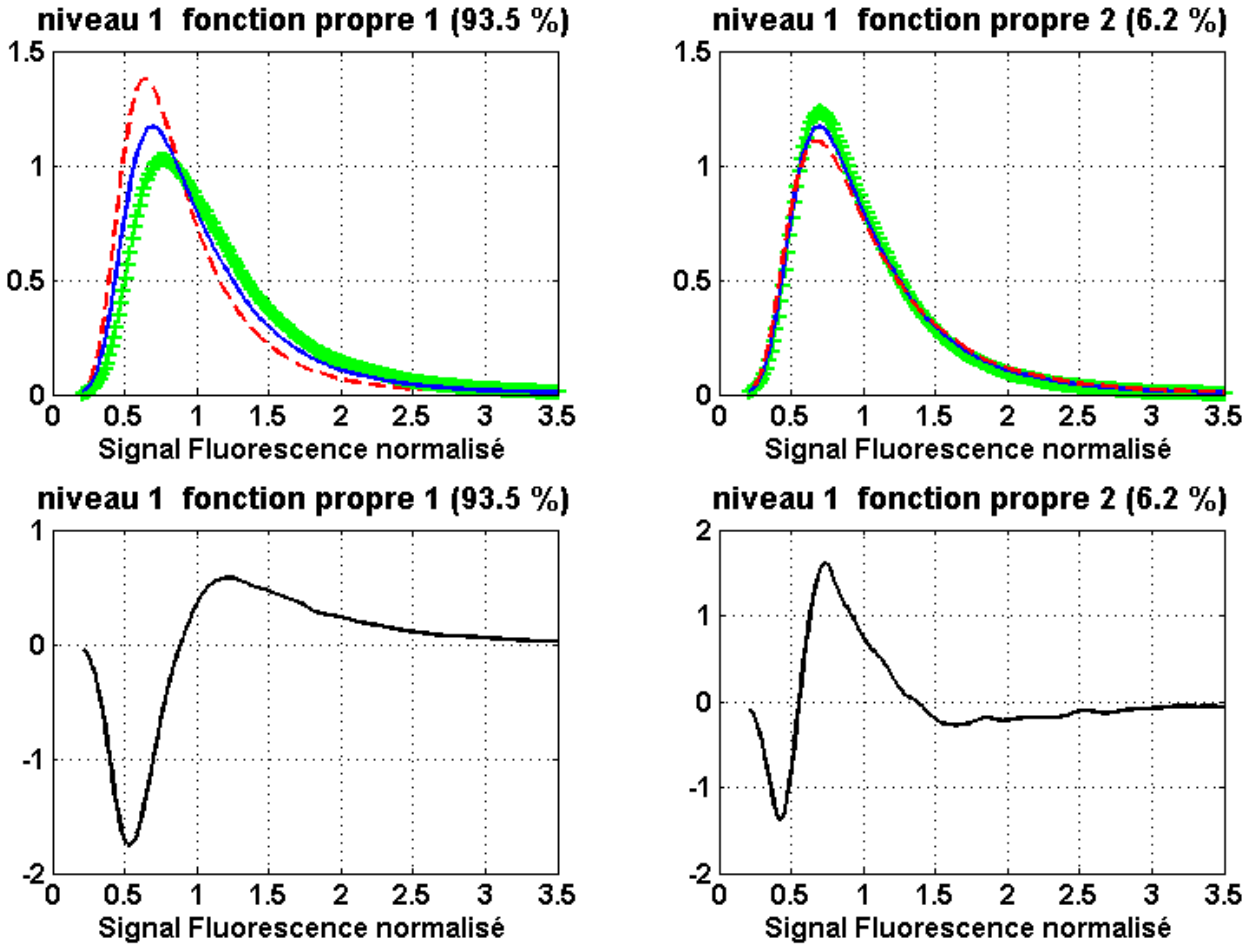

FIGURE 5.6 – (en bas) Les deux premières fonctions propres niveau 1. (en haut) L'effet de ces fonctions propres sur la densité moyenne (ici en bleu) de toutes les courbes de fluorescence. La courbe '-' (resp en '+') représente la densité moyenne à laquelle on retranche (resp on ajoute) un multiple de la fonction propre.

L'analyse en composantes principales fonctionnelles multiniveau permet une décomposition parcimonieuse des variations fonctionnelles observées. Elle permet de distinguer la part de variabilité qui relève de la spécificité individuelle de celle propre aux variabilités expérimentales avec en prime, une réduction dimensionnelle conséquente. En effet, dans notre cas, nous n'avons besoin que de deux espaces, de dimension deux chacun, pour décrire les modes de variabilités individuelles et expérimentales.

La structure géométrique de ces données aide à mieux comprendre le mécanisme de variabilité individuelle de la distribution des longueurs de télomères en le résumant qualitativement et quantitativement à deux mécanisme simples : variabilité télomères courts-longs

et variabilité télomères extrêmes/intermédiaires.

La prise en compte de la variabilité intra-individuelle est souvent négligée en pratique notamment en raison de l'usage de la longueur moyenne individuelle. Nous avons ainsi constaté que l'impact des procédures de normalisations usuelles via le signal de fluorescence de lignées cellulaires était surestimé dans l'approche multivariée utilisant des doses moyennes et gommant ainsi de riches informations sur les modes de variabilité des données en présence.
Ainsi, l'usage des scores de niveau 1 issus de l'analyse en composantes principales fonctionnelles paraît, grâce à leurs interprétations simples, une alternative plus rigoureuse à la notion de longueur moyenne car directement destinée à la description du niveau individu débarrassé du bruit expérimental.

La limite majeure de ce travail réside bien évidement dans le nombre de sujets restreint avec l'absence notable de sujets dans la tranche d'âge 56-63 ans. Ainsi, les résultats de cette première étude sont à prendre avec précaution et doivent être confirmés par une prochaine analyse à la fin du projet EpiRadBio qui visera à inclure une centaine de patients, avec pour chacun, une double mesure de fluorescence.
Cependant, et dans un cadre de démarche méthodologique, l'analyse de données fonctionnelles multiniveau se pose malgré tout comme une alternative sérieuse aux méthodes multivariées classiques basées sur les longueurs moyennes de télomères.

Conclusion générale

L'objectif de ce travail de thèse a été de développer des méthodes capables de répondre à des questions spécifiques aux deux aspects, dosimétriques et statistiques, intervenant dans la modélisation du risque de survenue d'événements iatrogènes radio-induits.

Le calcul de la dose hors-champ est devenu une préoccupation majeure des professionnels de la radiothérapie, principalement en raison de l'augmentation régulière de l'espérance de vie après un premier traitement par radiothérapie.

Exceptées les modélisations stochastiques de type Monte Carlo, assez peu de développements se sont focalisés sur l'estimation de la dose à distance du champ d'irradiation. Une limitation majeure des méthode Monte Carlo, outre le temps de calculs encore trop longs et incompatibles avec la routine clinique, est la disponibilité des détails techniques de construction de la machine, notamment sur la composition des matériaux en interaction avec les faisceaux. Ces informations sont souvent considérées comme des secrets industriels qui ne sont pas divulgués par les constructeurs.

Notre contribution dans ce domaine est d'avoir mis au point un outil analytique précis de calcul de la composante machine-dépendante de la dose à distance capable de s'adapter à tout type d'appareils.

Nous avons pour cela adapté l'approche multi-sources, déjà utilisée dans le champ pour l'étude de l'influence de l'ouverture collimateur sur la dose à l'axe, à la problématique plus complexe de la dose à distance ayant nécessité de repenser le concept de plan source.
En le déplaçant au niveau du système de collimation, celui-ci est composé d'autant de faces émettrices que nécessaire pour traduire la forme du champ en présence. Ainsi, la structure du système de collimation, relégué à l'intérieur du champ d'irradiation à un simple outil d'intersection géométrique, devient la source de rayonnement principal de la dose diffusé-collimateur, dotée d'une énergie et d'une intensité d'émission construite à

partir de lois d'interactions de la physique des particules. La calibration se faisant par la suite sur la base de mesures métrologiques.

La souplesse de ce modèle vis-à-vis du type d'appareil d'irradiation considéré ainsi que sa vitesse de calcul en font donc une alternative sérieuse à l'approche stochastique Monte Carlo.

Un des apports de ce travail de thèse du point de vue de la modélisation statistique du risque d'événements iatrogènes radio-induits est d'avoir illustré la capacité des outils de statistique fonctionnelle à intégrer de manière flexible l'ensemble de l'information disponible qu'elle soit dosimétrique (distribution de la dose à l'organe) ou radiobiologique (distibution intra-individuelle de longueurs de télomères), tout en les rendant compatibles avec les modèles linéaires généralisés ou de survie utilisés en épidémiologie des radiations. Il était en effet important de proposer à la fois un outil de modélisation innovant tout en s'assurant de la possibilité d'interpréter simplement, en terme de hasard ratio ou de Odds Ratio, les résultats obtenus.

Le modèle NTCP construit par régression sur composantes principales fonctionnelles offre la possibilité d'exprimer l'effet volume de l'organe via une fonction de risque qui rend compte aussi bien la radiosensibilité de l'organe vis-à-vis de certaines gammes de dose que de la technique technique d'irradiation en présence. Ce résultat, on l'a vu, constitue une source d'informations plus riche que celle obtenue par les autres modèles NTCP. Nous projetons d'appliquer ce modèle à différents plans de traitements pour un même organe afin d'en extraire à chaque fois le paramètre dose-volume fonctionnel.
A une époque de progrès technologiques importants, de nombreux plans de traitements concurrents d'une même localisation tumorale voient le jour régulièrement et notre méthode pourrait permettre une appréciation plus fine de la balance bénéfice/risque propre à chaque technique.

Nous avons également proposé dans le domaine de la radiobiologie, science expérimentale par essence, des méthodes de statistique fonctionnelle capables de quantifier la part de la variabilité expérimentale dans la variabilité des signaux mesurés.
Nous avons donc montré que le concept de longueur moyenne surestimait la qualité de la reproductibilité expérimentale en raison de la non-prise en compte de la variabilité intra-individuelle. De plus, l'analyse en composantes principales fonctionnelles multiniveau permet d'isoler la partie propre à l'individu des densités de probabilité estimées à partir des

résultats expérimentaux afin d'en explorer les modes de variabilités les plus dominants. Nous avons ainsi établi que la variabilité inter-individuelle de la longueur des télomères était essentiellement un espace à deux dimensions, à savoir la corrélation contraire entre télomères "courts"/télomères "longs" ainsi que celle des télomères "extrêmes"/télomères "intermédiaires". Cette analyse permet également de proposer des seuils de longueurs télomériques pour définir chacun de ces termes.
Grâce à une nouvelle analyse qui sera menée sur plus de cent cinquante patients à la fin du projet EpiRadBio, ces valeurs seuils vont être affinées et les proportions associées seront proposées comme nouvelles métriques alternatives au concept de longueur télomérique moyenne.

Ce travail de thèse a également proposé une méthode nouvelle et originale pour modéliser, via un modèle de Cox flexible, la relation dose-effet entre une covariable continue et un événement d'intérêt.
Contrairement aux méthodes de régression spline classiques qui reposent souvent sur une représentation en puissances tronquées, notre méthode est basée sur la représentation B-spline et tire profit de leur propriété de support minimal afin de déterminer de manière spatio-adaptative une localisation optimale des nœuds intérieurs.
Là encore, les outils de la statistique fonctionnelle (analyse en composantes principales fonctionnelles, rotation Varimax fonctionnelle) ont été appliqués sur des fonctions résumant, au voisinage de chaque valeur de la covariable, la propriété de support minimal. Comme nous l'avons vu, cette méthode modifie le lien initialement très rigide qui reliait le nombre de nœuds intérieurs au nombre de paramètres à estimer au cours d'une régression spline.
En réalité, cette approche est une sorte de régression spline pénalisée dont le paramètre de lissage est le nombre de composantes principales que l'on choisit d'inclure dans le modèle. Contrairement aux techniques splines pénalisées classiques par essence non paramétriques, la nôtre est au contraire entièrement paramétrique et donc immédiatement compatible avec des outils et logiciels destinés initialement à la statistique multivariée.

L'application de cette méthode à la modélisation de la relation dose-effet entre la dose de radiation à la thyroïde et le risque de tumeur radio-induite a permis d'établir l'allure unimodale de cette relation uniquement à l'aide d'un seul paramètre alors que le modèle paramétrique linéaire exponentiel en nécessite deux et qu'un modèle de régression spline classique en compterait au moins quatre s'il est construit à partir d'un unique nœud intérieur.

Ainsi, la régression spline, longtemps considérée comme trop couteuse en paramètres à estimer dans un contexte d'analyse de survie à forte censure, pourrait, grâce au travail proposé dans ce mémoire arriver à un meilleur compromis entre flexibilité et complexité du modèle. A l'avenir, Nous envisageons d'appliquer la méthode proposée aussi bien dans un contexte d'analyse de survie pour estimer des relation dose-effet d'autres pathologies radio-induites mais aussi dans un contexte de splines des moindres carrés afin d'investiguer sa capacité à modéliser des formes de courbes plus complexes que l'on rencontre dans les domaines de la santé et de l'industrie.

Perspectives

L'objectif principal de l'équipe dans le domaine de la physique médicale est l'achèvement et la validation clinique d'un nouveau logiciel spécifique pour l'estimation des doses à distance des champs de radiothérapie externe. Ce logiciel est développé dans un cadre lui garantissant à la fois une diffusion internationale importante, l'assurance d'une évolution, une valorisation industrielle et une utilisation en routine dans les services de radiothérapie. Le principe général est de tirer profit des ressources disponibles sur les réseaux de radiothérapie et des algorithmes de calculs existant déjà dans les logiciels de planification de traitement par radiothérapie (TPS), et de les étendre à l'ensemble du corps humain, après simulation de celui-ci.
L'architecture logicielle reposerait dans un premier temps sur la modélisation de la fluence de rayonnements ionisants dans l'air hors des champs de radiothérapie, en fonction des caractéristiques de chaque appareil, auquel contribue ce travail de thèse.
Ensuite, après une reconstruction du corps entier (via des fantômes voxélisés) à partir de la partie du corps déjà scannée pour les besoins de la radiothérapie, le calcul du TPS pourra être étendu à tout le reste du corps. Il sera ainsi possible d'obtenir des distributions de dose précise même pour des tissus sains irradiés à faibles doses et à distance du champ, du fait des rayonnements résiduels (photons diffusés, électrons et neutrons de contamination) inévitables pendant la radiothérapie.

L'épidémiologie des radiations doit se doter des mêmes outils que la radiothérapie moderne pour l'étude des risques de pathologies radio-induites à long terme comme les cancers secondaires et les pathologies cardiovasculaires. Ceci dans le but de produire des estimations du risque à long terme directement utilisables en clinique, aidant ainsi les radiothérapeutes et les physiciens médicaux à planifier leurs stratégies de soins de façon optimale.
Ces outils portent à la fois sur les indicateurs dosimétriques et la méthodologie statistique utilisée.

La reconstruction de plans de traitements et la distribution de la dose à l'organe d'intérêt en constitue une des étapes clé et c'est une compétence que maîtrise et perfectionne l'équipe 3 du CESP INSERM1018 depuis plusieurs années. En effet, le développement des procédures de reconstruction de la dose périphérique et à distance a commencé à partir du milieu des années 80 et notre équipe a procédé à des améliorations continues portant à la fois sur l'anatomie de fantôme et les calculs de dose.
Ceci a valu une reconnaissance internationale à notre équipe qui s'est vu confier la charge de la reconstruction dosimétrique de nombreux projets européens de premier plan tels que les projets Cerebrad/Procardio (FP7 EURATOM/ 2011-2015) avec la reconstruction de la dose aux structures et artères cerébrales des 150 cas et 150 témoins ainsi que la dose au cœur des 300 cas et 300 témoins de l'étude et le projet PanCareSurfUp (FP7 HEALTH / 2011-2015) avec la reconstruction des doses d'exposition de tous les patients recrutés pour l'étude qui comprend des études cas-témoins sur les pathologies cardiaques (600 cas et 600 témoins), les carcinomes urogénitaux et digestifs (300 cas et 300 témoins) et les sarcomes osseux et tissus mous (300 cas et 300 témoins) après cancer de l'enfant.

La reconstruction dosimétrique effectuée dans le cadre de ces projets fournira donc une quantité très importante de distributions de doses pour différents organes (cerveau, cœur, organes génitaux et digestifs) qui seront exploitées selon la méthodologie statistique proposée dans ce travail de thèse, à savoir la statistique fonctionnelle, et ce, dans le but de les exploiter de façon optimale afin d'estimer et de prédire des pathologies à long terme (cancers secondaires, pathologies cardiaque et cérébrales).

Par ailleurs, l'exploitation de la richesse des reconstructions dosimétriques voxélisées de plan de traitement que réalisent l'équipe 3 CESP INSERM1018 peut être très significativement optimisée en intégrant explicitement la distribution spatiale de la dose à l'organe dans l'estimation du risque de complications.
En effet, les HDVs bien que riches en informations et d'usage très répandu en clinique souffrent d'un inconvénient majeur : la perte de l'information spatiale. Or, décrire la relation entre la dose locale par voxel et la toxicité peut se révéler très instructif et peut potentiellement jouer un grand rôle dans l'identification des sous-régions à risque plus élevé fournissant ainsi des contraintes doses-organes plus précises.

Nous avons donc pour objectif de développer un modèle spatial capable de tenir compte non seulement de l'hétérogénéité de la distribution de dose mais aussi de la localisation spatiale des points voxels. Des généralisations aux données tridimensionnelles des mé-

thodes fonctionnelles proposées ici existent et un tel modèle pourrait à terme proposer une cartographie anatomique du risque de pathologies radio-induites en lieu et place des courbes dose-effets actuelles.

Bibliographie

[Ahnesjo, 1994] Ahnesjo, A. (1994). Analytic modeling of photon scatter from flattening filters in photon therapy beams. *Med. Phys.*, 21(8) :1227.

[Ahnesjo, 1995] Ahnesjo, A. (1995). Modeling transmission and scatter for photon beam attenuators. *Med. Phys.*, 22(11) :1711.

[Allard et al., 2010a] Allard, A., Haddy, N., Deley, M. C. L., Rubino, C., Lassalle, M., Samsaldin, A., Quiniou, E., Chompret, A., Lefkopoulos, D., Diallo, I., and de Vathaire F. (2010a). Role of radiation dose in the risk of secondary leukemia after a solid tumor in childhood treated between 1980 and 1999. *International journal of radiation oncology, biology, physics*, 78(5) :1474–82.

[Allard et al., 2010b] Allard, A., Haddy, N., Le Deley, M., C, R., Lassalle, M., Samsaldin, A., Quiniou, E., Chompret, A., Lefkopoulos, D., Diallo, I., and de Vathaire, F. (2010b). Role of radiation dose in the risk of secondary leukemia after a solid tumor in childhood treated between 1980 and 1999. *Int J Radiat Oncol Biol Phys*, 78(5) :1474–82.

[Altekruse et al., 2007] Altekruse, S. F., Kosary, C. L., Krapcho, M., Neyman, N., Aminou, R., Waldron, W., Ruhl, J., Howlader, N., Tatalovich, Z., Cho, H., Mariotto, A., Eisner, M. P., Lewis, D. R., Cronin, K., Chen, H., Feuer, E., Stinchcomb, D., and BK, E. (2007). Seer cancer statistics review, 1975-2007, national cancer institute. http ://seer.cancer.gov/csr/1975_2007/.

[Andreassen et al., 2012] Andreassen, C., Dikomey, E., Parliament, M., and West, C. (2012). Will snps be useful predictors of normal tissue radiosensitivity in the future ? *Radiother Oncol*, 105 :283–288.

[Baladandayuthapani et al., 2008] Baladandayuthapani, V., Mallick, B., Young, H., Lupton, J., Turner, N., and Carroll, R. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics*, 64 :64 – 73.

[Bapna et al., 2008] Bapna, R., Jank, W., and Shmueli, G. (2008). Price formation and its dynamics in online auctions. *Decis Support Syst*, 44 :641 – 656.

[Beeharry and Broccoli, 2005] Beeharry, N. and Broccoli, D. (2005). Telomere dynamics in response to chemotherapy. *Curr Mol Med*.

[BEIR, 2006] BEIR (2006). *Health Risks from Exposure to Low Levels of Ionizing Radiation BEIR VII, Phase 2*. Biological Effects of Ionizing Radiation (BEIR).

[Berardinelli et al., 2013] Berardinelli, F., Antoccia, A., Buonsante, R., Gerardi, S., Cherubini, R., Nadal, V., Tanzarella, C., and Sgura, A. (2013). The role of telomere length modulation in delayed chro- mosome instability induced by ionizing radiation in human primary fibroblasts. *Environ. Mol. Mutagen*, 54 :172–179.

[Berardinelli et al., 2011] Berardinelli, F., Antoccia, A., Cherubini, R., Nadal, V. D., Gerardi, S., Tanzarella, C., and Sgura, A. (2011). Telomere alterations and genomic instability in long-term cultures of normal human fibroblasts irradiated with x rays and protons. *Radiat. Prot.Dosim*, 143 :274–278.

[Berardinelli et al., 2012] Berardinelli, F., Nieri, D., Sgura, A., Tanzarella, C., and Antoccia, A. (2012). Telomere loss, not average telomere length, confers radiosensitivity to tk6-irradiated cells. *Mutat. Res*, 740 :13–20.

[Binzoni et al., 2006] Binzoni, T., Leung, T., AH, G., Rüfenacht, D., and Delpy, D. (2006). Comment on 'the use of the henyey–greenstein phase function in monte carlo simulations in biomedical optics'. *Phys. Med. Biol.*, 51(22) :L39.

[Bjornstad et al., 1998] Bjornstad, O., Chr, S., Saitoh, T., and OC, O. (1998). Mapping the regional transition to cyclicity in clethrionomys rufocanus : spectral densities and functional data analysis. *Res Pop Ecol*, 40 :77 – 84.

[Bouffler et al., 2001] Bouffler, S., Blasco, M., Cox, R., and Smith, P. (2001). Telomeric sequences, radiation sensitivity and genomic instability. *Int. J. Radiat. Biol*, 77 :995–1005.

[Breiman and Friedman, 1985] Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. with discussion and with a reply by the authors. *J. Amer. Statist. Assoc*, 80 :580–619.

[Buckner et al., 2004] Buckner, R., Head, D., Parker, J., Fotenos, A., Marcus, D., Morris, J., and Snyder, A. (2004). A unified approach for morphometric and functional data

analysis in young, old, and demented adults using automated atlas-based head size normalization : reliability and validation against manual measurement of total intracranial volume. *NeuroImage*, 23 :724–738.

[Burman et al., 1991] Burman, C., Kutcher, G., Emami, B., and Goitein, M. (1991). Fitting of normal tissue tolerance data to an analytic function. *Int J Radiat Oncol Biol Phys*, 21(1) :123–35.

[Cardot et al., 1999] Cardot, H., Ferraty, and Sarda, P. (1999). Functional linear model. *Statist. and Prob. Letters*, 45 :11–22.

[Cardot et al., 2003] Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13 :571–91.

[Cardot and Sarda, 2005] Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *J. Multivariate Anal.*, 92(1) :24–41.

[Castella et al., 2007] Castella, M., Puerto, S., Creus, A., Marcos, R., and Surralles, J. (2007). Telomere length modulates human radiation sensitivity in vitro. *Toxicol. Lett.*, 17 :29–36.

[Catmull, 1984] Catmull, E. (1984). An analytic visible surface algorithm for independent pixel processing. *SIGGRAPH Proc.*, page 109.

[Cech, 2004] Cech, T. (2004). Beginning to understand the end of the chromosome. *Cell*, 116 :273–9.

[Chaney et al., 1994] Chaney, E., Cullip, T., and Gabriel, T. (1994). A monte carlo study of accelerator head scatter. *Med. Phys.*, 21(9) :1383.

[Chapados and Levitin, 2008] Chapados, C. and Levitin, D. (2008). Cross-modal interactions in the experience of musical performances : physiological correlates. *Cognition*, 108 :639–651.

[Chofor et al., 2012] Chofor, N., Harder, N., Willborn, K., and Poppe, B. (2012). Internal scatter, the unavoidable major component of the peripheral dose in photon-beam radiotherapy. *Phys. Med. Biol.*, 57(6) :1733.

[Compton, 1923] Compton, A. (1923). A quantum theory of the scattering of x-rays by light elements. *Phys. Rev.*, 21(5) :483.

[Cox, 1972] Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34 :187–220.

[Crambes, 2007] Crambes, C. (2007). Functional linear model. *Comptes Rendus de l'Académie des Sciences*, 345(9) :519–522.

[Crane et al., 2010] Crane, E., Cassidy, R., Rothman, E., and Gerstner, G. (2010). Effect of registration on cyclical kinematic data. *J Biomech*, 43 :2444–2447.

[Cristy, 1981] Cristy, M. (1981). Active bone marrow distribution as a function of age in humans. *Phys Med Biol*, 26(3) :389–400.

[Dasu et al., 2005] Dasu, A., Toma-Dasu, T., Olofsson, J., and Karlsson, M. (2005). The use of risk estimation models for the induction of secondary cancers following radiotherapy. *Acta Oncol.*, 44(4) :339.

[Dauxois et al., 1982] Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function : Some applications to statistical inference. *Journal of Multivariate Analysis*, 12 :136–54.

[Dawson et al., 2005] Dawson, L., Biersack, M., and Lockwood, G. (2005). Use of principal component analysis to evaluate the partial organ tolerance of normal tissues to radiation. *Int J Radiat Oncol Biol Phys*, 62(3) :829–837.

[de Boor, 1978] de Boor, C. (1978). *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer, New York.

[de Vathaire et al., 1999a] de Vathaire, F., Hardiman, C., Shamsaldin, A., Campbell, S., Grimaud, E., Hawkins, M., Raquin, M., Oberlin, O., Diallo, I., Zucker, J., Panis, X., Lagrange, J., Daly-Schveitzer, N., Lemerle, J., Chavaudra, J., Schlumberger, M., and Bonaïti, C. (1999a). Thyroid carcinomas after irradiation for a first cancer during childhood. *Arch Intern Med*, 159(22) :2713–9.

[de Vathaire et al., 1999b] de Vathaire, F., Hardiman, C., Shamsaldin, A., Campbell, S., Grimaud, E., Hawkins, M., Raquin, M., Oberlin, O., Diallo, I., Zucker, J. M., Panis, X., Lagrange, J. L., Daly-Schveitzer, N., Lemerle, J., Chavaudra, J., Schlumberger, M., and Bonaïti, C. (1999b). Thyroid carcinomas after irradiation for a first cancer during childhood. *Archives of Internal Medicine*, 159(22) :2713–9.

[Defraene et al., 2012] Defraene, G., den Bergh, L. V., Al-Mamgani, A., Haustermans, K., Heemsbergen, W., den Heuvel, F. V., and Lebesque, J. (2012). The benefits of

including clinical factors in rectal normal tissue complication probability modeling after radiotherapy for prostate cancer. *Int J Radiat Oncol Biol Phys*, 82(1) :1233–42.

[Deng et al., 2000] Deng, J., Jiang, S., Kapur, A., Li, J., Pawlicki, T., and Ma, C. (2000). Photon beam characterization and modelling for monte carlo treatment planning. *Phys. Med. Biol.*, 45(2) :411.

[Derreumaux et al., 1995] Derreumaux, S., Chavaudra, J., Bridier, A., Rossetti, V., and Dutreix, A. (1995). A european quality assurance network for radiotherapy : dose measurement procedure. *Phys. Med. Biol.*, 40(7) :1191.

[Deville, 1974] Deville, J. (1974). *Méthodes statistiques et numériques de l'analyse harmonique*, volume 15. Paris, France.

[Di et al., 2009] Di, C., Crainiceanu, C., Caffo, B., and Punjabi, N. (2009). Multilevel functional principal component analysis. *The Annals of Applied Statistics*, 3(1) :458–488.

[Diaconis and Shashahani, 1984] Diaconis, P. and Shashahani, M. (1984). On nonlinear functions of linear combinations. siam. *J. Sci. Statist. Comput*, 5 :175–191.

