Estimation du délai de guérison statistique chez les patients atteints de cancer

Gaëlle Romain

Thèse de doctorat

de l'établissement Université Bourgogne Franche-Comté

Ecole doctorale n°554

Ecole Doctorale Environnements-Santé

Doctorat de médecine, santé publique, environnement et société

Par Mlle ROMAIN Gaëlle

Estimation du délai de guérison statistique chez les patients atteints de cancer

Thèse présentée et soutenue à Dijon, le 10 décembre 2019

Composition du Jury :

Mr LEPAGE Côme, Professeur des Universités-Praticien Hospitalier.....Président
Unité INSERM UMR1231, Faculté de médecine de Dijon

Mr DELPIERRE Cyrille, Directeur de recherche INSERM.....Rapporteur
Unité INSERM UMR1027, Faculté de médecine de Toulouse

Mme LEGRAND Catherine, Professeure des Universités.....Rapporteur
Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain

Mme BOUVIER Anne-Marie, Directrice de recherche INSERM.....Examinatrice
Unité INSERM UMR1231, Faculté de médecine de Dijon

Mr COLONNA Marc, Directeur du Registre du Cancer de l'Isère.....Examinateur
Centre Hospitalier Universitaire de Grenoble

Mr FOUCHER Yohann, Maître de Conférences des Universités.....Examinateur
Unité INSERM UMR1246, Université de Nantes

Mme JOOSTE Valérie, Ingénieure de recherche, HDR.....Directrice de thèse
Unité INSERM UMR1231, Faculté de médecine de Dijon

Titre : Estimation du délai de guérison statistique chez les patients atteints de cancer.

Mots clés : survie, cancer, modèle de guérison, délai de guérison, données de registre, étude de simulations

Résumé : Trois millions de personnes vivent en France avec un antécédent personnel de cancer et ont des difficultés d'accès à l'emprunt et à l'assurance. Depuis 2016, la loi de « modernisation de notre système de santé » a fixé le « droit à l'oubli » (délai au-delà duquel les demandeurs d'assurance ayant eu un antécédent de cancer n'auront plus à le déclarer) à 10 ans après la fin des traitements. D'un point de vue statistique, on peut considérer ce délai comme le délai au-delà duquel la surmortalité liée au cancer (taux de mortalité en excès) s'annule durablement, ce qui se traduit sur les courbes de survie nette par un plateau correspondant à la proportion de patients guéris. La vérification de l'hypothèse de guérison repose sur deux critères : un taux de mortalité en excès négligeable et la confirmation graphique de l'existence d'un plateau. Une nouvelle définition du délai de guérison a été proposée pour ce travail comme le temps à partir duquel la probabilité d'appartenir au groupe des guéris atteint 95%.

Le premier objectif de cette thèse était de fournir des estimations du délai de guérison à partir des données des registres de cancer du réseau FRANCIM pour chaque localisation de cancer selon le sexe et l'âge. Le délai de guérison est inférieur à 12 ans pour la majorité des localisations vérifiant l'hypothèse de guérison. Il est notamment inférieur ou égal à 5 ans, voire nul pour certaines classes d'âge, pour le mélanome de la peau, le cancer du testicule et de la thyroïde. Les critères pour la vérification de la guérison sont subjectifs et le délai de guérison ne repose pas sur une estimation directe par les modèles de guérison préexistants. Un nouveau modèle de guérison a été développé, incluant le délai de guérison comme paramètre à estimer afin de répondre objectivement à la question de l'existence d'une guérison statistique et de permettre une estimation directe du délai de guérison.

Le second objectif de la thèse était de comparer, dans des situations contrôlées pour lesquelles le taux de mortalité en excès devenait nul, les performances de ce nouveau modèle à celles de deux autres modèles de guérison. La survie nette et la proportion de guéris estimées par les modèles ont été comparées aux valeurs théoriques utilisées pour simuler les données. Le nouveau modèle permet, avec des conditions strictes d'application, d'estimer directement le délai de guérison avec des performances aussi satisfaisantes que celles des autres modèles.

Title: Estimation of statistical time-to-cure in cancer patients.

Keywords: survival, cancer, cure model, time-to-cure, population-based data, simulation study

Abstract: Three million people are living in France with a personal past of cancer and undergo difficulties in accessing loans and insurance. Since 2016, the French law « modernisation de notre système de santé » set the "right to be forgotten" (time beyond which insurance applicants with a past of cancer will no longer have to declare it) at 10 years after the end of treatment. From a statistical point of view, this delay can be considered as the time from which mortality due to cancer (excess mortality) disappears. After this time, the net survival curves reach a plateau corresponding to the proportion of cured patients. The verification of this hypothesis is based on two criteria: a negligible excess mortality rate and a graphic confirmation of the existence of a plateau. We proposed a new definition of the time-to-cure as the time from which the probability of belonging to the cured group reaches 95%.

The first aim of this thesis was to estimate time-to-cure for each cancer site by sex and age using population-based data from the FRANCIM registries network. Time-to-cure was lower than 12 years in most sites complying with the cure hypothesis. It was less than 5 years, or even null in some age groups, for skin melanoma, testicular and thyroid. Criteria to verify the cure hypothesis are subjective and time-to-cure is not directly estimated in pre-existing cure models. A new model has been developed including time-to-cure as a parameter to address the question of statistical cure and to allow direct estimation of time-to-cure.

The second objective of this thesis was to compare, in controlled situations in which the excess mortality rate became null, the performances of this new model with that of two other cure models. Estimated net survival and cure fraction have been compared to the theoretical values used to simulate the data. Direct estimation of time-to-cure is possible under strict conditions.

Remerciements

Parce qu'une thèse n'est pas un travail individuel, je tiens à remercier toutes les personnes qui ont participé de près ou de loin à l'aboutissement de cette thèse.

Je remercie Madame Anne-Marie Bouvier, Directrice du Registre Bourguignon des Cancers Digestifs d'avoir accepté que je me lance dans l'aventure d'un doctorat en plus de mon activité au registre. Vos conseils ainsi que vos relectures de mes différents articles et de cette thèse furent d'une aide précieuse.

Je remercie Madame Valérie Jooste, qui a soutenu son HDR pour devenir officiellement ma directrice de thèse. Je te remercie d'avoir accepté que cette thèse s'inclue dans tes différents projets sur la guérison statistique : un domaine aussi passionnant que complexe. Je te remercie également pour ton soutien et tes conseils avisés à chaque instant de cette thèse, ainsi que pour ta positivité dans chaque moment de doute.

Je remercie Madame Catherine Legrand et Monsieur Cyrille Delpierre d'avoir accepté d'être les rapporteurs de cette thèse. Je vous remercie pour l'intérêt que vous portez à ce travail et pour vos commentaires qui m'ont permis d'améliorer ce manuscrit.

Je remercie également les membres du jury Monsieur Côme Lepage, Monsieur Yohann Foucher et Monsieur Marc Colonna d'avoir accepté de juger ce travail. Je remercie particulièrement Monsieur Yohann Foucher pour ses cours de master sur la survie et d'avoir accepté de me suivre tout au long de cette thèse en faisant partie de mon comité de thèse. Nos discussions et tes remarques furent d'une grande aide pour l'avancée de cette thèse. Je remercie également Monsieur Marc Colonna avec qui ce fût un réel un plaisir de collaborer et d'apprendre sur les différents projets.

Je remercie Olayidé Boussari, pour ses idées de génie. Je te remercie pour ce nouveau modèle sur lequel j'ai pu me tirer les cheveux, ainsi que pour ta patience et ta pédagogie quand il s'agissait de me faire des cours particuliers « d'étude de simulations ».

Je remercie l'ensemble des membres du registre. Je remercie Stéphanie Normand, collègue de bureau en première ligne pour absorber toutes les joies et les doutes qu'a pu apporter cette thèse. Je te remercie pour ta patience et ton aide plus qu'utile lorsque j'étais en perte d'inspiration. Je remercie Ludivine Garrier et Alexandra Morey, de me faire partager vos connaissances sur les cancers digestifs. Je vous remercie aussi toutes les trois pour votre bonne humeur et pour votre soutien dans les hauts comme dans les bas. Je remercie Gregory Viénot d'avoir toujours tenté de me trouver des solutions informatiques pour accélérer les temps de calculs lors des simulations de données ou pour copier mes énormes fichiers. Je remercie l'ensemble des « filles du CIC », pour votre bonne humeur pendant les repas du midi, c'est important de pouvoir rire et décompresser dans les moments les plus intenses.

Enfin je remercie également ma famille et mes ami(e)s qui ont su être compréhensifs lorsque j'avais peu de temps à leur accorder et qui ont su malgré tout être présents tout au long de cette thèse.

Table des matières

Liste des tableaux

Table des figures

Valorisation scientifique

Publications scientifiques et communications liées à la thèse

Publications

❖ **Romain G**, Boussari O, Bossard N, Remontet L, Bouvier AM, Mounier M, Iwaz J, Colonna M, Jooste V; French Network of Cancer Registries (FRANCIM). *Time-to-cure and cure proportion in solid cancers in France. A population based study*. Cancer Epidemiol. 2019 Jun;60:93-101. doi: 10.1016/j.canep.2019.02.006. Epub 2019 Mar 30. PubMed PMID: 30933890.

❖ Boussari O, **Romain G**, Remontet L, Bossard N, Mounier M, Bouvier AM, Binquet C, Colonna M, Jooste V. *A new approach to estimate time-to-cure from cancer registries data*. Cancer Epidemiol. 2018 Apr;53:72-80. doi: 10.1016/j.canep.2018.01.013. Epub 2018 Feb 4. PubMed PMID: 29414635.

Article en cours de révision

❖ Boussari O, Bordes L, **Romain G**, Colonna M, Bossard N, Remontet L, Jooste V, *Modeling excess hazard with time-to-cure as a parameter*, Biometrics (Soumis, version 2 en relecture)

Article soumis au numéro spécial « ISCB » de Biometrical Journal

❖ **Romain G**, Boussari O, Colonna M, Jooste V, *Comparing performances of three cure models including a new model with time-to-cure as a parameter*

Communications orales

❖ **Gaëlle Romain**, Olayidé Boussari, Laurent Remontet, Nadine Bossard, , Morgane Mounier, Alice Gagnaire, Marc Colonna, Valérie Jooste, *Existence of cure, estimation of time to cure and cure fraction. A FRANCIM population based study on 27 cancer sites,* Workshop Groupe des registres de langue latine (GRELL), 2017, Bruxelles (Belgique)

❖ **Gaëlle Romain**, Olayidé Boussari, Marc Colonna, Valérie Jooste, *Comparing performances of three cure models including a new model with time-to-cure as a parameter*, 40th Annual Conference of the International Society for Clinical Biostatistics (ISCB40), 2019, Louvain (Belgique)

❖ **Gaëlle Romain**, Trois définitions du délai de guérison statistique, à partir d'un modèle de guérison en survie nette, Séminaire de méthodologie, 2017, Dijon (France)

❖ **Gaëlle Romain**, Olayidé Boussari, Nadine Bossard, Morgane Mounier, Alice Gagnaire, Marc Colonna, Valérie Jooste, *Estimation du délai de guérison et de la proportion de guéris à partir des données FRANCIM*, Journée d'échange - modèle de guérison, 2017, Institut National du Cancer (INCa), Paris (France)

Contribution à d'autres publications scientifiques dans le domaine

❖ **Romain G**, Mariet AS, Jooste V, Duloquin G, Thomas Q, Durier J, Giroud M, Quantin C, Béjot Y. *Long-term relative survival in stroke patients: the Dijon Stroke Registry*. Neuroepidemiology. 2019 (Accepted). doi: 10.1159/000505160

❖ Drouillard A, Bouvier AM, Boussari O, **Romain G**, Manfredi S, Lepage C, Faivre J, Jooste V. *Net survival in recurrence-free colon cancer patient*s. Cancer Epidemiol. 2019 Aug;61:124-128. doi: 10.1016/j.canep.2019.06.001. Epub 2019 Jun 15. PubMed PMID: 31212224.

❖ Willem H, Jooste V, Boussari O, **Romain G**, Bouvier AM. Impact of absence of consensual cutoff time distinguishing between synchronous and metachronous metastases: illustration with colorectal cancer. Eur J Cancer Prev. 2019 May;28(3):167-172. doi: 10.1097/CEJ.0000000000000450. PubMed PMID: 29738323.

❖ Lopez A, Bouvier AM, Jooste V, Cottet V, **Romain G**, Faivre J, Manfredi S, Lepage C. *Outcomes following polypectomy for malignant colorectal polyps are similar to those following surgery in the general population*. Gut. 2019 Jan;68(1):111-117. doi: 10.1136/gutjnl-2016-312093. Epub 2017 Oct 26. PubMed PMID: 29074726.

❖ Colonna M, Boussari O, Cowppli-Bony A, Delafosse P, **Romain G**, Grosclaude P, Jooste V; French Network of Cancer Registries (FRANCIM). *Time trends and short term projections of cancer prevalence in France*. Cancer Epidemiol. 2018 Oct;56:97-105. doi: 10.1016/j.canep.2018.08.001. Epub 2018 Aug 17. Erratum in: Cancer Epidemiol. 2018 Dec;57:158-159. PubMed PMID: 30125884.

Abréviations

ALD	Affections de Longue Durée
AREAS	s'Assurer et Emprunter avec un Risque Aggravé de Santé
CER	Comité d'Evaluation des Registres
CIF	Cumulative Incidence Function (en français, Founctio d'incidence cumulée)
CIRC	Centre International de la Recherche sur le Cancer
CISS	Collectif Inter-associatif Sur la Santé
CNR	Comité National des Registres
CONCORD	Global surveillance of cancer survival
EUROCARE	EUROpean CAncer REgistry
FRANCIM	France-Cancer-Incidence et Mortalité
HCL	Hospices Civils de Lyon
IACR	Association Internationale des Registres du Cancer
INCa	Institut National du Cancer
PMSI	Programme de Médicalisation des Systèmes d'Information
RCS	splines cubiques restreintes
RCS	Restricted Cubic Spline (en français, spline cubique restreinte)
RNIPP	Répertoire National d'Identification des Personnes Physiques
SMR	Standardised Mortality Ratio (en français, ratio standardisé de mortalité)

I Introduction générale

Il existe près de 200 types de cancers différents. Les cancers sont caractérisés par la prolifération anormale de cellules dans l'organisme. Les cellules cancéreuses envahissent les organes qui les entourent, constituant la tumeur primitive, et peuvent également se propager dans d'autres organes, créant des métastases. Les cancers sont l'une des premières causes de mortalité dans le monde et en France. Le Centre International de la Recherche sur le Cancer (CIRC) estime que 18,1 millions de nouveaux cas de cancer ont été diagnostiqués dans le monde en 2018 et 9,6 millions de personnes sont décédées d'un cancer. Le continent européen concentre à lui seul 23,4 % des cas de cancer[1]. Pour la même année en France, 382 000 personnes ont été diagnostiquées et 157 400 sont décédées de leur cancer. En France, les cancers les plus fréquents sont le cancer de la prostate chez l'homme et du sein chez la femme, devant les cancers du poumon et colorectal [2].

Les deux principaux indicateurs qui permettent l'évaluation de l'impact des pratiques médicales au niveau d'une population sont l'incidence et la survie. L'incidence est le rapport entre le nombre de nouveaux cas d'une maladie et la population générale, pour une période donnée et une population déterminée. En cancérologie, l'étude de la survie est l'étude du temps entre le diagnostic du cancer et le décès. La survie peut se décliner sous deux formes : la survie observée ou la survie nette. La survie observée exprime, en tenant compte de la censure, la proportion de patients ayant survécu 1, 2, 5, 10 ans… après le diagnostic de leur maladie. Cependant pour mesurer le réel impact du cancer sur la survie d'une population il est nécessaire de pouvoir s'affranchir des décès qui ne sont pas liés à la maladie. La survie nette est définie comme la survie que l'on observerait dans un monde hypothétique où la seule cause possible de décès serait la maladie[3]. N'étant pas influencée par les différences de mortalité due aux autres causes de décès, la survie nette est un indicateur permettant de comparer le pronostic des patients entre différentes périodes ou différents pays.

Pour de nombreuses localisations cancéreuses, le risque de décéder du cancer diminue au cours du temps écoulé après le diagnostic jusqu'à parfois devenir proche de zéro. Bien qu'il soit quasiment impossible de vérifier au niveau d'une cohorte populationnelle que des individus atteints de cancer sont cliniquement guéris de leur maladie, il est raisonnable d'envisager que certains ne décèderont jamais de leur cancer. Autrement-dit, on peut définir

la guérison statistique dans une population de patients comme l'absence de risque de décéder du cancer.

Plus de trois millions de personnes vivent en France avec un antécédent personnel de cancer. Les progrès médicaux ont permis une augmentation de la durée de vie de ces personnes. Une attention particulière a été plus récemment portée à l'impact du cancer sur la vie personnelle. Pour cela, l'une des actions des politiques de santé a été d'instaurer en 2015 le « droit à l'oubli » pour faciliter l'accès à l'emprunt et à l'assurance[4]. Le « droit à l'oubli » est le délai au-delà duquel les personnes ayant une histoire personnelle de cancer n'ont plus à le déclarer aux banques et aux assurances. Ce délai est actuellement fixé à 10 ans à partir de la fin des traitements (Code de santé publique art. L. 1141-5, al. 4).

D'un point de vue statistique, on peut considérer le délai du « droit à l'oubli » comme le délai au-delà duquel la population de patients retrouve la même probabilité de décéder que la population générale : le délai de guérison. La proportion de patients guéris et le délai pour atteindre la guérison, après un diagnostic de cancer, sont des indicateurs importants pour les patients, les cliniciens et les acteurs de santé publique.

II Contexte épidémiologique

II.1 Registre des cancers

II.1.1 Définition

Un Registre de population est une structure qui réalise « un recueil continu et exhaustif de données nominatives intéressant un ou plusieurs événements de santé dans une population géographiquement définie, à des fins de recherche et de santé publique, par une équipe ayant les compétences appropriées » (arrêté du 6 novembre 1995 relatif au Comité National des Registres). Au-delà de leur apport unique pour la veille sanitaire, les registres représentent l'outil idéal pour évaluer les pratiques médicales et l'efficacité globale du système de soin, notamment à travers des études de survie[i] en population générale. Les registres de population permettent le recueil de données exhaustives provenant de toutes les filières de soins sur des zones géographiques bien définies et avec des procédures de recueil et de codification standardisées. Ils sont des outils de recherche épidémiologique et en Santé Publique[5]. Les séries hospitalières (ou registres hospitaliers) et, *a fortiori*, les essais thérapeutiques recueillent seulement les informations sur les cas suivis dans un hôpital ou pour un essai. Ils présentent des biais de sélection et de recrutement non contrôlables qui empêchent la généralisation de leurs résultats à la population. Des différences très importantes apparaissent entre les résultats statistiques issus des essais randomisés ou des registres hospitaliers et les résultats issus de registres de population.

II.1.2 Historique

Avant la création des registres de cancers, les seules informations sur les cas de cancer provenaient des dossiers médicaux et des rapports d'autopsie, et à l'échelle de la population, des données relatives aux causes de décès. Lorsque l'issue de la maladie était très souvent

[i] Survie : Etude de la probabilité de décéder au cours du temps

fatale, les taux d'incidence[ii] pouvaient être approchés par les taux de mortalité[iii]. Avec les premiers succès thérapeutiques, les taux de mortalité ne reflétaient plus les taux d'incidence. Il alors est devenu impératif de consigner chaque nouveau cas de cancer diagnostiqué dans une zone géographique donnée afin d'avoir une représentation juste de la maladie dans une population.

En France, les premiers registres de cancers ont été créés à partir des années 1975 sur des initiatives individuelles. Les registres de cancers français ont été inscrits dans une politique nationale de santé publique grâce à la création, en 1986, du Comité National des Registres (CNR, devenu depuis 2014 le Comité d'Evaluation des Registres, CER).

II.1.3 Les registres des cancers dans le monde

Le Centre International de Recherche sur le Cancer (CIRC) et l'Association Internationale des Registres du Cancer (IACR) ont pour objectifs de promouvoir et de standardiser les méthodes de recueil et de codage de données de l'ensemble des registres validés à travers le monde. Les estimations de l'incidence des cancers dans le monde entier sont publiées tous les 5 ans au sein des monographies « Cancer Incidence in Five Continents » (CI5, http://ci5.iarc.fr/Default.aspx). Le dernier volume regroupe les données des cas diagnostiqués entre 2008 et 2012 à partir de 343 registres issus de 65 pays[6].