[Diallo et al., 1996] Diallo, I., Lamon, A., Shamsaldin, A., Grimaud, E., de Vathaire, F., and Chavaudra, J. (1996). Estimation of the radiation dose delivered to any point outside the target volume per patient treated with external beam radiotherapy. *Radiother. Oncol.*, 38(3) :269.

[Dunscombe and Nieminen, 1992] Dunscombe, P. and Nieminen, J. (1992). On the field size dependence of relative output from a linear accelerator. *Med. Phys.*, 19(6) :1441.

[Eilers and Marx, 1996] Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing using b-splines and penalized likelihood (with comments and rejoinder). *Statistical Science*, 11(2) :89–121.

[Epanechnikov, 1969] Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications*, 14 :153–158.

[Erbas et al., 2010a] Erbas, B., Akram, M., Gertig, M., English, D., Hopper, J., Kavanagh, A., and Hyndman, R. (2010a). Using functional data analysis models to estimate future time trends in age-specific breast cancer mortality for the united states and england-wales. *J Epidemiol*, 20 :159 – 165.

[Erbas et al., 2010b] Erbas, B., Akram, M., Gertig, M., English, D., Hopper, J., Kavanagh, A., and Hyndman, R. (2010b). Using functional data analysis models to estimate future time trends in age-specific breast cancer mortality for the united states and england-wales. *J Epidemiol*, 20 :159 – 165.

[Escabias et al., 2004] Escabias, M., Aguilera, A., and Valderrama, M. (2004). Principal component estimation of functional logistic regression : discussion of two dierent approaches. *J. Nonparametr. Stat*, 16 :365–384.

[Escabias et al., 2005] Escabias, M., Aguilera, A., and Valderrama, M. (2005). Modeling environmental data by functional principal component logistic regression. *Environmetrics*, 16 :95–107.

[Fernet and Hall, 2004] Fernet, M. and Hall, J. (2004). Genetic biomarkers of therapeutic radiation sensitivity. *DNA Repair*, 3 :1237–1243.

[Ferraty and Vieu, 2002] Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Comput. Statist*, 17(4) :545–564.

[Ferraty and Vieu, 2006] Ferraty, F. and Vieu, P. (2006). Nonparametric functional data analysis. theory and practice. *Springer Science*, 5 :175–191.

[Ferraty et al., 2007] Ferraty, F., Vieu, P., and Viguier-Pla, S. (2007). Factor-based comparison of groups of curves comp. *Stat. and Data Anal*, 51(10) :4903–4910.

[Ferreira et al., 2000] Ferreira, I., Dutreix, A., Bridier, A., Chavaudra, J., and Svensson, H. (2000). The estro-quality assurance network (equal). *Radiother. Oncol.*, 55(3) :273.

[Fiorino et al., 2008] Fiorino, C., Fellin, G., Rancati, T., Vavassori, V., Bianchi, C., Borca, V., Girelli, G., Mapelli, M., Menegotti, L., Nava, S., and Valdagni, R. (2008). Clinical and dosimetric predictors of late rectal syndrome after 3d-crt for localized prostate cancer : preliminary results of a multicenter prospective study. *Int J Radiat Oncol Biol Phys*, 70(4) :1130–7.

[Fowler, 1992] Fowler, J. (1992). Brief summary of radiobiological principles in fractionated radiotherapy. *Semin Radiat Oncol*, 2(1) :16–21.

[Francois et al., 1988] Francois, P., Beurtheret, C., and Dutreix, A. (1988). Calculation of the dose delivered to organs outside the radiation beams. *Med. Phys.*, 15(6) :879.

[Frank and Friedman, 1993] Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35 :109–135.

[Friedman, 1991] Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics*, 19 :1–141.

[Gao, 2007] Gao, H. (2007). Day of week effects on diurnal ozone/nox cycles and transportation emissions in southern california. *Transp Res Part D Transp Envr*, 12 :292–305.

[Gauss, 1809] Gauss, C. (1809). *Theoria motus corporum celestium.*

[Genesca et al., 2006] Genesca, A., Martin, M., Latre, L., Soler, D., Pampalona, J., and Tusell, L. (2006). Telomere dysfunction : a new player in radiation sensitivity. *Bioessays*, 28 :1172–1180.

[Gilbert et al., 2003] Gilbert, E., Stovall, M., Gospodarowicz, M., Leeuwen, V. F., Andersson, M., Glimelius, B., Joensuu, T., Lynch, C. F., Curtis, R. E., Holowaty, E., Storm, H., Pukkala, E., van't Veer, M. B., Fraumeni, J. F., Boice, J., Clarke, E. A., and LB, T. (2003). Lung cancer after treatment for hodgkin's disease : focus on radiation effects. *Radiation Research*, 159(2) :161–73.

[Gilson and Londono-Vallejo, 2007] Gilson, E. and Londono-Vallejo, A. (2007). Telomere length profiles in humans : all ends are not equal. *Cell Cycle*, 6 :2486–2494.

[Gorbunova et al., 2002] Gorbunova, V., Seluanov, A., and Pereira-Smith, O. (2002). Expression of human telomerase (htert) does not prevent stress-induced senescence in normal human fibroblasts but protects the cells from stress-induced apoptosis and necrosis. *J. Biol.Chem*, 277 :38540–38549.

[Gramatges et al., 2014] Gramatges, M., Liu, Q., Yasui, Y., Okcu, M., Neglia, J., LC, S., Armstrong, G., Robison, L., and S, B. (2014). Telomere content and risk of second malignant neoplasm in survivors of childhood cancer : a report from the childhood cancer survivor study. *Clin Cancer Res*, 20(4) :904–11.

[Grambsch et al., 1995] Grambsch, P., Randall, B., Bostick, R., Potter, J., and Louis, T. (1995). Modeling thelabeling index distribution : An application of functional data analysis. *J Am Stat Assoc*, 90 :813–821.

[Green and Silverman, 1994] Green, P. and Silverman, B. (1994). *Nonparametric regression and generalized linear models : A roughness penalty approach.* Chapman and Hall, London.

[Greene et al., 1983] Greene, D., Chu, G., and Thomas, D. (1983). Dose levels outside radiotherapy beams. *Br. J. Radiol*, 56 :543–50.

[Guibout et al., 2005a] Guibout, C., Adjadj, E., Rubino, C., Shamsaldin, A., Grimaud, E., Hawkins, M., Mathieu, M., Oberlin, O., Zucker, J., Panis, X., Lagrange, J., Daly-Schveitzer, N., Chavaudra, J., and de Vathaire, F. (2005a). Malignant breast tumors after radiotherapy for a first cancer during childhood. *J Clin Oncol.*, 23(1) :197–204.

[Guibout et al., 2005b] Guibout, C., Adjadj, E., Rubino, C., Shamsaldin, A., Grimaud, E., Hawkins, M., Mathieu, M. C., Oberlin, O., Zucker, M. J., Panis, X., Lagrange, J. L., Daly-Schveitzer, N., Chavaudra, J., and de Vathaire, F. (2005b). Malignant breast tumors after radiotherapy for a first cancer during childhood. *Journal of clinical oncology*, 23(1) :197–204.

[Guérin et al., 2003] Guérin, S., Dupuy, A., Anderson, H., Shamsaldin, A., Svahn-Tapper, G., Moller, T., Quiniou, E., Garwicz, S., Hawkins, M., Avril, M., Oberlin, O., Chavaudra, J., and de Vathaire, F. (2003). Radiation dose as a risk factor for malignant melanoma following childhood cancer. *Eur J Cancer*, 39 :2379–86.

[H Martens and Naes, 1989] H Martens, H. and Naes, T. (1989). *Multivariate calibration*. Chapman and Hall, London, Chapman and Hall, London.

[Haddy et al., 2009a] Haddy, N., Andriamboavonjy, T., Paoletti, C., Dondon, M., Mousannif, A., Shamsaldin, A., Doyon, F., Labbé, M., Robert, C., Avril, M., Fragu, P., Eschwege, F., Chavaudra, J., Schvartz, C., Lefkopoulos, D., Schlumberger, M., Diallo, I., and de Vathaire, F. (2009a). Thyroid adenomas and carcinomas following radiotherapy for a hemangioma during infancy. *Radiother Oncol*, 93 :377–82.

[Haddy et al., 2009b] Haddy, N., Andriamboavonjy, T., Paoletti, C., Dondon, M., Mousannif, A., Shamsaldin, A., Doyon, F., Labbé, M., Robert, C., Avril, M., Fragu, P., Eschwege, F., Chavaudra, J., Schvartz, C., Lefkopoulos, D., Schlumberger, M., Diallo, I., and de Vathaire, F. (2009b). Thyroid adenomas and carcinomas following radiotherapy for a hemangioma during infancy. *Radiother Oncol*, 93(2) :377–82.

[Haddy et al., 2010] Haddy, N., Dondon, M., Paoletti, C., Rubino, C., Mousannif, A., A, S., Doyon, F., Labbé, M., Robert, C., Avril, M., Demars, R., Molinie, F., Lefkopoulos, D., Diallo, I., and de Vathaire, F. (2010). Breast cancer following radiotherapy for a hemangioma during childhood. *Cancer Causes Control*, 21(11) :1807–16.

[Haddy et al., 2012a] Haddy, N., El-Fayech, C., Guibout, C., Adjadj, E., Thomas-Teinturier, C., end C Veres, O. O., Pacquement, H., Jackson, A., Munzer, M., N'Guyen, T. D., Bondiau, P. Y., end A Laprie, D. B., Bridier, A., Lefkopoulos, D., Schlumberger,

M., Rubino, C., Diallo, I., and de Vathaire, F. (2012a). Thyroid adenomas after solid cancer in childhood. •, 84(2) :e209–15.

[Haddy et al., 2012b] Haddy, N., El-Fayech, C., Guibout, C., Adjadj, E., Thomas-Teinturier, C., Oberlin, O., Veres, C., Pacquement, H., Jackson, A., Munzer, M., N'guyen, T., Bondiau, P., Berchery, D., Laprie, A., Bridier, A., Lefkopoulos, D., Schlumberger, M., Rubino, C., Diallo, I., and de Vathaire, F. (2012b). Thyroid adenomas after solid cancer in childhood. *Int J Radiat Oncol Biol Phys*, 84(2) :e209–15.

[Haddy et al., 2006] Haddy, N., Le Deley, M., Samand, A., Diallo, I., Guérin, S., Guibout, C., Oberlin, O., Hawkins, M., Zucker, J., and de Vathaire, F. (2006). Role of radiotherapy and chemotherapy in the risk of secondary leukaemia after a solid tumour in childhood. *Eur J Cancer*, 42 :2757–64.

[Haddy et al., 2012c] Haddy, N., Mousannif, A., Paoletti, C., Dondon, M., Shamsaldin, A., Doyon, F., Avril, M., Fragu, P., Labbé, M., Lefkopoulos, D., Chavaudra, J., Robert, C., Diallo, I., and de Vathaire, F. (2012c). Radiotherapy as a risk factor for malignant melanoma after childhood skin hemangioma. *Melanoma Res*, 22(1) :77–85.

[Hall et al., 2001] Hall, P., Poskitt, D., and Presnell, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics*, 43 :1 – 24.

[Harrison et al., 2007] Harrison, A., Ryan, W., and Hayes, K. (2007). Functional data analysis of joint coordination in the development of vertical jump performance. *Sports Biomech*, 6 :199 – 214.

[Harville, 1977] Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72 :320–340.

[Hastie and Mallows, 1993] Hastie, T. and Mallows, C. (1993). Discussion of "a statistical view of some chemometrics regression tools." by frank, i.e. and friedman, j.h. *Technometrics*, 35 :140–143.

[Hastie and Tibshirani, 1980] Hastie, T. and Tibshirani, R. (1980). Generalized additive models. with discussion. *Statist. Sci*, 1 :297–318.

[Hastie and Tibshirani, 1990] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

[Hauguel and Bunz, 2003] Hauguel, T. and Bunz, F. (2003). Haploinsufficiency of htert leads to telomere dysfunction and radiosensitivity in human cancer cells. *Cancer Biol Ther*, 2 :679–684.

[Helland, 1990] Helland, I. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, 17 :97–114.

[Henderson, 2006a] Henderson, B. (2006a). Day of week effects on diurnal ozone/nox cycles and transportation emissions in southern california. *Environmetrics*, 12 :65–80.

[Henderson, 2006b] Henderson, B. (2006b). Exploring between site differences in water quality trends : a functional data analysis approach. *Environmetrics*, 17 :65–80.

[Henyey and Greenstein, 1941] Henyey, L. and Greenstein, J. (1941). Diffuse radiation in the galaxy. *Astrophys. J.*, 93 :70.

[HG Müller and Stadtmüller, 2005] HG Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, 33 :774–805.

[Hilbert, 1912] Hilbert, D. (1912). *Grundz uge einer allgemeinen Theorie der linearen Integralgleichungen.*

[Hoerl and Kennard, 1981] Hoerl, A. and Kennard, R. (1981). Ridge regression : advances, algorithms and applications. *american Journal of Mathematical Management Sciences*, 1 :5–83.

[Hörmann and Kokoszka, 2013] Hörmann, S. and Kokoszka, P. (2013). Consistency of the mean and the principal components of spatially distributed functional data. *Bernoulli*, 19 :1535—-58.

[Hotelling, 1933] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6) :417–441.

[Howell et al., 2010] Howell, R., Scarboro, S., Kry, S., and Yaldo, Z. (2010). Accuracy of out-of-field dose calculations by a commercial treatment planning system. *Phys. Med. Biol.*, 55(23) :6999.

[ICRU, 2010] ICRU (2010). Prescribing, recording, and reporting intensity-modulated photon-beam therapy (imrt). Technical Report 83, International Commission on Radiation Units and Measurements.

[Ikeda et al., 2008] Ikeda, T., Dowd, M., and Martin, J. (2008). Application of functional data analysis to investigate seasonal progression with interannual variability in plankton abundance in the bay of fundy, canada. *Estuar Coast Shelf Sci*, 78 :445 – 455.

[Ilyenko et al., 2011] Ilyenko, I., Lyaskivska, O., and Bazyka, D. (2011). Analysis of relative telomere length and apoptosis in humans exposed to ionising radiation. *Exp Oncol*, 33 :235–8.

[Ivanescu et al., 2012] Ivanescu, A., Staicu, A., Scheipl, F., and Greven, S. (2012). Penalized function-on-function regression. Technical report, Hopkins University, Dept. of Biostatistics. Technical Report 254.

[James, 2002] James, G. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B*, 64 :411–432.

[James and Sugar, 2003] James, G. and Sugar, C. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98 :397–408.

[Jian et al., 2001] Jian, S. B., Boyer, A. L., and Ma, C. M. (2001). Modeling the extrafocal radiation and monitor chamber backscatter for photon beam dose calculation. *Med. Phys.*, 28(1) :55.

[Jiang et al., 2009] Jiang, C., Aston, J., and Wang, J. (2009). Smoothing dynamic positron emission tomography time courses using functional principal components. *NeuroImage*, 47 :184 – 193.

[Jones et al., 1996] Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433) :401–407.

[Joosten et al., 2011] Joosten, A., Bochud, F., Baechler, S., Levi, F., Mirimanoff, R., and Moeckli, R. (2011). Variability of a peripheral dose among various linac geometries for second cancer risk assessment. *Phys. Med. Biol.*, 56(16) :5131.

[Karhunen, 1947] Karhunen, K. (1947). Über lineare methoden in der wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys*, 37 :1–79.

[Kase et al., 1983] Kase, K., Svensson, G., Wolbarst, A., and Marks, M. (1983). Measurements of dose from secondary radiation outside a treatment field. *Int. J. Radiat. Oncol. Biol. Phys.*, 9(8) :1177.

[Kneip and Utikal, 2001] Kneip, A. and Utikal, K. (2001). Inference for density families using functional principal components analysis. *Journal of the American Statistical Association*, 96 :519–31.

[Koenig et al., 2008] Koenig, L., Lucero, J., and Perlman, E. (2008). Speech production variability in fricatives of children and adults : results of functional data analysis. *J Acoust Soc Am*, 124 :3158–3170.

[Kooperberg et al., 1995] Kooperberg, C., Stone, C., and Truong, Y. (1995). Hazard regression. *Journal of the American Statistical Association*, 90 :78–94.

[Kovalchik et al., 2013] Kovalchik, S., Ronckers, C., Veiga, L., Sigurdson, A., Inskip, P., de Vathaire, F., Sklar, C., Donaldson, S., Anderson, H., Bhatti, P., Hammond, S., Leisenring, W., Mertens, A., Smith, S., Stovall, M., Tucker, M., Weathers, R., Robison, L., and Pfeiffer, R. (2013). Absolute risk prediction of second primary thyroid cancer among 5-year survivors of childhood cancer. *J Clin Oncol*, 31 :119–127.

[Kovalenko et al., 2010] Kovalenko, O., Kaplunov, J., Herbig, U., Detoledo, S., Azzam, E., and Santos, J. (2010). Expression of (nes-)htert in cancer cells delays cell cycle progression and increases sensitivity to genotoxic stress. *PLoS ONE*, 5 :e10812.

[Kry et al., 2007] Kry, S., Titt, U., Followill, D., Pönisch, F., Vassiliev, O., White, R., Stovall, M., and Salehpour, M. (2007). A monte carlo model for out-of-field dose calculation from high-energy photon therapy. *Med. Phys.*, 34(9) :3489.

[Kry et al., 2006] Kry, S., Titt, U., Pönisch, andON Vassiliev, D. F., White, R., Mohan, R., and Salehpour, M. (2006). A monte carlo model for calculating out-of-field dose from a varian 6-mv beam. *Med. Phys.*, 33(11) :4405.

[Kurvinen et al., 2006] Kurvinen, K., Rantanen, V., Syrjanen, S., and Johansson, B. (2006). Radiation-induced effects on telomerase in gynecological cancer cell lines with different radiosensitivity and repair capacity. *Int. J. Radiat. Biol*, 82 :859–867.

[Kutcher et al., 1991] Kutcher, G., Burman, C., Brewster, L., and M Goitein, R. M. (1991). Histogram reduction method for calculating complication probabilities for three-dimensional treatment planning evaluations. *Int J Radiat Oncol Biol Phys*, 21(1) :137–46.

[Latre et al., 2003] Latre, L., Tusell, L., Martin, M., Miro, R., Egozcue, J., Blasco, M., and Genesca, A. (2003). Shortened telomeres join to dna breaks interfering with their correct repair. *Exp. Cell Res*, 287 :282–288.

[Laukaitis, 2005] Laukaitis, A. (2005). Functional data analysis for clients segmentation tasks. *Eur J Oper Res*, 163 :2010–2016.

[Laukaitis, 2008] Laukaitis, A. (2008). Functional data analysis for cash flow and transactions intensity continuous-time prediction using hilbert-valued autoregressive processes. *Eur J Oper Res*, 185 :1607 – 1614.

[Le Vu et al., 1998] Le Vu, B., de Vathaire, F., Shamsaldin, A., Hawkins, M., Grimaud, E., Hardiman, C., Diallo, I., Vassal, G., Bessa, E., Campbell, S., Panis, X., Daly-Schveitzer, N., Lagrange, J., Zucker, J., Eschwège, F., Chavaudra, J., and Lemerle, J. (1998). Radiation dose, chemotherapy and risk of osteosarcoma after solid tumours during childhood. *Int J Cancer*, 77 :370–77.

[Lee et al., 2006] Lee, S., Byrd, D., and KrivokapicÌ, J. (2006). Functional data analysis of prosodic effects on articulatory timing. *J Acoust Soc Am*, 119 :1666–1671.

[Legendre, 1809] Legendre, A. (1809). *Uber lineare Funktionalgleichungen.*

[Li et al., 2012] Li, P., Hou, M., Lou, F., Bjorkholm, M., and Xu, D. (2012). Telomere dysfunction induced by chemotherapeutic agents and radiation in normal human cells. *Int. J. Biochem. Cell Biol*, 44 :1531–1540.

[Little et al., 1998] Little, M., de Vathaire, F., Shamsaldin, A., Oberlin, O., Campbell, S., Grimaud, E andChavaudra, J., Haylock, R., and Muirhead, C. (1998). Risks of brain tumour following treatment for cancer in childhood : modification by genetic factors, radiotherapy and chemotherapy. *Int J Cancer*, 78 :269–75.

[Loader, 1999] Loader, C. (1999). Bandwidth selection : Classical or plug-in ? *The Annals of Statistics*, 27(2) :415–138.

[Long et al., 2005] Long, C., Brown, E., Triantafyllou, C., Aharon, I., Wald, L., and Solo, V. (2005). Nonstationary noise estimation in functional mri. *NeuroImage*, 28 :890 – 903.

[Loève, 1945] Loève, M. (1945). Functions aleatoire de second ordre. *C R Academie des Sciences*, 37 :1–79.

[Lucero, 2005] Lucero, J. (2005). Comparison of measures of variability of speech movement trajectories using synthetic records. *J Speech Lang Hear Res*, 48 :336 – 344.

[Lundell et al., 1999] Lundell, M., Mattsson, A., Karlsson, P., Holmberg, E., Gustafsson, A., and Holm, L. E. (1999). Breast cancer risk after radiotherapy in infancy : a pooled analysis of two swedish cohorts of 17,202 infants. *Radiation Research*, 151(5) :626–32.

[Lyman, 1985] Lyman, J. (1985). Complication probability as assessed from dose-volume histograms. *Radiat Res Suppl*, 8 :S13––S19.

[Maeda et al., 2013] Maeda, T., Nakamura, K., Atsumi, K., Hirakawa, M., Ueda, Y., and Makino, N. (2013). Radiation- associated changes in the length of telomeres in peripheral leukocytes from inpatients with cancer. *Int. J. Radiat. Biol*, 89 :106–109.

[Malloy et al., 2009] Malloy, E. J., Spiegelman, D., and Eisen, E. A. (2009). Comparing measures of model selection for penalized splines in cox models. *Comput Stat Data Anal*, 53(7) :2605–2616.

[Manté et al., 2005a] Manté, C., Durbec, J., and Dauvin, J. (2005a). A functional data-analytic approach to the classification of species according to their spatial dispersion. application to a marine macrobenthic community from the bay of morlaix (western english channel). *J Appl Stat*, 32 :831 – 840.

[Manté et al., 2005b] Manté, C., Durbec, J., and Dauvin, J. (2005b). A functional data-analytic approach to the classification of species according to their spatial dispersion. application to a marine macrobenthic community from the bay of morlaix (western english channel). *J Appl Stat*, 32 :831 – 840.

[Marre et al., 2000] Marre, D., Ferreira, I., Bridier, A., Björeland, A., Svensson, H., Dutreix, A., and Chavaudra, J. (2000). Energy correction factors of lif powder tlds irradiated in high-energy electron beams and applied to mailed dosimetry for quality assurance networks. *Phys. Med. Biol.*, 45(12) :3657.

[Maslova et al., 2010] Maslova, I., Kokoszka, P., Sojka, J., and Zhu, L. (2010). Statistical significance testing for the association of magnetometer records at high-, mid- and low latitudes during substorm days. *Planet Space Sci*, 88 :437 – 445.

[Massy, 1965] Massy, W. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60 :234–256.

[McAdams, 2004] McAdams, S. (2004). Influences of large-scale form on continuous ratings in response to a contemporary piece in a live concert setting. *Music Percept*, 22 :297 – 350.

[McParland and Fair, 1991] McParland, B. and Fair, H. (1991). A method of calculating peripheral dose distributions of photon beams below 10 mv. *Med. Phys.*, 19(2) :283.

[Menu-Branthomme et al., 2004a] Menu-Branthomme, A., Rubino, C., Shamsaldin, A., Hawkins, M., Grimaud, E., Dondon, M., Hardiman, C., Vassal, G., Campbell, S., Panis, X., Daly-Schveitzer, N., Lagrange, J., Zucker, J., Chavaudra, J., Hartman, O., and de Vathaire, F. (2004a). Radiation dose, chemotherapy and risk of soft tissue sarcoma after solid tumours during childhood. *Int J Cancer*, 110 :87–93.

[Menu-Branthomme et al., 2004b] Menu-Branthomme, A., Rubino, C., Shamsaldin, A., Hawkins, M., Grimaud, E., Dondon, M. G., Hardiman, C., Vassal, G., Campbell, S., Panis, X., Daly-Schveitzer, N., Lagrange, J. L., Zucker, J. M., Chavaudra, J., Hartman, O., and de Vathaire, F. (2004b). Radiation dose, chemotherapy and risk of soft tissue sarcoma after solid tumours during childhood. *International journal of cancer*, 110(1) :87–93.

[Michalski et al., 2010] Michalski, J., Gay, H., Jackson, A., Tucker, S., and JO, D. (2010). Radiation dose-volume effects in radiation-induced rectal injury. *Int J Radiat Oncol Biol Phys*, 76(3) :S123-S129. Supplement.

[M'Kacher et al., 2007] M'Kacher, R., Bennaceur-Griscelli, A., Girinsky, T., Koscielny, S., Delhommeau, F., Dossou, J., and et al (2007). Telomere shortening and associated chromosomal instability in peripheral blood lymphocytes of patients with hodgkin's lymphoma prior to any treatment are predictive of second cancers. *Int J Radiat Oncol Biol Phys*, 68 :465–71.

[M'kacher et al., 2014] M'kacher, R., Girinsky, T., Colicchio, B., Ricoul, M., Dieterlen, A., Jeandidier, E., Heidingsfelder, L., Cuceu, C., Shim, G., Frenzel, M., Lenain, A., Morat, L., Bourhis, J., WM, H., Koscielny, S., Paul, J., Carde, P., and Sabatier, L. (2014). Telomere shortening : a new prognostic factor for cardiovascular disease post-radiation exposure. *Radiat Prot Dosimetry*.

[Müller et al., 2008] Müller, H., Chiou, J., and Leng, X. (2008). Inferring gene expression dynamics via functional regression analysis. *BMC Bioinformatics*, 33(1) :9–60.

[Nakamura et al., 2005] Nakamura, M., Masutomi, K., Kyo, S., Hashimoto, M., Maida, Y., Kanaya, T., M.Tanaka, Hahn, W., and Inoue, M. (2005). Efficient inhibition of human telomerase reverse transcriptase expression by rna interference sensitizes cancer cells to ionizing radiation and chemotherapy. *Hum. Gene Ther*, 16 :859–868.

[NCRP, 2012] NCRP (2012). *Report No 170 Second primary cancers and cardiovascular disease after radiation therapy.* National Council on Radiation Protection and Measurements (NCRP).

[Neglia et al., 2006] Neglia, J. P., Robison, L. L., Stovall, M., Liu, Y., Packer, R. J., Hammond, S., Yasui, Y., Kasper, C. E., Mertens, A. C., Donaldson, S. S., Meadows, A. T., and Inskip, P. D. (2006). New primary neoplasms of the central nervous system in survivors of childhood cancer : a report from the childhood cancer survivor study. *Journal of the National Cancer Institute*, 98(21) :1528–37.

[Niemierko, 1997] Niemierko, A. (1997). Reporting and analyzing dose distributions : a concept of equivalent uniform dose. *Med Phys*, 24(1) :103–10.

[Nieri et al., 2013] Nieri, D., Berardinelli, F., Sgura, A., Cherubini, R., Nadal, V. D., Gerardi, S., Tanzarella, C., and Antoccia, A. (2013). Cyogenetics effects in ag01522 human primary fibro- blasts exposed to low-doses of radiations with different quality. *Int. J. Radiat.Biol*, 89 :698–707.

[Ogden and Greene, 2010] Ogden, R. and Greene, E. (2010). Wavelet modeling of functional random effects with application to human vision data. *J Stat Plan Inference*, 140 :3797 – 3808.

[Parker and Wen, 2009] Parker, B. and Wen, J. (2009). Predicting microrna targets in time-series microarray experiments via functional data analysis. *BMC Bioinforma*, 10 :S32.

[Pawitan, 2001] Pawitan, Y. (2001). *In All Likelihood : Statistical Modelling and Inference Using Likelihood.* Oxford University Press.

[Pearson, 1901] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6) :559–572.

[Peeters et al., 2006] Peeters, S., Hoogeman, M., Heemsbergen, W., Hart, A., Koper, P., and Lebesque, J. (2006). Rectal bleeding, fecal incontinence, and high stool frequency after conformal radiotherapy for prostate cancer : normal tissue complication probability modeling. *Int J Radiat Oncol Biol Phys*, 66(1) :11–9.