Ces bases de données et la mise en place de réseaux actifs de collaboration ont permis de concevoir de vastes projets permettant de comparer la prise en charge des cancers entre les pays. Au niveau européen, le projet EUROCARE (EUROpean CAncer REgistry), par exemple, existe depuis 1989 sur une initiative Italienne. L'objectif initial de ce projet était de comparer la survie des patients en Europe. Il est maintenant élargi à comparaison de la prévalence[iv] et des schémas thérapeutiques en Europe. La base de données EUROCARE inclue plus de 20 millions de cas de cancers provenant de 116 registres dans 30 pays européens. En 2008, le programme CONCORD s'est mis en place avec la volonté d'étendre les comparaisons de survie au niveau international. Actuellement, le projet CONCORD-3 rassemble les données de 322 registres de cancers issues de 71 pays. L'objectif de ce projet est

ii Incidence : Rapport entre le nombre de nouveaux cas d'une maladie et la population générale, pour période donnée et une population déterminée.
iii Taux de mortalité : Rapport entre le nombre de décès et la population générale, pour période donnée et une population déterminée.
iv Prévalence : Nombre de cas d'une maladie dans une population à un moment donné, englobant aussi bien les cas nouveaux que les cas anciens.

de décrire et d'expliquer les inégalités de survie entre pays et au regard des caractéristiques socio-économiques et ethniques des populations.

II.1.4 En France, réseau FRANCIM

L'ensemble des registres des cancers, qualifiés par le CNR puis le CER, sont regroupés en association au sein du réseau français des registres de cancer, FRANce-Cancer-Incidence et Mortalité (FRANCIM) (Figure II.1). Les registres de cancers peuvent être généraux ou spécialisés :

- Les registres de cancers généraux recueillent les informations sur toutes les localisations de cancers
- Les registres de cancers spécialisés recueillent les informations sur des localisations particulières (par exemple : appareil digestif, hémopathies malignes, sein, thyroïde, …) ou sur des populations particulières (enfants)

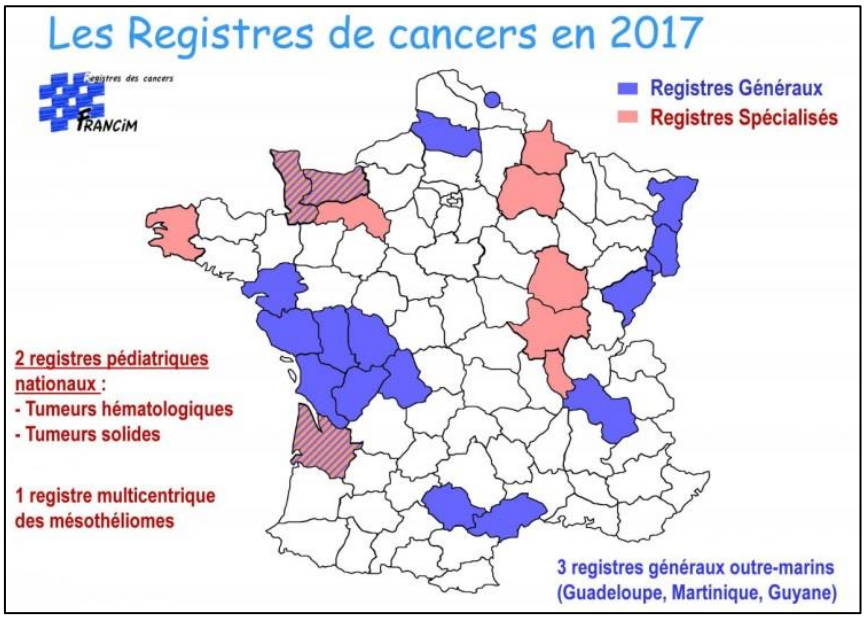


Figure II.1 - Carte des départements couverts par un registre des cancers général ou spécialisé.

L'objectif du réseau FRANCIM est « d'harmoniser les pratiques d'enregistrement, de coordonner et de faciliter les travaux réalisés par les registres de cancer existants, de fournir à la communauté les indicateurs épidémiologiques utiles à la connaissance et à la prise en charge des cancers en lien avec différents partenaires institutionnels ».

Afin de mener une politique de santé publique commune dans le cadre de l'épidémiologie descriptive du cancer, une base commune des registres de cancers a été

constituée. Elle regroupe l'ensemble des données recueillies par les registres depuis leur création. Vingt-huit registres de cancers contribuent à cette base, qui regroupe plus d'1,2 millions de tumeurs enregistrées depuis 1975. Elle permet d'étudier l'évolution de l'incidence et de la survie au cours du temps en France, et d'estimer la prévalence et les projections d'incidence nationale. La base commune est hébergée dans le service de biostatistiques des Hospices Civils de Lyon (HCL). En 2008, FRANCIM, le service de biostatistiques des HCL, l'Institut de Veille Sanitaire (aujourd'hui Santé Publique France) et l'Institut National du Cancer (INCa) ont établi un « programme scientifique de travail partenarial » quadriennal régulièrement reconduit.

II.1.5 Sources et principales données recueillies

Le recours à plusieurs sources d'information est nécessaire pour recenser tous les cas de cancer. Les principales sources sont les laboratoires d'anatomopathologie et les bases administratives telles que les PMSI des établissements de soins (centres hospitaliers universitaires ou périphériques, centres de lutte contre le cancer, cliniques privées) ; ou les listes des Affections Longue Durée (ALD).

Les données communes aux registres généraux et aux registres spécialisés indispensables pour l'enregistrement des cas sont : le sexe, la date de naissance, le lieu de résidence, la date de diagnostic (également appelée date d'incidence), la morphologie et la topographie du cancer, le statut vital et la date de dernière nouvelle du patient (correspondant à la date de décès si le patient est décédé). Le recueil du statut vital se fait de façon active en interrogeant de différentes sources d'information et suit une procédure standardisée. La principale source permettant de mettre à jour le statut vital est le Répertoire National d'Identification des Personnes Physiques (RNIPP).

Les registres sont évalués tous les 4 ans par le CER sur la qualité et l'exhaustivité du recueil des cas et les travaux de recherche entrepris.

II.2 Droit à l'oubli

Les personnes vivants avec un antécédent personnel de cancer font face à différentes difficultés, dont sociales. Du fait de leur antécédent de cancer elles présentent ce que les assurances considèrent comme « un risque aggravé de santé » et, jusqu'à récemment, pouvaient : 1) se voir refuser tout contrat d'assurance, 2) devoir payer des surprimes ou 3) supporter des assurances avec exclusions de garantie. En 2012, ces restrictions concernaient respectivement 32%, 28% et 35% des personnes en arrêt longue maladie (d'après le Baromètre des droits des malades 2012, CISS).

Depuis 2003, la lutte contre le cancer s'est structurée en France, avec la mise en place des différents Plans Cancer nationaux. L'une des ambitions du troisième Plan Cancer (2014-2019) est de limiter les conséquences sociales et économiques de la maladie, notamment en améliorant l'accès aux assurances et au crédit. Pour cela, le « droit à l'oubli » a été instauré comme le « délai au-delà duquel les demandeurs d'assurance ayant eu un antécédent de cancer n'auront plus à le déclarer » (Plan Cancer 2014-2019, action 9.13). L'Etat, les professionnels de l'assurance et les associations représentant les personnes malades avaient signé en 2006 la convention AREAS (s'Assurer et Emprunter avec un Risque Aggravé de Santé, http://www.aeras-infos.fr), afin d'améliorer l'assurabilité des personnes présentant un risque de santé aggravé. Le troisième Plan cancer a permis d'introduire, en 2015, le « droit à l'oubli » dans la Convention AREAS : les informations médicales relatives aux pathologies cancéreuses ne pouvaient désormais plus être recueillies par les organismes bancaires et les assureurs après un délai de 15 ans à partir de la fin du protocole thérapeutique (amendement du 2 septembre 2015). Dans le cadre de la loi de « modernisation de notre système de santé » du 26 janvier 2016, l'Etat a abaissé le délai du « droit à l'oubli » à 10 ans[v], quels que soient la localisation et le type de cancer (Code de santé publique art. L. 1141-5, al. 4). Dans le même temps, la convention AERAS a mis en place « une Grille de Référence » (https://www.previssima.fr/files/previssima/documents_pdf/autres/grille-aeras-de-reference-16juillet-2018.pdf) permettant des aménagements du « droit à l'oubli » pour certains cancers ayant un type histologie et un stade pré-thérapeutique bien défini. Par exemple, le délai du droit à l'oubli est de 1 an pour les mélanomes de la peau *in situ* pur sans

[v] Pour les personnes ayant eu un diagnostic de cancer avant 18 ans le délai du droit à l'oubli est de 5 ans.

caractère micro-infiltrant ou de niveau I de Clark, avec exérèse complète et absence de syndrome des nævi dysplasiques.

D'un point de vue statistique, on peut considérer ce délai comme le délai au-delà duquel la surmortalité liée au cancer s'annule durablement, autrement dit, que la population de patients retrouve la même probabilité de décéder que la population générale : c'est le délai de guérison. Cette définition statistique de la guérison s'applique au niveau populationnel et n'implique pas que tous les patients soient cliniquement guéris. Connaitre le délai de guérison pour chaque localisation de cancers permettrait une utilisation davantage personnalisée et donc de faire évoluer le droit à l'oubli selon les localisations de cancers. Cette thèse s'inclue dans une étude FRANCIM, codirigée par Valérie Jooste et impliquant trois autres équipes : le Registre Général de l'Isère (dont le directeur Marc Colonna codirige l'étude), le Registre des Hémopathies Malignes de Côte d'Or et le Service de Biostatistiques des HCL (financement INCa N° 2014-087).

II.3 Objectifs de la thèse

L'objectif principal de la thèse était d'estimer le délai de guérison et la proportion de guéris pour les localisations qui le permettent parmi les 30 principales localisations de cancer, en utilisant les modèles statistiques appropriés et la base de données nationale commune des registres FRANCIM. L'estimation du délai de guérison a été faite à partir d'une nouvelle méthode que nous avons proposée et comparée à deux autres existantes.

Le second objectif était de comparer les performances des modèles de guérison existant à un nouveau modèle développé par l'équipe. En effet, les méthodes statistiques existantes pour étudier la guérison ayant montré certaines limites, un nouveau modèle de guérison a été développé au sein de l'équipe par Olayidé Boussari en collaboration avec Laurent Bordes (LMA, UPPA). Nous avons comparé les performances de ces modèles sur des données simulées et nous les avons appliqués à des données réelles.

III Analyse de survie

L'analyse de survie consiste à étudier le temps d'apparition d'un évènement. La variable étudiée est la variable aléatoire appelée « durée de vie » ou « temps de survie » car généralement le temps étudié en médecine est celui jusqu'au décès. Dans la réalité le décès des patients n'est pas toujours observé avant la fin de l'étude. Les temps de survie sont dits censurés. La censure à droite est le type de censure le plus couramment rencontré dans les études de survie en population générale. Dans ce cas-là, il n'est pas possible d'étudier les temps de survie comme une variable quantitative et des méthodes d'analyses statistiques adaptées à la censure doivent être utilisées.

III.1 Données censurées

L'analyse des temps d'apparition jusqu'au décès, nécessite de connaître différentes informations :

- La date d'origine (t_0) : date à partir de laquelle le patient est observé. Dans le cadre de cette thèse la date d'origine correspond à la date du diagnostic du cancer ;
- La date de dernière nouvelle : date la plus récente pour laquelle le statut vital du patient est connu. Lorsque le patient est décédé cette date correspond à la date du décès ;
- La date de point : date jusqu'à laquelle l'information sur le statut vital du patient est recueillie. Au-delà de cette date, l'information n'est pas prise en compte. La date de point est commune à tous les individus inclus dans l'étude.

Le temps écoulé entre le diagnostic du cancer (t_0) et le décès du patient i, appelé temps de survie T_i, est une variable aléatoire positive. Le temps de censure du patient i, C_i, est une autre variable aléatoire positive ou nulle. Le temps de participation à l'étude X_i (ou temps de suivi) du patient i est donc le minimum entre le temps de censure et le temps de survie, soit : $X_i = min\{C_i, T_i\}$ avec l'indicatrice d'évènement δ_i égale à 1 si le patient est décédé et 0 s'il est

censuré. Le temps de censure correspond à différents temps selon le mécanisme de censure à droite (Figure III.1) :

- Perdu de vue : le patient n'est pas suivi jusqu'à la fin de l'étude et est vivant à la date de dernière nouvelle. C_i est le temps écoulé entre le diagnostic et la date de dernière nouvelle du patient. Pour ce patient i, $C_i < T_i$, donc $X_i = C_i$ et $\delta_i = 0$;
- Censure administrative : le patient est vivant à la date de point et son temps de suivi est inférieur au temps de suivi maximal t_{max}. C_i est le temps écoulé entre le diagnostic et la date de point. Pour ce patient i, $C_i < T_i$, donc $X_i = C_i$ et $\delta_i = 0$;
- Censure au temps maximal de suivi prévu par l'étude, t_{max} : le patient est suivi jusqu'à t_{max} et est vivant à la fin du suivi. C_i est le temps écoulé entre le diagnostic et t_{max}. Pour ce patient i, $C_i < T_i$, donc $X_i = C_i$ et $\delta_i = 0$.

Si le patient i est décédé avant la fin de l'étude, son temps de survie n'est pas censuré : $C_i \geq T_i$, donc $X_i = T_i$ et $\delta_i = 1$.

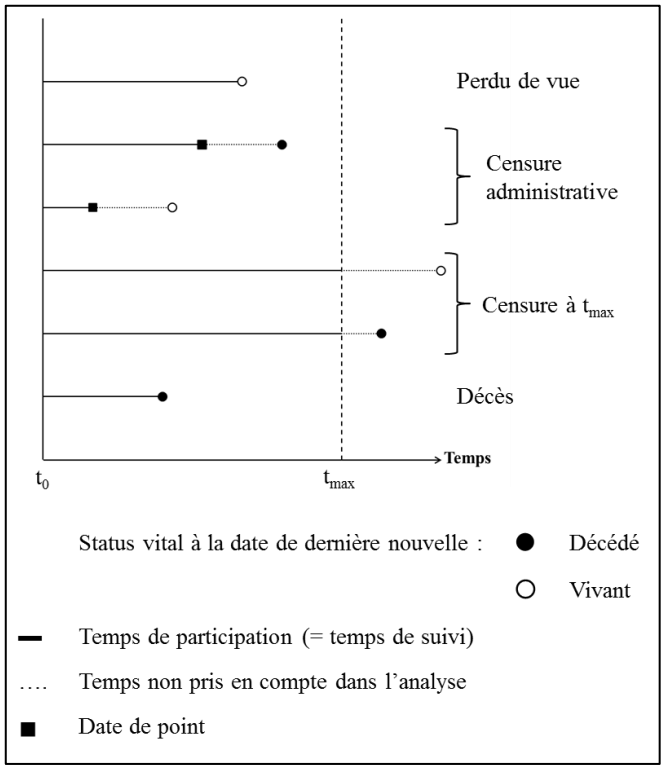


Figure III.1 - Représentation des différentes possibilités de suivi au cours d'une étude de survie.

Les analyses de survie supposent que le temps de censure et le temps de survie sont indépendants. En d'autres termes, les individus censurés ont les mêmes caractéristiques que les individus décédés. Dans le cas où la cause de la censure et l'évènement étudié sont dépendants, un biais de sélection peut apparaitre impliquant une surestimation de la

survie[7]. L'hypothèse d'indépendance entre les patients décédés et ceux censurés administrativement ou à t_{max} est naturelle car la censure est déterminée par une date de point ou un temps de suivi maximal définie pour l'analyse, indépendamment des caractéristiques des individus ; en revanche, cela n'est pas le cas pour les patients perdus de vue. Il est important d'avoir un suivi actif et rigoureux des patients afin de limiter le nombre de perdus de vue et de s'assurer que la censure soit indépendante, autrement dit que T_i et C_i sont indépendants. Les registres du réseau FRANCIM suivent une procédure standardisée faisant appel au RNIPP pour la recherche du statut vital des patients.

L'hypothèse d'indépendance peut également être remise en cause en présence de risques concurrents. Un risque concurrent est un évènement qui empêche la survenue de l'évènement étudié. Par exemple, si l'évènement étudié est le décès par cancer, le décès « autre cause » peut empêcher la survenue du décès dû au cancer. Si ces deux temps de survie sont affectés par une même covariable, le fait de censurer les décès « autres causes » induit une censure dite « informative ». En effet le risque de décéder d'autres causes (autres que le cancer) augment avec l'âge, donc les individus censurés sont plus particulièrement des personnes âgées. Dans ce cas-là, la censure doit donc être prise en compte lors de l'estimation de la survie.

III.2 Fonctions de base d'estimation de la survie

Le délai de survenue du décès à partir du diagnostic du cancer est représenté par la variable aléatoire T, continue et positive. L'analyse de survie se base sur différentes fonctions, liées entre elles, pour décrire la loi de probabilité de T. Soit t, le temps auquel la survie est mesurée :

- La fonction de survie $S(t)$ est la probabilité d'être vivant à t, autrement dit la probabilité de décéder après t :

$$S(t) = Prob(T > t) \tag{III.1}$$

La fonction de survie est une fonction monotone décroissante vérifiant les conditions suivantes : $S(0) = 1$ et $lim_{t\to\infty} S(t) = 0$

- La fonction de répartition $F(t)$ est la probabilité de décéder avant t :

$$F(t) = Prob(T \leq t) = 1 - S(t) \tag{III.2}$$

F(t) est associée à la fonction de densité de probabilité $f(t)$, telle que :

$$F(t) = \int_0^t f(u)du \text{ et sa dérivée : F'}(t) = f(t) = -S'(t) \quad \text{(III.3)}$$

- La fonction de densité de probabilité $f(t)$ est la limite de la probabilité de décéder dans un petit intervalle de temps après *t*, c.à.d. entre t et $t + \Delta t$. Plus communément, la fonction $f(t)$ est considérée comme la probabilité de décéder à *t* :

$$f(t) = \lim_{\Delta t \to 0^+} \frac{Prob(t \leq T < t + \Delta t)}{\Delta t} \quad \text{(III.4)}$$

- La fonction de risque instantané $\lambda(t)$, également appelé taux instantané de mortalité, correspond à la limite de la probabilité de décéder entre t et $t + \Delta t$ sachant que le décès n'est pas survenu avant *t* :

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{Prob(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = -\ln[S(t)]' \quad \text{(III.5)}$$

- La fonction de risque cumulé $\Lambda(t)$, également appelée taux de mortalité cumulé, est la somme des risque instantanés jusqu'à *t*, à partir du diagnostic :

$$\Lambda(t) = \int_0^t \lambda(u)du \quad \text{(III.6)}$$

A partir de l'ensemble des fonctions précédentes, la fonction de survie *S* peut donc être exprimée en fonction du risque instantané et cumulé, telle que :

$$S(t) = exp\left[-\int_0^t \lambda(u)du\right] = exp[-\Lambda(t)] \quad \text{(III.7)}$$

III.3 Les concepts de la survie

III.3.1 Survie observée

La survie observée (brute ou globale) au temps *t* correspond à la proportion de patients vivant au temps *t* lorsque tous les décès observés au cours de l'étude sont pris en compte, quelle qu'en soit la cause. La survie observée englobe donc les deux types de mortalité : la mortalité due au cancer et la mortalité due à d'autres causes (par exemple : accident de voiture, autres maladies, vieillesse, ...). De ce fait, la survie observée peut être très influencée par les décès « autres causes » notamment si la population étudiée est âgée. Si l'évènement d'intérêt est le décès dû au cancer alors les méthodes d'estimation de la survie observée ne sont pas adaptées. La survie observée ne permet pas de mettre en évidence la part de la maladie dans la mortalité.

III.3.2 Survie nette

Dans les études de survie après un cancer, l'évènement d'intérêt est le décès dû au cancer. La survie nette est définie comme la survie qu'on observerait dans un monde hypothétique où la seule cause possible de décès serait le cancer[8]. Elle est d'autant plus intéressante lorsque la population observée est âgée étant donné que le risque de décès dû à d'autres causes que le cancer augmente avec l'âge. Du fait de l'élimination de la part de la mortalité due à d'autres causes, la survie nette permet de comparer la mortalité due uniquement à la maladie entre différents pays ou différentes périodes[9].