[Pernot et al., 2012] Pernot, E., Hall, J., Baatout, S., Benotmane, M., Blanchardon, E., Bouffler, S., H. El Saghireand M. Gomolka, A. G., Harms-Ringdahl, M., Jeggo, P., Kreuzer, M., D.Laurier, Lindholm, C., Mkacher, R., Quintens, R., Rothkamm, K., Sabatier,

L., Tapio, S., de Vathaire, F., and Cardis, E. (2012). Ionizing radiation biomarkers for potential use in epidemiological studies. *Mutat. Res.*, 751 :258–286.

[Pirzio et al., 2004] Pirzio, L., Freulet-Marriere, M., Bai, Y., Fouladi, B., Murnane, J., Sabatier, L., and Desmaze, C. (2004). Human fibroblasts expressing htert show remarkable karyotype stability even after exposure to ionizing radiation. *Cytogenet. Genome Res.*, 104 :87–94.

[Preston et al., 1993] Preston, D. L., Lubin, J. H., and Pierce, D. A. (1993). *EPICURE User's guide*. Seattle, WA, HiroSoft Inter- national Corp.

[Ramsay, 2000] Ramsay, J. (2000). Functional components of variation in handwriting. *J Am Stat Assoc*, 95 :9–15.

[Ramsay, 2007] Ramsay, J. (2007). Introduction to functional data analysis. *Can Psychol*, 48 :135 – 155.

[Ramsay et al., 1996] Ramsay, J., Munhall, K., Gracco, V., and Ostry, D. (1996). Functional data analyses of lip motion. *J Acoust Soc Am*, 99 :3718 – 3727.

[Ramsay and Ramsey, 2002] Ramsay, J. and Ramsey, B. (2002). Functional data analysis of the dynamics of themonthly index of nondurable goods production. *J Econ*, 107 :327 – 344.

[Ramsay and Silverman, 2002] Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis : Methods and Case Studies*.

[Ramsay et al., 2011] Ramsay, J. O., Ramsay, T., and Sangalli, L. M. (2011). *Recent Advances in Functional Data Analysis and Related Topics*, chapter Spatial Functional Data Analysis, pages 269–75. PhysicaVerlag HD Springer.

[Ramsay and Silverman, 2005] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York, 2nd edition.

[Rao, 1958] Rao, C. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, 14 :1–17.

[Ratcliffe et al., 2002a] Ratcliffe, S., Heller, G., and Leader, L. (2002a). Functional data analysis with application to periodically stimulated foetal heart rate data. ii : logistic regression. *Stat Med*, 21 :1115 – 1127.

[Ratcliffe et al., 2002b] Ratcliffe, S., Leader, L., and Heller, G. (2002b). Functional data analysis with application to periodically stimulated foetal heart rate data. i : functional regression. *Stat Med*, 21 :1103 – 1114.

[Reiss et al., 2010] Reiss, P., Huang, and Mennes, M. (2010). Fast function-on-scalar regression with penalized basis expansions. *The International Journal of Biostatistics*, 6 :28.

[Reiss and Ogden, 2009] Reiss, P. and Ogden, R. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 71 :505–523.

[Reiss and Ogden, 2010] Reiss, P. and Ogden, R. (2010). Functional generalized linear models with images as predictors. *Biometrics*, 66 :61–69.

[Rice and Wu, 2001] Rice, J. and Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57 :253–259.

[Richardson and Wing, 2007] Richardson, D. B. and Wing, S. (2007). Leukemia mortality among workers at the savannah river site. *American journal of epidemiology*, 166(9) :1015–22.

[Riesz, 1918] Riesz, F. (1918). Nonparametric functional data analysis. theory and practice. *Acta Mathematica*, 41(1) :71–98.

[Ron et al., 1995] Ron, E., Lubin, J. H., Shore, R. E., Mabuchi, K., Modan, B., Pottern, L. M., Schneider, A. B., Tucker, M. A., and Boice, J. D. (1995). Thyroid cancer after exposure to external radiation : A pooled analysis of seven studies. *Radiation Research*, 141(3) :259–77.

[Ronckers et al., 2006] Ronckers, C. M., Sigurdson, A. J., Stovall, M., Smith, S. A., Mertens, A. C., Liu, Y., Hammond, S., Land, C. E., Neglia, J. P., Donaldson, S. S., Meadows, A. T., Sklar, C. A., Robison, L. L., and Inskip, P. D. (2006). Thyroid cancer in childhood cancer curvivors : a detailed evaluation of radiation dose response and its modifiers. *Radiation Research*, 166(4) :618–628.

[Rosenblatt et al., 2013] Rosenblatt, E., Izewska, J., Anacak, Y., Pynda, Y., Scalliet, P., and end PAutier P., M. B. (2013). Radiotherapy capacity in european countries : an analysis of the directory of radiotherapy centres (dirac) database. *Lancet Oncology*, 14(2) :e79–86.

[Royston et al., 2005] Royston, P., Altman, D. G., and Sauerbrei, W. (2005). Dichotomizing continuous predictors in multiple regression : a bad idea. *Statistics in Medicine*, 25(1) :127–141.

[Ruben et al., 2011] Ruben, J., Lancaster, C., Jones, P., and Smith, R. (2011). A comparison of out-of-field dose and its constituent components for intensity-modulated radiation therapy versus conformal radiation therapy : implications for carcinogenesis. *Int. J. Radiat. Oncol. Biol. Phys.*, 81(5) :1458.

[Ruben et al., 2014] Ruben, J. D., Smith, R., Lancaster, C. M., Haynes, M., Jones, P., and Panettieri, V. (2014). Constituent components of out of field scatter dose for 18 mv intensity modulated radiation therapy versus 3 dimensional conformal radiation therapy : A comparison with 6 mv and implications for carcinogenesis. *Int. J. Radiat. Oncol*, 90 :645–53.

[Rubino et al., 2002] Rubino, C., Cailleux, A., De Vathaire, F., and Schlumberger, M. (2002). Thyroid cancer after radiation exposure. *Eur J Cancer*, 38(5) :645–7.

[Rubio et al., 2004] Rubio, M., Davalos, A., and Campisi, J. (2004). Telomere length mediates the effects of telomerase on the cellular response to genotoxic stress. *Exp. Cell Res*, 298 :17–27.

[Rubio et al., 2002] Rubio, M., Kim, S., and Campisi, J. (2002). Reversible manipulation of telomerase expres- sion and telomere length. implications for the ionizing radiation response and replicative senescence of human cells. *J. Biol. Chem*, 277 :28609–28617.

[Ruppert, 2002] Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4) :735–757.

[Ruppert et al., 2003] Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

[Sauder et al., 2002] Sauder, C., Cardot, H., Disenhaus, C., and Cozler, Y. L. (2002). Non-parametric approaches to the impact of holstein heifer growth from birth to insemination on their dairy performance at lactation one. *The Journal of Agricultural Science*, 151 :578–589.

[Scarboro et al., 2011] Scarboro, S., Followill, D., Howell, D., and Kry, S. (2011). Variations in photon energy spectra of a 6 mv beam and their impact on tld response. *Med. Phys.*, 38(5) :2619.

[Schneider et al., 2005] Schneider, U., Zwahlen, D., Ross, D., and Kaser-Hotz, B. (2005). Estimation of radiation-induced cancer from three-dimensional dose distributions : concept of organ equivalent dose. *Int. J. Radiat. Oncol. Biol. Phys.*, 61(5) :1510.

[Schroder et al., 2001] Schroder, C., Wisman, G., de Jong, S., van der Graaf WT, MH, R., Mulder, N., and et al. (2001). Telomere length in breast cancer patients before and after chemotherapy with or without stem cell transplantation. *Br J Cancer*, 84 :1348–53.

[Schwartz et al., 2014] Schwartz, B., Benadjaoud, M., Cléro, E., Haddy, N., El-Fayech, C., Guibout, C., Teinturier, C., Oberlin, O., Veres, C., Pacquement, H., Munzer, M., N'guyen, T., Bondiau, P., Berchery, D., Laprie, A., Hawkins, M., Winter, D., D, L., Chavaudra, J., Rubino, C., Diallo, I., Bénichou, J., and de Vathaire, F. (2014). Risk of second bone sarcoma following childhood cancer : role of radiation therapy treatment. *Radiat Environ Biophys*, 53 :381–90.

[Scott, 1992] Scott, D. W. (1992). *Multivariate Density Estimation : Theory, Practice, and Visualization*. New York.

[Sheather and Jones, 1991] Sheather, S. and Jones, M. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc.B*, 53(3) :683–690.

[Shim et al., 2014] Shim, G., Ricoul, M., Hempel, W., Azzam, E., and Sabatier, L. (2014). Crosstalk between telomere maintenance and radiation effects : A key player in the process of radiation-induced carcinogenesis. *Mutat Res Rev Mutat Res*, pages 1–17.

[Silverman, 1985] Silverman, B. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal Royal Statistical Society*, 47(1) :1–52.

[Skwarchuk et al., 2000] Skwarchuk, M., Jackson, A., Zelefsky, M., Venkatraman, E., Cowen, D., Levegrün, S., Burman, C., Fuks, Z., Leibel, S., and Ling, C. (2000). Late rectal toxicity after conformal radiotherapy of prostate cancer (i) : multivariate analysis and dose-response. *Int J Radiat Oncol Biol Phys*, 47(1) :103–13.

[Söhn et al., 2007] Söhn, M., Alber, M., and Yan, D. (2007). Principal component analysis-based pattern analysis of dose-volume histograms and influence on rectal toxicity. *Int J Radiat Oncol Biol Phys*, 69(1) :230–9.

[Soler et al., 2009] Soler, D., Pampalona, J., Tusell, L., and Genesca, A. (2009). Radiation sensitivity increases with proliferation-associated telomere dysfunction in nontransformed human epithelial cells. *Aging Cell*, 8 :414–425.

[Sørensen et al., 2013] Sørensen, H., Goldsmith, J., and Sangalli, L. (2013). An introduction with medical applications to functional data analysis. *Statistics in Medecine*, 32 :5222–5240.

[Stier et al., 2004] Stier, A., Stein, J., Schwaiger, M., and Heidecke, C. (2004). Modeling of esophageal bolus flow by functional data analysis of scintigrams. *Dis Esophagus*, 17 :51 – 57.

[Stone, 1980a] Stone, C. (1980a). Additive regression and other nonparametric models. *Ann. Statist*, 13 :689–705.

[Stone, 1980b] Stone, C. (1980b). Optimal rates of convergence for nonparametric estimators. *Ann. Statist*, 8 :1348–1360.

[Stone, 1982] Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist*, 10 :1040–1053.

[Stone, 1983] Stone, C. (1983). *Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. Recent advances in statistics.* New York. p393-406.

[Stone et al., 1997] Stone, C., Hansen, M., Kooperberg, C., , and Truong, Y. (1997). The use of polynomial splines and their tensor products in extended linear modeling (with discussion). *Annals of Statistics*, 25(4) :1371–1425.

[Stovall et al., 1995] Stovall, M., Blackwell, C., Cundiff, J., Novack, D., Palta, J., Wagner, L. K., Webster, E. W., and Shalek, R. (1995). Fetal dose from radiotherapy with photon beams : Report of aapm radiation therapy committee task group. *Med. Phys*, 36 :22–63.

[Stovall et al., 1989] Stovall, M., Smith, S., and Rosenstein, M. (1989). Tissue doses from radiotherapy of cancer of the uterine cervix. *Med. Phys.*, 16 :726.

[Sylvestre and Abrahamowicz, 2008] Sylvestre, M. P. and Abrahamowicz, M. (2008). Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medecine*, 27(14) :2618–34.

[Taylor and Yu, 2002] Taylor, J. M. G. and Yu, M. G. (2002). Bias and efficiency loss due to categorizing an explanatory variable. *Journal of Multivariate Analysis*, 83 :248–63.

[Thames and Hendry, 1987] Thames, H. and Hendry, J. (1987). *Fractionation in radiotherapy*. Taylor and Francis, London.

[Tomita et al., 2013] Tomita, N., Soga, N., Ogura, Y., Hayashi, N., Shimizu, H., Kubota, T., Ito, J., Hirata, K., Ohshima, Y., Tachibana, H., and Kodaira, T. (2013). Preliminary analysis of risk factors for late rectal toxicity after helical tomotherapy for prostate cancer. *J Radiat Res*, 54(1) :919–24.

[Tucker, 1958] Tucker, L. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika*, 23 :19–23.

[Ullah and Finch, 2010a] Ullah, S. and Finch, C. (2010a). Functional data modelling approach for analysing and predicting trends in incidence rates-an application to falls injury. *Osteoporos Int*, 21 :2125 – 2134.

[Ullah and Finch, 2010b] Ullah, S. and Finch, C. (2010b). Functional data modelling approach for analysing and predicting trends in incidence rates-an application to falls injury. *Osteoporos Int*, 21 :2125 – 2134.

[Ullah and Finch, 2013] Ullah, S. and Finch, C. (2013). Applications of functional data analysis : A systematic review. *BMC Medical Research Methodology*, 13 :43.

[UNSCEAR, 2006] UNSCEAR (2006). *Effects of ionizing radiation*. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR).

[Van der Giessen, 1994] Van der Giessen, P. (1994). Calculation and measurement of the dose at points outside the primary beam for photon energies of 6, 10, and 23 mv. *Int. J. Radiat. Oncol. Biol. Phys.*, 30(5) :1239.

[Van der Giessen and Hurkmans, 1993] Van der Giessen, P. and Hurkmans, C. (1993). Calculation and measurement of the dose in points outside the primary beam for 60 co gamma radiation. *Int. J. Radiat. Oncol. Biol. Phys.*, 27(3) :717.

[Veiga et al., 2012] Veiga, L. H., Lubin, J. H., Anderson, H., de Vathaire, F., Tucker, M., Bhatti, P., Schneider, A., Johansson, R., Inskip, P., Kleinerman, R., Shore, R., Pottern, L., Holmberg, E., Hawkins, M. M., Adams, M. J., Sadetzki, S., Lundell, M., Sakata, R., Damber, L., Neta, G., and Ron, E. (2012). A pooled analysis of thyroid cancer incidence following radiotherapy for childhood cancer. *Radiation Research*, 178(4) :365–76.

[Vesprini et al., 2011] Vesprini, D., Sia, M., Lockwood, G., Moseley, D., Rosewall, T., Bayley, A., Bristow, R., Chung, P., Meńard, C., Milosevic, M., Warde, P., and Catton,

C. (2011). Role of principal component analysis in predicting toxicity in prostate cancer patients treated with hypofractionated intensity-modulated radiation therapy. *Int J Radiat Oncol Biol Phys.*

[Vines et al., 2005] Vines, B., Nuzzo, R., and Levitin, D. (2005). Analysing temporal dynamics in music. *Music Percept*, 23 :137–152.

[Viviani et al., 2005] Viviani, R., Gron, G., and M, M. S. (2005). Functional principal component analysis of fmri data. *Hum Brain Mapp*, 24 :109 – 129.

[Wahba, 1990] Wahba, G. (1990). *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics. Philadelphia.

[Wand and Jones, 1995] Wand, M. and Jones, M. (1995). *Kernel Smoothing.* Chapman and Hall.

[Wasserman, 2006] Wasserman, L. (2006). *All of Nonparametric Statistics.* Berlin.

[Wesbuer et al., 2010] Wesbuer, S., Lanvers-Kaminsky, C., Duran-Seuberth, I., Bolling, T., Schafer, K., Braun, Y., Willich, N., and Greve, B. (2010). Association of telomerase activity with radio- and chemosensitivity of neuroblastomas. *Radiat. Oncol*, 5 :66.

[West et al., 2007] West, R., Harris, K., Gilthorpe, M., Tolman, C., and Will, E. (2007). Functional data analysis applied to a randomized controlled clinical trial in hemodialysis patients describes the variability of patient responses in the control of renal anemia. *J Am Soc Nephrol*, 18 :2371 – 2376.

[Wold, 1975] Wold, H. (1975). Soft modelling by latent variables : the non-linear iterative partial least squares (nipals) approach. perspectives in probability and sta- tistics (papers in honour of m. s. bartlett on the occasion of his 65th birthday). *Applied Probability*, pages 117–142.

[Wu and Muller, 2010] Wu, P. and Muller, H. (2010). Functional embedding for the classification of geneexpression profiles. *Bioinformatics*, 26 :509 – 517.

[Xu et al., 2008] Xu, X., B, B. B., and Paganetti, H. (2008). A review of dosimetry studies on external-beam radiation treatment with respect to second cancer induction. *Phys. Med. Biol.*, 53(13) :R193.

[Yang et al., 2002] Yang, Y., Xing, L., Boyer, A., Song, Y., and Hu, Y. (2002). A three-source model for the calculation of head scatter factors. *Med. Phys.*, 29(9) :2024.

[Yao et al., 2005a] Yao, F., Müller, H., and Wang, J. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100 :577–590.

[Yao et al., 2005b] Yao, F., Müller, H., and Wang, J. (2005b). Functional linear regression analysis for longitudinal data. *Annals of Statistics*, 33(1) :2873–2903.

[Yu and Sloboda, 1996] Yu, M. and Sloboda, R. (1996). Analytical representation of head scatter factors for shaped photon beams using a two-component x-ray source model. *Med. Phys.*, 23(6) :973.

[Zhao and Kolonel, 1992] Zhao, L. P. and Kolonel, L. N. (1992). Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. *American journal of epidemiology*, 136(4) :464–74.

[Zhou et al., 2010] Zhou, F., Xiong, J., Luo, Z., Dai, J., Yu, H., Liao, Z., Lei, H., Xie, C., and Zhou, Y. (2010). cdna expression analysis of a human radiosensitive-radioresistant cell line model identifies telomere function as a hallmark of radioresistance. *Radiat. Res*, 174 :550–557.

Annexe A

Estimation de densité de probabilité

Sommaire

A.1 Introduction 156

A.1.1 Définitions 156

A.2 Estimation non paramétrique d'une densité de probabilité . . 157

A.3 Différentes méthodes non paramétriques d'estimation de la densité de probabilité 158

A.3.1 Histogramme de densité 159

A.3.2 Estimateur à noyau d'une densité 161

A.1 Introduction

Nous procédons dans ce chapitre à quelques rappels de statistique concernant *la fonction de densité de probabilité* ainsi que de leurs méthodes d'estimation non paramétriques.

A.1.1 Définitions

La *fonction de répartition* F_X d'une variable aléatoire réelle X est la fonction $F_X : \mathbb{R} \to [0,1]$ définie par :

$$F_X(x) = \mathbb{P}(X \leq x) \tag{A.1}$$

Les fonctions de répartition étant définies à partir d'une mesure de probabilité, nous pouvons établir les propriétés suivantes :

- F_X est croissante et continue à droite.
- $lim_{x \to +\infty} F_X(x) = 1$ et $lim_{x \to -\infty} F_X(x) = 0$

On dit qu'une fonction continue par morceaux : $f : \mathbb{R} \to \mathbb{R}$ est une *densité de probabilité* lorsqu'elle vérifie :

$$f \geq 0 \quad \text{et} \quad \int_{-\infty}^{+\infty} f(x)\ dx = 1$$

Il est important à ce stade de rappeler la définition de la fonction de répartition empirique construite à partir des réalisations $X_1, ..., X_n$ d'une variable aléatoire X définie pour tout réel $a \in \mathbb{R}$:

$$\widehat{F_n}(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{]-\infty, a]}(X_i) \tag{A.2}$$

En d'autres termes, cette fonction comptabilise le pourcentage des observations $\leq a$.

On dit qu'une variable aléatoire réelle X possède une loi de densité f lorsque pour tout intervalle $I \subset \mathbb{R}$:

$$\mathbb{P}(X \in I) = \int_I f(x)\ dx \tag{A.3}$$

Ainsi, La fonction de répartition F_X est la primitive de f valant 1 en $+\infty$. Voici quelques exemples de lois à densité :

- loi uniforme sur $[a,b]$: $f(x) = \frac{1}{b-a} \times \mathbf{1}_{[a,b]}(x)$ avec $a < b$.
- loi exponentielle sur $\mathbb{R}$: $f(x) = \lambda \times \exp(-\lambda x) \times \mathbf{1}_{\mathbb{R}_+}(x)$ avec $\lambda > 0$
- loi de Weibull à deux paramètres sur $[0, +\infty[$: $f(x) = \frac{k}{\lambda} \times \left(\frac{x}{\lambda}\right)^{k-1} \times \exp\left(-(x/\lambda)^k\right)$
- loi normale (ou gaussienne) $f(x) = \frac{1}{\sqrt{2\pi}} \times \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$ avec $m \in \mathbb{R}$ et $\sigma > 0$

A.2 Estimation non paramétrique d'une densité de probabilité

A présent, supposons que nous avons à notre disposition un échantillon de données observées supposée issues d'une fonction de densité de probabilité inconnue. L'estimation de densité de probabilité est la construction d'une estimation de la fonction de densité de probabilité à partir de données observées.
Plus précisément, notons $\{X_1, ..., X_n\}$ n observations indépendantes issues d'une loi de probabilité à densité inconnue notée f.
Pour estimer la densité f, deux approches sont possibles : l'estimation paramétrique et l'estimation non paramétrique.

L'estimation paramétrique repose sur une forme *a priori* de la densité de probabilité sous-jacente à nos observations. Dans ce cas, la densité f est inconnue uniquement parce qu'elle dépend d'un nombre fixé à l'avance de paramètre(s) inconnu(s).

Par exemple, si l'on suppose que nos observations $X,...,,X_n \sim N(m, \sigma^2)$, alors la densité :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \times \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

peut être estimée par *substitution* ou *plug-in* :

$$\widehat{f}(x) = \frac{1}{\hat{\sigma}\sqrt{2\pi}} \times \exp\left(-\frac{(x-\hat{m})^2}{2\hat{\sigma}^2}\right)$$

avec $\hat{m} = \overline{X_n} = \frac{X_1+...+X_n}{n}$ et $\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$ les estimateurs usuels du maximum de vraisemblance.

En résumé, l'approche paramétrique consiste à spécifier une forme de densité ramenant ainsi le problème d'estimation de densité à celui d'estimation de ses paramètres.

Les risques de biais encourus lorsqu'on effectue une estimation paramétrique d'une densité de probabilité est double : l'inadéquation de la forme *a priori* choisie pour f et l'estimation du paramètre correspondant.
En effet, même s'il est courant de retrouver les lois de probabilité usuelles dans des résultats asymptotiques généraux de la théorie statistique, on rencontre très souvent dans la pratique des lois observées plus "exotiques" et difficiles à résumer, du moins en première approche, à une formule analytique explicite.

L'avantage d'une approche *non paramétrique* du problème est le fait qu'aucune spécification a priori de la loi de densité de probabilité des $X_1, ..., X_n$ n'est requise évitant ainsi le biais lié à l'erreur de spécification.

A.3 Différentes méthodes non paramétriques d'estimation de la densité de probabilité

Dans cette question nous allons voir comment il est possible d'estimer une densité de probabilité uniquement à partir des données et sans avoir recours à une hypothèse supplémentaire quant à la forme analytique de celle-ci.

Le point de départ est le théorème de **Glivenko-Cantelli** qui relie la fonction de répartition empirique définie par l'équation (A.2) à la fonction de répartition "vraie" F à l'origine de l'échantillonnage des donnée observées.
En effet, on démontre que :

$$\sup_a \left| \widehat{F_n}(a) - F(a) \right| \underset{n \to +\infty}{\longrightarrow} 0$$

Ce résultat montre donc que pour accéder à une loi de probabilité, il suffit d'accumuler les observations générées à partir de celle-ci de façon à ce que l'écart *maxima* entre la fonction de répartition empirique et la fonction de répartition sous-jacente tende vers zéro.

Reste à présent à définir ce qu'est un "bon" estimateur d'une densité de probabilité. Le critère le plus couramment utilisé est l'erreur quadratique moyenne (MSE) de l'estimateur $\widehat{f}(x)$ de $f(x)$:

$$\begin{aligned} MSE &= E\left(\widehat{f}(x) - f(x)\right)^2 \\ &= E\left(\widehat{f}(x) - E(\widehat{f}(x)) + E(\widehat{f}(x)) - f(x)\right)^2 \\ &= \left(E(\widehat{f}(x)) - f(x)\right)^2 + E\left(\widehat{f}(x) - E(\widehat{f}(x))\right)^2 \\ &= \left[Biais\left(\widehat{f}(x)\right)\right]^2 + Var\left(\widehat{f}(x)\right) \end{aligned}$$

Ce qui se traduit donc par le compromis biais/variance de l'estimateur de densité $\widehat{f}$ en x. Ce critère n'évalue donc que la qualité d'estimation de $\widehat{f}$ au voisinage de w. Afin d'évaluer de façon plus globale la qualité de l'approximation sur toutes les valeurs x, nous définissons la notion d'erreur quadratique moyenne intégrée (MISE) :

$$\begin{aligned} MISE &= E\int\left(\widehat{f}(x) - f(x)\right)^2 \, dx \\ &= \int\left[Biais\left(\widehat{f}(x)\right)\right]^2 \, dx + \int Var\left(\widehat{f}(x)\right) \, dx \end{aligned}$$

Nous allons à présent passer en revue les différentes méthodes d'estimation non paramétriques en commençant par la plus simple : l'Histogramme de densité.

A.3.1 Histogramme de densité

Définir un histogramme de densité requiert deux paramètres : un point d'origine $t_0$et une longueur de classe h ($h > 0$).
Les classes sont définies pour tous les entiers $k = 0, \pm 1, \pm 2, ...$ comme suit :

$$B_k = [\ t_0 + k \times h\ ,\ t_0 + (k+1) \times h\ ]$$

l'histogramme de la densité f est alors défini par :

$$\begin{aligned} \widehat{f}(x) &= \frac{le\ nombre\ d'observations\ dans\ la\ classe\ qui\ contient\ x}{nh} \\ &= \frac{1}{nh}\sum_{i=1}^{n} \mathbf{1}_{B_k}(x_i) \quad pour\ x \in B_k \end{aligned} \tag{A.4}$$

L'histogramme de densité est une fonction étagée, et donc discontinue. L'estimateur $\widehat{f}$ dépend donc de deux paramètres : le point d'origine t_0 et la largeur de classe h. Ces deux paramètres peuvent avoir une influence importante sur l'histogramme (Il faut un exemple graphique).

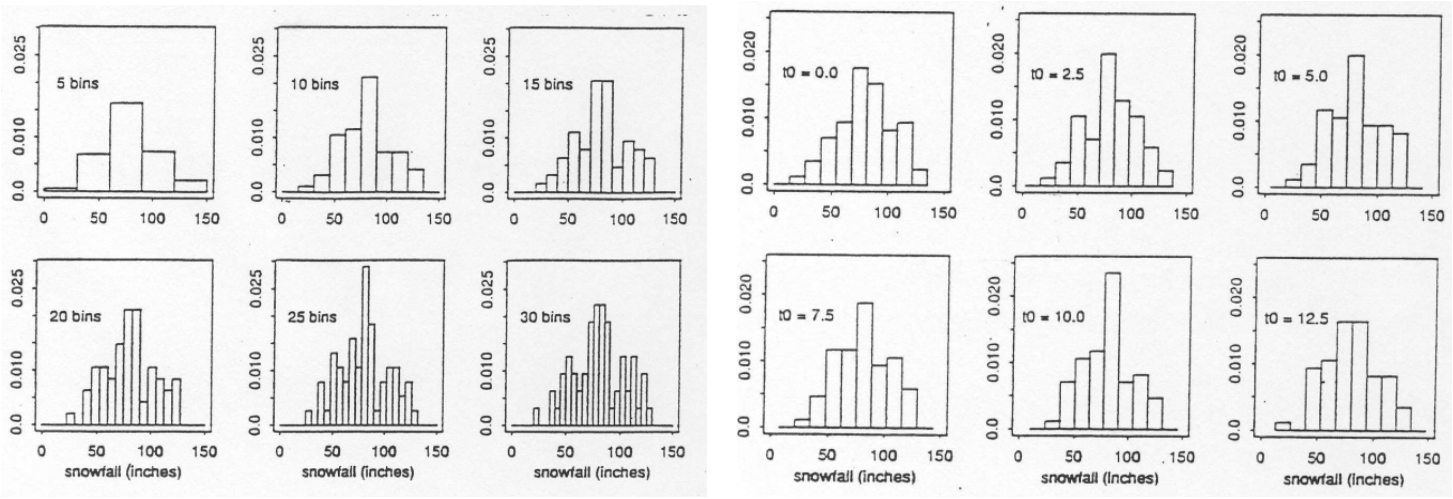

FIGURE A.1 – Des exemples de la sensibilité des histogrammes aux modifications des paramètres de fenêtres h avec point d'origine $t_0 = 0$ (gauche) et aux modifications du point d'origine t_0 avec $h = 13.5$ (droite). *Figure tirée de [Scott, 1992]*

En effet, pour tout $x \in B_k$, $\widehat{f}(x)$ est égal à $\frac{1}{h}$ fois la proportion de données appartenant à B_k. Le paramètre h apparaît alors comme un paramètre de lissage : en augmentant h on a tendance à *lisser* la courbe obtenue mais en le diminuant, la courbe devient plus irrégulière.

Du coté des performances statistiques, on montre que [Wasserman, 2006] :

$$\int \left[Biais\left(\widehat{f}(x)\right)\right]^2 \, dx = \frac{h^2}{12} \int \left(f'(x)\right)^2 \, dx + o(h^2) \tag{A.5}$$

$$\int Var\left(\widehat{f}(x)\right) \, dx = \frac{1}{nh} + o\left(\frac{1}{n}\right) \tag{A.6}$$

Soit en combinant les équations (A.5) et (A.6) :

$$MISE = \frac{h^2}{12} \int \left(f'(x)\right)^2 \, dx + \frac{1}{nh} + o(h^2) + o\left(\frac{1}{n}\right) \tag{A.7}$$

La première remarque est donc que l'histogramme de densité est consistant lorsque

$h \to 0$ mais $nh \to +\infty$ lorsque $n \to +\infty$. Nous devons donc réduire la largeur de classe à mesure que la taille des données augmente mais à une vitesse plus faible que $\frac{1}{n}$.

Pour avoir une idée de la vitesse de réduction optimale de la largeur de la fenêtre, il faut déterminer pour quelle valeur de h la dérivée de la formule (A.7) s'annule. Cela donne :

$$\frac{h_{opt}}{6}\int \left(f'(x)\right)^2 \, dx = \frac{1}{nh_{opt}^2} \implies h_{opt} = \left(\frac{6}{\int \left(f'(x)\right)^2 \, dx}\right)^{1/3} n^{-\frac{1}{3}} \tag{A.8}$$

Nous avons donc besoin d'une largeur de classe réduite si la densité change rapidement de variations et d'une largeur de fenêtre plutôt large dans le cas contraire.
Quoi qu'il en soit, h doit décroître à une vitesse $n^{-1/3}$.
En remplaçant dans l'équation A.7, cela donne une vitesse de convergence du $MISE$ de $n^{-2/3}$.

A.3.2 Estimateur à noyau d'une densité

L'estimateur à noyau a le double avantage de s'affranchir d'une part du choix d'un point d'origine t_0 et à remplacer l'estimateur discontinu qu'est l'histogramme par des fonctions plus régulières.

Commençons tout d'abord par présenter autrement l'estimateur histogramme de densité afin de dégager l'idée de noyau. La densité de probabilité f d'une variable aléatoire X étant la dérivée de sa fonction de répartition F, nous pouvons écrire pour tout x, point de continuité de f :

$$f(x) = lim_{h\to 0}\frac{\mathbb{P}\left(x-h \leq X \leq x+h\right)}{2h} = lim_{h\to 0}\frac{F(x+h)-F(x-h)}{2h} \tag{A.9}$$

Or en vertu du théorème de **Glivenko-Cantelli**, la quantité $\dfrac{F(x+h)-F(x-h)}{2h}$ peut être estimée, pour une taille d'échantillon n assez grande, par $\dfrac{\widehat{F}_n(x+h)-\widehat{F}_n(x-h)}{2h}$

Nous pouvons alors considérer l'estimateur :

$$\begin{aligned}\widehat{f}(x) &= \frac{1}{2nh}\sum_{i=1}^{n}\mathbb{1}_{[x-h\leq x_i\leq x+h]} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}_{[x-h\leq x_i\leq x+h]}}{2h} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-xi}{h}\right) \end{aligned} \tag{A.10}$$

avec K le noyau défini par : $K(u) = \frac{1}{2}\times\mathbb{1}_{|u|\leq 1}$.

Là encore, le paramètre h est un paramètre de lissage ou *bandwith* en anglais.

La figure A.2 illustre l'effet de différents choix du paramètre de lissage pour une estimation de densité à noyaux :

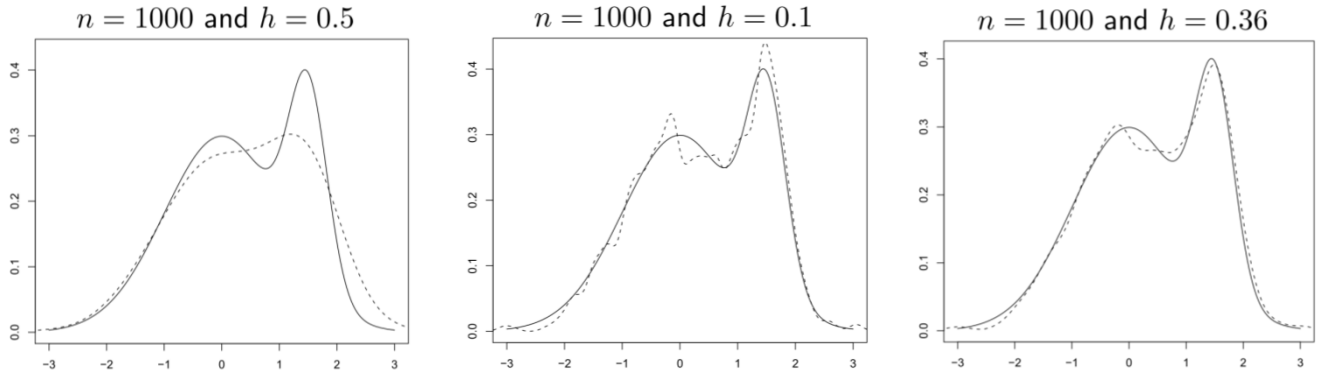

FIGURE A.2 – Exemple d'estimation à noyau (en pointillées) d'une densité de probabilité (en ligne solide) pour différentes valeurs du paramètre de lissage h.

Rien qu'avec cette nouvelle formulation, nous venons de nous affranchir de la première contrainte de l'histogramme de densité à savoir le choix d'un point origine t_0.
De plus, en optant pour un noyau K plus régulier, nous pouvons en plus obtenir un estimateur de densité continue et dérivable voire infiniment dérivable !

Dans la suite, nous ne considérerons que des noyaux symétriques, positifs, et possédant un moment d'ordre 2. Nous supposerons de plus que leurs intégrales sur $\mathbb{R}$ valent 1.

Intéressons-nous à présent aux propriétés de convergence de l'estimateur à noyau. On montre que [Wasserman, 2006] :

$$Biais\left(\widehat{f}(x)\right) = \frac{h^2\sigma_K^2 f''(x)}{2} + o\left(h^2\right) \tag{A.11}$$

$$Var\left(\widehat{f}(x)\right) = \frac{f(x)}{nh}\int K^2(u)\ du + O\left(\frac{1}{n}\right) \tag{A.12}$$

Avec $\sigma_K^2 = \int u^2 K(u)\ du$.

Les remarques sur les tendances de h et n sont donc les mêmes que pour les histogrammes de densités : à mesure que la taille de l'échantillon n croît, le paramètre de lissage h tend vers zéro mais à une vitesse plus lente que $1/n$ de façon à ce que $nh \to +\infty$.

En combinant les équations (A.11) et (A.12) après intégration on obtient :

$$MISE = \frac{h^4\sigma_K^4}{4}\int \left(f''(x)\right)^2\ dx + \frac{\int K^2(u)\ du}{nh} + o(h^2) + O\left(\frac{1}{n}\right) \tag{A.13}$$

Pour avoir une idée de la vitesse de réduction optimale du paramètre de lissage h, il faut déterminer pour quelle valeur de h la dérivée de la formule (A.13) s'annule. Cela donne :

$$h_{opt}^3\sigma_K^4\int \left(f''(x)\right)^2\ dx = \frac{\int K^2(u)\ du}{nh_{opt}^2} \implies h_{opt} = \left(\frac{\int K^2(u)\ du}{\sigma_K^4\int \left(f''(x)\right)^2 dx\ dx}\right)^{1/5} n^{-1/5} \tag{A.14}$$

Cette fois, h doit décroître à une vitesse $n^{-1/5}$ et en remplaçant dans l'équation A.13, cela donne une vitesse de convergence du $MISE$ de $n^{-4/5}$.

Cette vitesse de convergence est donc meilleure que celles des densités d'histogramme bien que pénalisée par la courbure de la densité de probabilité inconnue f via le terme $\int \left(f''(x)\right)^2\ dx$.

Choix optimal du noyau et du paramètre de lissage

Choix du noyau

Si nous insérons l'expression du paramètre optimal de lissage donné dans l'équation (A.14) dans l'expression du MISE de l'équation (A.13) on trouve :

$$MISE\left(\widehat{f}\right) = \frac{5}{4}\left(\sigma_K^4 R^4(K)\right)^{1/5}\left(R\left(f''\right)\right)^{1/5} n^{-4/5} \tag{A.15}$$

Avec $R(g) = \int g^2(u)\, du$.

Il s'ensuit que la dépendance de $MISE\left(\widehat{f}\right)$ en fonction du noyau s'exprime à travers le terme $\left(\sigma_K^4 R^4(K)\right)^{1/5}$. La minimisation sous contrainte par rapport à K de cette quantité mène à la solution [Epanechnikov, 1969] :

$$K_{opt}(x) = \frac{3}{4}\left(1 - x^2\right)\mathbb{1}_{|x| \leq 1} \tag{A.16}$$

L'effet sur le $MISE\left(\widehat{f}\right)$ de l'utilisation d'un autre noyau K en lieu et place de celui d'Epanechnikov est souvent évalué par l'efficacité du noyau K par rapport à K_{opt} :

$$\text{eff(K)} = \left(\frac{C(K_{opt})}{C(K)}\right)^{5/4} \tag{A.17}$$

Interprétation : c'est *le rapport de taille d'échantillons* pour K_{opt} nécessaire pour atteindre le même $AMISE$ qu'avec le noyau K.

Kernel K	eff(K)
Epanechnikov $\frac{3}{4}(1-u^2)I(\lvert u\rvert \leq 1)$	1.000
Quartic (biweight) $\frac{15}{16}(1-u^2)^2 I(\lvert u\rvert \leq 1)$	0.994
Triweight $\frac{35}{32}(1-u^2)^3 I(\lvert u\rvert \leq 1)$	0.987
Gaussian $\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$	0.951
Uniform $\frac{1}{2}I(\lvert u\rvert \leq 1)$	0.930
Triangle $(1-\lvert u\rvert)I(\lvert u\rvert \leq 1)$	0.986

TABLE A.1 – Tableau des efficacités relatives de plusieurs noyaux

Choix du paramètre de lissage

Comme nous venons de le voir, les formules théoriques des paramètres de lissage optimaux des histogrammes ou des estimateurs à noyaux dans les équations (A.8) et (A.14) respectivement sont importantes dans le sens où elles fournissent une vitesse de décroissance du paramètre de lissage en fonction de la taille des données. Cependant, elles ont l'inconvénient de dépendre explicitement de la densité de probabilité inconnue f ce qui rend leur utilisation directe impossible en pratique.
C'est pourquoi, de nombreuses méthodes ont été proposées pour déterminer le paramètre de lissage optimal.

La règle de référence à une distribution gaussienne (*Rule-of-thumb* bandwidth selector)

Une solution simple pour une première approche consiste à déterminer h_{opt} sous l'hypothèse que les données sont générées par une densité de probabilité f gaussienne $\mathcal{N}(m, \sigma^2)$. Sous cette hypothèse, $R(f'') = \dfrac{3}{8\sqrt{\pi}\sigma^5}$ et le bandwidth du type *gaussien de référence* est défini par :

$$\widehat{h_{GR}} = \left(\frac{8\sqrt{\pi}R(K)}{3\sigma_K^4}\right)^{1/5} \widehat{\sigma} n^{-1/5}$$

où $\widehat{\sigma}$ est un estimateur de l'écart-type σ qui peut être par exemple égal à :

$$\widehat{\sigma} = min\left\{\widehat{s}_n, \frac{Q_3 - Q_1}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)}\right\} \tag{A.18}$$

avec $\widehat{s}_n$, Q_1, Q_3 respectivement l'écart type, le premier et le troisième quartile de l'échantillon des valeurs observées et Φ^{-1} l'inverse de la fonction de répartition d'une gaussienne $\mathcal{N}(m, \sigma^2)$.
Signalons enfin que bien qu'optimales pour les distributions gaussiennes, il est en général admis que les valeurs des paramètres de lissage issues de cette méthode ont tendance à fournir des estimateurs de densités trop lisses [Jones et al., 1996].

La méthode de la validation croisée

L'idée de cette méthode est d'estimer d'abord le $MISE\left(\widehat{f}\right)$ et de minimiser cet estimateur en h.

$$\begin{aligned} MISE\left(\widehat{f}\right) &= E\int\left(\widehat{f}(x)-f(x)\right)^2\,dx \\ &= E\int\widehat{f}^2(x)\,dx - 2E\int\widehat{f}(x)f(x)\,dx + \int f^2(x)\,dx \end{aligned} \tag{A.19}$$

La quantité $\int f^2(x)\,dx$ ne dépendant pas de h, l'optimisation ne portera donc que sur les deux premiers termes de l'équation (A.19).

On montre par la méthode des moments que le terme $E\int\widehat{f}(x)f(x)\,dx + \int f^2(x)\,dx$ peut être estimé sans biais par l'estimateur *leave-one-out* défini par :

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{-i}\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\frac{1}{h}K\left(\frac{X_i-X_j}{h}\right) \tag{A.20}$$

Ainsi le paramètre de lissage optimal au sens de la méthode *validation croisée moindre carrés* est donné par :

$$\widehat{h}_{CV} = \underset{h}{\operatorname{argmin}}\int\widehat{f}^2(x)\,dx - 2\frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{-i}\left(X_i\right) \tag{A.21}$$

Les critiques formulées à l'égard du paramètre de lissage sont sa forte variabilité et le fait qu'il sous-estime souvent le paramètre de lissage optimal et donc susceptible de produire de fausses modalités [Jones et al., 1996].

La méthode de substitution *Plug-in*

Rappelons l'expression du paramètre de lissage optimal de la méthode d'estimation par noyaux d'une densité de probabilité :

$$h_{opt} = \left(\frac{\int K^2(u)\ du}{\sigma_K^4 \int \left(f''(x)\right)^2\ dx} \right)^{1/5} n^{-1/5}$$

Cette expression n'étant pas exploitable directement en raison de la dépendance en f (la densité de probabilité inconnue), la Méthode Plug-in [Sheather and Jones, 1991] propose de substituer (d'où le nom de la méthode) le terme $\int \left(f''(x)\right)^2\ dx$ par un autre estimateur, appelé estimateur *pilote*, et qui se base sur un choix de paramètre de lissage pilote. Ce paramètre de lissage pilote h_{Pilote} est différent du paramètre de lissage recherché h_{opt} car il est avant tout destiné à estimer l'intégrale du carré de la dérivée seconde et non de la fonction de densité elle-même.

La méthode Plug-in directe consiste alors à choisir h_{Pilote} en assumant une forme paramétrique de f par exemple gaussienne. Ceci est motivé par le fait que, plus généralement, si l'on désigne par $f^{(r)}$ la $r^{ème}$ dérivée de f, alors le biais du paramètre de lissage servant à estimer $\int \left(f^{(r)}\right)^2$ diminue s'affranchissant ainsi progressivement de la dépendance de l'hypothèse de normalité. La contre-partie étant, comme souvent, l'augmentation de la variance. La dérivée seconde est souvent un bon compromis rendant cette méthode particulièrement indiquée pour l'estimation de h_{opt}.

La méthode *Plug-in* la plus efficace est la méthode *Plug-in par résolution d'équation.*

L'idée est d'exprimer le paramètre h_{Pilote} comme fonction g du paramètre de lissage recherché, le problème se formulant alors en terme d'équation de type *point fixe* :

$$h = \left(\frac{\int K^2(u)\ du}{\sigma_K^4 \int \left(\widehat{f}''_{g(h)}(x)\right)^2\ dx} \right)^{1/5} n^{-1/5} \tag{A.22}$$

Si on remarque que le terme $R\left(f^r\right) = \int \left(f^{(r)}\right)^2$ peut s'écrire $\int f^{(r)} \times f^{(r)}$ alors une succession d'intégrales par parties (prendre les primitives successives du membre de droite du produit et les dérivées successives de celui de gauche) donne :

$$\begin{aligned} R\left(f^{(r)}\right) &= (-1)^r \int f^{(2r)(x)} \times f(x)\, dx \\ &= (-1)^E \left(f^{(2r)}\left(X\right)\right) \\ &= lim_{n\to+\infty} \widehat{\Phi}_{2r,h_{Pilote,r}} \end{aligned}$$

avec $\widehat{\Phi}_{2r,h_{Pilote,r}} = \frac{1}{N}\sum_{i=1}^{N} f^{(2r)}\left(x_i\right) = \frac{1}{Nh^{r+1}}\sum_{i=1}^{N}\sum_{j=1}^{N} K^{(2r)}\left(\frac{x_i - x_j}{h_{Pilote,r}}\right)$ un estimateur à noyau de l'intégrale du carré de la $r^{ième}$ dérivée de la densité f. On peut montrer [Wand and Jones, 1995] que le paramètre de lissage optimal $h_{Pilote,r}$ de $\widehat{\Phi_{2r}}$ est donné par :

$$h_{Pilote,r} = \left(\frac{-2K^{(2r)}(0)}{N\sigma_K^2 \widehat{\Phi}_{2r+2}}\right)^{\frac{1}{3+2r}} \tag{A.23}$$

Ce processus d'optimisation peut ainsi être poursuivi avec l'estimation du terme $\widehat{\Phi}_{2r+2}$ nécessitant le paramètre de lissage $h_{Pilote,r+1}$ qui à son tour dépendra de $\widehat{\Phi}_{2r+4}$ ..etc
En pratique, on s'arrête aux estimations $\widehat{\Phi}_6$ et $\widehat{\Phi}_8$ réalisant un compromis biais variance satisfaisant que l'on peut calculer explicitement sous l'hypothèse d'une densité de loi normale d'écart-type $\widehat{\sigma}$ (estimateur empirique) :

$$\widehat{\Phi}_6 = -\frac{15}{16\sqrt{\pi}}\widehat{\sigma}^{-7} \text{ et } \widehat{\Phi}_8 = -\frac{105}{32\sqrt{\pi}}\widehat{\sigma}^{-9} \tag{A.24}$$

On en déduit les paramètres de lissages pilotes suivants :

$$h_{Pilote,2} = \left(\frac{-6}{\sqrt{2\pi}\widehat{\Phi}_6 N}\right)^{1/7} \text{ et } h_{Pilote,3} = \left(\frac{30}{\sqrt{2\pi}\widehat{\Phi}_8 N}\right)^{1/9} \tag{A.25}$$

La fonction g dans l'équation (A.22) pouvant alors s'écrire :

$$g(h) = \left(\frac{-6\sqrt{2}\widehat{\Phi}_{4,h_{Pilote,2}}}{\widehat{\Phi}_{6,h_{Pilote,3}}}\right)^{1/7} h^{1/7} \tag{A.26}$$

Comparaison des vitesses de convergence

Les résultats théoriques de convergence des estimateurs du paramètre de lissage unidimensionnel montrent que la méthode du *Plug-in* posssède de meilleures propriétés asymptotiques :

$$\frac{\widehat{h}_{CV}}{h_{MISE}} - 1 = O_p\left(n^{-1/10}\right)$$
$$\frac{\widehat{h}_{PI}}{h_{MISE}} - 1 = O_p\left(n^{-5/14}\right)$$

La notation O_p signifiant bornée "en probabilité".

Cependant, les méthodes sophistiquées n'ont pas pour autant clos le débat comme on peut le constater par exemple dans les travaux de Loader [Loader, 1999] qui relativise la supériorité des méthodes *Plug-in* sur les méthodes dites plus classiques. En réalité, aucune de ces méthodes n'a donné pleine satisfaction pour tout type de distributions paramétriques, particulièrement pour les densités à convergence lentes telles que les lois log-normales ou exponentielles.

Annexe B

Régression Spline

Sommaire

B.1 Pourquoi les splines ? 172

B.2 Splines : Définition et premières propriétés 174

B.2.1 La base de puissances tronquée 174

B.2.2 Base B-splines 175

B.3 Splines des moindres carrés 178

B.4 Splines de lissage 181

B.4.1 Définition 181

B.4.2 Une variante : les splines pénalisées 182

B.1 Pourquoi les splines ?

Notons $x = (x_i)_{1 \leq i \leq n}$ et $y = (y_i)_{1 \leq i \leq n}$ issus du modèle :

$$E(Y|X) = f(X) \tag{B.1}$$

Le traditionnel modèle de régression linéaire est défini par :

$$\widehat{f}(x) = \beta_0 + \beta_1 x \tag{B.2}$$

avec les paramètres β_0 et β_1 estimés par la méthode des moindres carrés.
Ce type de modèle n'est approprié que dans le cas où la fonction f dans l'équation (B.1) est (approximation) linéaire.
Une extension naturelle de l'estimateur (B.2) est de considérer la classe des estimateurs polynomiaux :

$$\widehat{f}(x) = \sum_{k=0}^{p} \beta_k x^k \tag{B.3}$$

La Figure B.1 illustre un exemple de régression polynomiale.

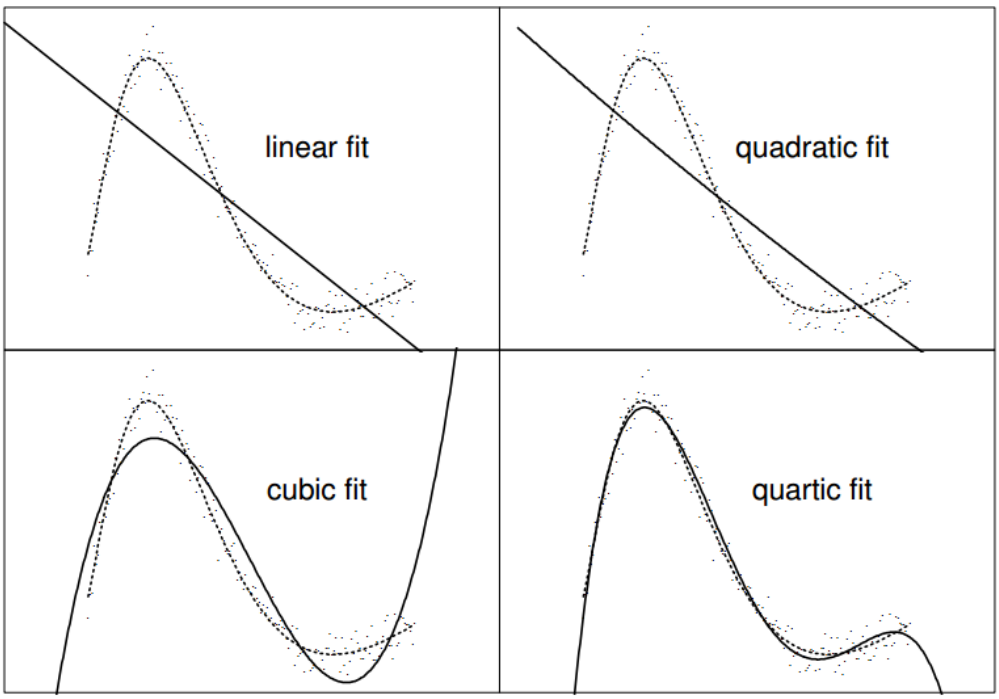

FIGURE B.1 – Exemple de régression polynomiale à plusieurs degrés. Les pointillés représentent le nuage de points généré à partir de la fonction f en tirets.

Nous remarquons un nombre de problèmes importants dans la régression polynomiale particulièrement celle à haut degré en x. Ceci est essentiellement dû à deux raisons.
La première, c'est que la matrice de régression de ce type de modèle se heurte au problème d'inversion de matrices "pleines", opération qui peut se révéler très délicate numériquement. En effet, l'estimateur des moindres carrés des coefficients de la régression (B.3) est défini par :

$$\widehat{\beta} = \left(\widehat{\beta}_0\ \widehat{\beta}_1 ...\ \widehat{\beta}_p\right)^T = \left(B^T B\right)^{-1} B^T y \tag{B.4}$$

avec $B = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{pmatrix}$ et B^T la transposée de la matrice B.
Cet estimateur nécessite donc l'inversion de la matrice "pleine" $B^T B$.
Ce problème peut être résolu à l'aide des polynômes orthogonaux qui transforment la matrice d'intérêt en matrice diagonale dont l'inversion est immédiate.

La seconde raison, peut-être la plus importante, est que l'effet de la régression polynomiale est global et non local. Ceci est visible dans les deux figures du bas de la Figure B.1 où les courbes ont de forts effets de bord aux extrémités des données (en d'autres termes, les estimateurs polynomiaux possèdent une large variance en ces points).

Une façon d'obtenir un estimateur aux propriétés plus locales tout en évitant les problèmes des polynômes de degrés élevés est de recourir à des estimateurs polynomiaux de **bas degré** (au maximum de degré trois) **par morceaux**.

B.2 Splines : Définition et premières propriétés

Les splines peuvent essentiellement être définies comme des polynômes par morceaux connectés entre eux en des points nommés *nœuds*.
Dans cette section, nous allons d'abord comprendre pourquoi de telles fonctions peuvent être vues comme des combinaisons linéaires de fonctions puissances tronquées et nous introduirons par la suite les fonctions de base B-splines plus pratiques numériquement pour implémenter les splines.

B.2.1 La base de puissances tronquée

Notons P un polynôme de degré p.

$$P(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_p x^p \tag{B.5}$$

Alors désignons par S la fonction :

$$S(x) = P(x) + \beta_{p+1} (x-t)_+^p = \begin{cases} P(x) & \text{si} \quad x \leq t \\ P(x) + \beta_{p+} (x-t)^p & \text{si} \quad x > t \end{cases} \tag{B.6}$$

Avec la notation $(a)_+ = a$ si $a > 0$ et 0 sinon.

La fonction S est donc un polynôme de degré p dans chaque région $x \leq t$ et $x > t$ et donc, un polynôme par morceaux de degré p.
Remarquons que ce polynôme par morceaux est correctement défini par $p+2$ coefficients contre $p+1$ pour un polynôme de degré p classique. Le nombre de coefficients nécessaires pour entièrement définir la fonction S est alors appelé *degrés de liberté*. Plus généralement, le polynôme par morceaux $S(x) = P(x) + \beta_{p+1} (x-t_1)_+^p + ... + \beta_{p+m+1} (x-t_m)_+^p$ est un polynôme par morceaux de degré p avec $p+m+1$ degrés de libertés.

Une importante propriété des splines héritée des polynômes est sa régularité en terme de continuité et dérivabilité : un polynôme étant infiniment dérivable en tout réel, il en est de même pour les splines en tout réel excepté les nœuds où elle n'est que p-1 continument dérivable. Par exemple, une fonction spline cubique est deux continument dérivable ce qui est souvent suffisant pour obtenir des estimations lisses satisfaisantes dans la plupart des situations.
En résumé, nous pouvons écrire n'importe quel polynôme par morceaux de degré p sous

la forme [de Boor, 1978] :

$$S(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_p x^p + \beta_{p+1} (x - t_1)^p + ... + \beta_{p+m} (x - t_m)^p \qquad (B.7)$$

En d'autres termes, $\{1, x, x^2, ..., x^p, (x - t_1)^p, ..., (x - t_m)^p\}$ est une base de l'espace vectoriel des splines de degré p définies à partir des nœuds $\{t_1, ..., t_m\}$.