III.4 Estimation de la survie nette

III.4.1 Cause de décès connue : méthode survie spécifique

Dans certains cas où la cause de décès est connue et l'évènement d'intérêt étudié est le décès par cancer, les décès « autre cause » peuvent être censurés. Les méthodes d'estimation de la survie spécifique peuvent alors être les mêmes que pour l'estimation de la survie observée[10] : estimateur non paramétrique de Kaplan-Meier[11, 12] ou le modèle semi-paramétrique de Cox[13]. Cependant, ceci n'est applicable que dans des conditions très spécifiques, le décès « autre cause » empêchant le plus souvent l'observation du décès dû à la

maladie, il devrait être considéré comme un risque compétitif. Par exemple, le modèle de régression en risque compétitif de Fine et Gray[14] peut être utilisé pour estimer la fonction d'incidence cumulée (CIF) comme une fonction de risque cumulé.

L'estimation de la survie nette par cette méthode peut être considérée fiable lorsque la cause de décès est connue et fiable. Cependant, dans le cadre d'étude de la survie à long terme, le taux de mauvaise classification des décès augmente avec le temps de suivi et avec l'âge des patients[15]. L'âge étant l'un des principaux facteurs pouvant affecter les deux types de mortalité (dû au cancer et « autres causes »), chez les patients âgés, le cancer est plus souvent défini à tort comme étant la cause de décès. Dans ce cas-là, la méthode de survie spécifique fournit des estimations biaisées de la survie nette[10].

III.4.2 Cause de décès non connue : méthode de survie relative

III.4.2.a Survie relative

Dans le cadre des registres de cancers, l'information sur la cause de décès est souvent inconnue ou peu fiable[16, 17]. L'analyse de la survie nette nécessite donc d'utiliser des méthodes appropriées permettant d'estimer la survie sans avoir besoin de connaître les causes des décès. La survie nette est alors estimée par la méthode de survie relative. C'est cette méthode qui a été utilisée dans le cadre de cette thèse. Elle consiste à corriger la survie observée dans une cohorte de patients par la survie attendue dans la population générale de mêmes caractéristiques démographiques que la cohorte. Dans le cadre de la survie relative, le taux instantané de mortalité observé λ_{obs} se base sur une relation additive[8, 9] (Remarque III.1). Il est égal à la somme du taux instantané de mortalité attendue dans la population générale λ_{pop} (mortalité « autres causes ») et du taux de mortalité en excès dans la population de patients (mortalité liée directement ou indirectement au cancer) λ_{exc} :

$$\lambda_{obs}(t) = \lambda_{pop}(t_{pop}) + \lambda_{exc}(t) \qquad \text{(III.8)}$$

Avec t, le temps écoulé entre depuis le diagnostic et t_{pop} est le temps écoulé depuis la naissance ($t_{pop} = t + a$, avec a l'âge au diagnostic).
En intégrant de 0 à t l'équation (III.8), le taux de mortalité observé cumulé est égal à :

$$\Lambda_{obs}(t) = \Lambda_{pop}(t_{pop}) + \Lambda_{exc}(t) \qquad \text{(III.9)}$$

Avec Λ_{pop}, le taux de mortalité attendu cumulé et Λ_{exc}, le taux de mortalité en excès cumulé.

A partir de l'équation (III.9), la fonction de survie s'écrit telle que :

$$S_{obs}(t) = S_{pop}(t_{pop}) \times S_n(t) \quad \text{(III.10)}$$

Avec S_{obs} et S_{pop}, correspondant respectivement à la survie observée dans la population de patient et attendue dans la population générale. S_n correspond à la fonction de survie nette associée au taux de mortalité en excès λ_{exc}, telle que :

$$S_n(t) = exp[-\Lambda_{exc}(t)] = exp\left[-\int_0^t \lambda_{exc}(u)du\right] \quad \text{(III.11)}$$

De façon générale, les taux de mortalité attendus sont approximés par les taux de mortalité observés dans la population générale de mêmes caractéristiques. Pour les études FRANCIM pour chaque sexe, département et année calendaire, nous utilisons les taux de mortalité lissés de la population générale fournis par l'Institut National de la Statistique et des Études Économiques (INSEE). Les tables de mortalité de la population générale tiennent compte de tous les décès (décès dus au cancer et décès « autres causes »). Cependant la part de la mortalité due à la maladie étudiée dans l'ensemble de la population générale étant faible par rapport à la part de la mortalité toutes causes, cela n'a pas de réel impact sur l'estimation de la mortalité[18].

Remarque III.1 : Dans certaines études, le taux instantané de mortalité observée est décrit par une relation multiplicative entre le taux de mortalité attendu et le taux de mortalité dû à la maladie[19-21]. Le taux de mortalité dû à la maladie est appelé taux de mortalité relatif, noté λ_{rel}*. Dans ce cas, le taux de mortalité observé est égal à :* $\lambda_{obs}(t) = \lambda_{pop}(t_{pop}) \times \lambda_{rel}(t)$*. Le taux de mortalité relatif est, généralement, lié au calcul du ratio standardisé de mortalité (SMR)[22]. Différentes études ont montré que la relation additive est plus satisfaisante, d'un point de vue biologique, dans les études pronostiques du cancer en population[9, 23, 24]. Dans cette thèse, nous nous intéresserons uniquement au taux de mortalité en excès.*

Il existe plusieurs approches pour estimer la survie nette d'un groupe., Danieli *et al.*[25] ont montré que seules deux d'entre elles fournissent des estimations non biaisées de la survie nette : l'estimateur non paramétrique proposé par Pohar-Perme *et al.*[26] et la modélisation du risque en excès basée sur une modélisation ajustée sur les variables démographiques de la table de mortalité[25].

III.4.2.b Méthode non-paramétrique : Estimateur de Pohar-Perme

Pohar-Perme *et al.* ont développé un estimateur de la survie nette non-biaisé[26]. Cet estimateur tient compte de la présence de censure « informative » (Section III.1) contrairement à l'estimateur d'Ederer II, historiquement utilisé dans les études sur base de population[25, 26] (Remarque III.2). L'estimateur de Pohar-Perme *et al.* corrige le biais dû à la censure informative en pondérant l'estimateur d'Ederer II par l'inverse de la probabilité de censure[11, 27]. D'où le taux cumulé en excès de Pohar-Perme :

$$\widehat{\Lambda}_{exc}(t) = \int_0^t \frac{dN^w(u)}{Y^w du} - \int_0^t \frac{\sum_{i=1}^n Y_i^w(u)\mathrm{d}\Lambda_{pop,i}(u)}{Y^w(u)}$$

$$\text{Avec : } N^w(t) = \sum_{i=1}^n N_i^w(t) = \sum_{i=1}^n \frac{N_i(t)}{S_{pop,i}(t)} \; ; \; Y^w(t) = \sum_{i=1}^n Y_i^w(t) = \sum_{i=1}^n \frac{Y_i(t)}{S_{pop,i}(t)} \qquad \text{(III.12)}$$

$\Lambda_{pop,i}$: taux de mortalité attendu au temps t pour le patient i

L'estimateur de Pohar-Perme pondère, à chaque temps t, le nombre de décès par cancer ($N_i(t)$) et le nombre d'individus à risque ($Y_i(t)$) par w, l'inverse de la survie attendue dans la population générale de mêmes caractéristiques (S_{pop}). Cette méthode tient compte de l'effet de sélection pouvant apparaitre au cours du temps : le nombre de patients susceptibles de sortir de la population à risque de décéder du cancer augmente au cours du temps et avec l'âge car ils ont plus de risque de décéder d'autres causes (mortalité attendue plus élevée).

Cet estimateur est largement utilisé par les registres de cancer car il est non biaisé. Il s'agit d'un estimateur non-paramétrique fournissant des estimations ponctuelles de la survie nette, il ne permet donc pas d'étudier plusieurs facteurs simultanément. De plus, il peut présenter une grande variabilité dans l'estimation de la survie nette à long-terme lorsqu'il y a peu d'information dans les données, notamment si la cohorte est très âgée (pondération très forte car le risque de décès « autres causes » est élevé)[26, 28].

La commande `stns` a été implémenté dans Stata permettant de calculer la survie nette tel que proposé par Pohar *et al.*[26, 29].

Remarque III.2 : Auparavant, afin de s'affranchir des causes de décès, la survie nette été estimée comme le ratio de la survie observée et attendue : méthode dites « Ratio-Estimate ». Trois estimateurs non paramétriques, se basant sur cette méthode, ont été proposés : l'estimateur d'Ederer I[30], l'estimateur d'Ederer II[31] et l'estimateur d'Hakulinen[32]. Ces estimateurs diffèrent par l'estimation de la survie attendue[25]. Cependant, ces estimateurs ont tendance à surestimer la survie nette, car ils ne tiennent pas compte de la présence de la censure informative, notamment chez les personnes âgées[25, 26].

III.4.2.c Méthode paramétrique : Modèle paramétrique flexible de survie relative

Plusieurs modèles de régression ont été développés modélisant le taux instantané de mortalité en excès : les modèles proposés par Estève *et al.*[9] et Dickman *et al.*[3] modélisent les taux instantanés de mortalité en excès en considérant le taux de base[vi] constant par intervalle de temps, l'hypothèse de proportionnalité[vii] et l'hypothèse de log-linéarité[viii]. Afin de s'affranchir de ces hypothèses pouvant être contraignantes, des extensions à ces modèles ont été proposées par la suite. Il s'agit de modèles paramétriques flexibles modélisant le taux de mortalité en excès par des splines (polynômes fractionnaires, B-splines, splines cubiques restreintes, …) : le modèle flexible de Giorgi *et al.*[33], le modèle flexible additif de Remontet *et al.*[34], le modèle flexible multiplicatif de Mahboubi *et al.*[35] ou encore le modèle paramétrique flexible de Nelson *et al.*[36]. Parmi les modèles paramétriques flexible de survie nette précités, seul celui proposé par Nelson *et al.* a été utilisé dans cette thèse et est développé par la suite.

[vi] Taux de base : Taux de mortalité lorsque toutes les variables explicatives prennent pour valeur la modalité de référence.
[vii] Hypothèse de proportionnalité : le risque de décès entre deux groupes est supposé constant au cours du temps.
[viii] Hypothèse de log-linéarité : pour une variable continue, le risque de décès est supposé log-linéaire lorsqu'il y a une variation d'une unité de la variable.

Nelson *et al.*[36] ont étendu à la survie relative le modèle paramétrique flexible de Royston et Parmar[37, 38] qui était lui-même une alternative au modèle de Cox qui permettait d'estimer la survie observée en incluant des effets dépendants du temps. Ce modèle paramétrique flexible de survie nette utilise des splines cubiques restreintes[39, 40] pour modéliser le logarithme du taux de mortalité en excès cumulé. Le modèle proportionnel s'écrit tel que :

$$\ln[\Lambda_{exc}(t;z)] = \ln[\Lambda_0(t)] + \beta z = s(x;\gamma_0) + \beta z \qquad \text{(III.13)}$$

Avec : $x = \ln(t)$, le logarithme du temps de suivi ; $\Lambda_0(t)$ est le taux en excès cumulé de base égal à $s(x;\gamma_0)$, une fonction spline cubique restreinte du taux de base dont le vecteur de paramètre est γ_0 ; β est le vecteur de paramètres permettant d'estimer les covariables z.

Les splines cubiques sont des fonctions mathématiques définies par plusieurs polynômes de degrés 3 reliés par des « nœuds » et dont les dérivées première et seconde doivent être continues, afin d'avoir une fonction lissée. Les splines cubiques restreintes (RCS) sont des splines cubiques contraintes à être linéaire avant le premier nœud et après le dernier nœud. Une fonction RCS avec *K* nœuds signifie qu'elle inclue *K-1* fonctions. Soit, *K* nœuds de k_1 *à* k_K, la fonction RCS de l'équation (III.13) est définie telle que :

$$s(x;\gamma_0) = \gamma_{00} + \gamma_{01}.\upsilon_{01}(x) + \gamma_{02}.\upsilon_{02}(x) + \cdots + \gamma_{0K-1}.\upsilon_{0K-1}(x) \qquad \text{(III.14)}$$

Pour $j = 1, \ldots, K-1$, $\upsilon_{0j}(x)$ est la j[ème] fonction de base définie telle que :

$$\upsilon_{0j}(x) = \begin{cases} x & ,pour\ j = 1 \\ (x-k_j)_+^3 - \phi_j(x-k_1)_+^3 - (1-\phi_j)(x-k_K)_+^3 & ,pour\ j = 2, \ldots, K-1 \end{cases} \qquad \text{(III.15)}$$

Où : $u_+ = u$ si $u > 0$ et $u_+ = 0$ si $u \leq 0$. Les nœuds externes, k_1 et k_K correspondent, respectivement, à la position du premier et du dernier nœuds ; et k_j correspond à la position du j[ème] nœud. $\phi_j = \frac{k_K - k_j}{k_K - k_1}$

Dans les études de cancer sur des données de population, les effets des covariables sont généralement dépendants du temps (effet non proportionnel). Cet effet est inclus dans le modèle par une interaction entre les covariables et le temps, l'équation (III.13) devient :

$$\ln \Lambda_{exc}(t;z) = s(x;\gamma_0) + \beta z + \sum_{\boldsymbol{d=1}}^{\boldsymbol{D}} \boldsymbol{s(x;\gamma_d)z_d} \qquad \text{(III.16)}$$

Avec : $s(x;\gamma_0)$ une fonction RCS du log du temps t ; D est le nombre de d'effets de covariables dépendants du temps, z_d et $s(x;\gamma_d)$ est la fonction RCS pour le d[ème] effet

dépendant du temps. La RCS $s(x;\gamma_d)$ s'écrit de la même manière que les équations (III.14) et (III.15).

A partir des équations (III.13) ou (III.16), notons, $\eta(t) = \ln \Lambda_{exc}(t;z)$. La fonction de survie relative et le taux instantané de mortalité en excès associé, correspondent alors respectivement à :

$$S_n(t;z) = exp\{-exp[\eta(t)]\}\ ;\ \lambda_{exc}(t;z) = \frac{d}{dt} exp[\eta(t)] \qquad \text{(III.17)}$$

Le modèle paramétrique flexible de survie relative a été implémenté dans Stata avec la commande `stpm2`[28, 41]. Les paramètres du modèle sont estimés par la méthode de la maximisation de la vraisemblance (Section III.4.3).

Le modèle de Nelson *et al.* [36] présente plusieurs avantages : il permet de modéliser différents facteurs pronostiques de la survie nette et d'inclure des effets dépendants du temps complexes. Les splines cubiques restreintes, utilisées ici, permettent d'avoir une plus grande flexibilité de la forme du taux de mortalité en excès comparé aux modèles paramétriques standards. De plus, contrairement aux autres modèles paramétriques flexibles développés (précité au début de cette section), les fonctions de survie et de mortalité peuvent être obtenues analytiquement et les données n'ont pas besoin d'être groupées par temps de suivi. De ce fait, le modèle de Nelson *et al.* permet d'utiliser des temps de suivis continus et les temps de calculs sont plus rapides avec ce modèle qu'avec les autres[36]. Enfin, de par l'utilisation des splines cubiques restreintes, il existe un développement de ce modèle permettant d'étudier la guérison (Section IV.2.3).

III.4.3 Maximum de vraisemblance en survie nette

La méthode du maximum de vraisemblance permet d'estimer les paramètres d'un modèle dont la distribution est connue. La vraisemblance (L) correspond à la probabilité d'observer un échantillon. La probabilité d'observer indépendamment chaque individu *i*, également appelée contribution individuelle et notée l_i, est égale à f(t_i) si le décès est observé (δ_i = 1) ou à S(t_i) si le décès n'est pas observé (δ_i = 0). La vraisemblance est donc le produit des contributions individuelles :

$$L = \prod_{i=1}^{N} f_{obs}(t_i; z_i)^{\delta_i} S_{obs}(t_i; z_i)^{1-\delta_i} \tag{III.18}$$

avec : *N,* le nombre de patients ; δ_i, l'indicatrice de décès qui vaut 1 si le patient *i* est décédé ou 0 s'il est censuré ; f_{obs} et S_{obs} sont respectivement, la fonction de densité et la survie observée du patient *i* au temps t_i et pouvant dépendre de covariables z_i. En utilisant les relations liant les fonctions de densité (f_{obs}), de survie (S_{obs}) et de risque instantané (λ_{obs}) (Section III.2) on obtient :

$$L = \prod_{i=1}^{N} \lambda_{obs}(t_i; z_i)^{\delta_i} S_{obs}(t_i; z_i) \tag{III.19}$$

La méthode consiste à maximiser la vraisemblance (L) par rapport aux paramètres inconnus (à estimer). De façon générale, il est plus facile de maximiser la fonction log-vraisemblance (Remarque III.3), qui est la somme des contributions des patients *i* :

$$\ln L = \sum_{i=1}^{N} \delta_i \ln[\lambda_{obs}(t_i; z_i)] + \ln[S_{obs}(t_i; z_i)] \tag{III.20}$$

D'après les équations (III.8) et (III.10), la fonction log-vraisemblance en survie nette est donc :

$$ln\, L = \sum_{i=1}^{N} \delta_i \, ln\big[\lambda_{pop}(t_{pop,i}; z_{pop,i}) + \lambda_{exc}(t_i; z_i)\big] + ln\big[S_{pop}(t_{pop,i}; z_{pop,i})\big] + ln[S_n(t_i; z_i)] \tag{III.21}$$

Avec : z_{pop}, le vecteur de covariables prises en compte pour le calcul du taux instantané de mortalité et de la survie attendu. $S_{pop}(t_{pop,i}; z_{pop,i})$ ne contribue pas au calcul de la

log-vraisemblance car elle ne dépend pas des paramètres à estimer. Donc, la fonction log-vraisemblance utilisée est :

$$ln\, L = \sum_{i=1}^{N} \delta_i \, ln\left[\lambda_{pop}(t_{pop,i}; z_{pop,i}) + \lambda_{exc}(t_i; z_i)\right] + ln[S_n(t_i; z_i)] \quad \text{(III.22)}$$

$t_{pop,i}$ est le temps de suivi du patient i depuis la naissance ($t_{pop,i} = t_i + a_i$, avec a_i l'âge au diagnostic).

Remarque III.3 : La fonction logarithme étant strictement croissante, la vraisemblance et la log-vraisemblance atteignent donc leur maximum au même point. De plus la recherche du maximum de vraisemblance nécessite de calculer la dérivée de la vraisemblance, et cela est beaucoup plus simple avec la log-vraisemblance, car il est plus aisé de dériver une somme de termes qu'un produit.

IV Guérison statistique

Les modèles de guérison ont été initialement développés dans le cadre de la survie observée lorsque la survie observée à long terme atteignait un plateau. Les deux principaux types de modèles de guérison développés sont : les modèles de mélange et les modèle de non-mélange. Les premiers modèles ont été développés par Boag[42] puis par Berkson et Gage[43] afin d'estimer la proportion de patients atteints de cancer qui ne décèderont pas suite à leur traitement. De nombreux modèles de mélanges en survie observée ont été proposés par la suite[44-47]. Le modèle de guérison de non-mélange ont été développés, à l'origine, pour modéliser la récidive du cancer en supposant qu'après un traitement, un individu a encore des cellules tumorales et que celles-ci peuvent évoluer jusqu'à la métastase[48-50]. Par la suite, les modèles de guérison ont été étendus à la survie nette. Dans cette section, nous nous intéresserons uniquement aux modèles de guérison en survie nette.

IV.1 Définition

Pour certaines localisations de cancer, tous les patients ne décèderont pas de leur maladie. La mortalité due au cancer (taux de mortalité en excès) est généralement élevée les premières années suivant le diagnostic puis diminue au cours du temps. Le moment à partir duquel le taux de mortalité en excès devient nul est appelé point de guérison. Avant que le point de guérison ne soit atteint, la mortalité observée dans la population de patient correspond à l'équation (III.8) : $\lambda_{obs}(t) = \lambda_{pop}(t_{pop}) + \lambda_{exc}(t)$. A partir du point de guérison, la mortalité en excès ($\lambda_{exc}(t)$) étant nulle, la mortalité observée dans la population de patients ($\lambda_{obs}(t)$) revient au même niveau que la mortalité dans la population générale ($\lambda_{pop}(t_{pop})$) (ayant les mêmes caractéristiques que les patients). Autrement dit, à partir du point de guérison, on a :

$$\lambda_{exc}(t) = 0 \text{ et } \lambda_{obs}(t) = \lambda_{pop}(t_{pop}) \qquad \text{(IV.1)}$$

Graphiquement, la courbe de survie nette atteint un plateau à partir du point de guérison (Figure IV.1). Ce plateau correspond donc à la proportion de patients qui ne décèderont pas de leur cancer, « statistiquement guéris »[42]. Ceci illustre le fait que la fonction de survie

nette est en réalité une fonction de survie impropre (elle n'atteint pas nécessairement 0 quand t tend vers l'infini) contrairement à la survie observée.

Les patients encore en vie au point de guérison sont considérés comme « statistiquement guéris ». Il s'agit d'une définition statistique de la guérison, au niveau populationnel et cela n'implique pas que tous les patients sont guéris cliniquement.

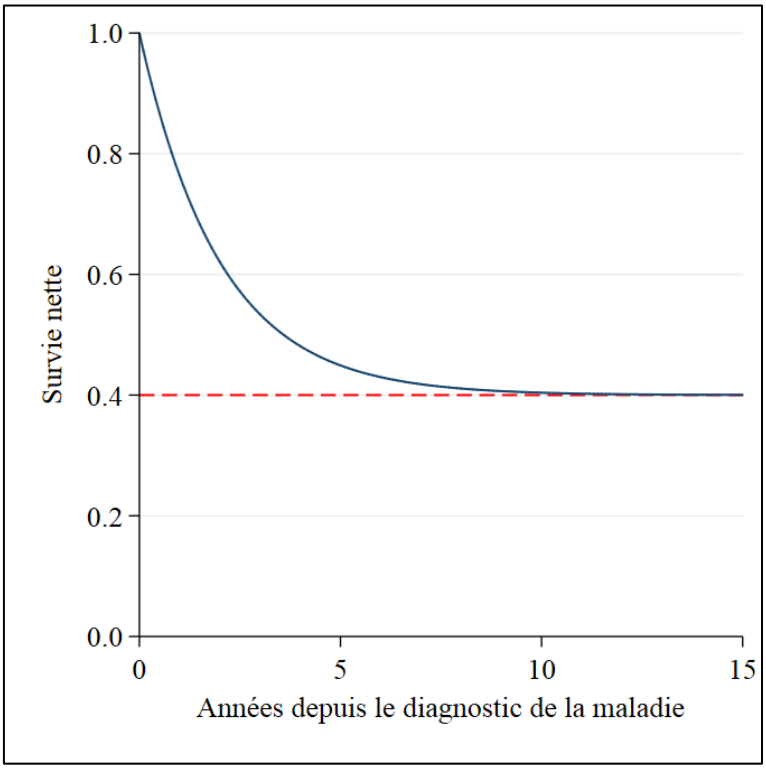


Figure IV.1 - Exemple fictif de courbe de survie nette estimé dans une population de patients où la guérison statistique est atteinte.

IV.2 Modèles de guérison paramétriques

IV.2.1 Modèles de guérison de mélange

Le principe des modèles de mélange est de considérer la population de patients comme un mélange de 2 sous-populations :

- Une proportion, π, de patients non à risque ou guéris avec une distribution de survie $S_c(t)$ égal à 1, quel que soit le temps t. Ces patients ne décèderont jamais de leur cancer.

- Une proportion, $1-\pi$, de patients à risque ou non-guéris avec une distribution de survie $S_u(t)$. En l'absence de censure, ces patients décèderont de leur cancer.

En survie nette, la fonction de survie d'un modèle de de guérison de mélange, s'écrit telle que :

$$S_n(t) = \pi . S_c(t) + (1-\pi)S_u(t) = \pi + (1-\pi)S_u(t) \quad \text{(IV.2)}$$

Et le taux de mortalité en excès associé, est défini tel que :

$$\lambda_{exc}(t) = \frac{(1-\pi)f_u(t)}{1+(1-\pi)S_u(t)} \quad \text{(IV.3)}$$

Avec : $f_u(t)$, la fonction de densité de probabilité associé à la survie des « non-guéris » $S_u(t)$. Une distribution paramétrique de Weibull ou exponentielle est généralement choisie pour modéliser la survie nette des « non-guéris »[51, 52]. Dans les équations (IV.2) et (IV.3), la survie des « non-guéris » peut s'écrire :

- $S_u(t) = exp(-kt)$: fonction de distribution exponentielle de paramètre k
- $S_u(t) = exp[-kt^m]$: fonction de distribution de Weibull de paramètres k et m

Verdecchia *et al.*[52] ont proposé un modèle de mélange utilisant des données groupées, ce qui ne permet pas d'introduire des covariables pour modéliser π, k et m. Les modèles de mélange de De Angelis *et al.*[51] et de Lambert *et al.*[53] utilisent, quant à eux, des données individuelles permettant d'inclure des covariables (z) pour modéliser π, k et m et de spécifier une ou plusieurs distributions paramétriques pour la fonction de survie des « non-guéris ».

Différentes fonctions de liens peuvent être utilisées lorsque la proportion de guéris dépend de covariables z, avec β le coefficient de régression :

- Lien identité : $\pi = \beta z$. L'effet de la covariable est dans la même unité que la proportion de guéris (il faut alors noter que π n'est pas contraint à être compris entre 0 et 1),
- Lien logistique : $\ln\left[{}^{\pi}/_{(1-\pi)}\right] = \beta z$. L'effet de la covariable s'interprète comme dans une régression logistique.
- Liens log(-log) : $\ln[-\ln(\pi)] = \beta z$. Les effets de la covariable s'interprètent comme le log ratio de taux de mortalité en excès.

Lambert *et al.*[54] ont implémenté la commande *strsmix* dans Stata, permettant de modéliser la survie nette avec un modèle de guérison de mélange avec des covariables et de spécifier la fonction de distribution de survie des « non-guéris » et la fonction de lien. Les

paramètres du modèle sont estimés en utilisant la méthode de maximisation de la vraisemblance (Section III.4.3). La log-vraisemblance d'un modèle de mélange est :

$$ln\,L = \sum_{i=1}^{N} \delta_i\, ln\left\{\lambda_{pop}(t_{pop}; z_{pop,i}) + \frac{[1-\pi(z_i)]f_u(t;z_i)}{1+[1-\pi(z_i)]S_u(t;z_i)}\right\} + ln\{\pi(z_i) + [1-\pi(z_i)]S_u(t;z_i)\} \quad \text{(IV.4)}$$

IV.2.2 Modèles de guérison de non-mélange

Le second grand type de modèle de guérison développé est le modèle de guérison de non-mélange, étendu à la survie nette par Lambert *et al.* [53]. La fonction de survie nette, dans le cadre de modèle de non-mélange, est :

$$S_n(t) = \pi^{F_y(t)} = \pi^{1-S_y(t)} = exp\left[ln(\pi) - S_y(t)\, ln(\pi)\right] \quad \text{(IV.5)}$$

Avec : π, la proportion de patients guéris et $F_y(t)$, une fonction de répartition et $S_y(t)$ une fonction de distribution ayant les même propriétés qu'une fonction de survie.
Et le taux de mortalité en excès associé, est défini tel que :

$$\lambda_{exc}(t) = -\,ln(\pi) f_y(t) \quad \text{(IV.6)}$$

$f_y(t)$ étant la fonction de densité de probabilité associée à la fonction de répartition $F_y(t)$.
On peut noter que d'après l'équation (IV.2) :

$$S_u(t) = \frac{S_n(t)-\pi}{1-\pi} = \frac{\pi^{1-S_y(t)} - \pi}{1-\pi} \quad \text{(IV.7)}$$

Donc, le modèle de guérison de non-mélange peut être écrit comme un modèle de mélange :

$$S_n(t) = \pi + (1-\pi)\left(\frac{\pi^{1-S_y(t)} - \pi}{1-\pi}\right) \quad \text{(IV.8)}$$

Comme pour les modèles de mélange, la fonction de distribution de Weibull est souvent utilisée pour exprimer $S_y(t)$. La proportion de guéris et la survie des non-guéris peuvent dépendre de covariables (z). Les paramètres du modèle sont estimés en utilisant la méthode de maximisation de la vraisemblance (Section III.4.3).

La log-vraisemblance d'un modèle de non-mélange est :

$$ln\, L = \sum_{i=1}^{N} \delta_i \, ln\{\lambda_{pop}(t_{pop}; z_{pop,i}) - ln[\pi(z_i)] f_y(t_i; z_i)\} + ln[\pi(z_i)] - S_y(t_i; z_i)\, ln[\pi(z_i)] \quad \text{(IV.9)}$$

La commande `strsnmix` a été implémentée dans Stata, permettant d'estimer la survie nette avec un modèle de non-mélange en spécifiant différentes distributions pour la fonction $S_y(t)$[54]. Cependant contrairement au modèle de mélange, une seule fonction de distribution peut être spécifiée.

Les modèles de guérison précédents sont des modèles paramétriques dont il faut spécifier la forme de la fonction de distribution de survie des « non-guéris ». Cependant, lorsque la proportion de guéris est très élevée ou au contraire lorsque la mortalité en excès est élevée au cours de la première année suivant le diagnostic Lambert *et al.* ont montré que les estimations pouvaient être biaisées, et que des problèmes de convergence des modèles de guérison pouvaient survenir[55]. Ils concluaient que les modèles précédents n'étaient pas assez flexibles. En 2011, Andersson *et al.*[56] ont proposé un modèle de guérison paramétrique flexible permettant de d'affranchir de la nécessité de spécifier une distribution particulière pour la survie des « non-guéris ». Ce modèle est développé dans la Section ci-dessous.

IV.2.3 Modèle de guérison paramétrique flexible

Le modèle de guérison paramétrique flexible d'Andersson *et al.*[56] est une extension au modèle paramétrique flexible de survie nette[28] présenté en Section III.4.2.c.La fonction $S_y(t)$ d'un modèle de non-mélange (équation (IV.5)), est modélisée par des splines cubiques restreintes (RCS). Ce modèle se base sur le fait qu'à partir du point de guérison, le taux instantané de moralité en excès devient nul, et donc que le taux de mortalité cumulé en excès devient constant.

Dans le modèle paramétrique flexible classique, le taux de mortalité cumulé est forcé à être linéaire après le dernier nœud (Section III.4.2.c). Dans le cas du modèle de guérison flexible, le taux de mortalité cumulé est, en plus, forcé à être constant après le dernier nœud

(pente nulle). Pour que cette condition soit plus facile à forcer, Andersson *et al.*[56, 57] contraignent le premier paramètre à être nul ($\gamma_{01} = 0$) et traitent les nœuds des RCS du modèle de guérison paramétrique flexible en ordre inverse. Dans le cadre du modèle avec guérison la fonction RCS définie dans l'équation (III.14) avec *K* nœuds de k_1 à k_K, devient alors :

$$s(x;\gamma_0) = \gamma_{00} + \gamma_{02}\upsilon_{02}(x) + \cdots + \gamma_{0K-1}\upsilon_{0K-1}(x) \qquad \text{(IV.10)}$$

Et la j[ème] fonction de base $\upsilon_{0j}(x)$, est :

$$\upsilon_{0j}(x) = (x - k_{K-j})_+^3 - \phi_j(k_K - x)_+^3 - (1 - \phi_j)(k_1 - x)_+^3 \quad , pour j = 2, \ldots, K-1 \qquad \text{(IV.11)}$$

Où : $x = \ln(t)$ et $u_+ = u$ si $u > 0$ et $u_+ = 0$ si $u \leq 0$; les nœuds externes, k_1 et k_K correspondent, respectivement, à la position du premier et du dernier nœud ; k_j correspond à la position du j[ème] nœud ; $\phi_j = \frac{k_{K-j} - k_j}{k_K - k_1}$

Donc, l'expression de fonction de la survie nette à partir du modèle de guérison paramétrique flexible est :

$$S_n(t) = exp\{-exp[\gamma_{00} + \gamma_{02}\upsilon_{02}(x) + \cdots + \gamma_{0K-1}\upsilon_{0K-1}(x)]\} \qquad \text{(IV.12)}$$

Avec : $x = \ln(t)$, le logarithme du temps de suivi ; $\Lambda_0(t)$ est le taux en excès cumulé de base égal à $s(x;\gamma_0)$, une fonction spline cubique restreinte du taux de base ; β est le vecteur de paramètres permettant d'estimer l'effet des covariables z.

La fonction de survie nette peut également s'écrire, telle que :

$$S_n(t) = exp[-exp(\gamma_{00})]^{exp\left[\sum_{j=2}^{K-1} \gamma_{02}\upsilon_{0j}(x)\right]} \qquad \text{(IV.13)}$$

Cette expression est similaire à l'expression de la fonction de survie nette d'un modèle de non-mélange (équation (IV.5)) :

$$S_n(t) = \pi^{F_y(t)} = \pi^{exp\left[\sum_{j=2}^{K-1} \gamma_{0j}\upsilon_{0j}(x)\right]} \qquad \text{(IV.14)}$$

Où : $\pi = exp[-exp(\gamma_{00})]$ et $F_y(t) = 1 - S_y(t) = exp\left[\sum_{j=2}^{K-1} \gamma_{0j}\upsilon_{0j}(x)\right]$:

Le modèle de guérison paramétrique flexible est donc un cas particulier des modèles de non-mélange. Ce modèle permet également d'étudier la guérison en introduisant des covariables. Soit, le vecteur de paramètres β permettant d'estimer l'effet des covariables z :

$$S_n(t) = exp[-exp(\gamma_{00} + \beta z)]^{exp[\sum_{j=2}^{K-1} \gamma_{0j} v_{0j}(x) + \sum_{d=1}^{D} s(x; z_d) z_d]} \quad \text{(IV.15)}$$

Avec : D est le nombre de d'effets de covariables dépendants du temps et $s(x; \gamma_d)$ est la fonction RCS pour le d[ème] effet dépendant du temps. La proportion de guéris π est modélisée par les paramètres γ_{00} et β et la fonction $F_y(t)$ est modélisée par les paramètres dépendants du temps.

Bien que la plupart des décès aient lieu au début du suivi, il est important que les nœuds des RCS soient placés tout au long du suivi pour que le modèle s'ajuste correctement à la fin du suivi. De plus, la proportion de guéris étant estimée au dernier nœud, Andersson *et al.* recommandent que le dernier nœuds soit alors placé au temps du dernier décès ou plus tard[56].

Les approches les plus communes pour modéliser la guérison statistique ont été comparées par Yu *et al.*[58]. Le modèle de guérison paramétrique flexible a notamment été comparé aux modèles de mélange dont celui de De Angelis *et al.* et Lambert *et al.* (Section IV.2.1). Les deux types de modèles de guérison (mélange et non-mélange) fournissent des estimations de proportion de guéris (π) est de survie nette ($S_n(t)$) assez proches. Le principal avantage des modèles de mélange est que la modélisation de π et $S_n(t)$ peuvent inclure des covariables différentes. Il apparaît que les performances du modèle de guérison paramétrique flexible sont équivalentes voire meilleures que celles des autres approches[58] et il converge plus souvent. C'est ce modèle que nous avons appliqué aux données réelles dans l'article présenté dans la Section VI.

Le package `stpm2` de Stata, qui inclue le modèle paramétrique flexible de survie nette a été adapté afin de modéliser le modèle de guérison flexible paramétrique[59].

IV.3 Vérification de l'hypothèse de guérison

Les modèles de guérison actuels ne permettent pas de tester statistiquement l'hypothèse de guérison et ,par conséquent, les indicateurs de guérison peuvent être estimé même lorsque l'hypothèse de guérison n'est pas raisonnable. Des tests de l'hypothèse de guérison existent[45, 60, 61] mais ils ne sont pas adaptés aux modèles flexibles de survie relative. Donc des critères ont été établis afin de s'assurer que l'hypothèse de guérison est respectée :

- Taux instantané de mortalité en excès faible,
- Adéquation aux données des modèles de guérison.

Dans cette thèse, le taux instantané de mortalité en excès est considéré comme suffisamment petit lorsqu'il tend vers zéro et que sa valeur à la fin du suivi est inférieure à 0,05. L'adéquation des modèles de guérison a été évaluée graphiquement, en vérifiant que les courbes de survie nette estimées par le modèle de guérison et par un modèle de survie nette standard se superposent [58].

Remarque IV.1 : Le modèle de guérison paramétrique flexible et le modèle paramétrique flexible de survie nette sont des modèles emboîtés et donc peuvent être comparé en utilisant le test du rapport de vraisemblance[56]. Cependant, cette comparaison ne peut être considérée comme un test de l'hypothèse de guérison. En effet la vraisemblance des modèles est calculée en utilisant l'ensemble du suivi et pas uniquement la fin du suivi, or l'hypothèse de guérison est vérifiable uniquement sur la fin du suivi.

IV.4 Indicateurs de guérison

IV.4.1 La proportion de patients guéris

La proportion de patients guéris, notée π, est un indicateur important donnant une information sur la survie à long-terme des patients qui ne mourront pas du cancer étudié. La survie nette atteint un plateau à partir du point de guérison. π correspond donc à l'asymptote de la courbe de survie nette (Figure IV.2).

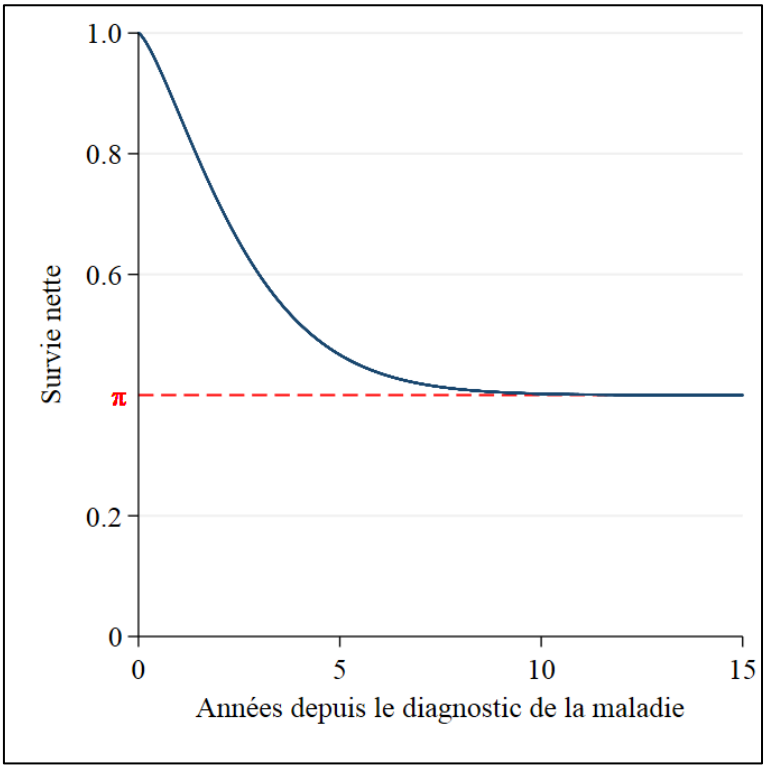


Figure IV.2 - Représentation graphique de la survie nette et de la proportion de guéris (π) estimées.

IV.4.2 Le temps de survie médian du groupe de patients « non-guéris »

Le second indicateur de guérison largement utilisé est le temps de survie médian du groupe de patients « non-guéris », noté T_{50}[52]. C'est le temps écoulé depuis le diagnostic jusqu'à ce que la moitié des patients « non-guéris » soient décédés : $S_u(T_{50}) = 0{,}5$ (Figure IV.3). T_{50} permet d'avoir une information sur la durée de vie patients qui décèderont du cancer étudié.