En ajoutant un terme de d'erreur ϵ à l'équation (B.7), nous obtenons un modèle de régression spline $y = S(x) + \epsilon$ en fonction du predicteur x.
L'avantage d'un tel modèle est le fait que la régression s'effectue dans une classe de fonctions linéaires par rapport aux paramètres à estimer ce qui est particulièrement intéressant numériquement car celà repose sur des méthodes bien connues telles que l'estimation des moindres carrés.

Même si nous avons réussi à apporter une réponse en terme de propriété locale, l'estimateur des moindres carrés construit à partir des puissances tronquées souffre malheureusement des mêmes lacunes que celles des estimateurs des moindres carrés polynomiaux car les équations normales qui en résultent constituent souvent un problème mal-posé.
Comme évoqué plus haut, ce problème peut être résolu dans le cas polynomiale en faisant appel aux polynômes orthogonaux d'où l'idée de construire une nouvelle base splines *quasi-orthogonale* dans un sens que nous préciserons dans la prochaine section.

B.2.2 Base B-splines

Fixons une séquence de nœuds $\{t_1, ..., t_m\}$ dans l'intervalle $[a, b]$.
Une premier exemple très simple est celui des polynômes constants par morceaux de nœuds $\{t_1, ..., t_m\}$ et engendré par la base $\{1, \mathbb{1}_{x>t_1}, ..., \mathbb{1}_{x>t_m}\}$.

Remarquons que les supports des fonctions $\mathbb{1}_{x>t_1}, ..., \mathbb{1}_{x>t_m}$ se chevauchent énormément avec comme conséquence un estimateur des moindres carrés contenant une matrice difficile à inverser comme expliqué plus haut.
Le cas des fonctions constantes par morceaux de nœuds $\{t_1, ..., t_m\}$ offre une alternative naturelle : elles peuvent s'écrire comme une combinaison linéaire des fonctions indicatrices $\{\mathbb{1}_{a \leq x < t_1}, \mathbb{1}_{t_1 \leq x < t_2}, ..., \mathbb{1}_{t_m \leq t < b}\}$.
Ainsi, sous la condition d'avoir au moins un x_i dans chaque intervalle de la partition $[a, t_1[\cup [t_1, t_2[... \cup [t_m, b[$ de $[a, b]$, Le problème de régression des moindres carrés est alors un problème bien-posé puisque les supports des fonctions $\{\mathbb{1}_{a \leq x < t_1}, \mathbb{1}_{t_1 \leq x < t_2}, ..., \mathbb{1}_{t_m \leq t < b}\}$

sont disjoints les rendant ainsi orthogonales : chacune de ces fonctions est non nulle uniquement dans l'intervalle $[t_i, t_{i+1}[$. Remarquons enfin que la regression dans cette base requière exactement le même degré de liberté que la base puissance tronquée.
Notons à présent :

$$B_{i,1}(x) = \mathbb{1}_{t_i \leq x < t_{i+1}} \tag{B.8}$$

que l'on appellera **les B-splines d'ordre 1** associées aux nœuds $\{t_1, ..., t_m\}$.
La seule contrainte sur ces fonctions est qu'elles forment une **partition de l'unité** i.e $\sum_i B_{i,1}(x) = 1$ pour tout x.

On construit les B-splines d'ordres supérieurs par récurrence selon un procédé définissant chaque B-spline d'ordre supérieur comme un élément de l'enveloppe convexe de deux B-splines d'ordres inférieurs.

Les B-splines d'ordre 2, qui sont des des fonctions linéaires par morceaux, sont obtenus de la façon suivante :

$$B_{i,2} = \omega_{i,2} B_{i,1} + (1 - \omega_{i+1,2})\, B_{i+1,1} \tag{B.9}$$

avec pour $w_{ik} = \dfrac{x - t_i}{t_{i+k-1} - t_i}$ les coefficients de la combinaison linéaire convexe.
La fonction $B_{i,2}$ est donc nulle en dehors de l'intervalle $[t_i, t_{i+2}[$ en raison de la présence dans son expression des fonctions indicatrices $\mathbb{1}_{t_i \leq x < t_{i+1}}$ et $\mathbb{1}_{t_{i+1} \leq x < t_{i+2}}$. Elle est formée de deux fonctions linéaires en x, raison pour laquelle les fonctions $(B_{i,2})$ sont appelées **les B-splines linéaires**.

Pour obtenir les B-splines d'ordre 3, on applique le même procédé que précédemment :

$$\begin{aligned} B_{i,3} &= \omega_{i,3} B_{i,2} + (1 - \omega_{i+1,3})\, B_{i+1,2} \\ &= \omega_{i,3}\omega_{i,2} B_{i,1} + \left[\omega_{i,3}\,(1 - \omega_{i+1,2}) + \omega_{i+1,2}\,(1 - \omega_{i+1,3})\right] B_{i+1,1} \\ &\quad + (1 - \omega_{i+1,3})\,(1 - \omega_{i+2,2})\, B_{i+2,1} \end{aligned} \tag{B.10}$$

Cette fois, la fonction $B_{i,3}$ est la jonction de trois polynômes du second degré formant une fonction continument dérivable et nulle en dehors de l'intervalle $[t_i, t_{i+3}[$ en raison de la présence des fonctions $B_{i,1}, B_{i+1,1}$ et $B_{i+2,1}$.

Après $k-1$ étapes de récurrence, l'expression de $B_{i,k}$ pour tout k s'écrit sous la forme :

$$B_{i,k} = \sum_{j=i}^{i+k-1} b_{jk} \times B_{j,1} \tag{B.11}$$

avec $b_{j,k}$ un polynôme en x de degré $k-1$ en tant que produit de $k-1$ polynômes linéaires en x.
Ainsi, la fonction B-spline $B_{i,k}$ est un polynôme par morceaux de degré $k-1$ s'annulant en dehors de l'intervalle $[t_i, t_{i+k}[$.

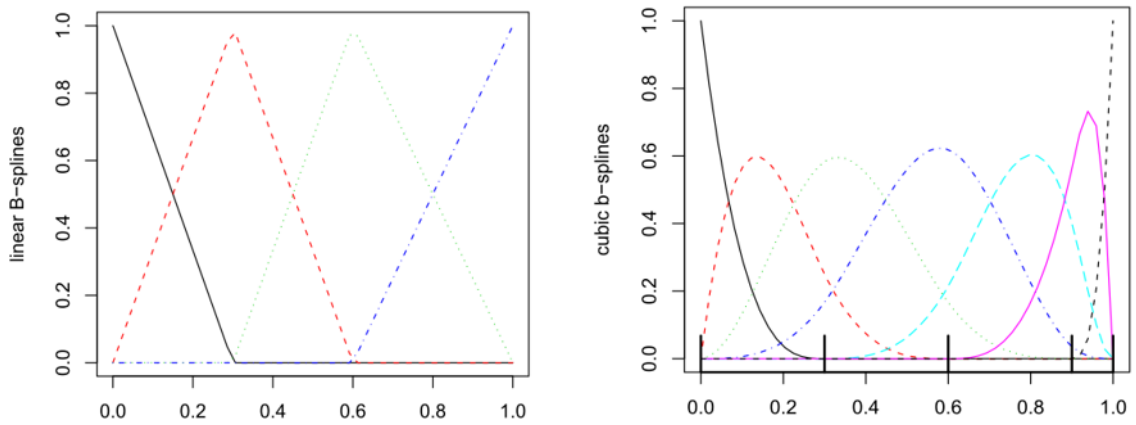

FIGURE B.2 – Exemple de quatre B-splines linéaires (à gauche) de nœuds $\{0.3, 0.6\}$ et sept B-splines cubiques (à droite) de nœuds $\{0.3, 0.6, 0.9\}$

La Figure B.2 montre par exemple que la $i^{ème}$ B-spline cubique est non nulle uniquement dans l'intervalle $[t_i, t_{i+4}[$. En général, cette propriété de **support minimal** assure que la $i^{ème}$ et la $i+j+1^{ème}$ B-splines sont orthogonales pour $j \geq p$, p étant l'ordre de la B-splines. Celles dont les supports se chevauchent restent néanmoins linéairement indépendantes.

Cette propriété de **support minimal** aura un rôle centrale dans cette thèse quand il sera question de modélisation d'une relation dose-effet dans un modèle de Cox.

B.3 Splines des moindres carrés

Revenons à présent à notre problème d'estimation des splines de moindres carrés. Un estimateur spline peut donc avoir deux représentations : la représentation en combinaison linéaire de puissances tronquées ou celle en B-splines.

Puissances tronquées

Le modèle de régression s'écrit pour tout $1 \leq j \leq n$:

$$y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + ... + \beta_p x_j^p + \beta_{p+1} (x_j - t_1)^p + ... + \beta_{p+m} (x_j - t_m)^p + \epsilon_j \quad \text{(B.12)}$$

En écriture matricielle, le modèle devient :

$$y = T\beta + \epsilon \quad \text{(B.13)}$$

où T est une matrice $n \times (p + m + 1)$ avec les p première colonnes correspondent à la matrice du modèle de régression polynomiale de degré p et les éléments d'indices $(j, p + i + 1)$ égaux à $(x_j - t_i)_+^p$.
L'estimateur des moindres carrés vecteur de coefficients β du modèle (B.13) est alors égal à :

$$\widehat{\beta} = \left(T^T T\right)^{-1} T^T y \quad \text{(B.14)}$$

Ainsi, avec la méthodologie des moindres carrés à notre disposition, il est également possible d'en déduire les écart-type, p-value et autres intervalles de confiance autours des estimations de $\widehat{\beta}$.
La seule difficulté rappelons concerne les propriétés numériques médiocres de la base puissance tronquée menant ainsi à des imprécisions dans l'estimation de $\widehat{\beta}$. C'est pour cette raison que les bases B-splines ont été introduites.

B-splines

Le modèle de régression s'écrit pour tout $1 \leq j \leq n$:

$$y_j = \beta_0 + \sum_{i=0}^{p+m} \beta_i B_{i,p} (x_i) + \epsilon_j \quad \text{(B.15)}$$

En écriture matricielle, le modèle devient :

$$y = B\beta + \epsilon \tag{B.16}$$

où B est une matrice $n \times (p + m + 1)$ des éléments d'indices (j, i) égaux à $B_{i,p}(x_j)$. L'estimateur des moindres carrés du vecteur de coefficients β de (B.16) est alors égal à :

$$\widehat{\beta} = \left(B^T B\right)^{-1} B^T y \tag{B.17}$$

La *quasi-orthogonalité* des B-splines discutée dans la section précédente assure une forme de matrice à bandes pour la matrice $B^T B$ avec des propriétés numériques bien meilleures que la matrice $T^T T$ et des estimations bien plus précises pour $\widehat{\beta}$. En revanche, l'interprétation intuitive des coefficients de régression des puissances tronquées est perdue puisque la construction de chaque fonction B-spline fait intervenir non pas un mais plusieurs nœuds.

Emplacement des nœuds

Finissons cette section en analysant, via la Figure B.3, l'effet du nombre et de l'emplacements des nœuds sur la qualité de la régression spline des moindres carrés.

En haut à gauche se trouve l'estimateur spline optimale de la moyenne conditionnelle du modèle selon un procédé de sélection de nœuds par élimination. il réalise le bon compromis en terme de nombre et d'emplacements des nœuds, les deux étant cruciaux pour aboutir à une estimation de bonne qualité comme l'illustrent les trois autres représentation graphiques de la figure : en haut à droite, l'estimateur construit à partir de la séquence de nœuds optimaux mais en nombre trop réduit ce qui omet des détails importants de la structure de la vraie moyenne. En bas à gauche, l'estimation obtenue à partir de la sequence optimale mais en utilisant un trop grand nombre de nœuds ce qui produit des oscillations sans rapport avec la vraie moyenne. En bas à droite, l'estimateur utilise le même nombre de nœuds que le résultat optimal mais pour des emplacements équidistant qui donne des résultats très éloignés de la vérité.

La méthode usuelle pour déterminer l'emplacement et le nombre de nœuds dans une régression spline exploite le cadre des moindres carrés en via des techniques de sélection de variables de type élimination descendante ou *backward*. Ces méthodes commencent par un ensemble de nœuds intérieurs initiaux basé comme les quantiles et s'appuient sur la représentation en puissance tronquée permettant de travailler sur des modèles emboités

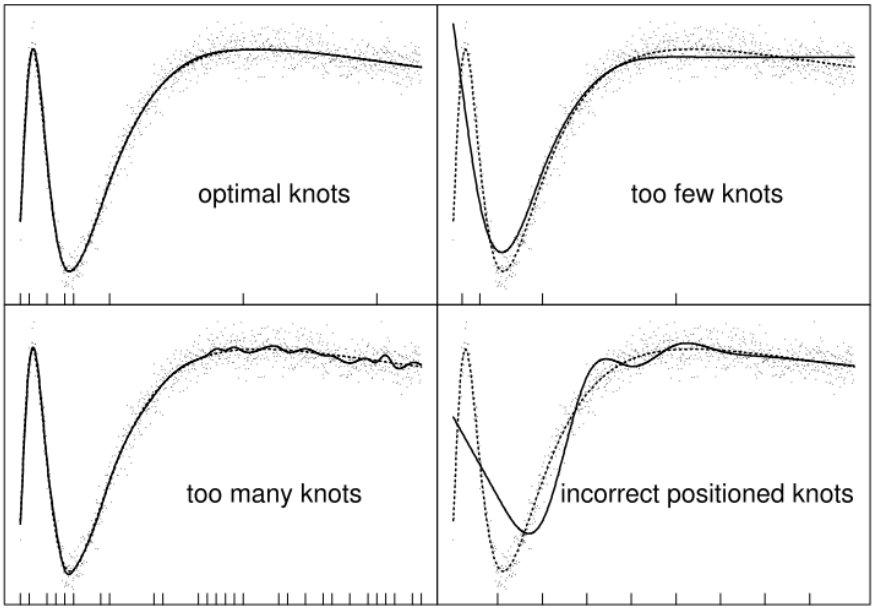

FIGURE B.3 – Effet du nombre et de l'emplacement des nœuds sur la qualité d'une regression spline. Le nuage de points représente les données simulées à partir de la vraie moyenne (en tirets). la courbe en ligne continue est l'estimateur spline de la moyenne

.

où à chaque étape, le terme puissance tronquée le moins corrélé avec la variable réponse est éliminé, éliminant au même temps le nœud intérieur associé. Le modèle réduit est alors comparé avec le précédent via un critère de type AIC ou BIC afin d'obtenir le modèle le plus parcimonieux.

B.4 Splines de lissage

B.4.1 Définition

Nous venons de le voir, le problème du choix des nœuds intérieurs est délicat en régression spline.
Des résultats analogues au théorème d'approximation de Weierstrass pour les polynômes existent pour les splines : ils établissent que toute fonction lisse peut être assez bien approximée par une fonction spline construite à partir de suffisamment de nœuds.

L'approche consistant à utiliser un nombre important de nœuds ne suffit pas à mener à bien la régression car plus il ya de nœuds et plus le risque *overfitting* augmente une fois que $p+m+1 > n$. Pour garder une modèle suffisamment stable, nous appliquons un terme de pénalisation au critère des moindres carrés classique. Le terme de pénalisation permet de garder un contrôle sur la courbature de l'estimateur splines S obtenu en minimisant sur toutes les fonctions continument dérivables deux fois la quantité :

$$\sum_{j=1}^{n} \left(y_j - S\left(x_j\right)\right)^2 + \lambda \int_a^b \left(S''(x)\right)^2 \, dx \tag{B.18}$$

La solution de ce problème de minimisation est étonnamment simple puisqu'il s'agit d'une fonction spline cubique avec pour nœuds toutes les observations distinctes.

Plus précisément, les coefficients dans la base B-spline de cette fonction peut être obtenus via la relation :

$$\widehat{\beta} = \left(B^T B + \lambda D^T D\right)^{-1} B^T y \tag{B.19}$$

Avec D une matrice liée à la dérivée seconde de la fonction [de Boor, 1978].

Ainsi, le problème du choix des nœuds est remplacé par celui du choix du paramètre de lissage λ les deux cas extrêmes étant $\lambda \to 0$ synonyme d'interpolation de tous les points observés et $\lambda \to +\infty$ avec comme résultat la spline réduite à une droite.En d'autres termes, plus la valeur du paramètre de lissage augmente et plus la courbe devient lisse.

B.4.2 Une variante : les splines pénalisées

Eilers et Marx [Eilers and Marx, 1996] ainsi que Ruppert, Wand et Carroll [Ruppert et al., 2003] ont proposé d'autres formes de pénalisation à appliquer au critère des moindres carrés.
Les splines pénalisées de ces auteurs possèdent deux modification importantes :

- Les nœuds ne sont plus forcément situés aux points d'observation mais sont en nombre plus réduit soit équi-distribués soit situés aux quantiles des points observés.

- Le terme de pénalité basé sur la dérivée seconde est remplacé par un autre basé soit sur des différences finies i.e $\left(\Delta^2\beta\right)_j = \beta_j - 2\beta_{j-1} + \beta_{j-2}$ avec , soit sur des normes de type *Ridge* i.e norme partielle du vecteur coefficient β.

Dans les deux cas, on se retrouve avec une expression de l'estimateur voisine de celle de l'équation (B.19) à ceci près que la matrice D dépend du type de pénalisation envisagée.

Les travaux de Ruppert [Ruppert, 2002] ont montré que, au-delà d'un nombre minimum de nœuds, le problème essentiel des splines pénalisées réside dans le choix du paramètre de lissage λ. Pour reprendre ses termes : *It has generally been believed that a P-spline can have too few knots but not too many knots.*

La critères classiques à minimiser pour choisir le paramètre de lissage optimal λ sont la validation croisée ou la validation croisée généralisée qui donnent souvent des résultats voisins en pratique [Eilers and Marx, 1996] :

$$CV\left(\lambda\right) = \sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_i}{1 - H_{ii}}\right)^2 \tag{B.20}$$

$$GCV\left(\lambda\right) = \sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_i}{1 - tr\left(H\right)}\right)^2 \tag{B.21}$$

avec $H = B\left(B^TB + \lambda D^TD\right)^{-1}B^T$ le *Lisseur* ou matrice *"hat" (chapeau)* dont la trace, appelée **degré de liberté effectif**, généralise la notion de degrés de liberté des splines de régression.

Annexe C

Publications

IOP Publishing Physics in Medicine and Biology
Phys. Med. Biol. **57** (2012) 7725–7739 doi:10.1088/0031-9155/57/22/7725

A multi-plane source model for out-of-field head scatter dose calculations in external beam photon therapy

Mohamed Amine Benadjaoud[1,2,3], Jérémi Bezin[1,2,3], Attila Veres[4], Dimitri Lefkopoulos[2], Jean Chavaudra[2], André Bridier[2], Florent de Vathaire[1,2,3] and Ibrahima Diallo[1,2,3]

[1] Inserm, CESP Centre for Research in Epidemiology and Population Health, U1018, Radiation Epidemiology Team, F 94807, Villejuif, France
[2] Institut Gustave Roussy, Villejuif, F-94805, France
[3] Université Paris XI, Villejuif, F-94800, France
[4] Equal-Estro Laboratory, Villejuif, F-94805, France

E-mail: ibrahim.diallo@igr.fr

Received 19 June 2012, in final form 26 September 2012
Published 2 November 2012
Online at stacks.iop.org/PMB/57/7725

Abstract
Our purpose was to assess the out-of-field dose component related to head scatter radiation in high-energy photon therapy beams and then derive a multisource model for this dose component. For scattered photons, several planar sources have been defined, with number, location and tilt depending on the complexity of the field shape. In the absence of precise knowledge of out-of-field scattering characteristics, several assumptions are made to derive emission spectra and radiation intensity from measurements. Among these, the Compton formula is used to evaluate scattered photon energy and the Henyey–Greenstein phase function is used to evaluate the scattered photon angular distribution. For measured doses under out-of-field conditions, the average local difference between the calculated and measured photon dose is 10%, including doses as low as 0.01% of the maximum dose on the beam axis. This study demonstrates that the multi-plane source approach is suitable for accurate analytical modeling of the out-of-field dose component related to head scatter radiation. These results should be taken into account when evaluating doses to the remaining volume at risk in external beam radiotherapy planning.

(Some figures may appear in colour only in the online journal)

1. Introduction

The basic purpose of high energy external beam photon therapy (EBPT) is to deliver the prescribed dose to the planned target volume (PTV) while sparing healthy tissues (ICRU 2010). Compared to conventional techniques, it is currently possible, using modern techniques, to reduce the normal tissue dose without affecting the tumor dose. This results in improved

tumor control and fewer complications in normal tissue. Nevertheless, EBPT is unavoidably associated with irradiation, at lower doses, of large volumes of normal tissue away from the beam path.

According to the latest recommendations of the International Commission on Radiation Units and Measurements (ICRU) concerning the remaining volume at risk (RVR), the search for means of more accurately determining such doses is of renewed clinical interest. Indeed, according to ICRU Report 83 (ICRU 2010), all normal tissues that could potentially be irradiated should be included in the RVR, and the absorbed dose in the RVR might be useful for estimating risk of later effects such as carcinogenesis.

Several important experimental studies have provided useful information on the relative contributions of the different components of the dose away from the beam path (Kase *et al* 1983, McParland and Fair 1991, Van Der Giessen and Hurkmans 1993, Van Der Giessen 1994, Ruben *et al* 2011). Other authors reported analytical models (Francois *et al* 1988, McParland and Fair 1991, Van Der Giessen and Hurkmans 1993, Van Der Giessen 1994, Diallo *et al* 1996) or Monte Carlo simulation codes (Mazonakis *et al* 2006, Kry *et al* 2006, 2007, Rijkee *et al* 2006, Joosten *et al* 2011, Chofor *et al* 2012) aimed at determining the absorbed dose away from the beam path in EBPT patients.

In essence, the out-of-field dose arises from three main sources: (1) leakage from the treatment unit; (2) scatter from the treatment unit head and from beam modifiers such as wedges and blocks; and (3) internal scatter originating in the patient. Only the first two sources depend on machine design and/or additional beam modifiers placed in the path of the beam; they must be measured for each individual treatment machine, while radiation scattered inside the patient depends only on energy and can be used irrespective of the machine under consideration. It also appears that only at a short distance does the patient scatter show a substantial contribution; at a longer distance, leakage and head scatter are the major contributors. At 6 and 15 MV for example, it has recently been established that, starting from 20–25 cm off axis at 10 cm depth, the contribution of leakage and head scatter is greater than 50% of the total out-of-field dose for field sizes up to to 20 $\times$ 20 cm^2 open fields and prevails in the far periphery (Joosten *et al* 2011, Chofor *et al* 2012).

The aim of this work was to experimentally assess, for several therapy machines, out-of-field absorbed dose components due to external scattering radiation in EBPT and then derive a multisource model suitable for computation of related absorbed dose distributions away from the beam path.

2. Materials and methods

2.1. Experimental data acquisition

Dose measurements for the present study were performed on three treatment machines: one cobalt unit, Alcyon II, and two linear accelerators, Clinac 2300 C/D operated at 6 and 20 MV and Novalis Tx operated at 6 MV (Varian Medical Systems, Palo Alto, CA, USA). Photon doses were measured with powder-type TLD-700 thermoluminescent dosimeters (Harshaw Chemical Company, Solon, OH). Dosimeters were read out in a PCL-3 (Fimel, Velizy, France) automated TLD reader.

A 100 $\times$ 50 $\times$ 30 (cm^3) water tank was used for absorbed dose-to-water measurements. The water tank was filled to a depth of 20 cm. A special TLD holder was designed in which capsules could be inserted at various depths and distances outside the beam path, enabling reproducible positioning of the capsules. The source-to-surface distance (SSD) was set at 80 and 100 cm for the cobalt unit and linear accelerators, respectively.

Five different field size settings were tested for the cobalt unit: 5 × 5, 10 × 10, 15 × 15, 20 × 20 and 30 × 30 cm^2; points of measurement for the absorbed dose were set, respectively, in the range of 15 to 70 cm, 15 to 70 cm, 15 to 70 cm and 25 to 70 cm distance from the beam axis at 10 cm depth in the water tank.

Four different field size settings were tested for the linear accelerators: 5 × 5, 10 × 10, 15 × 15 and 30 × 30 cm^2. Points of measurement for the absorbed dose were set in the range of a 15 to 70 cm, 15 to 70 cm, 15 to 70 cm, 20 to 70 cm and 25 to 70 cm distance from the beam axis at 10 cm depth in the water tank.

All TLDs were prepared and read by Equal-Estro Laboratory (Equal-Estro, Villejuif, France), which has longstanding experience in TLD use and analysis (Derreumaux *et al* 1995, Marre *et al* 2000, Ferreira *et al* 2000). Calibration for dose dependence was performed using a ^{60}Co gamma beam. Linearity correction was determined using interpolation of multiple known exposures over the range of doses received by experimental TLDs. Correction for energy dependence was determined by comparing the signal of known exposure with each treatment energy to known exposure with ^{60}Co. No fading correction was necessary, as reference TLDs were irradiated and read at the same time as test TLDs.

2.2. Semi-empirical method for evaluation of external scattering dose

The total absorbed dose $D(f,\ r,\ z,\ \varphi)$ to a site of interest located outside the beam path is generally described as the following sum of three main separate constituent components equation (1):

$$D(f,\ r,\ z,\ \varphi) = D_{\mathrm{IS}}(f,\ r,\ z,\ \varphi) + D_{\mathrm{ES}}(f,\ r,\ z,\ \varphi) + D_{\mathrm{EL}}(f, r,\ z,\ \varphi), \tag{1}$$

where f, r, z and φ are the field size, distance from the beam central axis, depth and angular orientation, respectively. $D_{\mathrm{IS}}(f,\ r,\ z,\ \varphi)$ is the absorbed dose due to internal scattering in the patient, $D_{\mathrm{ES}}(f,\ r,\ z,\ \varphi)$ is the absorbed dose due to external scattering in the treatment machine's head and $D_{\mathrm{EL}}(f,\ r,\ z,\ \varphi)$ is the absorbed dose due to leakage also due to the treatment machine's head.

Since our objective was to develop a semi-empirical model, analysis of $D_{\mathrm{ES}}(f,\ r,\ z,\ \varphi)$ alone was initially required in order to experimentally select this component. The experimental set-up has been reported (Kase *et al* 1983, Ruben *et al* 2011) and enables $D_{\mathrm{IS}}(f,\ r,\ z,\ \varphi)$, $D_{\mathrm{ES}}(f, r, z, \varphi)$ and $D_{\mathrm{EL}}(f, r, z, \varphi)$ to be separated via direct measurements. The experimental setup (figure 1), designed to avoid any contribution from $D_{\mathrm{IS}}(f,\ r,\ z,\ \varphi)$, consists of positioning the water tank relative to the beam edge so that the primary beam does not pass through or interact with water in the tank. The resulting measured dose $D_M(f,\ r,\ z,\ \varphi)$ thus reflects the sum of $D_{\mathrm{ES}}(f,\ r,\ z,\ \varphi)$ and $D_{\mathrm{EL}}(f,\ r,\ z,\ \varphi)$ only:

$$D_M(f,\ r,\ z,\ \varphi) = D_{\mathrm{EL}}(f,\ r,\ z,\ \varphi) + D_{\mathrm{ES}}(f,\ r,\ z,\ \varphi). \tag{2}$$

On each off axis position (r, z, φ), the measurement data for field size settings of 5 × 5, 10 × 10 and 15 × 15 cm^2 were used to fit the following function:

$$D_M^{\mathrm{Norm}}(f,\ r,\ z,\ \varphi) = \lambda(r,\ z,\ \varphi) + \kappa(r,\ z,\ \varphi) \times \left\{e^{[\nu(r,\ z,\ \varphi) \times f]} - 1\right\} \tag{3}$$

where $D_M^{\mathrm{Norm}}(f,\ r,\ z,\ \varphi)$ denotes the doses measured using the experimental setup in figure 1 normalized with respect to the central maximum dose, at distance r, depth z, in direction φ, for field size f. $\lambda(r,\ z,\ \varphi)$, $\kappa(r,\ z,\ \varphi)$ and $\nu(r, z, \varphi)$ are theoretically-driven parameters.