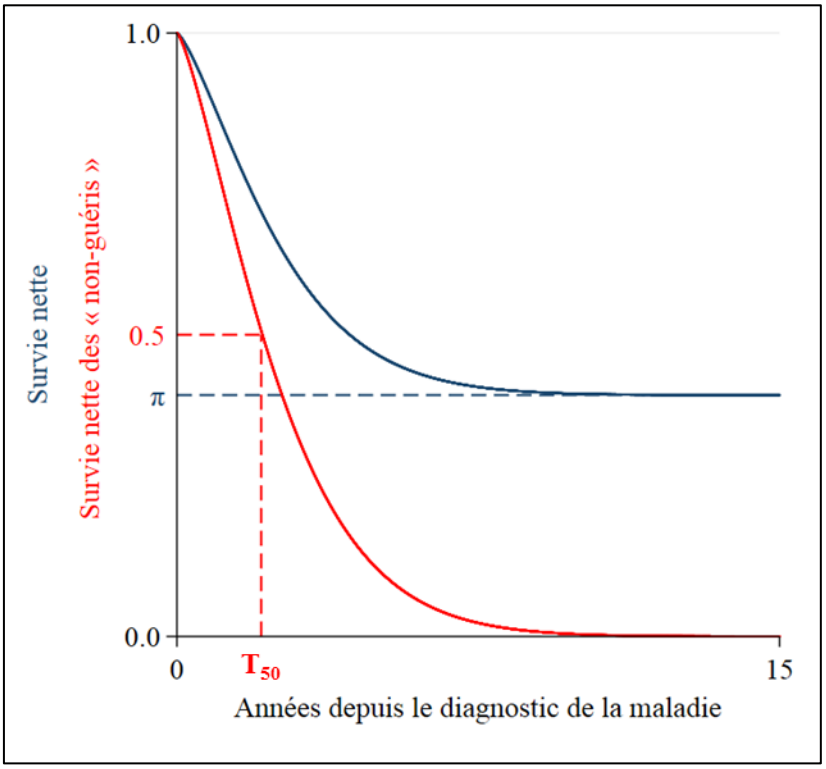


Figure IV.3 - Représentation graphique de la survie nette, de la survie des patients « non-guéris » et du temps de survie médian correspondant (T_{50}).

Ces deux premiers indicateurs sont largement utilisés dans les études de survie du cancer en population[62-64].

IV.4.3 La probabilité d'appartenir au groupe de patients guéris, au cours du temps

Plus rarement utilisée, ($P_i(t)$) la probabilité d'un individu i (et de caractéristiques $\underline{z_i}$) d'être guéris au temps t sachant qu'il est vivant au temps t peut aussi être estimée par les modèles de guérison[65] (Figure IV.4). Dès le diagnostic de leur cancer les patients sont répartis en deux groupes : les guéris et les non-guéris. Soit, Y l'indicatrice de guérison qui est égale à 1 si le patient appartient au groupe des guéris et 0 sinon. Par définition, pour un patient i, au temps t après le diagnostic, $P_i(t) = Prob(Y_i = 1|T_i > t)$, avec T_i le temps de survie observée à partir du diagnostic jusqu'au décès toutes causes. Au temps t, pour le patient i, on a donc :

$$Prob(Y_i = 1|T_i > t) = \frac{Prob(Y_i = 1, T_i \geq t)}{Prob(T_i \geq t)} = \frac{Prob(T_i \geq t|Y_i = 1) \times Prob(Y_i = 1)}{Prob(T_i \geq t)} \quad \text{(IV.16)}$$

La survie observée est définie comme : $S_{obs,i}(t) = Prob(T_i \geq t)$ (équation (III.1)). De plus, lorsque la probabilité d'être guéri ($Prob(Y_i = 1)$) vaut 1 cela signifie qu'il n'y a pas de mortalité due au cancer et donc que la survie observée dans la population de patients est égale à la survie attendue dans la population générale (équation (III.10)), d'où $Prob(T_i \geq t|Y_i = 1) = S_{pop,i}(t_{pop})$ avec : $t_{pop} = t + a_i$ est le temps écoulé depuis la naissance, a_i étant l'âge au diagnostic du patient i. Donc l'équation (IV.16) équivaut à :

$$Prob(Y_i = 1|T_i > t) = \frac{S_{pop,i}(t + a_i) \times \pi_i}{S_{obs,i}(t)} \quad \text{(IV.17)}$$

Autrement dit, d'après l'équation (III.10) de la survie nette $S_n(t)$:

$$Prob(Y_i = 1|T_i > t) = \frac{S_{pop,i}(t + a_i) \times \pi_i}{S_{pop,i}(t + a_i) \times S_{n,i}(t)} \quad \text{(IV.18)}$$

Donc, $P_i(t)$ est défini tel que :

$$P_i(t) = \frac{\pi_i}{S_{n,i}(t)} = \frac{\pi_i}{\pi_i + (1 - \pi_i)S_{u,i}(t)} \quad \text{(IV.19)}$$

Avec $S_{u,i}(t)$ la survie des patients non-guéris ayant les mêmes caractéristiques que le patient i.

Au diagnostic, la $S_{n,i}(t = 0) = 1$ donc la probabilité d'appartenir au groupe des guéris d'un patient i est égale à la proportion de guéris : $P_i(t = 0) = \pi_i$. La survie nette variant de 1 à π, cela implique que $P_i(t)$ varie de π à 1 au cours du temps.

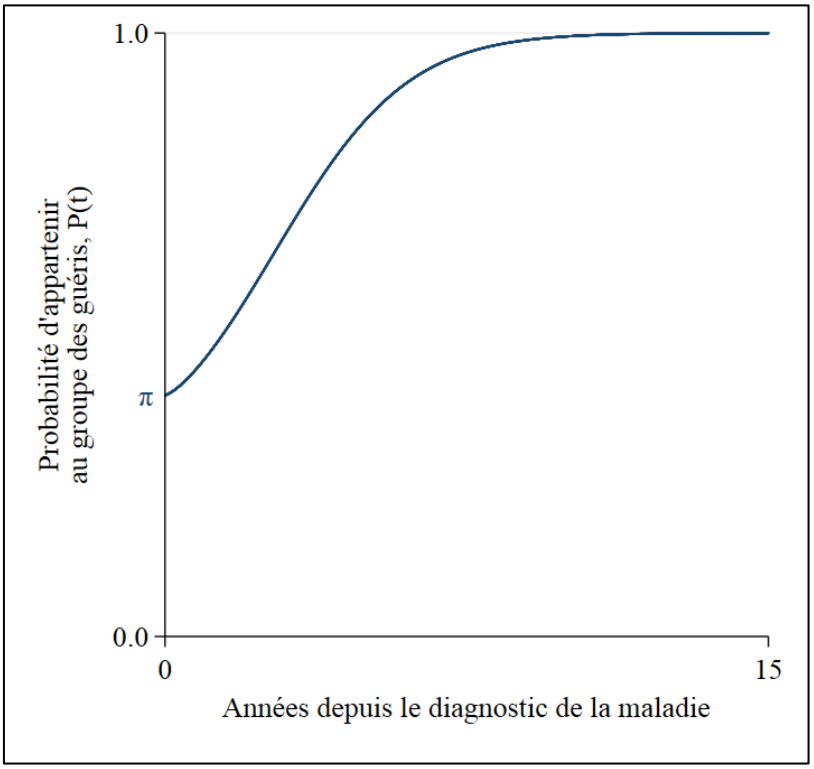


Figure IV.4 - Représentation graphique de la probabilité d'appartenir au groupe de patients guéris, au cours du temps (*P(t)*).

V Délai de guérison statistique

Les modèles de guérison existants ne permettent pas d'estimer directement le délai de guérison Le premier objectif de cette thèse était d'estimer le délai de guérison statistique après un cancer.

V.1 Définitions du délai de guérison statistique

V.1.1 A partir la survie nette du groupe de patients « non-guéris »

En 2009, Chauvenet *et al.*[66] ont estimé le délai de guérison à partir de la survie nette du groupe de patients « non-guéris ». Le point de guérison statistique est atteint lorsque presque tous les patients du groupe « non-guéris » sont morts, autrement dit, lorsque le nombre de décès dus au cancer devient négligeable. Chauvenet *et al.* ont estimé le délai de guérison comme le temps T_{95} à partir duquel 95% des patients « non-guéris » seraient décédés (Figure V.1), soit :

$$S_u(T_{95}) = 0{,}05 \tag{V.1}$$

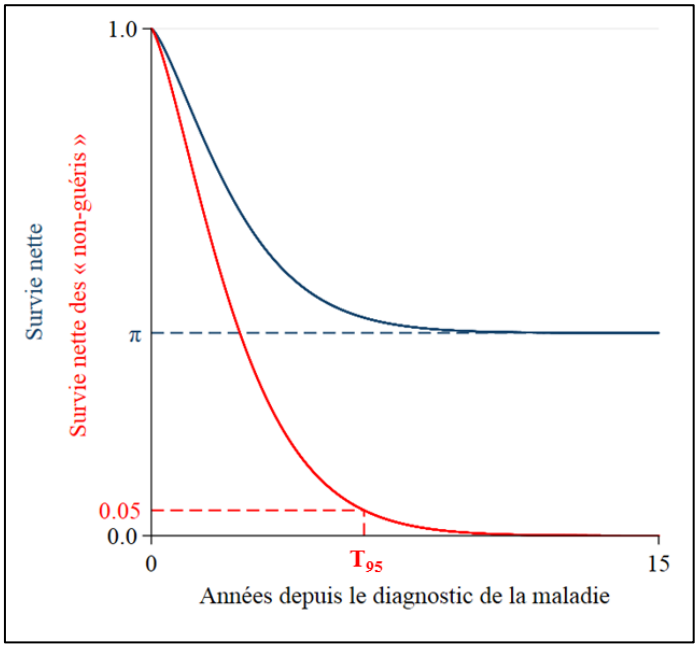


Figure V.1 – Définition du délai de guérison statistique de Chauvenet *et al.* basée sur la survie du groupe de patients « non-guéris » (T_{95}).

V.1.2 A partir de la survie nette conditionnelle à 5 ans

En 2012, Dal Maso *et al.*[67] ont estimé le délai de guérison à partir de la survie nette conditionnelle à 5 ans ($CNS_{t,5}$). Il s'agit de la probabilité de survivre cinq années de plus sachant que le patient est vivant t années après le diagnostic du cancer :

$$CNS_{t,5} = Prob(T > t + 5 | T > t) = \frac{Prob(T > t + 5)}{Prob(T > t} \quad \text{(V.2)}$$

Avec T le temps de survie. Donc, la survie nette conditionnelle à 5 ans se calcule comme le rapport entre la survie nette à t+5 et la survie nette à t :

$$CNS_{t,5} = \frac{S_n(t+5)}{S_n(t)} \quad \text{(V.3)}$$

Pour Dal Maso *et al.*, le point de guérison statistique est atteint lorsque la survie nette est constante pendant au moins 5 ans, soit, lorsque : $S_n(t+5) \approx S_n(t)$, soit $CNS_{t,5} \approx 1$. En utilisant un seuil de 95%, Dal Maso *et al.* estiment donc le délai de guérison statistique comme le temps T_{CNS} à partir duquel la survie nette conditionnelle à 5 ans atteint 95% :

$$CNS_{t,5} = \frac{S_n(T_{CNS} + 5)}{S_n(T_{CNS})} \geq 0{,}95 \quad \text{(V.4)}$$

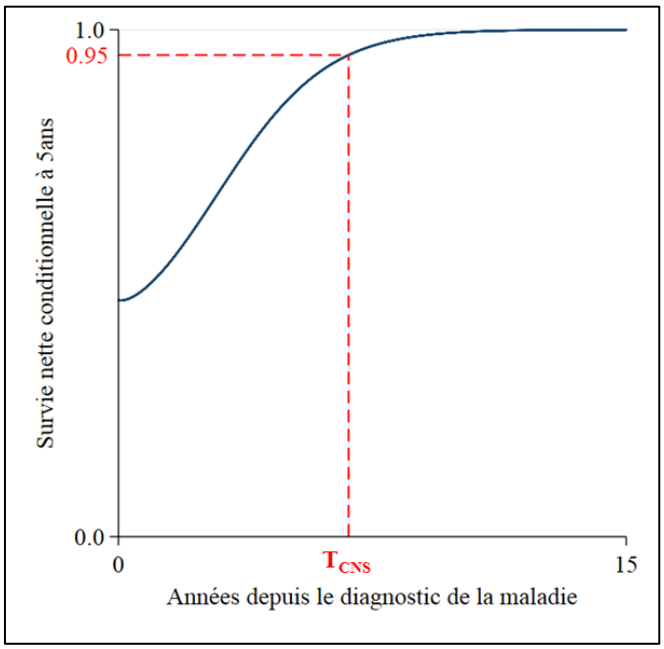


Figure V.2 - Définition du délai de guérison statistique de Dal Maso *et al.* basée sur la survie nette conditionnelle à 5 ans (T_{CNS}).

V.1.3 Nouvelle définition : à partir de la probabilité d'appartenir au groupe de patients guéris

En 2018, nous [68] avons proposé une nouvelle définition du délai de guérison se basant sur la probabilité d'un individu i d'être guéris au temps t sachant qu'il est vivant au temps t ($P_i(t) = \pi_i / S_{n,i}(t)$) (Section IV.4.3). Le délai de guérison est estimé comme le temps TTC_i à partir duquel la probabilité d'appartenir au groupe des patients guéris atteint 95% :

$$P_i(TTC_i) = \frac{\pi_i}{S_{n,i}(TTC_i)} \geq 0{,}95 \qquad \text{(V.5)}$$

Avec : π la proportion de patients guéris. Notons que si $\pi \geq 0{,}95$ alors $TTC = 0$.

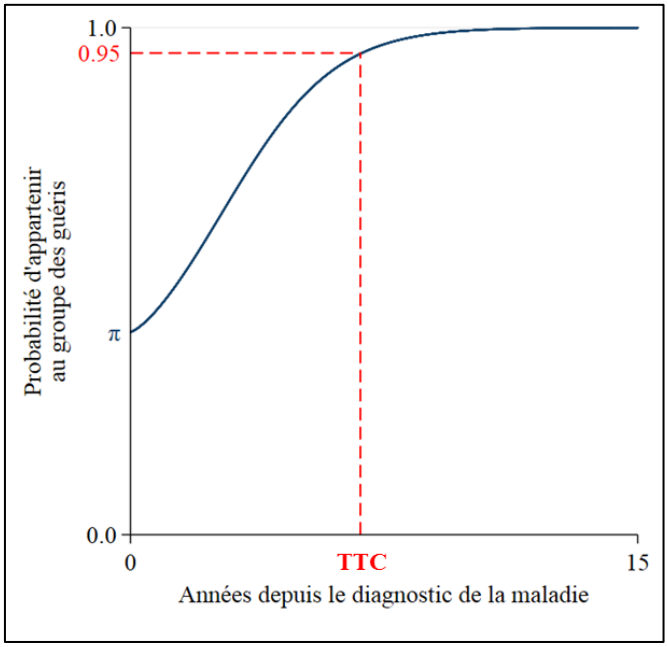


Figure V.3 - Définition du délai de guérison statistique de Boussari *et al.* basée sur probabilité d'appartenir au groupe de patients guéris au cours du temps (*TTC*).

J'ai réalisé une analyse de sensibilité du délai de guérison *TTC* pour différents seuils sur données réelles. Le délai *TTC* a été estimé pour P(t) = 90%, 95% et 99% sur quatre localisations de cancer illustratifs (colon-rectum, seins, pancréas et thyroïde) séparément pour chaque sexe et par classe d'âge au diagnostic. (Tableau V-1).

Selon le choix du seuil il peut y avoir de grandes variations sur l'estimation du délai de guérison. Pour un seuil variant entre 90% et 99 %, la variation du *TTC* était en moyenne de 6 ans pour le cancer colorectal et de 4 ans pour le cancer du pancréas chez les hommes et les

femmes. Chez les femmes atteintes d'un cancer du sein, la variation du *TTC* était d'environ 5 ans avant 45 ans et après 65 ans et de 7 ans dans les classes d'âge intermédiaires. Pour le cancer de la thyroïde, la variation du *TTC* était très hétérogène : de moins de 1 an à plus de 9 ans selon l'âge et le sexe.

La probabilité d'appartenir au groupe des guéris *P(t)* tend vers 1 quand *t* tend vers l'infini. Cela implique que plus *t* est grand plus la pente de la courbe de *P(t)* est faible. Le délai de guérison est estimé sur la fin de la courbe de *P(t)* (entre 90% et 99%), c'est pourquoi on observe de grandes variations du *TTC* pour les différentes valeurs de *P(t)*.

V.2 Comparaison des trois délais de guérison sur données réelles

La nouvelle définition du délai de guérison et la comparaison des trois délais de guérison (T_{95}, T_{CNS} et TTC) ont fait l'objet d'une publication dont je suis 2ème auteur (Section V.2.3).

V.2.1 Population étudiée et méthode

J'ai comparé les estimations des délais de guérison issues des trois définitions en utilisant les données de la base FRANCIM. Parmi l'ensemble des localisations de cancer disponibles et respectant l'hypothèse de guérison, nous avons choisi quatre localisations selon la dynamique du taux de mortalité en excès estimé par le modèle paramétrique flexible de Nelson *et al.*[36] (Section III.4.2.c). Les patients ayant eu un diagnostic de cancer du côlon-rectum, pancréas, sein ou de la thyroïde entre 15 et 74 ans sur la période 1995-2010 ont été inclus. Les courbes respectives du taux de mortalité en excès, sont présentées dans les Figures Figure V.4.A, Figure V.5.A, Figure V.6.A et Figure V.7.A.

Les analyses ont été faites séparément pour chaque sexe et localisation de cancer. L'âge au diagnostic a été inclus en catégories dans les modèles, comme suit : [15-45[, [45-55[, [55-65[, et [65-75[ans pour les cancers du côlon-rectum, sein et thyroïde ; [15-55[, [55-65[, et [65-75[pour les cancers du pancréas. Les patients ont été suivi 15 ans ou jusqu'au 30 juin 2013. La survie nette ($S_n(t)$), la survie des « non-guéris » ($S_u(t)$), la proportion de patients guéris (π), la probabilité d'appartenir au groupe des guéris (P(t)) et la survie nette conditionnelle à 5 ans (CNS(t)) ont été estimées à partir du modèle de guérison paramétrique flexible[56] (Section 0). Les estimations de TTC et T_{95} ont été obtenus avec l'algorithme de

Newton-Raphson et T_{CNS} a été obtenu à partir de $CNS(t)$ calculé pour chaque année t de 0 à 10 ans. Les valeurs de $P(t)$, $S_u(t)$ et $CNS(t)$ ont alors pu être comparées pour $t = TTC, T_{95}, T_{CNS}$ et sont présentées dans le Tableau 2 de l'article (Section V.2.3).

V.2.2 Conclusions

Ces résultats montrent les limites de l'estimation du délai de guérison à partir de la survie du groupe de patients « non-guéris » et à partir de la survie nette conditionnelle à 5 ans.

Le délai de guérison défini par T_{95} dépend uniquement de la survie des « non-guéris ». Or, un taux de mortalité en excès très élevé les premières années suivant le diagnostic implique une diminution brutale de la survie nette et de la survie des « non-guéris ». Le délai de guérison T_{95} a tendance à être très faible lorsque le taux de mortalité en excès est élevé près du diagnostic car la survie des « non-guéris » atteint rapidement 0,05. Par exemple, dans le cas du cancer du pancréas, la mortalité en excès proche du diagnostic est très élevée, $S_u(t)$ atteint donc le seuil de 0,05 au cours des cinq premières années suivant le diagnostic (T_{95}) alors que le taux de mortalité en excès n'est pas proche de 0 ($\lambda_{exc}(t \leq 5) > 0{,}1$). A l'inverse, dans le cas du cancer de la thyroïde chez les patients jeunes, le taux de mortalité en excès est très faible impliquant une diminution lente de $S_u(t)$. Dans ce cas le délai de guérison est estimé à plus de 10 ans après le diagnostic alors que le taux de mortalité en excès peut déjà être considéré comme négligeable au cours des cinq premières années.