We note that equation (3) is formed by the sum of two terms, the first is an independent field size term while the second varies according to the field size. Furthermore, previously published data (Kase *et al* 1983) revealed that, in general, the external leakage dose was virtually independent of field size, whereas the external scattering dose strongly depended

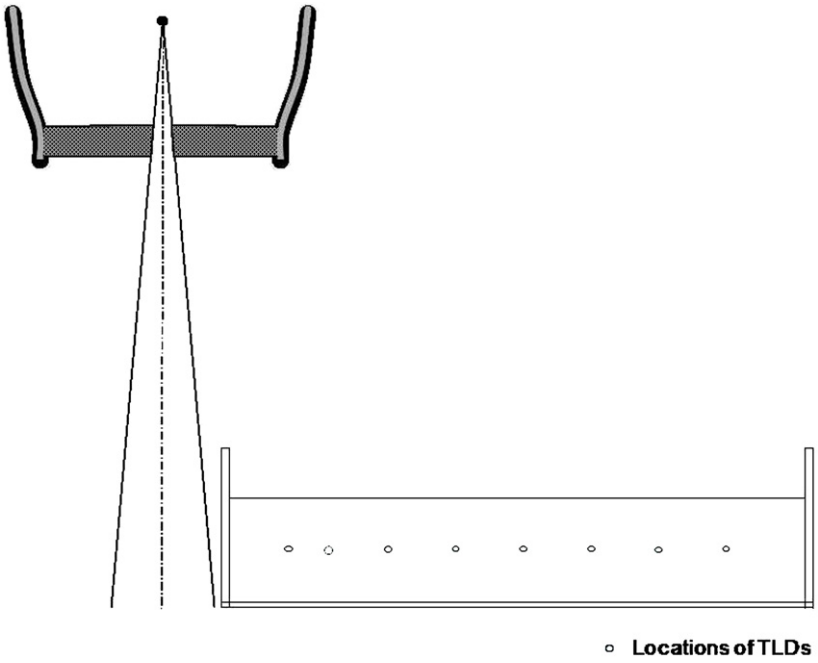

Figure 1. Experimental setup used to measure the dose reflecting the sum of the components related to external scattering and external leakage radiation, respectively.

on field size. It was therefore assumed, in this study, that the field size independent term $\lambda(r, z, \varphi)$ and varying term $\left(\kappa(r, z, \varphi) \times \left\{e^{[\nu(r,z,\varphi)\times f]} - 1\right\}\right)$ reflect at the fixed location (r, z, φ) respectively, the external leakage dose and the external scattering dose both normalized to the central axis dose maximum.

2.3. Multisource modeling

2.3.1. General concept. To describe $D_{ES}(f, r, z, \varphi)$, we chose multisource modeling introduced by Dunscombe and Nieminen (1992) and successfully used by several authors to study the field size dependence of relative output from linear accelerators (Dunscombe and Nieminen 1992, Yu and Sloboda 1995, Jian *et al* 2001, Yang *et al* 2002). Basically, in these models, a distributed source representing photons was scattered in the linear accelerator head. To model scattering on the beam central axis, it was generally assumed by previous authors that the dependence of photon fluence emanating from the distributed source on radial distance was Gaussian (Dunscombe and Nieminen 1992) or a combination of Gaussians (Yang *et al* 2002) or of different picked distributions (Yu and Sloboda 1995, Jian *et al* 2001). As depicted in figure 2, the model developed in the present study focused on scattering and leakage outside the beam path. It included a set of planar sources ($\boldsymbol{S}$) to represent photons reaching the measurement point after being scattered.

2.3.2. Scattering source. The scattering source $\boldsymbol{S}$ is not a physical source, but rather a virtual source modeled to include all scatter photons constituting the out-of-field scattering diffusion. For this, it is necessary to introduce, for $\boldsymbol{S}$, geometric and emission characteristics different from those previously reported. Hence, in the absence of precise knowledge about out-of-field scattering, several assumptions are made.

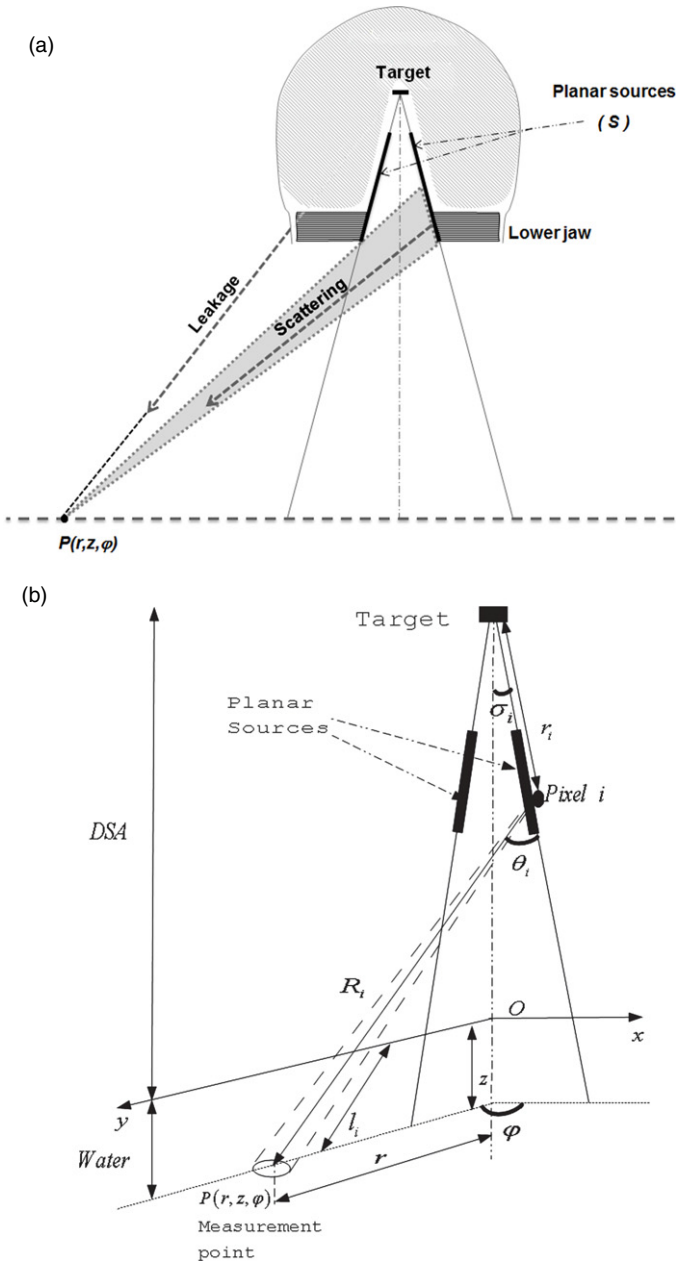

Figure 2. Basic principle of the multi-plane source model. (a) A schematic representation of the surface source ***S*** and of the linear accelerator head shielding and secondary collimator lower jaws is depicted and (b) indicates the geometric terms used in the final model formulation equation (7).

S is assumed to consist of a set of planar elementary sources emerging at the linear accelerator target and lining the inner surfaces of secondary collimator jaws and/or MLC elements and/or blocks, as appropriate, down to the bottom of the flattening filter (figure 2). Thus, number, location and tilt of surface element constituents of ***S*** depend on field shape.

Defined in this manner, this assembly of virtual sources is capable of modeling any complex field shape.

Furthermore, it is assumed that only the photons emanating from the region of $\boldsymbol{S}$ visible from the measurement point $P(r, z, \varphi)$, and denoted below as the aperture, contribute to $D_{\mathrm{ES}}(f, r, z, \varphi)$. On the other hand, we consider that a scattered photon from an invisible region of S contributes to the off axis dose as external leakage radiation taken into account by the term D_{EL} of the equation (2).

To facilitate computation of energy fluence distribution of photons emanating from $\boldsymbol{S}$, the aperture is divided into pixels, each measuring 5 mm $\times$ 5 mm. Software based on the z-buffer algorithm (Catmull 1984) is developed and used to identify pixels belonging to the aperture.

Source intensity distribution. The major portion of the in-field scatter radiation is generated by forward scattering in the flattening filter, whose scope is limited by the primary collimator (Ahnesjo 1994, Chaney *et al* 1994, Deng *et al* 2000). Therefore, we assume in this work that the planar sources of $\boldsymbol{S}$ are powered by the scattering photons originated from the bottom of the flattening filter.

Partly motivated by previous modeling works (Ahnesjo 1994, Yu and Sloboda 1995), the source intensity at the ith pixel is expressed analytically as an quasi triangular distribution:

$$\Omega_{\mathrm{ES}}(r_i, \sigma_i) = \begin{cases} (1 - \tau \times \sigma_i) \times \left(\dfrac{\mathrm{DSA}}{r_i}\right)^{\alpha} & r_i \leqslant r_0 \\ 1 - \tau \times \sigma_i & r_i \geqslant r_0 \end{cases}$$

where σ_i is the angle between the axis from the target to pixel i and the central axis, r_i is the distance between the target and pixel i and DSA is the distance source–axis assumed to be equal to 100 cm (80 cm for the cobalt unit), τ, r_0 and α are constants.

Energy distribution. The standard Compton expression (Compton 1923) is used to evaluate scattered photon energy:

$$E_{\theta_i} = \frac{E_0}{1 + \frac{E_0}{m_e c^2} \times (1 - \cos(\theta_i))} \tag{4}$$

where $E_{\theta i}$ is the energy of the scattered photon through angle θ_i; E_0 is a theoretically driven fit parameter reflecting the average energy of the incident photons; m_e is the mass of an electron at rest; and c is the speed of light.

Angular distribution. With high energy photon beams in high atomic number media (flattening filter, collimator, beam attenuators, etc), the generation of scattered photons is due to single or multiple Compton scattering as well as bremsstrahlung and annihilation radiation attributable to the secondary electrons and positrons of the photon beam.

Therefore, the Compton interaction must be corrected as proposed by authors such as Ahnesjo (1995).

We propose the use of the Henyey–Greenstein phase function which offers the advantage of a free parameter g which takes into account these complex phenomena after adjustment on experimental measurement data.

The Henyey–Greenstein phase function (Henyey and Greenstein 1941) is given by

$$p_{\mathrm{HG}}(\theta) = \frac{1}{4\pi} \times \frac{(1 - g^2)}{(1 + g^2 - 2 \times g \times \cos(\theta))^{\frac{3}{2}}} \tag{5}$$

where $\theta \in [0, \pi]$ (rad) is the angle existing between the direction of the photon before a scattering event and the direction after the scattering event. The parameter g is the Henyey–Greenstein asymmetry factor.

The scattering probability into the solid angle defined by an infinitesimal cone around the direction θ (see figure 2(b)) is actually given by (Binzoni *et al* 2006):

$$P_{\theta}^{g} = 2\pi \times p_{\mathrm{HG}}(\theta) \times \sin(\theta) = \frac{1}{2} \times \frac{(1-g^2) \times \sin(\theta)}{(1+g^2-2\times g \times \cos(\theta))^{\frac{3}{2}}}. \tag{6}$$

2.3.3 Multi plane source model. At point $P(r, z, \varphi)$ in the water tank, the external scattering dose normalized with respect to the maximum dose on the beam axis for field size f $D_{\mathrm{ES}}^{\mathrm{Norm}}(f, r, z, \varphi)$ related to photons emanating from the pixels of source $\boldsymbol{S}$ is expressed as:

$$D_{\mathrm{ES}}^{\mathrm{Norm}}(f, r, z, \varphi) = \omega \times \sum_{i=1}^{n(f,r,z,\varphi)} \Omega_{\mathrm{ES}}(r_i, \sigma_i) \times \frac{P_{\theta_i}^{g} \times E_{\theta_i} \times \exp(-\mu_{w,\theta_i} \times l_i)}{R_i^2} \tag{7}$$

where $\Omega_{\mathrm{ES}}(r_i, \sigma_i)$ is the source intensity at the pixel i, θ_i the angle between the axis from the target to pixel i and the axis from that pixel to the measurement point, $P_{\theta_i}^{g}$ is the probability of scattering through θ_i; $E_{\theta i}$ is the scattered photon energy; $\mu_{w,\theta i}$ is the linear attenuation coefficient in water for E_{θ_i}; l_i is the distance traveled in water by the photons emanating from pixel i of source S; R_i is the distance from that pixel to the measurement point; and ω is a theoretically driven fit parameter reflecting proportionality between estimated energy fluence and external scattering dose fraction of the central maximum dose. Summation is achieved over $n(f, r, z, \varphi)$ pixels corresponding to the region of $\boldsymbol{S}$ visible from the measurement point $P(r, z, \varphi)$ (see figure 2(b)).

Little information is needed on the unit treatment head (and can be obtained from the manufacturer's documentation) in order to fix the scattering source size and location: the distance between the target-bottom of the flattering filter, the target-jaws and/or target-multileaf collimator and leaf thickness.

Normally, the model parameters E_0, g, τ, r_0, α and ω depend on the field size since it interpolates doses normalized with respect to the maximum dose. However, we chose to fit the normalized doses all together to propose a model offering a compromise between all field sizes while also being compatible with the experimental results.

A least-square technique was then employed to determine values of E_0, g, τ, r_0, α and ω yielding the best fit to separate experimental data points available for ^{60}Co gamma rays, the two 6 MV RX beams and the 20 MV RX beams.

3. Results

Figure 3 shows estimations of the leakage component obtained when fitting our experimental data with equation (3). Depending on the distance from the beam axis, estimates of the leakage dose ranged from 0.013% to 0.055%, from 0.005% to 0.13% and from 0.01% to 0.18% of the maximum dose on the central axis, for the ^{60}Co, 6 and 20 MV beams, respectively. The mean standard error of the estimates is evaluated at 15% (min = 2%, max = 43%) for all beam energies used.

Table 1 summarizes parameters yielding the best fit to the measurement in multisource modeling. Figure 4 compares normalized dose profiles generated when using parameters from table 1 in the multisource model to measured profiles under a variety of conditions. The percent median difference in normalized local dose between measurements and calculations is 9%,

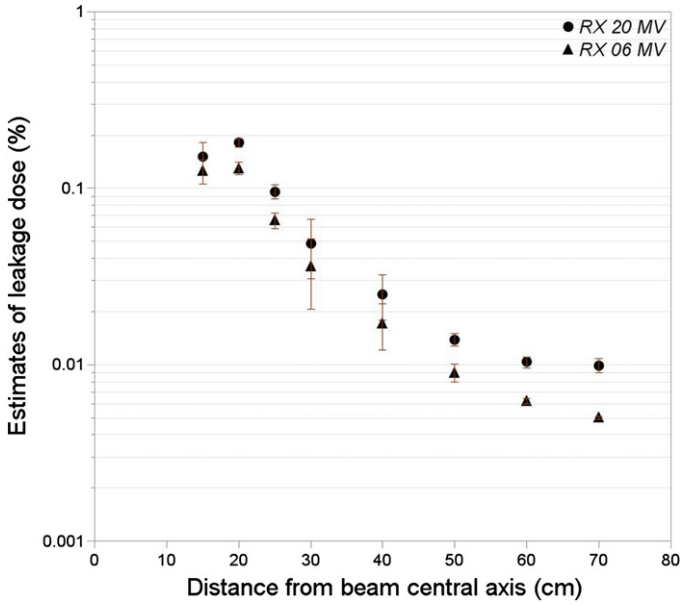

Figure 3. Estimates of leakage dose components (parameter $\lambda(r, z, \varphi)$, equation (3)) for the Varian linear accelerator Clinac 2300 C/D operated at 6 and 20 MV RX, respectively. The doses are given as percentages of the dose at the depth of the maximum dose on the beam axis.

Table 1. Values of energy factor E_0, the Henyey–Greenstein asymmetry factor g, the normalized dose–energy fluence proportionality coefficient ω and the source intensity constants κ, r_0 and α yielding the best fit of equation (7) to the separate experimental data points available in this study, for Alcyon II's ^{60}Co gamma beam, Varian Clinac 2300 C/D's 6 and 20 MV RX beams and Varian Novalis Tx's 6 MV RX beam.

	E_0 (MeV)	g	ω $\left(\text{cm}^2\ \text{MeV}^{-1}\right)$	τ (rad^{-1})	α	r_0 (cm)
Alcyon II	1	0.537	0.139	0[a]	1.7	15
6 MV Novalis Tx	2	0.223	0.314	3.5	2	20
6 MV Clinac 2300 C/D	2	0.418	0.319	3.8	2	25
20 MV Clinac 2300 C/D	6	0.103	0.303	3.8	2	25

[a] No flattening filter for the cobalt unit.

and three-quarters of the points on all fields and energies have a percent difference of less than 14%.

From this figure, it appears that the model breaks down for the smallest and the largest fields (14% for 5 × 5 cm^2 fields, 6% for 10 × 10, 8% for 15 × 15 cm^2 fields and 17% for 30 × 30 cm^2 fields). For example, the weakest agreement between calculations and measurements was observed for the Clinac's 5 cm × 5 cm (the model underestimated the normalized dose by 20% and 13% at off-axis distances ranging from 15 to 20 cm) and 30 cm × 30 cm (the model overestimated the normalized dose by 40% at 30 cm) field size.

Figures 5 and 6 illustrate the ability of present modeling to handle simple shapes (figures 5(a) and (b)) and multi-leaf-designed complex-shaped fields (figures 6(a) and (b)).

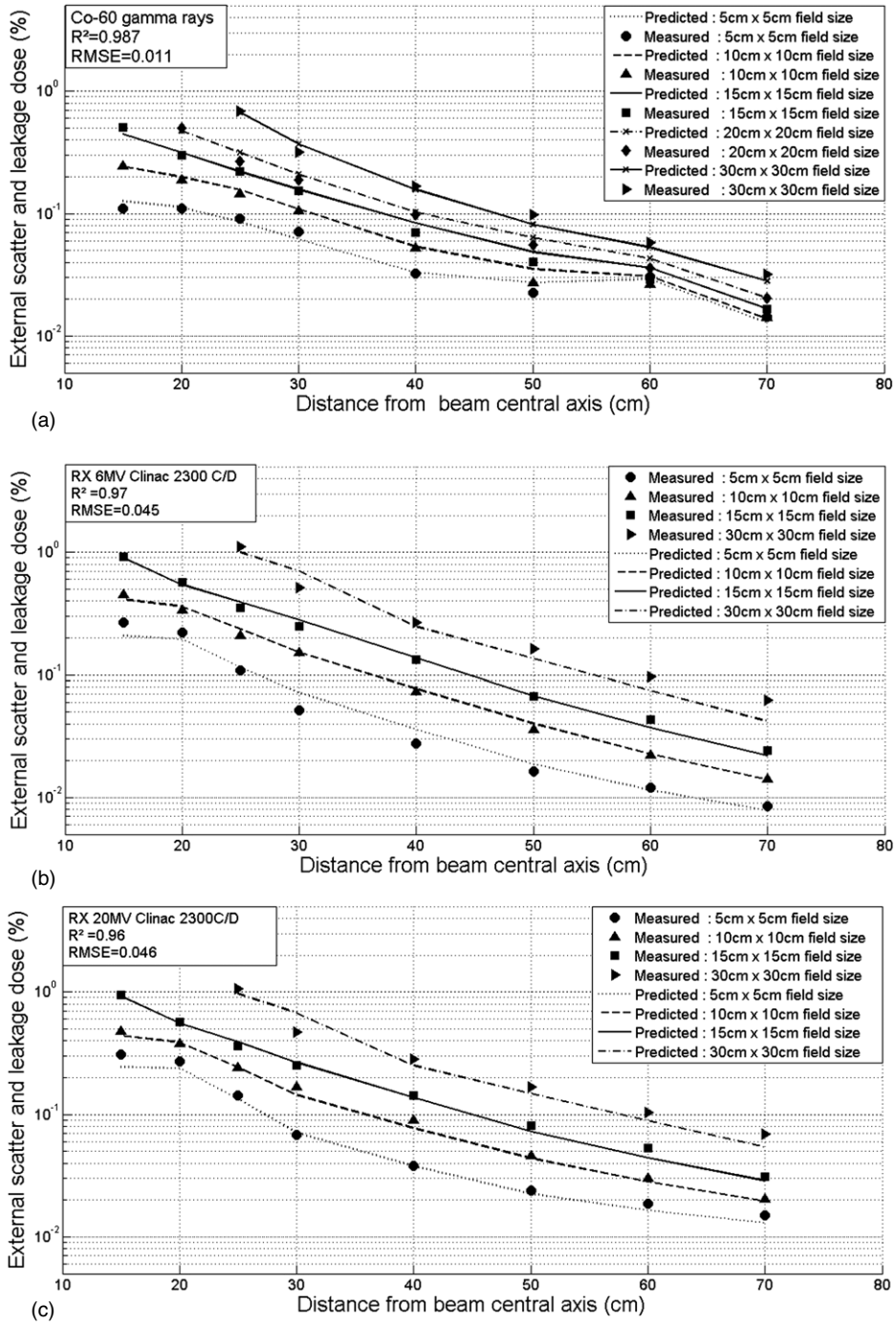

Figure 4. Variation, according to distance from the field center, of measured and calculated out-of-field doses at 10 cm depth in a water tank, for different field sizes (cm^2) and SSD set at 80 cm for cobalt treatment unit Alcyon II (a) and an SSD set at 100 cm for the Varian linear accelerators: Clinac 2300 C/D operated at 6 MV; (b) Clinac 2300 C/D operated at 20 MV RX; (c) and Novalis Tx operated at 6 MV (d). The doses are given as percentage of dose to dose at the depth of the maximal dose on the beam axis.

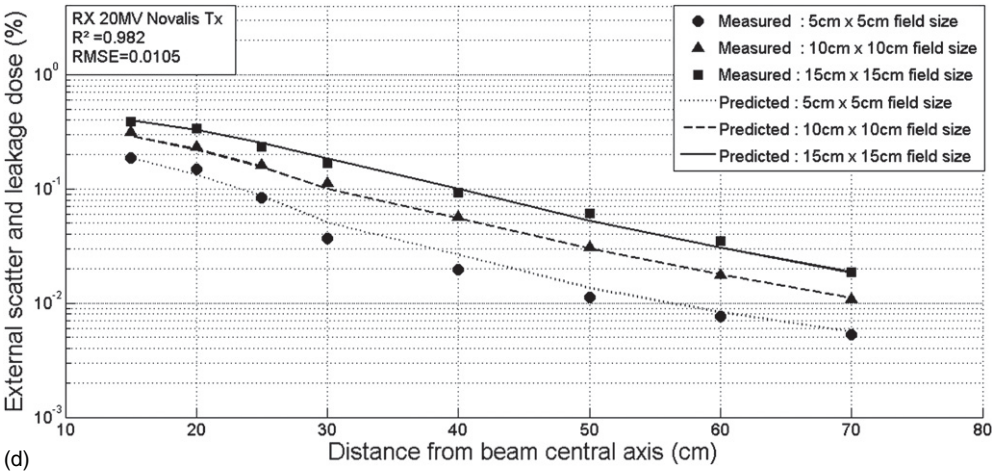

Figure 4. (Continued.)

4. Discussion

We attempted to assess the suitability of multisource modeling for computations of out-of field dose distributions related to linear accelerator head scatter and leakage radiation. This approach, introduced by Dunscombe and Nieminen (1992), has already been successfully used by several authors to study the field size dependence of relative output from linear accelerators (Dunscombe and Nieminen 1992, Yu and Sloboda 1995, Jian *et al* 2001, Yang *et al* 2002). By demonstrating that the mean difference between measurements and calculations may be less than 9% in out-of-field dose calculations, our work strongly suggests that multisource modeling could provide valuable assistance in developing modern out-of-field dose evaluation algorithms necessary for providing solutions to recent ICRU recommendations regarding the dose delivered to the RVR. This work also demonstrates the potential capacity of multi planar source modeling to accurately handle any complex shape field designed by modern medical linear accelerators equipped with multi-leaf collimators. It should also be noted that this approach can be adapted to a poly energetic beam by associating a different asymmetry parameter g_i with each energetic component E_i and consider the fluence weighted by the spectrum in equation (7). However, our results showed that the inclusion of a single energetic parameter E_0 was sufficient to obtain acceptable deviations from the experimental measurements.

Nonetheless, the validity of our semi-empirical modeling primarily depends on the accuracy of TLD measurements performed in order to establish estimates of the basic parameters g and ω. It had previously been established that overall uncertainty for TLD dose measurements in the Equal-Estro Laboratory using powder-type dosimeters under standard on-axis conditions was estimated as being in the range of 3.5 to 4.5% (2 SD) (Derreumaux *et al* 1995, Marre *et al* 2000, Ferreira *et al* 2000). In the present work, taking into account specific conditions due to off-axis measurements, additional uncertainties related to off-axis energy spectrum variation must be taken into account (Scarboro *et al* 2011). The overall uncertainty of dose measurements in off-axis conditions using a set of three TLDs was estimated to be about 10% (2 SD) for the lowest dose range. Uncertainty would probably be lower for higher doses.

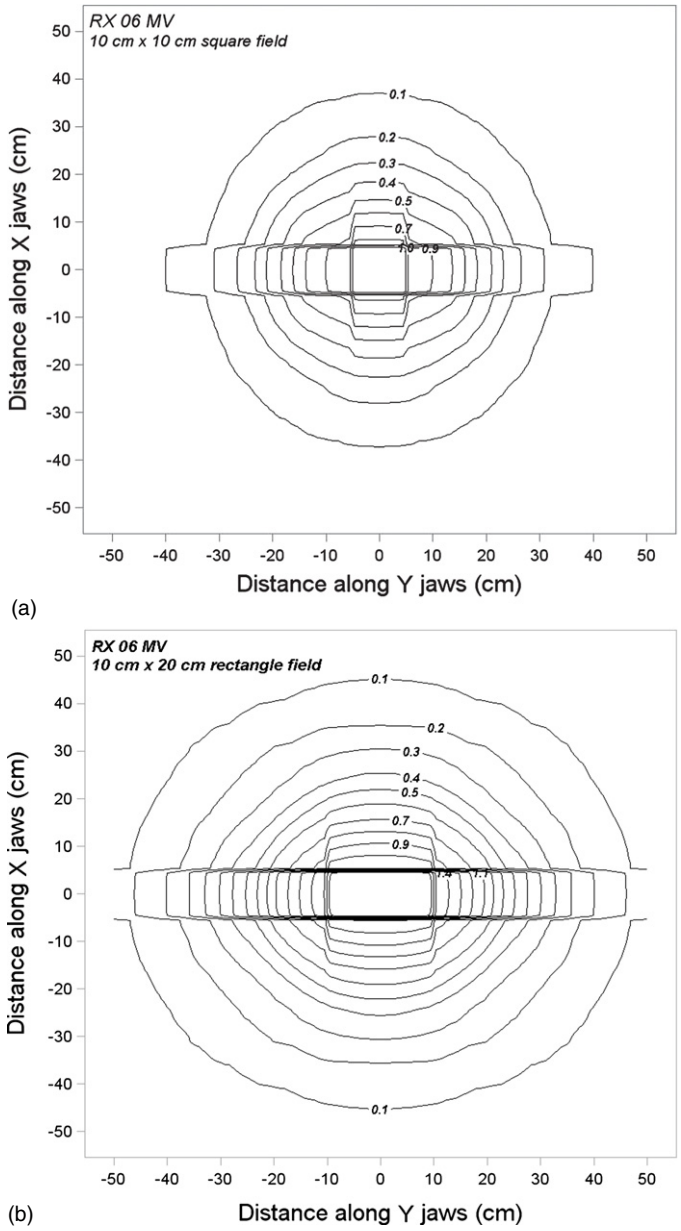

Figure 5. Calculated isodose for simple shaped fields (a) 10 × 10 cm^2 square field and (b) 10 × 20 cm rectangular field. Doses are given as percentage of the dose to the dose at the depth of the maximum dose on the beam axis. Calculations are made for Varian linear accelerator Clinac 2300 C/D operated at 6 MV RX.