Le délai de guérison défini par T_{CNS} peut être utilisé à condition que CNS(t) soit une fonction de temps monotone croissante, autrement dit que la dérivée de la fonction CNS(t) soit > 0. Or d'après la fonction dérivée de la survie nette conditionnelle à 5 ans :

$$CNS(t)' = CNS(t)[\lambda_{exc}(t) + \lambda_{exc}(t+5)] \qquad \text{(V.6)}$$

Si le taux de mortalité en excès n'est pas décroissant au cours du temps alors $CNS(t)'$ n'est pas positive donc CNS(t) n'est pas une fonction monotone croissante. T_{CNS} risque alors d'être sous-estimé. Par exemple, dans le cas du cancer de la thyroïde chez les femmes âgées, le taux de mortalité en excès présente un « rebond » : après avoir été proche de zéro les premières années suivant le diagnostic, il augmente de façon durable jusqu'à 12 ans. $CNS(t)$ dépasse 95% à 1 an pour ensuite redescendre en dessous de 95% jusqu'après 10 ans. Le délai de guérison est alors estimé à 1 an avec cette méthode (vs. TTC = 10 ans). Le second inconvénient du délai de guérison T_{CNS} est qu'il nécessite un temps de suivi de 5 ans de plus

afin d'estimer la survie à t + 5 ans. Dans cette étude nous n'avons pas pu estimer la survie conditionnelle nette à 5 ans pour t ≥ 10 ans. Le suivi maximal observé étant de 15 ans il aurait fallu extrapoler la survie nette au-delà de 15 ans. Contrairement au délai T_{95}, T_{CNS} peut être utilisé pour déterminer le délai de guérison à condition que le taux de mortalité en excès soit monotone décroissant. De plus, le délai de guérison défini à partir de *CNS(t)*, a l'avantage de pouvoir être estimé à partir d'un estimateur ponctuel et donc sans avoir à modéliser la survie nette.

La nouvelle définition du délai de guérison du cancer, basée sur la probabilité d'appartenir au groupe des patients guéris, est une définition naturelle et intuitive. En effet, la probabilité *P(t)* d'être guéri conditionnellement au fait d'être en vie au temps *t* est une traduction directe du concept de guérison statistique. De plus, *P(t)* mesure directement la proximité entre la courbe de survie nette et la proportion de patients guéris.

V.2.3 Article

Cancer Epidemiology 53 (2018) 72–80

Contents lists available at ScienceDirect

Cancer Epidemiology

journal homepage: www.elsevier.com/locate/canep

A new approach to estimate time-to-cure from cancer registries data

Olayidé Boussari[a,b,c], Gaëlle Romain[a,b], Laurent Remontet[d,e,f,g], Nadine Bossard[d,e,f,g], Morgane Mounier[h], Anne-Marie Bouvier[a,b], Christine Binquet[b,i,j], Marc Colonna[k], Valérie Jooste[a,b,*]

[a] Dijon-Bourgogne University Hospital, Registre Bourguignon des Cancers Digestifs, Dijon F-21000, France
[b] INSERM, U1231, EPICAD team, Univ Bourgogne-Franche-Comté, UMR 1231, Dijon F-21000, France
[c] LabEX LipSTIC, ANR-11-LABX-0021, Dijon F-21000, France
[d] Service de Biostatistique-Bioinformatique, Hospices Civils de Lyon, Lyon F-69003, France
[e] Université de Lyon, Lyon F-69000, France
[f] Université Lyon 1, Villeurbanne F-69100, France
[g] CNRS UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Équipe Biostatistique Santé, Pierre-Bénite F-69310, France
[h] Dijon-Bourgogne University Hospital, Univ Bourgogne Franche Comté, Registre des Hémopathies Malignes de Côte d'Or, Dijon, France
[i] INSERM, CIC1432, Clinical Epidemiology Unit, Dijon F-21000, France
[j] Dijon-Bourgogne University Hospital, Clinical Investigation Centre, Clinical Epidemiology/Clinical Trials Unit, Dijon F-21000, France
[k] Registre du Cancer de l'Isère, Grenoble University Hospital, Grenoble F-38000, France

ARTICLE INFO

Keywords:
Net survival
Cure models
Probability of being cured
Time-to-cure

ABSTRACT

Background: Cure models have been adapted to net survival context to provide important indicators from population-based cancer data, such as the cure fraction and the time-to-cure. However existing methods for computing time-to-cure suffer from some limitations.
Methods: Cure models in net survival framework were briefly overviewed and a new definition of time-to-cure was introduced as the time TTC at which P(t), the estimated covariate-specific probability of being cured at a given time t after diagnosis, reaches 0.95. We applied flexible parametric cure models to data of four cancer sites provided by the French network of cancer registries (FRANCIM). Then estimates of the time-to-cure by TTC and by two existing methods were derived and compared. Cure fractions and probabilities P(t) were also computed.
Results: Depending on the age group, TTC ranged from to 8 to 10 years for colorectal and pancreatic cancer and was nearly 12 years for breast cancer. In thyroid cancer patients under 55 years at diagnosis, TTC was strikingly 0: the probability of being cured was > 0.95 just after diagnosis. This is an interesting result regarding the health insurance premiums of these patients. The estimated values of time-to-cure from the three approaches were close for colorectal cancer only.
Conclusions: We propose a new approach, based on estimated covariate-specific probability of being cured, to estimate time-to-cure. Compared to two existing methods, the new approach seems to be more intuitive and natural and less sensitive to the survival time distribution.

1. Introduction

In some survival studies, a fraction of subjects never experience the event under study whatever the length of the follow-up period. In such cases, the observed survival curve levels off at a non-zero value which corresponds to the proportion of subjects who are free of failing from the event of interest and who are considered "statistically cured". Since their first formulation by Boag [1], cure models have been widely developed to analyse survival data with cure fractions (e.g., [2–5]).

In cancer population-based (such as registry-based) studies, the cause of death is often unreliable or unknown. Specific methods use the excess mortality rate (mortality due to cancer only) to estimate net survival without needing the cause of death [6,7]. Net survival is the survival in the hypothetical world where cancer would be the only cause of death [8,9]. When the net survival curve ends by flattening to a non-zero value, it is a typical case where cure models are useful for cancer survival analyses [10,11]. Several authors have extended cure models to the net survival framework [11–15]; we propose in

Abbreviations: CNS, conditional net survival; P, cure fraction; P(t), estimated covariate-specific cure probability; $S_n(t)$, net survival; $S_u(t)$, net survival of uncured subjects; T_{95}, time at which $S_u = 0.05$; T_{CNS}, time at which 5-year CNS = 0.95; TTC, time-to-cure

* Corresponding author at: Registre Bourguignon des Cancers Digestifs, UFR des sciences de santé BP 87900, Dijon cedex 21079, France.
E-mail address: valerie.jooste@u-bourgogne.fr (V. Jooste).

https://doi.org/10.1016/j.canep.2018.01.013
Received 6 September 2017; Received in revised form 19 January 2018; Accepted 21 January 2018
Available online 04 February 2018

Supplementary file 1 an overview of these extensions. In these models, two quantities describe the net survival: i) the cure fraction P or proportion of cured subjects ($0 \leq P \leq 1$) and ii) the net survival time distribution $S_u(t)$ of uncured subjects described by a parametric or semi-parametric survival distribution function. Cure models assume that a given patient belongs to a given group (cured or uncured) since diagnosis. The probability that this patient belongs to the cured group is P at the time of diagnosis and increases as time elapses if the patient is still alive [16]. The time-to-cure, i.e. the time elapsed between diagnosis and cure is a useful indicator. Two definitions of the time-to-cure have been already proposed by the specialized literature [17,18] and are recalled in the next section.

In this work, we proposed a new definition for the time-to-cure, based on the patient's probability of being cured over time: This is a synthetic way to indicate the time at which patients can be reasonably confident to belong to the cured group and to attach a probability to such confidence. We applied flexible parametric cure models [19] to four cancer real datasets and derived the time-to-cure estimates by the new definition and by the two existing definitions. Then the results were compared and discussed.

2. Recall on the two existing approaches to estimate the time-to-cure

Chauvenet et al. [17] proposed to estimate the time-to-cure as the time at which "almost" all uncured patients would have died. From that time, the number of deaths attributable to the cancer of interest becomes negligible. It can be estimated as the time T_{95} at which 95% of the uncured would have died: i.e., $S_u(T_{95}) = 0.05$.

Dal Maso et al. [18] used the 5-year conditional net survival (CNS) to propose a definition for the time-to-cure: because it is considered that cure is reached when the net survival becomes nearly constant, it can be assumed that cure is reached when the conditional net survival becomes close to 1. Using the 95% cut-off, these authors defined the time-to-cure as the shortest time T_{CNS} after diagnosis at which the 5-year conditional net survival reaches 0.95; i.e. $S_n(T_{CNS} + 5)/S_n(T_{CNS}) = 0.95$.

3. A new definition for the time-to-cure

A work by Sposto [16] allows estimating for a given patient i (i.e. with characteristics x_i), the probability $P_i(t)$ of being cured at a given time t after diagnosis knowing that he/she was alive up to time t.

$P_i(t)$ = (probability of being cured and alive up to time t given x_i)/(probability of being alive up to time t given x_i)

We demonstrate in Appendix, that:

$$P_i(t) = P_i/S_{n,i}(t) = P_i/[P_i + (1 - P_i)S_{u,i}(t)]$$

where P_i (resp. $S_{u,i}$) denotes the proportion of cured patients (resp. the net survival function of the uncured patients) within the group of patients sharing the same characteristics as i.

This probability is an indicator that has been rarely used so far [20] but is interesting because: 1) as cure models assume that the two subpopulations (cured and uncured) are defined right at diagnosis, $P_i(t)$ corresponds to a dynamic prediction of a patient's probability of belonging to the cured group; 2) it is intuitive since, for a group of patients with same covariates x_i, it corresponds at time t, to the proportion of patients that belong to the cured group among still alive patients. Moreover, it is easy to see that $P_i(t = 0)$ is simply the cure proportion P_i and that $P_i(t)$ increases with t from P_i at diagnosis ($t = 0$) to 1.

We propose here to estimate the time-to-cure by the time $t = TTC_i$ from which $P_i(t) \geq 0.95$. From this definition it can be deduced that when $P_i \geq 0.95$, $TTC_i = 0$. In the sequel, 'i' will be omitted for easier reading.

4. Illustrative application of the new approach to real cancer data

4.1. Materials and methods

The French network of cancer registries (FRANCIM) and the Service de Biostatistique-Bioinformatique (Hospices Civils de Lyon, France) are currently maintaining a common database started in 1975 that counts currently more than a million patients diagnosed with any of thirty cancer types [21]. Quality controls are performed by each registry and on the whole database with tools provided by the International Agency for Research on Cancer.

We illustrated the new approach with data from four cancer sites (colon-rectum, pancreas, breast and thyroid, of which respectively 46 371, 8169, 71 947 and 8762 cases were included) showing different settings of the dynamics of the excess hazard and for which cure assumption could be accepted. These cancer sites were chosen because they illustrate both long-time survival cancer, such as thyroid cancer, and short survival cancer, such as pancreatic cancer. We included all patients aged 15 to 74 years at diagnosis and diagnosed between 1995 and 2010. They were followed-up until June 30, 2013.

The analyses were conducted separately in men and women and age was included as a categorical covariate. Age groups were: [15–45[, [45–55[, [55–65[, and [65–75[years for colorectal, thyroid, and breast cancers. For pancreatic cancer the number of cases aged [15–45[was low (357 cases), we combined this group with the [45–55[leading to [15–55[, [55–65[, and [65–75[age groups. Breast cancer was considered in women only. Table 1 summarizes the baseline data.

For each site, the net survival was estimated by fitting a flexible parametric survival model [22]: the logarithm of the baseline cumulative excess hazard is written as a restricted cubic spline of the log time with four internal knots placed at the 25th, 50th, 75th and 95th centiles of the observed death times and the boundary knots placed at the 1st centile of the observed death times and 17 years. We introduced age at diagnosis (using the previously defined age groups) as a covariate. We evaluated time dependent effect of age at diagnosis through an interaction term with spline of time (2 internal knots located at the 33rd and 67th centiles of the observed death times), using a likelihood ratio test with 0.05 significance level. We checked graphically that the excess hazard derived from this model approached zero (i.e., the associated net survival reached a plateau) and then the net survival was modelled using a flexible parametric non-mixture cure model [19], according to the recommendations of Yu et al. [23]. The splines considered for this model used the same knots as that in the previous model with an additive knot at the 99th centile of the observed death times. Note that separate models were fitted for each sex for eligible site. Estimates of the net survival flexible parametric non-mixture cure model were validated by comparing them with the non-parametric estimates obtained from the Pohar-Perme estimator [24]. Results of the graphical checking of the cure assumption and the validation of the net survival estimates by the cure model are provided in Supplementary file 2 (Figs. S1–S4). Cure assumption was accepted for colorectal, thyroid and pancreatic cancers. For women with breast cancer, although the plateau in net survival was not obvious, the excess mortality rate was very low and survival curves with and without cure assumption were very close to each other for all age groups except 65–74 years (Fig. S3). Cure assumption was accepted for women under 65 years at diagnosis.

Based on the regression coefficients from the flexible parametric non-mixture cure model we computed P, $S_u(t)$ and $S_n(t)$, and then derived $P(t) = P/S_n(t)$. Thereafter, estimates of TTC and T_{95} were obtained using a Newton-Raphson technique. T_{CNS} estimates were derived by evaluating $CNS(t) = S_n(t + 5)/S_n(t)$ at yearly intervals (starting from $t = 0$). For each cancer site considered, the values of P(t), $S_u(t)$, and CNS(t) were compared at $t = TTC$, T_{95}, and T_{CNS} to better examine the three definitions of the time-to-cure.

In order to check the sensitivity of TTC to the cut-off changes for the four illustrative cancer sites, TTC was estimated for 3 different cut-off

Table 1
Characteristics of patients aged 15–74 years diagnosed in France with cancer of larynx, colon and rectum, pancreas, breast or thyroid, between 1995 and 2010 (FRANCIM data) and associated 1-, 5-, 10- and 15-year net survival.

Cancer site	Age at diagnosis	Cases	% lost to FU[a]	Net survival (NS) estimated with flexible model (without cure)								Cure proportion (P)[b]	
				1-year NS	95% CI	5-year NS	95% CI	10-year NS	95% CI	15-year NS	95% CI	P	95% CI
Men													
Colon and Rectum	[15–45[	1117	3.2	0.89	0.87–0.90	0.67	0.64–0.70	0.60	0.57–0.63	0.58	0.54–0.61	57	54–60
	[45–55[	3742	2.8	0.89	0.88–0.90	0.66	0.64–0.67	0.58	0.57–0.60	0.56	0.53–0.58	55	53–57
	[55–65[	8608	2.1	0.87	0.87–0.88	0.64	0.63–0.65	0.55	0.54–0.57	0.52	0.50–0.54	52	51–54
	[65–75[	14,394	1.6	0.82	0.82–0.83	0.59	0.58–0.60	0.50	0.49–0.51	0.46	0.44–0.47	47	46–49
Pancreas	[15–55[	960	1.0	0.39	0.36–0.42	0.12	0.10–0.14	0.10	0.08–0.12	0.09	0.07–0.11	9	7–11
	[55–65[	1654	1.0	0.33	0.31–0.35	0.08	0.07–0.09	0.05	0.04–0.06	0.04	0.03–0.06	6	5–7
	[65–75[	2316	1.1	0.27	0.25–0.29	0.07	0.06–0.08	0.05	0.04–0.06	0.05	0.04–0.06	5	4–6
Thyroid	[15–45[	683	2.6	0.99	0.99–1	0.99	0.97–0.99	0.98	0.96–0.99	0.98	0.96–0.99	98	95–99
	[45–55[	504	3.4	0.97	0.96–0.98	0.95	0.92–0.96	0.92	0.88–0.95	0.92	0.87–0.95	91	87–94
	[55–65[	531	2.3	0.93	0.91–0.95	0.88	0.85–0.91	0.83	0.77–0.87	0.82	0.75–0.88	81	74–86
	[65–75[	342	2.0	0.86	0.82–0.89	0.76	0.71–0.81	0.66	0.58–0.73	0.65	0.54–0.74	63	53–72
Women													
Colon and Rectum	[15–45[	1128	4.3	0.92	0.91–0.93	0.70	0.67–0.72	0.64	0.61–0.67	0.62	0.59–0.65	62	58–65
	[45–55[	2925	4.5	0.91	0.90–0.92	0.68	0.66–0.69	0.61	0.59–0.63	0.59	0.57–0.61	59	57–61
	[55–65[	5414	4.2	0.89	0.88–0.90	0.67	0.66–0.68	0.60	0.59–0.62	0.58	0.56–0.60	59	57–60
	[65–75[	9043	3.1	0.84	0.83–0.85	0.62	0.61–0.64	0.55	0.54–0.57	0.53	0.51–0.54	54	53–55
Pancreas	[15–55[	548	1.1	0.52	0.48–0.55	0.21	0.18–0.24	0.16	0.13–0.19	0.14	0.12–0.17	16	13–19
	[55–65[	922	1.3	0.42	0.39–0.44	0.12	0.10–0.14	0.09	0.07–0.11	0.08	0.06–0.09	9	7–11
	[65–75[	1769	0.7	0.32	0.30–0.34	0.07	0.06–0.08	0.04	0.03–0.05	0.03	0.03–0.04	4	3–5
Breast	[15–45[	11,215	3.3	0.99	0.98–0.99	0.89	0.88–0.89	0.80	0.79–0.80	0.72	0.71–0.74	73	71–74
	[45–55[	20,692	3.7	0.99	0.98–0.99	0.91	0.91–0.92	0.85	0.85–0.86	0.81	0.80–0.81	80	79–81
	[55–65[	20,714	3.6	0.98	0.98–0.98	0.91	0.90–0.91	0.84	0.83–0.85	0.79	0.78–0.80	79	78–79
	[65–75[	19,326	3.0	0.97	0.97–0.98	0.89	0.88–0.89	0.80	0.80–0.81	0.72	0.70–0.73	NC	NC
Thyroid	[15–45[	2539	4.1	1.00	0.99–1.00	1.00	0.99–1.00	0.99	0.99–1.00	0.99	0.98–1.00	99	98–100
	[45–55[	1855	5.0	1.00	0.99–1.00	0.99	0.99–1.00	0.99	0.98–1.00	0.99	0.97–1.00	99	97–100
	[55–65[	1453	3.8	0.99	0.98–1.00	0.99	0.97–0.99	0.97	0.96–0.99	0.97	0.94–0.98	97	94–98
	[65–75[	855	3.2	0.95	0.94–0.97	0.93	0.90–0.94	0.88	0.84–0.91	0.84	0.77–0.89	83	76–89

[a] Lost to follow up within 15 years after diagnosis.
[b] estimated with flexible parametric cure model NC: No statistical cure.

values of P(t): 90%, 95% and 99%.

4.2. Results

In **colorectal cancer** patients, the cure proportion P ranged between 47% in the oldest men to 62% in the youngest women (Table 1). A woman of any age diagnosed with colorectal cancer and still alive about nine years later had > 95% chance of being cured (Fig. 1). Table 2 shows that, similarly, the survival of the uncured group reached 0.05 before 10 years and 5-year CNS reached 0.95 for women of all age groups. For women, time-to-cure estimates by the three approaches were coherent (< 2 years difference); they ranged between 8 and 9 years depending on age. All estimates of time-to-cure were within a year of 10 years after diagnosis in men with colorectal cancer.

In both men and women with **pancreatic cancer**, P was very low, ranging between 4% and 16% (Table 1). Regardless of age, a pancreatic cancer patient had > 95% chance of being statistically cured past 9 years after diagnosis in women and eight years in men (Fig. 2). Table 2 shows that, similarly, 5-year CNS reached 0.95 after 8 years in men and around 9 years in women. The values of TTC and T_{CNS} were very close (< 1 year difference) and ranged between 8 and 10 years depending on age and sex. However, the survival of the uncured presented a very sharp decrease and reached 0.05 before 4 years in men of all age groups and before 6 years in women of all age groups. The values of T_{95} therefore ranged between 3 and 5 years. The discrepancy between the T_{95} and the other two estimates is important. The values of P(t) and CNS(t) at $t = T_{95}$ were far from 0.95 (range: 0.47–0.81 depending on age and sex). For example, for a woman aged [65-75[and still alive at $T_{95} = 3.36$ years, the probability of belonging to the cured group is lower than the probability of belonging to the uncured group (0.47 vs 0.53).

Cure assumption was accepted for women with **breast cancer** under 65 years at diagnosis. The cure fraction P was high, ranging between 73% and 80% (Table 1). P(t) exceeded 95% eleven years after diagnosis in patients over 45 and twelve years after diagnosis in younger patients (Fig. 3), 5-year CNS did not reach 0.95 within 10 years after diagnosis and $S_u(t)$ reached 0.05 around 14 years (Table 2). The estimates of the time-to-cure by T_{95} and TTC therefore differed by a few years: TTC ranged between 10.9 and 12.1 years depending on age whereas T_{95} was rather stable around 13.8 years (Table 2). T_{CNS} was not reached within 10 years after diagnosis.