5. Conclusions

The present work has shown that the total absorbed dose outside the beam path can be experimentally assessed as the sum of three separate constituent components: (1) the absorbed dose due to internal scattering originating within the patient; (2) the absorbed dose due to

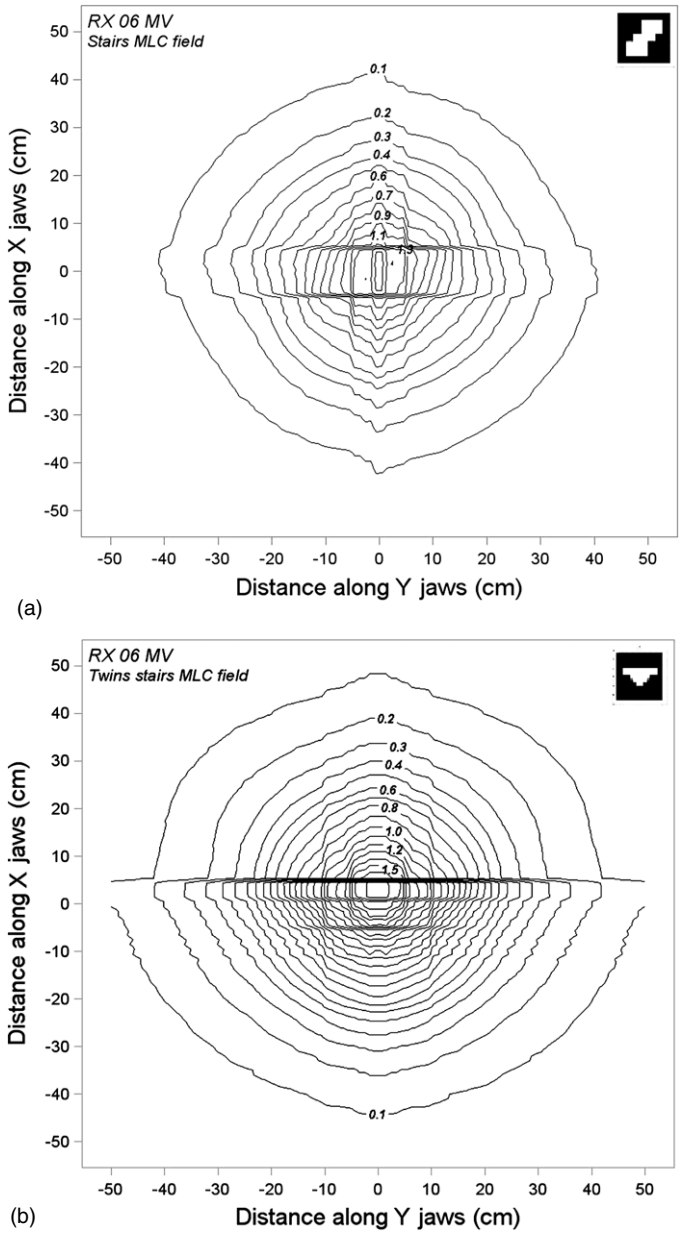

Figure 6. Calculated isodose curves for complex shaped fields. The shapes of the fields are inserted top left on the pictures. Doses are given as percentage of dose to dose at the depth of the maximal dose on the beam axis. Calculations are made for Varian linear accelerator Clinac 2300 C/D operated at 6 MV RX.

external scattering originating in the treatment machine's head and; (3) the absorbed dose due to external leakage also originating in the treatment machine's head. The latter two components can be successfully modeled using the multisource approach if proper design of the extrafocal sources is established.

The applicability of the method for complex field shapes is demonstrated as well as providing applications to the IMRT modalities, which are becoming an increasing trend. IMRT is most often accompanied by increased monitor units (Ruben *et al* 2011), which are most often delivered via small field segments for which the out-of-field dose by head scatter and leakage plays an even greater role.

6. Future directions

A recent work (Howell *et al* 2010) demonstrated that, for tissues located about ten centimeters from the treatment field border, most current commercial treatment planning systems (TPS) used in radiotherapy departments for patient treatment planning may underestimate out-of-field doses by more than 50%. According to ICRU Report 83 (ICRU 2010), absorbed dose in the RVR might be useful for estimating the risk of deterministic effect related for example to the dose to the ovary or crystalline and late effects such as cardiovascular diseases, pulmonary fibrosis, second cancers, etc. Therefore, providing solutions for better estimations of out-of-field doses will have increasing clinical significance (choosing between different techniques depending on the radiation dose delivered to RVR for example) and their integration in TPS will soon become widespread. This work may be relevant to the concerns of modern radiotherapy due to its flexibility with regard to the machine's geometry and its low computational cost.

Finally, several worldwide epidemiological publications (BEIR 2006, Xu *et al* 2008, NCRP report 170, De Vathaire 1999) have been devoted to the occurrence of second primary cancers among patients having received radiation therapy for primary cancers in childhood, adolescence, or as adults. There appear to be considerable uncertainties in the current second cancer risk models. To better assess the dose–response relationship, several studies have suggested that cancer risk is related to the inhomogeneous dose distribution across an organ rather than the average dose (Dasu *et al* 2005, Schneider *et al* 2005). However, these approaches require an accurate and patient-specific out-of-field dose reconstruction.

Our findings could then potentially impact the design of external beam radiotherapy patient-cohort-based epidemiological studies of radiation-induced late toxicity, notably for organs located at some distance from the beams.

Acknowledgments

This study was funded by the French National Institute of Cancer (INCa) (Award code: DGOS_2443) and by the European Commission, FP7-Health, PanCareSurFup project (grant agreement number 257505). The authors also would like to thank Mary Hittinger for editing.

References

Ahnesjo A 1994 Analytic modeling of photon scatter from flattening filters in photon therapy beams *Med. Phys.* **21** 1227–35

Ahnesjo A 1995 Modeling transmission and scatter for photon beam attenuators *Med. Phys.* **22** 1711–20

BEIR 2006 *Health Risks from Exposure to Low Levels of Ionizing Radiation BEIR VII, Phase 2* (Washington, DC: The National Academies Press, National Research Council, National Academy of Science)

Binzoni T, Leung T S, Gandjbakhche A H, Rüfenacht D and Delpy D T 2006 Comment on 'The use of the Henyey–Greenstein phase function in Monte Carlo simulations in biomedical optics' *Phys. Med. Biol.* **51** L39–41

Catmull E 1984 An analytic visible surface algorithm for independent pixel processing *SIGGRAPH Proc.* vol 18 pp 109–15

Chaney E L, Cullip T J and Gabriel T A 1994 A Monte Carlo study of accelerator head scatter *Med. Phys.* **21** 1383–90

Chofor N, Harder N, Willborn K C and Poppe B 2012 Internal scatter, the unavoidable major component of the peripheral dose in photon-beam radiotherapy *Phys. Med. Biol.* **57** 1733–43

Compton A H 1923 A quantum theory of the scattering of x-rays by light elements *Phys. Rev.* **21** 483–502

Dasu A, Toma-Dasu I, Olofsson J and Karlsson M 2005 The use of risk estimation models for the induction of secondary cancers following radiotherapy *Acta Oncol.* **44** 339–47

De Vathaire F *et al* 1999 Second malignant neoplasms after a first cancer in childhood: temporal pattern of risk according to type of treatment *Br. J. Cancer* **79** 1884–93

Deng J, Jiang S B, Kapur A, Li J, Pawlicki T and Ma C M 2000 Photon beam characterization and modelling for Monte Carlo treatment planning *Phys. Med. Biol.* **45** 411–27

Derreumaux S, Chavaudra J, Bridier A, Rossetti V and Dutreix A 1995 A European quality assurance network for radiotherapy: dose measurement procedure *Phys. Med. Biol.* **40** 1191–209

Diallo I, Lamon A, Shamsaldin A, Grimaud E, de Vathaire F and Chavaudra J 1996 Estimation of the radiation dose delivered to any point outside the target volume per patient treated with external beam radiotherapy *Radiother. Oncol.* **38** 269–71

Dunscombe P B and Nieminen J M 1992 On the field size dependence of relative output from a linear accelerator *Med. Phys.* **19** 1441–4

Ferreira IH, Dutreix A, Bridier A, Chavaudra J and Svensson H 2000 The ESTRO-QUALity assurance network (EQUAL) *Radiother. Oncol.* **55** 273–84

Francois P, Beurtheret C and Dutreix A 1988 Calculation of the dose delivered to organs outside the radiation beams *Med. Phys.* **15** 879–83

Henyey L G and Greenstein J L 1941 Diffuse radiation in the galaxy *Astrophys. J.* **93** 70–83

Howell R M, Scarboro S B, Kry S F and Yaldo Z 2010 Accuracy of out-of-field dose calculations by a commercial treatment planning system *Phys. Med. Biol.* **55** 6999–7008

International Commission on Radiation Units and Measurements (ICRU) 2010 Prescribing, Recording, and Reporting Intensity-Modulated Photon-Beam Therapy (IMRT) *ICRU Report* 83 *J. ICRU* **10**

Jian S B, Boyer A L and Ma C M 2001 Modeling the extrafocal radiation and monitor chamber backscatter for photon beam dose calculation *Med. Phys.* **28** 55–66

Joosten A, Bochud F, Baechler S, Levi F, Mirimanoff R-O and Moeckli R 2011 Variability of a peripheral dose among various linac geometries for second cancer risk assessment *Phys. Med. Biol.* **56** 5131–51

Kase K R, Svensson G K, Wolbarst A B and Marks M A 1983 Measurements of dose from secondary radiation outside a treatment field *Int. J. Radiat. Oncol. Biol. Phys.* **9** 1177–83

Kry S F, Titt U, Followill D, Pönisch F, Vassiliev O N, White R A, Stovall M and Salehpour M 2007 A Monte Carlo model for out-of-field dose calculation from high-energy photon therapy *Med. Phys.* **34** 3489–99

Kry S F, Titt U, Pönisch F, Followill D, Vassiliev O N, White R A, Mohan R and Salehpour M 2006 A Monte Carlo model for calculating out-of-field dose from a Varian 6-MV beam *Med. Phys.* **33** 4405–13

Marre D, Ferreira I H, Bridier A, Björeland A, Svensson H, Dutreix A and Chavaudra J 2000 Energy correction factors of LiF powder TLDs irradiated in high-energy electron beams and applied to mailed dosimetry for quality assurance networks *Phys. Med. Biol.* **45** 3657–74

Mazonakis M, Tzedakis A, Damilakis J, Varveris H, Kachris S and Gourtsoyiannis N 2006 Scattered dose to thyroid from prophylactic cranial irradiation during childhood: a Monte Carlo study *Phys. Med. Biol.* **51** N139–45

McParland B J and Fair H I 1991 A method of calculating peripheral dose distributions of photon beams below 10 MV *Med. Phys.* **19** 283–93

National Council on Radiation Protection and Measurements (NCRP) 2012 Second primary cancers and cardiovascular disease after radiation therapy *Report No 170* (NCRP)

Rijkee AG, Zoetelief J, Raaijmakers C P J, Van Der Marck S C and Van Der Zee W 2006 Assessment of induction of secondary tumours due to various radiotherapy modalities *Radiat. Prot. Dosim.* **118** 219–26

Ruben J D, Lancaster C M, Jones P and Smith R L 2011 A comparison of out-of-field dose and its constituent components for intensity-modulated radiation therapy versus conformal radiation therapy: implications for carcinogenesis *Int. J. Radiat. Oncol. Biol. Phys.* **81** 1458–64

Scarboro S B, Followill D S, Howell R M and Kry S F 2011 Variations in photon energy spectra of a 6 MV beam and their impact on TLD response *Med. Phys.* **38** 2619–28

Schneider U, Zwahlen D, Ross D and Kaser-Hotz B 2005 Estimation of radiation-induced cancer from three-dimensional dose distributions: concept of organ equivalent dose *Int. J. Radiat. Oncol. Biol. Phys.* **61** 1510–5

Van der Giessen P H 1994 Calculation and measurement of the dose at points outside the primary beam for photon energies of 6, 10, and 23 MV *Int. J. Radiat. Oncol. Biol. Phys.* **30** 1239–46

Van der Giessen P H and Hurkmans C W 1993 Calculation and measurement of the dose in points outside the primary beam for ^{60}Co gamma radiation *Int. J. Radiat. Oncol. Biol. Phys.* **27** 717–24

Xu X G, Bednarz B and Paganetti H 2008 A review of dosimetry studies on external-beam radiation treatment with respect to second cancer induction *Phys. Med. Biol.* **53** R193–241

Yang Y, Xing L, Boyer AL, Song Y and Hu Y 2002 A three-source model for the calculation of head scatter factors *Med. Phys.* **29** 2024–33

Yu M K and Sloboda R 1996 Analytical representation of head scatter factors for shaped photon beams using a two-component x-ray source model *Med. Phys.* **23** 973–84

International Journal of Radiation Oncology biology • physics

www.redjournal.org

Physics Contribution

Functional Data Analysis in NTCP Modeling: A New Method to Explore the Radiation Dose-Volume Effects

Mohamed Amine Benadjaoud, MS,[*,†,‡] Pierre Blanchard, MD, PhD,[†,§] Boris Schwartz, MS,[*,†,‡] Jérôme Champoudry, Dipl Phys,[||] Ryan Bouaita, MD,[¶] Dimitri Lefkopoulos, PhD,[**] Eric Deutsch, MD, PhD,[†,§,††] Ibrahima Diallo, PhD,[*,†,‡] Hervé Cardot, PhD,[‡‡] and Florent de Vathaire, PhD[*,†,‡]

Center for Research in Epidemiology and Population Health (CESP) INSERM 1018 Radiation, Epidemiology Group, Villejuif, France; †Université Paris sud, Le Kremlin-Bicêtre, France; ‡Institut Gustave Roussy, Villejuif, France; §Department of Radiation Oncology, Institut Gustave Roussy, Villejuif, France; ||Department of Radiation Oncology, CHU de la Timone, Marseille, France; ¶Department of Radiation Oncology, CHU Henri Mondor, Creteil, France; **Department of Radiation Physics, Institut Gustave Roussy, Villejuif, France; ††INSERM 1030, Molecular Radiotherapy, Villejuif, France; and ‡‡Institut de Mathématiques de Bourgogne, Université de Bourgogne, Dijon, France

Received Mar 24, 2014, and in revised form Jun 17, 2014. Accepted for publication Jul 9, 2014.

Summary

We propose a novel normal tissue complication probability model wherein the weights dose values in the generalized equivalent uniform dose equation are flexibly estimated using a functional data analysis tools.

Purpose/Objective(s): To describe a novel method to explore radiation dose-volume effects. Functional data analysis is used to investigate the information contained in differential dose-volume histograms. The method is applied to the normal tissue complication probability modeling of rectal bleeding (RB) for patients irradiated in the prostatic bed by 3-dimensional conformal radiation therapy.

Methods and Materials: Kernel density estimation was used to estimate the individual probability density functions from each of the 141 rectum differential dose-volume histograms. Functional principal component analysis was performed on the estimated probability density functions to explore the variation modes in the dose distribution. The functional principal components were then tested for association with RB using logistic regression adapted to functional covariates (FLR). For comparison, 3 other normal tissue complication probability models were considered: the Lyman-Kutcher-Burman model, logistic model based on standard dosimetric

Reprint requests to: Mohamed Amine Benadjaoud, MS, Radiation Epidemiology Group CESP - INSERM 1018, Institut Gustave Roussy - Bat B2M 114 rue Edouard Vaillant F-94805 Villejuif Cedex. Tel (+33) 0-1-42-11-55-73; E-mail: mohamedamine.benadjaoud@gustaveroussy.fr

Supported by the European Community's Seventh Framework Programme (FP7-EURATOM-FISSION) Research Infrastructures action under grant agreement No. FP7-269553 (EpeRadBio project).

Conflict of interest: none

Supplementary material for this article can be found at www.redjournal.org.

Int J Radiation Oncol Biol Phys, Vol. 90, No. 3, pp. 654–663, 2014
0360-3016/$ - see front matter
http://dx.doi.org/10.1016/j.ijrobp.2014.07.008

parameters (LM), and logistic model based on multivariate principal component analysis (PCA).

Results: The incidence rate of grade ≥2 RB was 14%. V_{65Gy} was the most predictive factor for the LM ($P = .058$). The best fit for the Lyman-Kutcher-Burman model was obtained with n = 0.12, m = 0.17, and TD50 = 72.6 Gy. In PCA and FLR, the components that describe the interdependence between the relative volumes exposed at intermediate and high doses were the most correlated to the complication. The FLR parameter function leads to a better understanding of the volume effect by including the treatment specificity in the delivered mechanistic information. For RB grade ≥2, patients with advanced age are significantly at risk (odds ratio, 1.123; 95% confidence interval, 1.03-1.22), and the fits of the LM, PCA, and functional principal component analysis models are significantly improved by including this clinical factor.

Conclusion: Functional data analysis provides an attractive method for flexibly estimating the dose-volume effect for normal tissues in external radiation therapy.

Introduction

The risk of normal tissue toxicity has always been a concern in the optimization of treatment plans by taking into account mainly normal tissue complication probability (NTCP). Variations in normal tissue dose-volume histograms (DVHs) from patient to patient are significantly greater than tumor DVH variations, and many NTCP models have been proposed to study the effect of this dose heterogeneity.

The most widely NTCP model is the Lyman-Kutcher-Burman (LKB) model (1-3) based on the reducing of an arbitrary nonuniform dose distribution to an equivalent uniform dose (EUD) and the use of a power law, with an exponent reflecting the volume effect (4).

One of the major limitations of this model is the noninclusion of predisposing clinical factors in its formulation, and some authors have proposed a method to overcome this problem at the expense of statistical power by stratifying the population (5) or by using dose-modifying factors (6, 7).

Multimetric NTCP models, for example based on logistic regression, have been proposed that attempt to explicitly take into account the clinical patient-specific parameters. In their simplest form, these models include only a few single points from the cumulative DVH (cDVH), which potentially lose important dose-volume information that might best discriminate between treatment plans at high or low risk for complications. More sophisticated models propose the use of principal component analysis (PCA) as a method of using all the information inherent in the cDVHs by quantifying the variability and segregating similar morphology (8, 9). Despite their attractive properties, the interpretation of PCA results is challenging, and the additional information provided in comparison with more standard DVH parameters is unclear (10).

The approach adopted here stems from a field of statistics known as functional data analysis (FDA) (11). The idea is to consider, for each patient, the dose values (calculated by the treatment planning system over a 3-dimensional [3D] matrix of voxels composed of critical structures within the patient anatomy) as realizations of random variable samples taken from an underlying probability density function (*pdf*). Although the normalized differential DVHs (dDVHs) provide a picture of the *pdf* shape, there are *pdf* estimators that have better mathematical properties than histograms (in terms of smoothness or convergence rate, for example) and the kernel density estimation is the most popular of them (12-14).

Consequently, the data for each patient are expressed as a single functional observation: the *pdf*. In this framework, the analysis of partial volume effects of normal tissues to radiation consists of describing a correlation between a functional predictor (*pdf*) and a scalar response (here binary response: toxicity yes/no).

Inasmuch as the objects of study are functional, ideally this should be the context of the data analysis, and FDA provides a generalization to functional data of various standard statistical methods like regression and PCA.

In the present study, FDA was applied to the rectal dDVHs of 141 patients irradiated in the prostatic bed by 4-field box, 3D conformal radiation therapy (3D CRT) technique. Correlation of *pdf*s and chronic rectal bleeding (RB) of grade ≥2 was then investigated with a extension of the logistic regression analysis to functional data. We compared on the same patient population the results of this approach with those given by 3 other NTCP models: the LKB model, the logistic model, and the PCA model.

Methods and Materials

Patient population and toxicity scoring

Between September 2005 and December 2010, 141 patients were irradiated in the prostatic bed for a rising prostate-

specific antigen after radical prostatectomy in the Gustave Roussy institute (Villejuif, France). The study was performed with the approval of an institutional review board.

All patients were treated using 3D CRT techniques with 18-MV or 20-MV photons including 3 dose levels: 66 Gy and 70 Gy consisting, respectively, of 33 or 35 prescribed daily fractions of 2 Gy, and 65 Gy consisting of 26 prescribed daily fractions of 2.5 Gy. Androgen deprivation was also permitted.

After RT, patients returned for follow-up visits with a digital rectal examination and serum prostate-specific antigen determination every 3 to 6 months for the first 5 years and yearly thereafter. The patients were followed up through December 31, 2013.

Toxicity data were collected retrospectively through follow-up records. The clinical variables recorded were age, diabetes mellitus, cardiovascular comorbidity, previous abdominal/pelvic surgery, the use of anticoagulants, antiaggregants, and antihypertensives. The type and duration of hormonal therapy, if prescribed, was also reported.

The endpoint considered was Radiation Therapy Oncology Group grade $\geq$2 RB. Grade 2 includes intermittent bleeding, grade 3 consists of bleeding requiring laser treatment or transfusion of packed cells, and grade 4 consists of necrosis/perforation fistula.

Radiation treatment planning and DVHs

All patients were scanned in a supine position with a slice thickness of 0.3 cm. The rectum was outlined from the anal verge to the rectosigmoid junction and considered as a solid organ (including rectal filling). Every patient was treated in a supine position with a 4-field box technique using parallel opposed anterior and posterior and parallel opposed lateral portals with individualized blocks, multileaf collimators, or both, derived from the beam's-eye-view. Patient positioning was checked on bony anatomy before treatment initiation and weekly hereafter. At the time of treatment planning, dose constraints were used to validate the appropriate plan: femoral heads ($V_{65Gy} < 10\%$), rectum ($V_{65Gy} < 50\%$), and bladder ($V_{65Gy} < 50\%$) while maintaining an adequate target volume coverage ($D_{95\%} > 95\%$ of the prescribed dose).

Differential absolute DVHs with dose bins of 0.33 Gy for the rectum were available as sum plan dose distribution. All the dDVHs were converted to linear-quadratic biologically effective dose (BED) delivered in 2-Gy fractions using $\alpha/\beta = 3$ Gy (15, 16).

Statistical analysis

Functional data analysis approach

Our FDA approach was inspired by Ramsay (17). The DVH analysis is a natural framework to deal with functional objects.

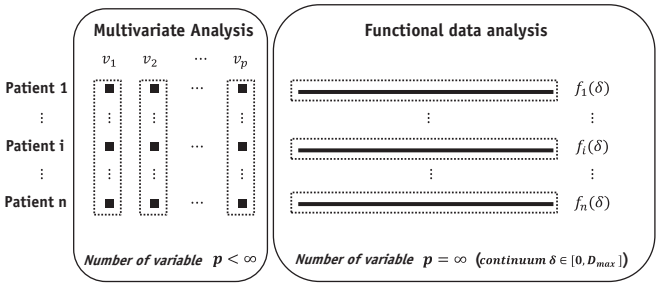

Fig. 1. Functional data analysis approach (right) compared with classic data matrix approach (left).

Figure 1 offers an approach to the concept of a functional datum. At the left we have the domain of the classic data matrix and its vertical approach: the i^{th} patients is paired with each of m variables $v_{i1},\ldots,v_{im}$ (percentage of the volume corresponding to the grid dose $d_1,\ldots,d_m$).

Let us now fix the number of subjects n and allow the size of the dose grid $d_1,\ldots,d_m$ to increase without limit, which reflects the ability of the actual treatment planning system to deliver a dDVH with a small bin width. This process can be extended to the situation in which where the dose grid points define a continuum. The right side of Figure 1 indicates the horizontal FDA approach: the data now become a *pdf* function $f_i(\delta)$ to designate the value assigned to each subject i at point δ on this continuum.

Functional principal component analysis

After estimating each *pdf* using the kernel density estimation (details in Appendix 1, available at www.redjournal.org), we turn to the functional principal component analysis (FPCA) method proposed by Kneip and Utikal (18) to succinctly describe the *pdf*'s dynamics that explain the most variability.

Following the Kneip and Utikal approach, we propose to represent each *pdf* $f_i(\delta)$ in terms of the Karhunen-Loève decomposition:

$$f_i(\delta) = \overline{f}(\delta) + \sum_{k=1}^{n} \theta_{ik} \times g_k(\delta) \qquad (1)$$

Where $\overline{f}(\delta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\delta)$ is the common mean, (g_k) the eigenfunctions, which exhibit, in an optimal way according to a variance criterion, the main modes of variation of the *pdfs* relative to $\overline{f}$. The scores (θ_{ik}) describe what these variation modes characterizes the i^{th} *pdf* f_i and their variances $\lambda_k(eigenvalue)$ can be interpreted as the contribution of k^{th} term in (1) to the total variance.

We estimate the constants (λ_k),(θ_k) and the (g_k) functions following the method described in Appendix 1 (available at www.redjournal.org).

Model formulation and interpretation

Our new FPCA NTCP model is based on a generalization of the multivariate logistic regression to functional covariates.

The objective of the functional logistic regression (FLR) is to explain a binary response (toxicity yes/no) in terms of a continuous functional predictor: ie, the *pdf*.

The FLR is formulated as follows:

$$logit(NTCP_i) = \alpha_0 + \sum_{j\geq 1} \alpha_j \times \text{clinical\_covariates}_{ij} + \int_0^{D_{max}} \beta(\delta) \times f_i(\delta) d\delta \quad (2)$$

Where $logit(NTCP_i) = \ln\left(\frac{NTCP_i}{1-NTCP_i}\right)$ is the logit transformation of the RB probability, α_0 (the intercept) and $(\alpha_j)_{j\geq 1}$ are real parameters, and $\beta(\delta)$ is a parameter function.

To estimate $\beta(\delta)$, we regress first on a reduced number K (in practice $K \leq 10$ to explain the high proportion of total variance) of functional principal component scores (FPCS):

$$logit(NTCP_i) = \alpha_0 + \sum_{j\geq 1} \alpha_j \times \text{clinical\_covariates}_{ij} + \sum_{k=1}^{K} \beta_k \times \theta_{ik} \quad (3)$$

Inasmuch as the model (2) has now become in (3) a standard logistic regression, we can use any multivariate model selection procedure to build a parsimonious model from a FPCS subset I_{FPCS}.

Then, the estimated coefficients $(\beta_k)_{k\in I_{FPCS}}$ lead to an estimation of the parameter function itself by:

$$\beta(\delta) = \sum_{k\in I_{FPCS}} \beta_k \times g_k(\delta) \quad (4)$$

Finally, the radiation effect odds ratio (OR) obtained from the FLR for the *pdf* $f_i(\delta)$ is given by:

$$OR_i^{\text{Radiation}} = \exp\left(\int_0^{D_{max}} \beta(\delta) \times f_i(\delta) d\delta\right) \quad (5)$$

Thus, it is very important to obtain an accurate estimation of $\beta(\delta)$ because its interpretation provides information about the volume effects in the probability of late RB.

Other NTCP models and model comparison

In the present analysis, the following models were used in addition to the FPCA NTCP model (description in Appendix 2, available at www.redjournal.org): the LKB model, the logistic model, and the PCA model. The likelihood ratio (LR) test was used to test significant differences between nested models, and the Akaike information criterion (AIC) compared the discriminating ability of all model fits (19). Finally, the area under the receiver operating characteristic curve (AUC) was used to quantify the accuracy of the predictive model.

Results

Incidence of rectal bleeding resume here

A total of 141 patients treated on protocol from September 2005 to December 2010 were analyzed in this study. The characteristics of all patients are shown in Table 1. The

Table 1 Patient and radiation therapy characteristics

Characteristic	No. of patients (% of total)	*P* value* grade ≥2	*P* value grade 3
Mean follow-up time	4 y (range, 0.5-8 y)		
Grade ≥2 rectal bleeding	20/141 (14%)		
Time to grade ≥2 rectal bleeding	Mean 17 mo (range, 5-37 mo)		
Grade 3 rectal bleeding	9/141 (6.4%)		
Time to grade 3 rectal bleeding	Mean 22 mo (range, 5-37 mo)		
Age	Mean 65 y (range, 51-80 y)	.01	.18
Abdominal surgery	35/141 (25%)	.28	>.5
Hypertension	39/141 (28%)	>.5	>.5
Hypercholesterolemia	18/141 (13%)	.28	>.5
Diabetes mellitus	11/141 (8%)	>.5	>.5
Use of drugs anticoagulants/ antiaggregants	14/141 (10%)	.5	.5
Hormonal therapy	34/141 (24%)	>.5	>.5
Dose fraction		.44	.37
2 Gy	68/141 (48%)		
Grade ≥2 rectal bleeding	8/68		
Grade 3 rectal bleeding	3/68		
2.5 Gy	73/141 (52%)		
Grade ≥2 rectal bleeding	12/73		
Grade 3 rectal bleeding	6/73		

* *P* value of the association with grade ≥2 or grade 3 rectal bleeding by univariate logistic regression.

median age at start of treatment was 65 years (range, 51-80 years), and the median follow-up time was 4 years (range, 6 months to 8 years).