In the younger **thyroid cancer** patients (aged < 45 in men and < 65 in women), the cure proportions were > 95% (Table 1), hence the P(t)s were > 95% just after diagnosis, they even reached 99% before five years of follow-up (Fig. 4). Table 2 shows that, in these age groups, the estimated TTC was 0. Similarly, the 5-year CNS was > 95% right after diagnosis, which resulted in $T_{CNS} = 0$. $S_u(t)$ decreased slowly until reaching 0.05 after 10 years. The estimated T_{95} was nearly 11 years in men and 13 years in women. The discrepancy between T_{95} and the two other estimates is major. In the other age groups, the probability of belonging to the cured group P(t) at $t = T_{CNS}$ was 0.90 in men and 0.87 in women. T_{95} was also around 11 years in men and 13 years in women but the other two estimates were very different: the T_{CNS} was much lower than the TTC in men aged [55–65[(2 vs. 7.65 years) and women

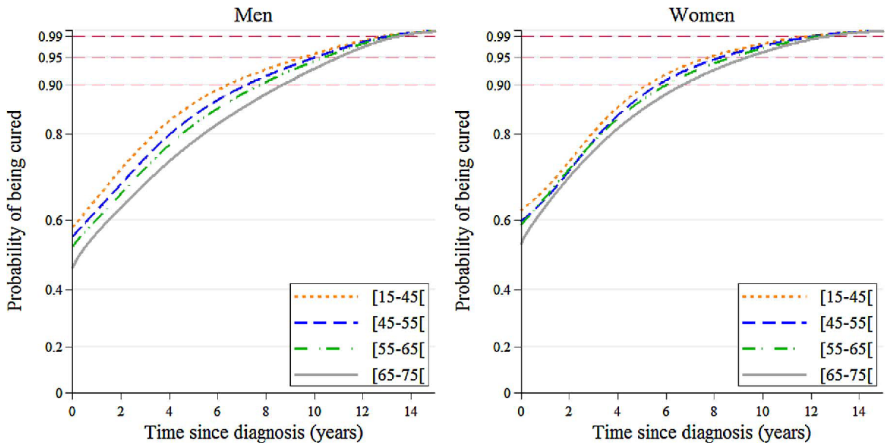


Fig. 1. Colorectal cancer. Estimated covariate-specific probability of being cured, conditional on survival to time t since diagnosis P(t) vs time since diagnosis (years). Horizontal dashed lines represent P(t) = 0.90, 0.95 and 0.99. All curves estimated separately for men and women by a flexible model with cure assumption for four groups of age at diagnosis (France, FRANCIM data, 1995–2010).

aged [65–75[(1 vs. 9.89 years). Indeed, in those 2 groups, the 5-year CNS reached 0.95 early then decreased before increasing again to 0.95.

Figs. 1–4 depict the changes in TTC estimation for 3 different cut-off values: 90, 95% and 99%. Results are detailed in Supplementary file 2, Table S1. When the cut-off ranged between 90 and 99%, the variation in TTC was approximately 6 years for colorectal cancer and 4 years for pancreatic cancer in both men and women. In women with breast cancer the variation in TTC was approximately 5 years in both oldest and youngest groups and 7 years in the intermediate groups. In thyroid cancer, the variation in TTC was heterogeneous and in women under 45 years, TTC was null for all cut-offs.

The proportions of cured patients P and the estimates of the time-to-cure by TTC with cut-off 95% are plotted in Fig. 5. The most favourable setting for patients is that of some thyroid cancer patients (men < 45 and women < 65 years): the cure fraction was > 95% and the time-to-cure was zero. In the remaining thyroid cancer patients, the proportions of cure were high (63 to 91%) but the TTC ranged widely from < 2 years to nearly 10 years. Breast cancer patients showed high cure rates but the longest TTC (nearly 11 years). In colorectal and pancreatic cancers, the TTC ranged between 8 and 11 years. However, whereas the cure fraction ranged from 47 to 62% in colorectal cancer, it was very low (under 10%) in pancreatic cancer patients of all age groups, except women < 55 years (16%). This calls into question the meaning of cure in pancreatic cancer. In patients with colorectal cancer, in women with pancreatic cancer and in men with thyroid cancer, the cure fraction decreased and the time-to-cure increased as the age increased.

5. Discussion

The present study proposes a new definition for the time-to-cure. Defined as the time at which the probability of belonging to the cured group exceeds 0.95, the proposed definition of time-to-cure (TTC) is a natural and intuitive definition for the time-to-cure from cancer. Indeed, the probability of being cured conditional to being alive is a direct translation of the notion of statistical cure. Moreover, P(t) is an interpretation of the asymptotic behaviour of the net survival curve in presence of cure and measures directly the proximity between the net survival curve and the cure fraction.

Using a real dataset, the proposed definition is compared to two existing definitions, which are based on the distribution of deaths from cancer (T_{95}) and on the conditional net survival (T_{CNS}). TTC has three main practical advantages over T_{95} and T_{CNS}:

First, a high early excess mortality may lead to an early sharp decrease in the net survival and, therefore, in the survival of the uncured. T_{95} is very influenced by the early excess mortality because it depends only on the survival of the uncured. Conceptually, time-to-cure should not be influenced by the mortality occurring shortly after diagnosis, during the peak of mortality due to cancer, but by the mortality occurring in the long term. By construction of T_{95}, a high early excess mortality brings the estimation of time-to-cure by T_{95} closer to diagnosis than a lower early excess mortality for an equivalent long term excess mortality. In contrast, TTC and T_{CNS} that depend on both the cure proportion and the survival of the uncured are less influenced by early excess mortality, which is a strong advantage. The discrepancy between T_{95} and both TTC and T_{CNS} observed in the illustrative pancreatic cancer data may be explained by the fact that the excess mortality rate being very high soon after diagnosis, $S_n(t)$ decreases very rapidly over time. By definition, $S_u(t) \leq S_n(t)$; therefore, $S_u(t)$ decreases also very rapidly and soon reaches 0.05 although the excess mortality rate is still non-negligible, and cure is therefore not reached yet. T_{95} is therefore an early estimate of the time-to-cure in pancreatic cancer. In young thyroid cancer patients, the excess mortality rate being very low since diagnosis and decreasing slowly, $S_u(t)$ decreases very slowly over time and only reaches 0.05 after a long time (nine years), P is high and very few patients belong to the uncured group. In this case the discrepancy between T_{95} and the both TTC and T_{CNS} is major. The more extreme are the excess mortality rates just after diagnosis, the more important is the difference between T_{95} and either TTC or T_{CNS}. The three estimates are very similar for colorectal and breast cancer, which present low to moderate excess mortality rates, but are very different for thyroid and pancreatic cancers which present extreme excess mortality rates.

Second, as shown by the example of thyroid cancer, T_{CNS} may underestimate the time-to-cure when the net survival curve flattens (temporary plateau) before decreasing again. The definition of T_{CNS} assumes that once the CNS reaches 0.95, it cannot decrease below this value. However, by construction, the CNS is not constrained to be an increasing function of time: the derivative function of the CNS shows that the monotony assumption implies that the excess hazard function has to be decreasing over time. When this assumption is violated, T_{CNS} may not correspond to the time-to-cure. Being calculated as $P/S_n(t)$ instead of $S_n(t+5)/S_n(t)$, P(t) has the advantage of being an increasing

Table 2
For each cancer site and age class (FRANCIM data), estimated value of time-to-cure t according to 3 definitions (TTC, T95 and TCNS), probability of belonging to the cured group P(t), proportion of uncured patients still alive Su(t) and 5-year conditional net survival CNS(t).

Cancer site/Age	Men					Women				
	Definition of time-to-cure t	t	P(t)	$S_u(t)$	CNS(t)[a]	Definition of time-to-cure t	t	P(t)	$S_u(t)$	CNS(t)[a]
Colon and Rectum										
[15–45[	TTC	9.84	*0.95*	0.07	0.95	TTC	8.29	*0.95*	0.08	0.95
	T_{95}	10.77	0.96	*0.05*	0.95	T_{95}	9.81	0.97	*0.05*	0.97
	T_{CNS}	10	0.95	0.07	*0.95*	T_{CNS}	8	0.95	0.09	*0.95*
[45–55[	TTC	10.09	*0.95*	0.06	0.95	TTC	8.55	*0.95*	0.08	0.96
	T_{95}	10.75	0.96	*0.05*	0.95	T_{95}	9.74	0.97	*0.05*	0.97
	T_{CNS}	10	0.95	0.07	*0.95*	T_{CNS}	9	0.96	0.06	*0.96*
[55–65[	TTC	10.31	*0.95*	0.06	0.95	TTC	8.47	*0.95*	0.07	0.95
	T_{95}	10.68	0.96	*0.05*	0.95	T_{95}	9.64	0.97	*0.05*	0.97
	T_{CNS}	10	0.94	0.06	*0.95*	T_{CNS}	9	0.96	0.06	*0.96*
[65–75[	TTC	10.63	*0.95*	0.05	0.94	TTC	8.76	*0.95*	0.06	0.96
	T_{95}	10.50	0.95	*0.05*	0.94	T_{95}	9.40	0.96	*0.05*	0.96
	T_{CNS}	10	0.94	0.06	*0.94*	T_{CNS}	9	0.95	0.06	*0.96*
Pancreas										
[15–55[	TTC	8.15	*0.95*	0.01	0.95	TTC	8.43	*0.95*	0.01	0.94
	T_{95}	3.69	0.67	*0.05*	0.73	T_{95}	5.03	0.79	*0.05*	0.81
	T_{CNS}	9	0.97	0.00	*0.97*	T_{CNS}	9	0.96	0.01	*0.96*
[55–65[	TTC	8.40	*0.95*	0.00	0.94	TTC	8.98	*0.95*	0.01	0.95
	T_{95}	3.22	0.57	*0.05*	0.57	T_{95}	4.26	0.66	*0.05*	0.66
	T_{CNS}	9	0.96	0.00	*0.96*	T_{CNS}	9	0.95	0.00	*0.95*
[65–75[	TTC	8.34	*0.95*	0.00	0.94	TTC	9.48	*0.95*	0.00	0.94
	T_{95}	2.88	0.52	*0.05*	0.58	T_{95}	3.36	0.47	*0.05*	0.47
	T_{CNS}	9	0.96	0.00	*0.97*	T_{CNS}	10	0.96	0.00	*0.96*
Breast										
[15–45[						TTC	12.08	*0.95*	0.86	0.92
						T_{95}	13.75	0.98	*0.95*	0.92
						T_{CNS}	> 10			*0.92*
[45–55[						TTC	10.94	*0.95*	0.79	0.94
						T_{95}	13.79	0.99	*0.95*	0.94
						T_{CNS}	> 10			*0.94*
[55–65[						TTC	11.21	*0.95*	0.81	0.94
						T_{95}	13.78	0.99	*0.95*	0.94
						T_{CNS}	> 10			*0.94*
[65–75[						TTC	NC	NC	NC	NC
						T_{95}	NC	NC	NC	NC
						T_{CNS}	NC	NC	NC	NC
Thyroid										
[15–45[	TTC	0	*0.98*	0	0.99	TTC	0	*0.99*	0.01	1
	T_{95}	11.13	1	*0.05*	1	T_{95}	13.03	1	*0.05*	1
	T_{CNS}	0	0.98	0	*0.99*	T_{CNS}	0	0.99	0.01	*1*
[45–55[	TTC	1.68	*0.95*	0.45	0.98	TTC	0	*0.99*	0.01	0.99
	T_{95}	11.11	1	*0.05*	0.99	T_{95}	13.03	1	*0.05*	1
	T_{CNS}	1	0.94	0.35	*0.97*	T_{CNS}	0	0.99	0.01	*0.99*
[55–65[	TTC	7.65	*0.95*	0.78	0.96	TTC	0	*0.97*	0.01	0.99
	T_{95}	10.95	0.99	*0.05*	0.98	T_{95}	13.02	1	*0.05*	0.99
	T_{CNS}	2	0.9	0.5	*0.95*	T_{CNS}	0	0.97	0.01	*0.99*
[65–75[	TTC	9.69	*0.95*	0.91	0.96	TTC	9.89	*0.95*	0.74	0.95
	T_{95}	10.83	0.97	*0.05*	0.96	T_{95}	12.90	0.99	*0.05*	0.95
	T_{CNS}	10	0.96	0.92	*0.96*	T_{CNS}	1	0.87	0.27	*0.96*

[a] $CNS(t) = S_n(t + 5)/S_n(t)$ NC: No statistical cure.

function of time, therefore TTC is not sensitive to the temporary plateau effect. In most cancer sites, this assumption of monotony is respected leading to similar values of TTC and T_{CNS}.

Third, using net survival at t + 5 to calculate the CNS at t, estimating T_{CNS} requires a 5-year longer follow-up than the other two methods, which is a clear disadvantage. In the illustrative breast cancer data, T_{CNS} was not reached. Although the net survival was modelled, estimations were not extrapolated after 15 years, which was the maximal observed time. Therefore, the CNS could not be estimated after 10 years. Although $T_{CNS} > 10$ is not incompatible with values of TTC and T_{95}, it cannot be considered as informative. T_{CNS} could be obtained with net survival point estimates (without modelling), which is an advantage of this method.

These three methods rely on the choice of a cut-off value, which influences the estimated value of the time-to-cure.

Data from cancer registries are widely used to analyse differences in cancer survival between populations and the comparisons often consider the net survival at 5 or 10 years after diagnosis [25–27]. When the follow-up period is long enough and the net survival curves show a plateau, cure assumption is reasonable [23]. Decision of accepting or

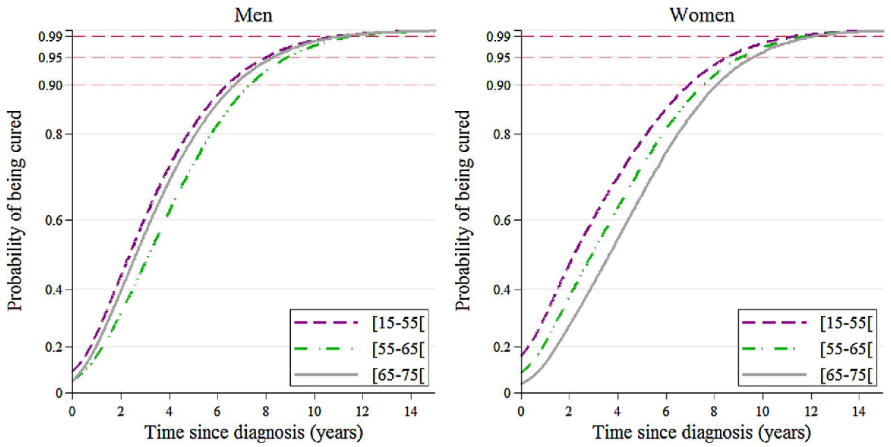


Fig. 2. Pancreatic cancer. Estimated covariate-specific probability of being cured, conditional on survival to time t since diagnosis P(t) vs time since diagnosis (years). Horizontal dashed lines represent P(t) = 0.90, 0.95 and 0.99. All curves estimated separately for men and women by a flexible model with cure assumption for four groups of age at diagnosis (France, FRANCIM data, 1995–2010).

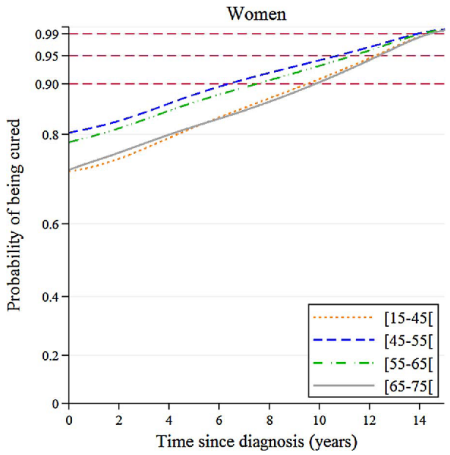


Fig. 3. Breast cancer. Estimated covariate-specific probability of being cured, conditional on survival to time t since diagnosis P(t) vs time since diagnosis (years). Horizontal dashed lines represent P(t) = 0.90, 0.95 and 0.99. All curves estimated separately for men and women by a flexible model with cure assumption for four groups of age at diagnosis (France, FRANCIM data, 1995–2010).

rejecting cure assumption based on graphical observation can be subjective. For instance, existence of statistical cure is still debated in breast cancer, some studies accepted cure assumption [18], others reported small but significant long term excess mortality [28]. In our study, although the plateau in net survival was not obvious, the excess mortality rate was very low. The reader can choose to reject it and discard the results presented in this article, or like the authors, choose to accept it for women under 65 years at diagnosis. When cure assumption is accepted, cure models can be fitted to provide useful indicators: i) the proportion of cured patients (i.e., those who will not die from cancer), which is an important information for practitioners and patients; ii) the median net survival time of patients who will die from cancer; iii) the estimated covariate-specific probability of being cured at any time since diagnosis; and, iv) the time-to-cure. As shown in the examples, estimating these indicators allows going beyond the classical 5-year survival. For instance, in breast cancer, the difference in 5-year net survival between age groups [15–45[and [45–55[is 3% (89 vs. 92%). Cure models provide two other indicators: the cure proportion is 7% lower in the youngest group (p = 73 vs. 80%) and the TTC is over a year later in the youngest (12.1 vs 10.9 years). Moreover, these indicators can be used to estimate relevant public health indicators such as the prevalence of cured and uncured patients (the latter is the disease burden requiring specialized care). They are also useful to set the time after which cancer should stop penalizing the access to bank credit or insurance. For example, already at diagnosis, a man aged < 55 or a woman aged < 65 with thyroid cancer has > 95% chance of belonging to the cured group; he/she should not pay extra insurance premiums.

Well-constructed cure models provide useful indicators for healthcare partners (improving communication) and policy makers (improving allocation of healthcare resources). The proposed new definition of time-to-cure, together with its confidence interval that could easily be computed, may be used to estimate the time after which people with history of cancer should access to insurance of bank loans without being overtaxed ("right to be forgotten") [29].

Funding

This work was supported by the Institut National du Cancer (INCa) [grant number 2014-087], by the Fondation ARC pour la recherche sur le cancer [personal grant for author O.B. number PDF20151203665] and by the French Government (Agence Nationale de la Recherche, "Investissements d'Avenir", ANR-11-LABX-0021). The funding sources had no involvement in the conduct of the research, the preparation of the article, the study design; the collection, analysis and interpretation of data, the writing of the report or in the decision to submit the article for publication.

Conflict of interest

None.

Authorship contribution statement

Olayidé Boussari contributed to the design of the study, the statistical analysis and the writing of the manuscript.

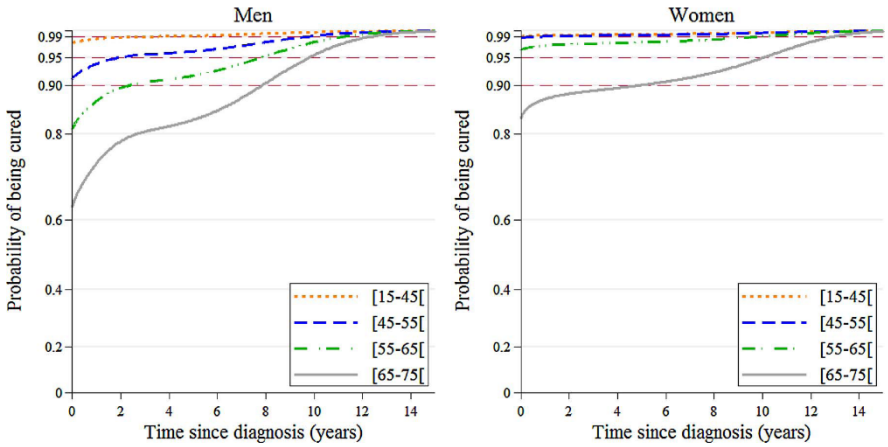


Fig. 4. Thyroid cancer. Estimated covariate-specific probability of being cured, conditional on survival to time t since diagnosis P(t) vs time since diagnosis (years). Horizontal dashed lines represent P(t) = 0.90, 0.95 and 0.99. All curves estimated separately for men and women by a flexible model with cure assumption for four groups of age at diagnosis (France, FRANCIM data, 1995–2010).