Twenty (14%) of the 141 patients experienced grade $\geq$2 RB, including 9 grade 3. Among these patients, 5 (25%) experienced RB between 6 months and 1 year, after radiation therapy, 11 (55%) between 1 and 2 years, and 4 (20%) after more than 2 years.

dDVH comparison for patients with and without bleeding

Figure 2A shows 2 examples of a kernel smoothing density estimation from a dDVH, and Figure 2B illustrates the amount of patient-to-patient variability in the estimated *pdfs* (left) and their average (right) for those with and without complication.

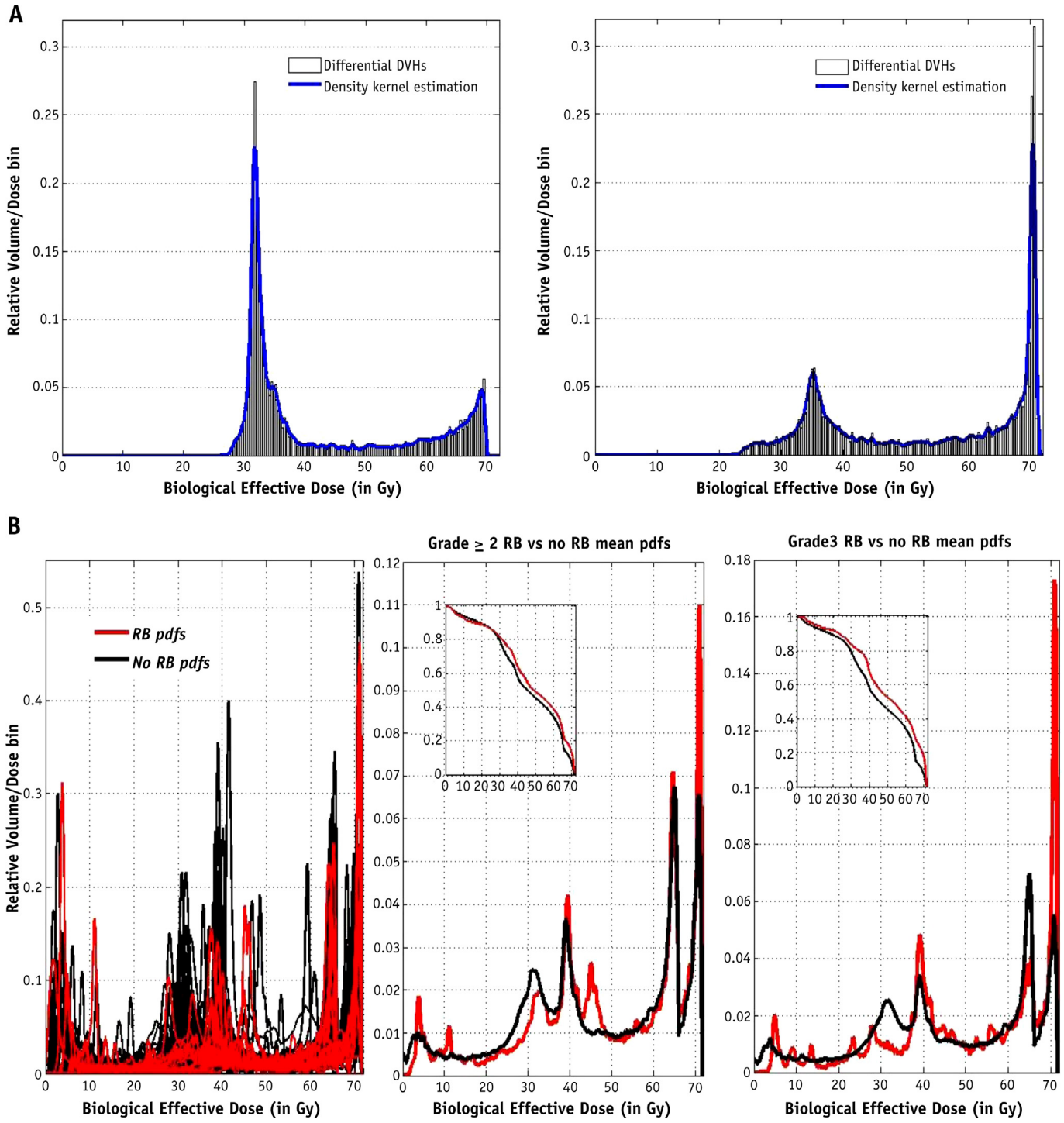

Fig. 2. (A) Examples of kernel smoothing density estimation. (B) Rectal bleeding (RB) (red) and no RB (black) rectum probability density functions (left) and their means (cumulative means also shown). DVH = dose-volume histogram. A color version of this figure is available at www.redjournal.org.

Patients who experienced grade ≥2 late RB had, on average, a more pronounced spike around the BED of their prescribed doses (66 Gy and 71.5 Gy, only around 71.5 Gy for grade 3), and their cumulative dose gradient was smaller in the range 20 to 40 Gy.

LKB model

The best LKB model fit for grade ≥2 and grade 3 RB resulted in the following model parameters (Table 2): TD50 = 72.6 Gy, m = 0.17, n = 0.12, and TD50 = 73.8 Gy, m= 0.16, n=0.24, respectively. The uncertainty in the fitted parameters was assessed by the profile likelihood method (20).

Logistic model

The results of the univariate analysis are summarized in Table 2. The best P values were obtained for the relative volumes exposed to the highest doses V_{35Gy} to V_{65Gy}. Patient age was the only significant clinical factor ($P=.01$).

Many of the DVH metrics are highly intercorrelated, but the correlation drops off sharply above 35 Gy (correlation matrix available in Appendix 3, available at www.redjournal.org). Therefore, V_{35Gy} to V_{55Gy} were not included in the multivariate analysis to prevent the problem of multicollinearity.

The addition of V_{60Gy} did not significantly improve the fit of the model based on V_{65Gy} (LR test $P=.59$).

PCA model

The first 3 eigenvectors resulting from PCA of the cDVHs are shown in Figure 3A. The first 5 PCs explain more than 96% of the rectum cumulative dose distribution (the first PC representing 61% and the second 16%). As shown in Table 2, the most parsimonious PCA model for grade ≥2 and grade 3 RB is based on PC2 and PC1, respectively.

FPCA model

In Figure 3B, the results of FPCA are illustrated for the first 2 functional principal components (FPC).

FPC1 is a "fraction-size" FPC because its scores according to their sign allows for a total discrimination between the patients who received 2 Gy/fraction and those treated with 2.5 Gy/fraction. FPC2 is a "4-field box treatment modality" FPC because it allows a picture of the reverse correlation between the rectum volume exposed to intermediate (2-field overlap region)/high doses (4-field overlap region).

With FPCA, 90% of the variability is explained by the first 16 FPCs, with 50% of the variability being explained by the first 2. We used the first 10 FPCs (80% of the explained variance) in the FLR.

The results of the univariate analysis in the FLR are summarized in Table 2. The best 2 P values were obtained by the FPC1 and FPC2, and the optimal FPCA models (according to the AIC and LRT) were based on FPC2 and FPC1 for grade ≥2 and grade 3 RB, respectively.

The parameter function β(δ) is displayed in Figure 4 for grade ≥2 and grade 3 RB. The parameter function leads to quantitatively estimate the volume effect's impact on the RB $OR^{Radiation}$: if its integral over a certain dose range is positive there will be interpreted as an increase in RB OR, whereas if it is negative there will be a decrease.

For example, let us compare the FLR ORs grade ≥2 RB of 2 patients, P1 and P2, using their *pdf*s $f_{P1}(left)$ *and* $f_{P2}(\text{right})$ represented in Figure 2A.

Then, according to equation 5, the ratio between the ORs is:

$$OR_{P1}^{Radiation} / OR_{P2}^{Radiation} = \exp\left(\int_0^{71.5} \beta_{Grade2}(\delta) \times (f_{P1}(\delta) - f_{P2}(\delta)) d\delta \right) = 0.57 \quad (6)$$

ie, the grade ≥2 RB OR is reduced by 43% in patient P1 compared with patient P2.

Model comparison

According to Table 2 and without including the clinical covariates, the AIC and AUC showed the FPCA model as the best among the tested models, particularly for grade 3 RB (AIC, 64.61 and AUC, 0.75). Except for LKB, the inclusion of age improved significantly the fit of RB grade ≥2 RB for all the models, and none of them provided a fit markedly better than another (AIC and AUC approximately 108 and 0.72, respectively).

Discussion

To our knowledge, this study represents the first application of the FDA to DVH analysis and NTCP modeling. The aim was to illustrate the feasibility and the advantage of the FDA by starting with a standard treatment plan such as 4-box 3D CRT planning before investigating more complex modalities.

We performed FPCA to explore how the *pdfs* vary between the patients and to see what are the most important modes of variation and how many of them seem to be substantially correlated with grade ≥2 RB.

The FPCA model can be seen as a nonparametric EUD NTCP model wherein the term $\int \beta(\delta) \times f_i(\delta) d\delta$ appears as a nonparametric weighted mean dose. The DVH reduction scheme is then performed by estimating the regression function parameter β(δ).

In opposition to the classic LKB EUD model, for example, which impose a power law with a volume exponent

Table 2 Univariate and multivariate analysis for various NTCP models

Univariate analysis					
Logistic model		PCA model		FPCA model	
Covariates	*P* value*	Covariates	*P* value*	Covariates	*P* value*
Mean dose	0.23 (.08)	PC1	.23 (.09)	FPC1	.11 (.017)
V_{10Gy}	>.5 (>.5)	PC2	.055 (.14)	FPC2	.065 (.036)
V_{15Gy}	>.5 (>.5)	PC3	.37 (.36)	FPC3	.78 (.32)
V_{20Gy}	>.5 (>.5)	PC4	>.5 (>.5)	FPC4	.35 (>.5)
V_{25Gy}	>.5 (>.5)			FPC5	>.5 (.2)
V_{30Gy}	>.5 (.40)			FPC6	.12 (.22)
V_{35Gy}	.13 (.09)			FPC7	.39 (>.5)
V_{40Gy}	.12 (.07)			FPC8	>.5 (>.5)
V_{45Gy}	.15 (.13)			FPC9	.14 (>.5)
V_{50Gy}	.23 (.10)			FPC10	>.5 (.34)
V_{55Gy}	.15 (.08)				
V_{60Gy}	.09 (.07)				
V_{65Gy}	.058 (.03)				

Multivariate analysis					
NTCP model	Parameters Nb†	Best model	AIC	*P* value LRT‡	AUC (95% CI)
Grade ≥2 RB					
Logistic Model	2	V_{65Gy}	111.95		0.62 (0.46-0.75)
	3	V_{65Gy} + V_{60Gy}	113.66	.59	
	3	V_{65Gy} + age	108.79	.023	0.70 (0.56-0.82)
PCA Model	2	PC2	111.83		0.62 (0.47-0.75)
	3	PC2 + PC1	112.29	.21	
	3	PC2 + age	108.20	.017	0.72 (0.59-0.82)
FPCA Model	2	FPC2	112.04		0.62 (0.48-0.74)
	3	FPC2 + FPC1	112.21	.17	
	3	FPC1 + age	107.87	.010	0.73 (0.59-0.84)
LKB Model	3	n=0.12 (0.03-0.36)§ m = 0.17 (0.11-0.27)§ TD50 = 72.6 Gy (67.2-82.3 Gy)§	113.01		0.63 (0.47-0.77)
Grade 3 RB					
Logistic model	2	V_{65Gy}	66.22		0.67 (0.40-0.88)
	3	V_{65Gy} + V_{60Gy}	68.17	.82	
	3	V_{65Gy} + age	67.06	.28	
PCA model	2	PC1	67.76		0.62 (0.39-0.83)
	3	PC1 + PC2	68.14	.21	
	3	PC1 + age	68.18	.21	
FPCA model	2	FPC1	64.67		0.75 (0.52-0.91)
	3	FPC1 + FPC2	64.71	.16	
	3	FPC1 + age	65.64	.31	
LKB model	3	n=0.24 (0.07-0.92)§ m = 0.16 (0.09-0.26)§ TD50 = 73.8 Gy (66.2-84.1 Gy)§	68.54		0.68 (0.43-0.87)

Abbreviations: AIC = Akaike's information criterion; AUC = area under the curve; CI = confidence interval; FPCA model = functional principal component analysis NTCP model; LKB = Lyman-Kutcher-Burman; NTCP = normal tissue complication probability; PCA model = principal component analysis NTCP model.

We propose a novel NTCP model wherein the weights dose values in the generalized equivalent uniform dose equation are flexibly estimated using a functional data analysis tools.

* *P* value of the univariate analysis grade ≥2RB (*P* value of the univariate analysis grade 3 RB between parentheses).

† Including the intercept for the models based on logistic regression.

‡ Likelihood ratio test *P* value (with respect to the simplest nested model, ie, 2 vs 3 parameters model).

§ 68% CI.

n, ie, $\left(\sum_{i=1}^{N} d_i^{\frac{1}{n}} \times v_i\right)^n \approx \left(\int \delta^{\frac{1}{n}} \times f_i(\delta) d\delta\right)^n$, the nonparametric approach has the advantage of avoiding major assumptions of a particular volume-effect relationship that may not reflect all situations encountered. Note also that compared with some other approaches, the FPCA NTCP model has the great computational advantage that no nonlinear optimization is involved to estimate the parameter function.

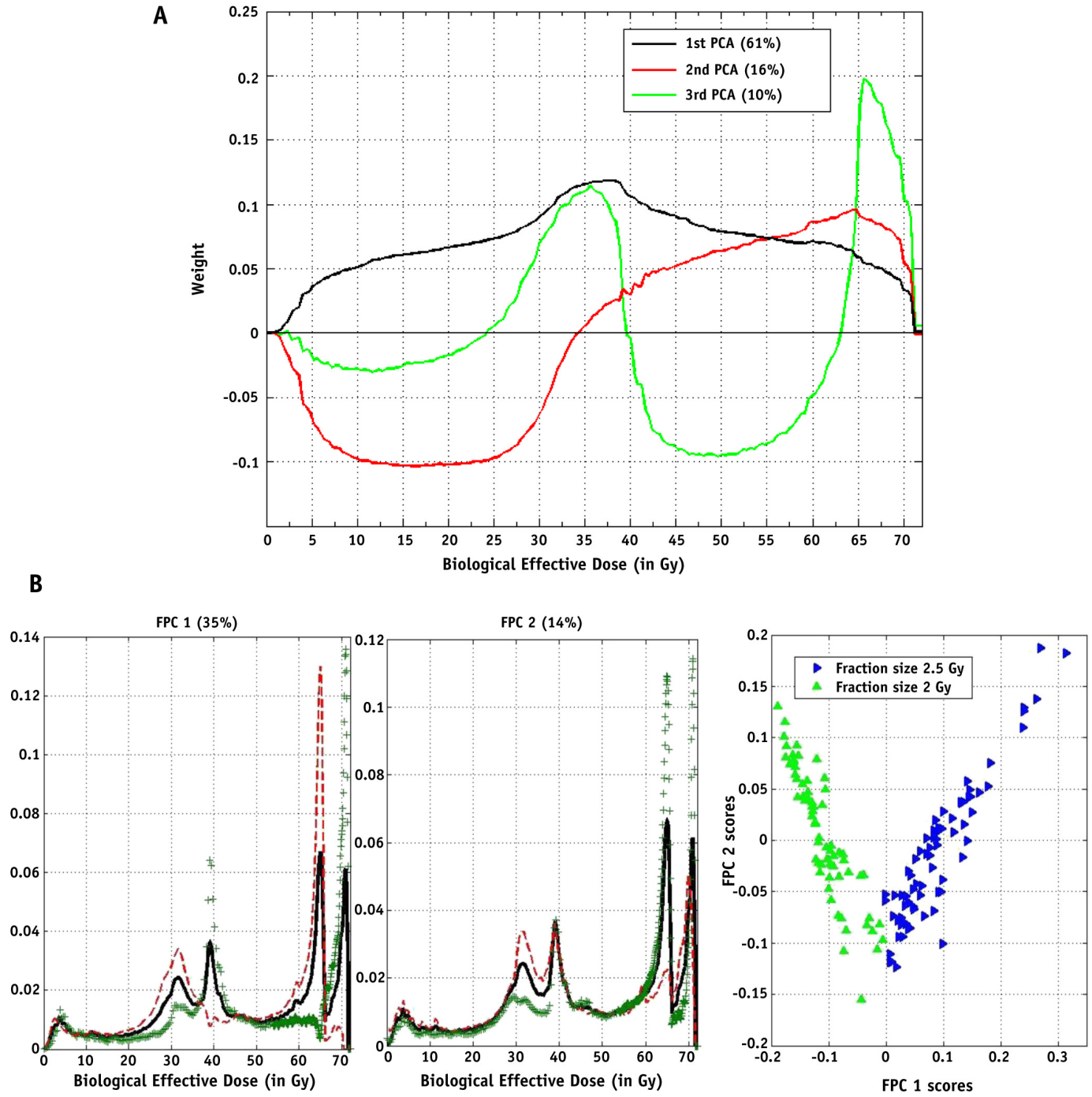

Fig. 3. (A) Eigenvectors resulting from principal component analysis (PCA) and percentage of explained variability. (B) Effect of the first 2 functional principal components (FPCs) of probability density functions (*pdfs*). In each graph, the solid curve is the overall mean, and the dashed curves are the mean ± a multiple (square root of the corresponding eigenvalue) of FPCs. The + and − signs show which curve is which. Percent variance described by eigenfunction is shown in brackets. Right, the scatterplot of FPC1 versus FPC2 scores.

Concerning the LKB model, the volume effect parameter was found to be 0.12 (68% CI, 0.03-0.36) for grade ≥2 RB, which was in agreement with the data published by QUANTEC (21): n = 0.09 (95% CI, 0.04-0.14) and emphasizes the role of high doses.

The specificity of the proposed model is to describe the dose-volume effect through the function $\beta(\delta)$ rather than a number. This provides richer information because it combines the radiobiological mechanistic message (role of the high doses) with the specificity of the 4-box 3D CRT by affecting a negative weight for the 2-field overlap region (25-40 Gy) and a positive one for the 4-field overlap region (around the BED of the prescribed doses). As mentioned by Michalski et al (21), "the volumes exposed to intermediate

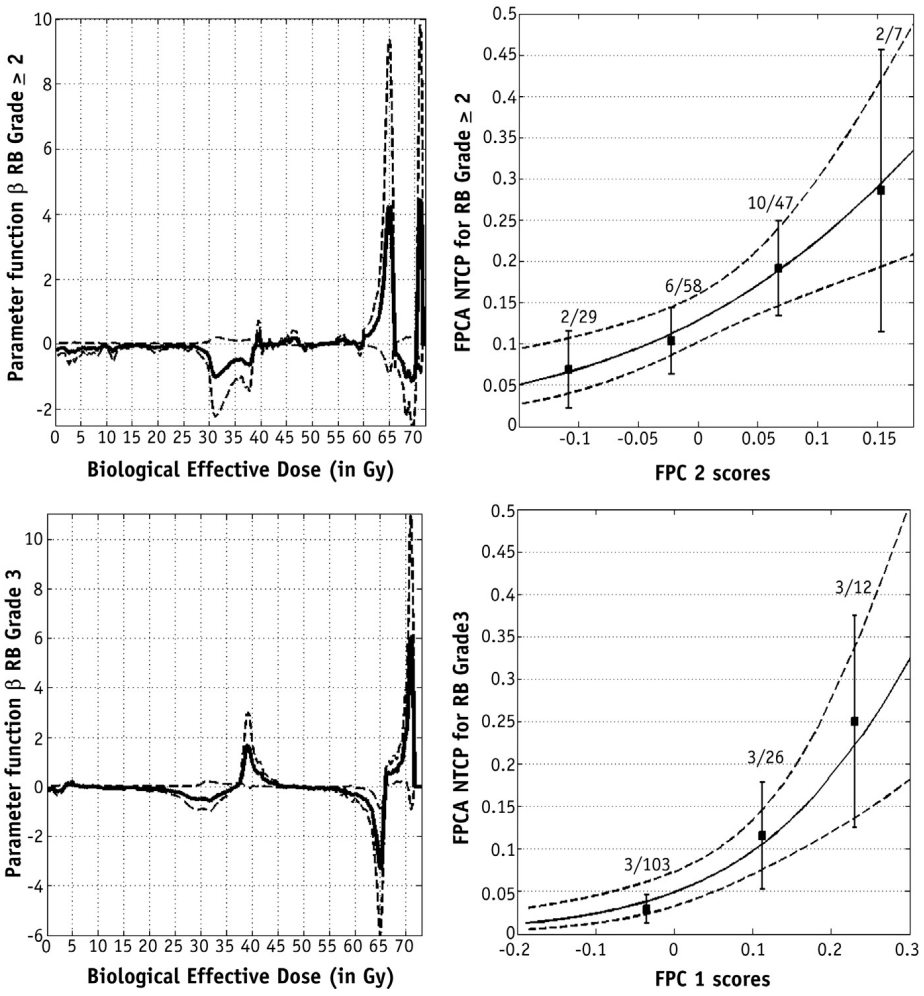

Fig. 4. Left, parameter function $\beta(\delta)$ for predicting grade $\geq$2 and grade 3 rectal bleeding (RB) after irradiation of prostate bed. The dashed lines indicate pointwise 95% confidence limits for values of $\beta(\delta)$ using 1000 bootstrap samples of the original data. Right, functional principal component analysis normal tissue complication probability curve as a function of functional principal component (FPC) scores for grade $\geq$2 and grade 3 RB; dashed lines limit the 68% confidence intervals. Solid squares represent observed complication rates with 68% error bars.

and high doses might both have biologic significance if, for example, the volumes exposed to intermediate doses play a role in the recovery of tissue exposed to the highest doses." It's exactly the information provided by $\beta(\delta)$ because reducing the rectal volume receiving intermediate doses (2-field overlap region) implies increasing the rectal volume exposed to high doses (4-field overlap region). This reverse variability is also depicted by the PC2 in the multivariate PCA, as pointed by Söhn et al (9). However, the cDVH PCA analysis was able to detect only the transition at 35 Gy without specifying the exact role of the upper and lower dose levels, which can limit interpretation of the PCA model results.

By pointing simultaneously to the role of high doses and fraction size 2.5 Gy, the proposed model outperformed statistically the other NTCP models for grade 3 RB. This could be explained by the unique aspect of FDA compared with multivariate analysis: in multivariate statistical analysis, the order of the components of observed random vectors is quite irrelevant, and any change in this order leads to the same result. By contrast, the FDA integrates the order as a supplementary dimension of the analysis and potentially better describes the correlation between the rectum volumes exposed to various doses.

The FPCA model also allows the addition of important nondosimetric prognostic variables, because even when optimal dose-volume constraints are applied, rectal complications can still occur as a result of these clinical factors. The results of our present study suggest that patients with advanced age are at risk of grade $\geq$ 2 rectal complications (OR, 1.123; 95% CI, 1.03-1.22). Abdominal surgery and hypercholesterolemia may be also associated, as shown in Table 1. These results are in line with previous reports (6, 22, 23).

The present analysis was made on 3D CRT plans but can be performed on intensity modulated radiation therapy plans in exactly the same way because it is based on dDVHs. It can also be applied to dose wall histograms if the rectum is considered as an empty organ. As has been previously proposed for PCA (24), a great potential of this

method is its extension to spatial functional data dose distribution, which would permit investigation of heterogeneous intraorgan radiosensitivity (25, 26).

References

1. Lyman J. Complication probability as assessed from dose-volume histograms. *Radiat Res Suppl* 1985;8:S13-S19.
2. Burman C, Kutcher G, Emami B, et al. Fitting of normal tissue tolerance data to an analytic function. *Int J Radiat Oncol Biol Phys* 1991;21:123-135.
3. Kutcher G, Burman C, Brewster L, et al. Histogram reduction method for calculating complication probabilities for three-dimensional treatment planning evaluations. *Int J Radiat Oncol Biol Phys* 1991;21:137-146.
4. Niemierko A. Reporting and analyzing dose distributions: A concept of equivalent uniform dose. *Med Phys* 1997;24:103-110.
5. Fiorino C, Fellin G, Rancati T, et al. Clinical and dosimetric predictors of late rectal syndrome after 3D-CRT for localized prostate cancer: Preliminary results of a multicenter prospective study. *Int J Radiat Oncol Biol Phys* 2008;70:1130-1137.
6. Peeters S, Hoogeman M, Heemsbergen W, et al. Rectal bleeding, fecal incontinence, and high stool frequency after conformal radiotherapy for prostate cancer: Normal tissue complication probability modeling. *Int J Radiat Oncol Biol Phys* 2006;66:11-19.
7. Defraene G, den Bergh LV, Al-Mamgani A, et al. The benefits of including clinical factors in rectal normal tissue complication probability modeling after radiotherapy for prostate cancer. *Int J Radiat Oncol Biol Phys* 2012;82:1233-1242.
8. Dawson L, Biersack M, Lockwood G. Use of principal component analysis to evaluate the partial organ tolerance of normal tissues to radiation. *Int J Radiat Oncol Biol Phys* 2005;620:829-837.
9. Söhn M, Alber M, Yan D. Principal component analysis-based pattern analysis of dose-volume histograms and influence on rectal toxicity. *Int J Radiat Oncol Biol Phys* 2007;69:230-239.
10. Vesprini D, Sia M, Lockwood G, et al. Role of principal component analysis in predicting toxicity in prostate cancer patients treated with hypofractionated intensity-modulated radiation therapy. *Int J Radiat Oncol Biol Phys* 2011;81:e415-e421.
11. Ramsay JO, Silverman BW. Functional Data Analysis. 2nd ed. New York: Springer; 2005.
12. Rosenblatt M. Remarks on some non-parametric estimates of a density function. *Ann Math Stat* 1956;27:832-837.
13. Parzen E. On estimation of a probability density function and mode. *Ann Math Stat* 1962;33:1065-1076.
14. Wand M, Jones M. Kernel Smoothing. New York: Chapman and Hall; 1995.
15. Thames H, Hendry J. Fractionation in radiotherapy. London: Taylor and Francis; 1987.
16. Fowler J. Brief summary of radiobiological principles in fractionated radiotherapy. *Semin Radiat Oncol* 1992;2:1621.
17. Ramsay JO. When the data are functions. *Psychometrika* 1982;47:379-396.
18. Kneip A, Utikal K. Inference for density families using functional principal components analysis. *J Am Stat Assoc* 2001;96:519-531.
19. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974;716-723. AC-19.
20. Pawitan Y. In: Likelihood All, editor. Statistical Modelling and Inference Using Likelihood. New York: Oxford University Press; 2001.
21. Michalski J, Gay H, Jackson A, et al. Radiation dose-volume effects in radiation-induced rectal injury. *Int J Radiat Oncol Biol Phys* 2010;76:S123-S129.
22. Skwarchuk M, Jackson A, Zelefsky M, et al. Late rectal toxicity after conformal radiotherapy of prostate cancer (I): Multivariate analysis and dose-response. *Int J Radiat Oncol Biol Phys* 2000;47:103-113.
23. Tomita N, Soga N, Ogura Y, et al. Preliminary analysis of risk factors for late rectal toxicity after helical tomotherapy for prostate cancer. *J Radiat Res* 2013;54:919-924.
24. Liang Y, Messer K, Rose BS, et al. Impact of bone marrow radiation dose on acute hematologic toxicity in cervical cancer: Principal component analysis on high dimensional data. *Int J Radiat Oncol Biol Phys* 2010;78:912-919.
25. Ramsay JO, Ramsay T, Sangalli LM. Recent Advances in Functional Data Analysis and Related Topics; chap. Spatial Functional Data Analysis. PhysicaVerlag HD Springer; 2011269-275.
26. Hörmann S, Kokoszka P. Consistency of the mean and the principal components of spatially distributed functional data. *Bernoulli* 2013;19:1535-1558.