Gaëlle Romain contributed to the statistical analysis and the critical reviewing of the manuscript.

Laurent Remontet contributed to the design of the study and the critical reviewing of the manuscript.

Nadine Bossard contributed to the design of the study and the critical reviewing of the manuscript.

Morgane Mounier contributed to the statistical analysis and the critical reviewing of the manuscript.

Anne-Marie Bouvier contributed the critical reviewing of the manuscript.

Christine Binquet contributed the critical reviewing of the manuscript.

Marc Colonna contributed to the design of the study and the critical reviewing of the manuscript.

Valérie Jooste contributed to the design of the study the statistical analysis and the writing of the manuscript.

Acknowledgements

The authors wish to thank the French network of cancer registries (FRANCIM) for providing the data for the study, Stéphanie Normand (Registre Bourguignon des Cancers Digestifs, France) for her technical assistance and Jean Iwaz (Hospices Civils de Lyon, France) for revising the manuscript.

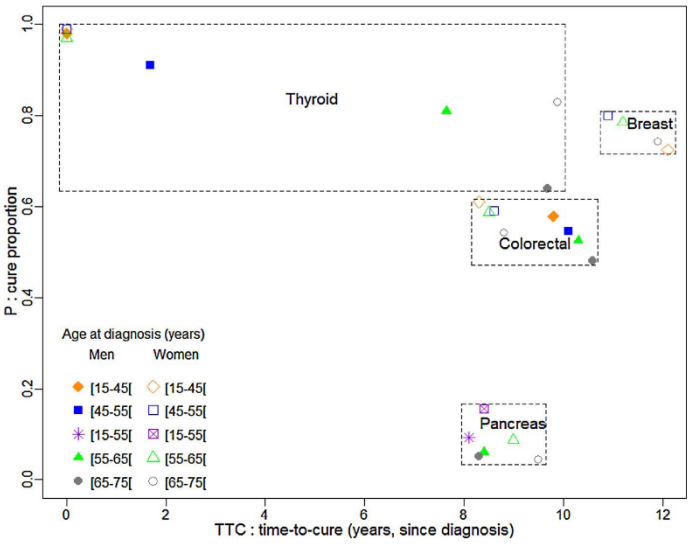


Fig. 5. Cure proportion vs. time to cure by sex and groups of age at diagnosis in patients diagnosed with colorectal, pancreatic, breast, or thyroid cancer (France, FRANCIM data, 1995–2010).

Appendix A. Description of Pi(t), the probability of belonging to the cured group

Let us denote by:

T the observed survival time in the study group (i.e., patients diagnosed with cancer) since diagnosis.

Y the indicator of cure that takes value 1 in case of cure and 0 otherwise (thus, $Prob(Y = 1) = P$ is the proportion of cured patients and $Prob(Y = 0) = 1 - P$ is the proportion of uncured patients).

S_o the observed survival from diagnosis in the study group and λ_o the corresponding hazard.

S_n the net survival in the study group and λ_e the corresponding hazard (the excess hazard).

S_p the expected survival in the general population and λ_p the corresponding hazard.

We are interested in $P_i(t)$ the probability for a patient i or a group of patients with the same characteristics as i to be in the 'cured group' at time t after diagnosis knowing that he/she is alive at time t.

For a given patient i aged a_i at diagnosis:

$$\lambda_{o,i}(t) = \lambda_{p,i}(t + a_i) + \lambda_{e,i}(t)$$

This implies that:

$$S_{o,i}(t) = S_{p,i}(t + a_i) \times S_{n,i}(t)$$

Knowing that $S_{n,i}(t) = P_i + (1 - P_i) \times S_{u,i}(t)$ where P_i denotes the proportion of cured patients and $S_{u,i}$ denotes the net survival function of the uncured patients in the group of patients that share the same characteristics as i, the observed survival may be written:

$$S_{o,i}(t) = S_{p,i}(t + a_i) \times [P_i + (1 - P_i) \times S_{u,i}(t)]$$

By definition $P_i(t) = Prob(Y_i = 1|T_i \geq t)$ and the second part of this equation may be obtained as follows:

$$Prob(Y_i = 1|T_i \geq t) = \frac{Prob(Y_i = 1, T_i \geq t)}{Prob(T_i \geq t)} = \frac{Prob(T_i \geq t|Y_i = 1) \times Prob(Y_i = 1)}{Prob(T_i \geq t)}$$

In the latter expression,

$Prob(T_i \geq t) = S_{o,i}(t)$; $Prob(Y_i = 1) = P_i$; $Prob(T_i \geq t|Y_i = 1) = S_{p,i}(t + a_i)$ (indeed, because $Y_i = 1$, i has no excess mortality due to cancer).

It follows that:

$$Prob(Y_i = 1|T_i \geq t) = \frac{S_{p,i}(t + a_i) \times P_i}{S_{o,i}(t)} = \frac{S_{p,i}(t + a_i) \times P_i}{S_{p,i}(t + a_i) \times S_{n,i}(t)} = \frac{P_i}{S_{n,i}(t)}$$

Hence:

$$Prob(Y_i = 1|T_i \geq t) = \frac{P_i}{P_i + (1 - P_i) \times S_{u,i}(t)}$$

And finally:

$$P_i(t) = \frac{P_i}{P_i + (1 - P_i) \times S_{u,i}(t)}$$

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.canep.2018.01.013.

References

[1] J. Boag, Maximum likelihood estimates of the proportion of patients cured by cancer therapy, J. R. Stat. Soc. Series B Stat. Methodol. 11 (1949) 15–53.

[2] J. Berkson, R. Gage, Survival curve for cancer patients following treatment, J. Am. Stat. Assoc. 47 (1952) 501–515.

[3] A. Yakovlev, B. Asselain, V. Bardou, A. Fourquet, T. Hoang, A. Rochefediere, A simple stochastic model of tumor recurrence and its application to data on pre-menopausal breast cancer, Biométrie et analyse de données spatio-temporelles 1993 (2016) 66–82.

[4] A. Yakovlev, A. Tsodikov, Stochastic models of tumor latency and their biostatistical applications, World Sci. 1996 (2016) 288.

[5] A.Y. Yakovlev, Parametric versus non-parametric methods for estimating cure rates based on censored survival data, Stat. Med. 13 (9) (1994) 983–986.

[6] J. Esteve, E. Benhamou, M. Croasdale, L. Raymond, Relative survival and the estimation of net survival: elements for further discussion, Stat. Med. 9 (5) (1990) 529–538.

[7] T. Hakulinen, L. Tenkanen, Regression analysis of relative survival rates, Appl. Stat. 36 (1987) 309–317.

[8] K.A. Cronin, E.J. Feuer, Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival, Stat. Med. 19 (13) (2000) 1729–1740.

[9] P.C. Lambert, P.W. Dickman, M.J. Rutherford, Comparison of different approaches to estimating age standardized net survival, BMC Med. Res. Methodol. 15 (2015) 64.

[10] T. Bejan-Angoulvant, A.M. Bouvier, N. Bossard, et al., Hazard regression model and cure rate model in colon cancer relative survival trends: are they telling the same story? Eur. J. Epidemiol. 23 (4) (2008) 251–259.

[11] A. Verdecchia, R. De Angelis, R. Capocaccia, et al., The cure for colon cancer: results from the EUROCARE study, Int. J. Cancer 77 (3) (1998) 322–329.

[12] T.M. Andersson, P.C. Lambert, A.R. Derolf, et al., Temporal trends in the proportion cured among adults diagnosed with acute myeloid leukaemia in Sweden 1973–2001, a population-based study, Br. J. Haematol. 148 (6) (2010) 918–924.

[13] R. De Angelis, R. Capocaccia, T. Hakulinen, B. Soderman, A. Verdecchia, Mixture models for cancer survival analysis: application to population-based data with covariates, Stat. Med. 18 (4) (1999) 441–454.

[14] J.W. Gamel, E.A. Weller, M.N. Wesley, E.J. Feuer, Parametric cure models of relative and cause-specific survival for grouped survival times, Comput. Methods Programs Biomed. 61 (2) (2000) 99–110.

[15] P.C. Lambert, J.R. Thompson, C.L. Weston, P.W. Dickman, Estimating and modeling the cure fraction in population-based cancer survival analysis, Biostatistics 8 (3) (2007) 576–594.

[16] R. Sposto, Cure model analysis in cancer: an application to data from the Children's Cancer Group, Stat. Med. 21 (2) (2002) 293–312.

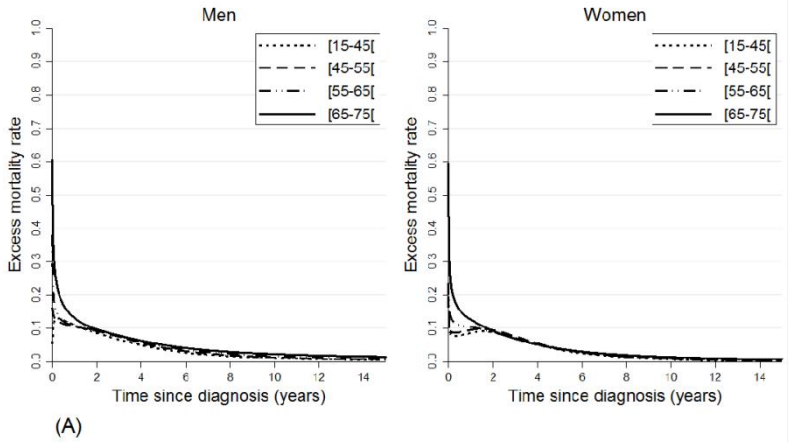
[17] M. Chauvenet, C. Lepage, V. Jooste, V. Cottet, J. Faivre, A.M. Bouvier, Prevalence of patients with colorectal cancer requiring follow-up or active treatment, Eur. J. Cancer 45 (8) (2009) 1460–1465.

[18] L. Dal Maso, S. Guzzinati, C. Buzzoni, et al., Long-term survival, prevalence, and cure of cancer: a population-based estimation for 818 902 Italian patients and 26 cancer types, Ann. Oncol. 25 (11) (2014) 2251–2260.

[19] T.M. Andersson, P.W. Dickman, S. Eloranta, P.C. Lambert, Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models, BMC Med. Res. Methodol. 11 (2011) 96.

[20] P. Lambert, P. Dickman, C. Weston, J. Thompson, Estimating the cure fraction in population-based cancer studies by using finite mixture models, J. R. Stat. Soc. C-Appl. Stat. 59 (2010) 35–55.
[21] A. Cowppli-Bony, Z. Uhry, L. Remontet, A.-V. Guizard, N. Voirin, A. Monnereau, Survie des personnes atteintes de cancer en France métropolitaine, 1989–2013, Partie 1–Tumeurs Solides, Institut de veille sanitaire, Saint-Maurice, 2016, p. 274.
[22] C.P. Nelson, P.C. Lambert, I.B. Squire, D.R. Jones, Flexible parametric models for relative survival, with application in coronary heart disease, Stat. Med. 26 (30) (2007) 5486–5498.
[23] X.Q. Yu, R. De Angelis, T.M. Andersson, P.C. Lambert, D.L. O'Connell, P.W. Dickman, Estimating the proportion cured of cancer: some practical advice for users, Cancer Epidemiol. 37 (6) (2013) 836–842.
[24] M.P. Perme, J. Stare, J. Esteve, On estimation in relative survival, Biometrics 68 (1) (2012) 113–120.
[25] C. Allemani, H.K. Weir, H. Carreira, et al., Global surveillance of cancer survival 1995–2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2), Lancet 385 (9972) (2015) 977–1010.
[26] R. De Angelis, M. Sant, M.P. Coleman, et al., Cancer survival in Europe 1999–2007 by country and age: results of EUROCARE–5-a population-based study, Lancet Oncol. 15 (1) (2014) 23–34.
[27] S. Rossi, P. Baili, R. Capocaccia, et al., The EUROCARE-5 study on cancer survival in Europe 1999–2007: database, quality checks and statistical analysis methods, Eur. J. Cancer (2015).
[28] M.L.G. Janssen-Heijnen, L.N. van Steenbergen, A.C. Voogd, V.C.G. Tjan-Heijnen, P.H. Nijhuis, P.M. Poortmans, J.W.W. Coebergh, D.J. van Spronsen, Small but significant excess mortality compared with the general population for long-term survivors of breast cancer in the Netherlands, Ann. Oncol. 25 (1) (2014) 64–68.
[29] Plan Cancer 2014–2019, guérir et prévenir les cancers: donnons les mêmes chances à tous, partout en France. Documents institutionnels/Plan cancer ed: Institut National du Cancer; 2014. English summary available at: http://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/Summary-Plan-cancer-2014-2019.

Graphical checking of the cure assumption and validation of the net survival estimates by the cure model



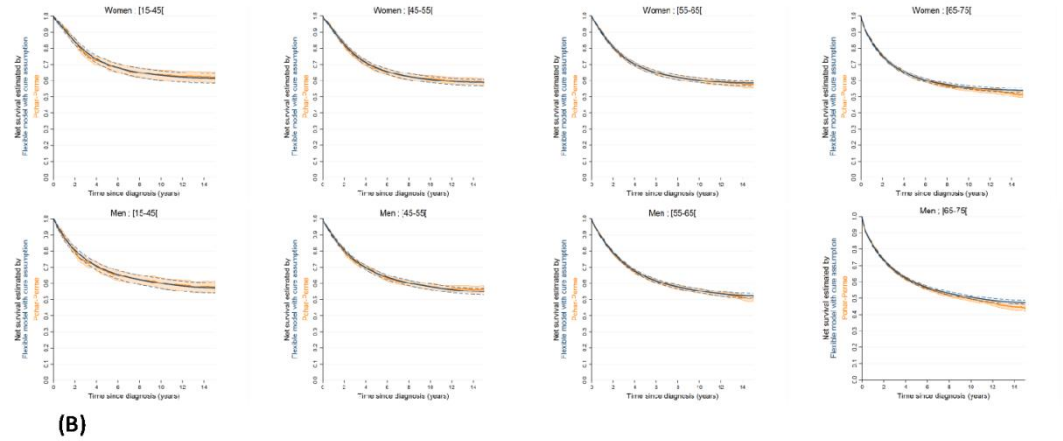
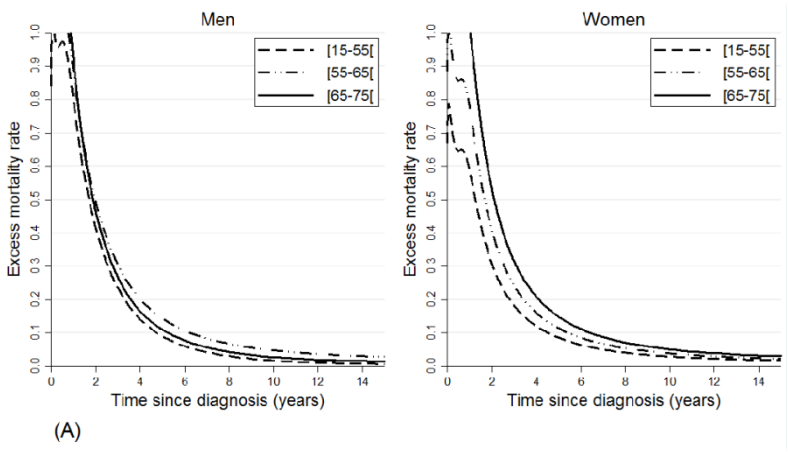
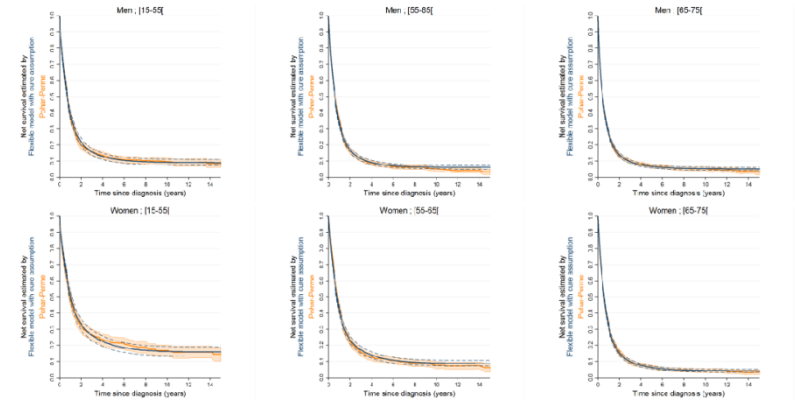


Figure S1 - Colorectal cancer. Panels A: Excess mortality rates as estimated by a flexible model without cure assumption. Panels B: Net survivals as estimated by a flexible model with cure assumption and Pohar-Perme approach. The curves concern four age categories of patients diagnosed with colorectal cancer (France, FRANCIM data, 1995 to 2010).

Supplementary file 2 from "A new approach to estimate time-to-cure with application to real cancer data from the French network of cancer registries (FRANCIM)" by Boussari *et al.*

Figure V.4 – Courbes pour la vérification de l'hypothèse de guérison dans le cancer colorectal. (A) Taux de mortalité en excès ; (B) Survie nette





(B)

Figure S2 - Pancreatic cancer. Panels A: Excess mortality rates as estimated a by a flexible model without cure assumption. Panels B: Net survivals as estimated by a flexible model with cure assumption and Pohar-Perme approach. The curves concern three age categories of patients diagnosed with pancreatic cancer (France, FRANCIM data, 1995 to 2010).

Figure V.5 - Courbes pour la vérification de l'hypothèse de guérison dans le cancer du pancréas. (A) Taux de mortalité en excès ; (B) Survie nette

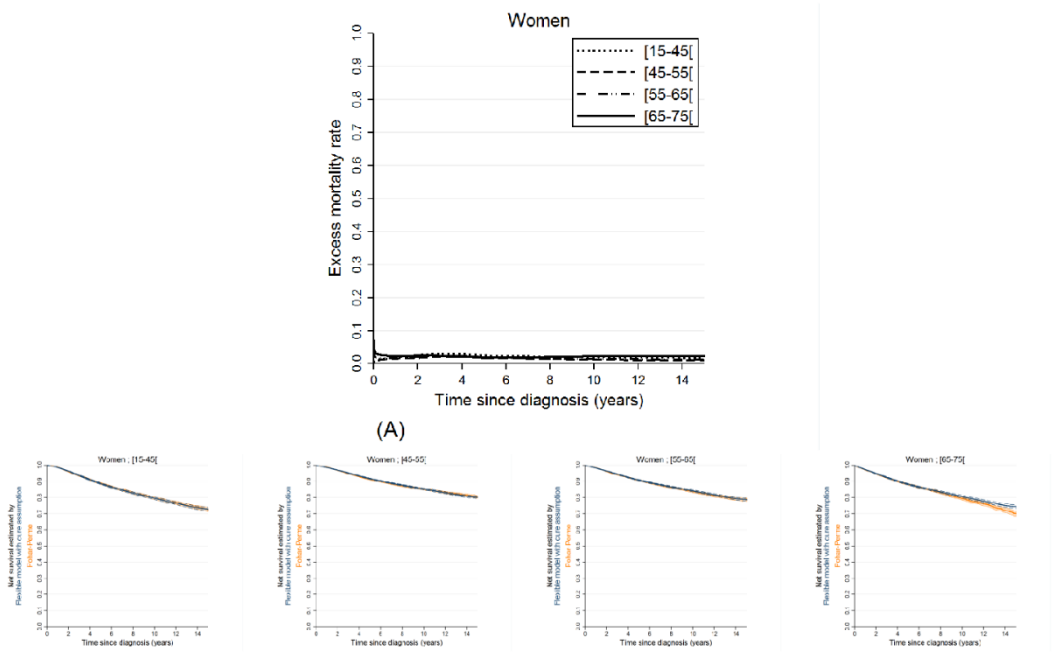


Figure S3 - Breast cancer. Panel A: Excess mortality rates as estimated by a flexible model without cure assumption. Panel B: Net survivals as estimated by flexible model with cure assumption and Pohar-Perme approach. The curves concern four age categories of women diagnosed with breast cancer (France, FRANCIM data, 1995 to 2010).

Figure V.6 - Courbes pour la vérification de l'hypothèse de guérison dans le cancer du sein. (A) Taux de mortalité en excès ; (B) Survie nette

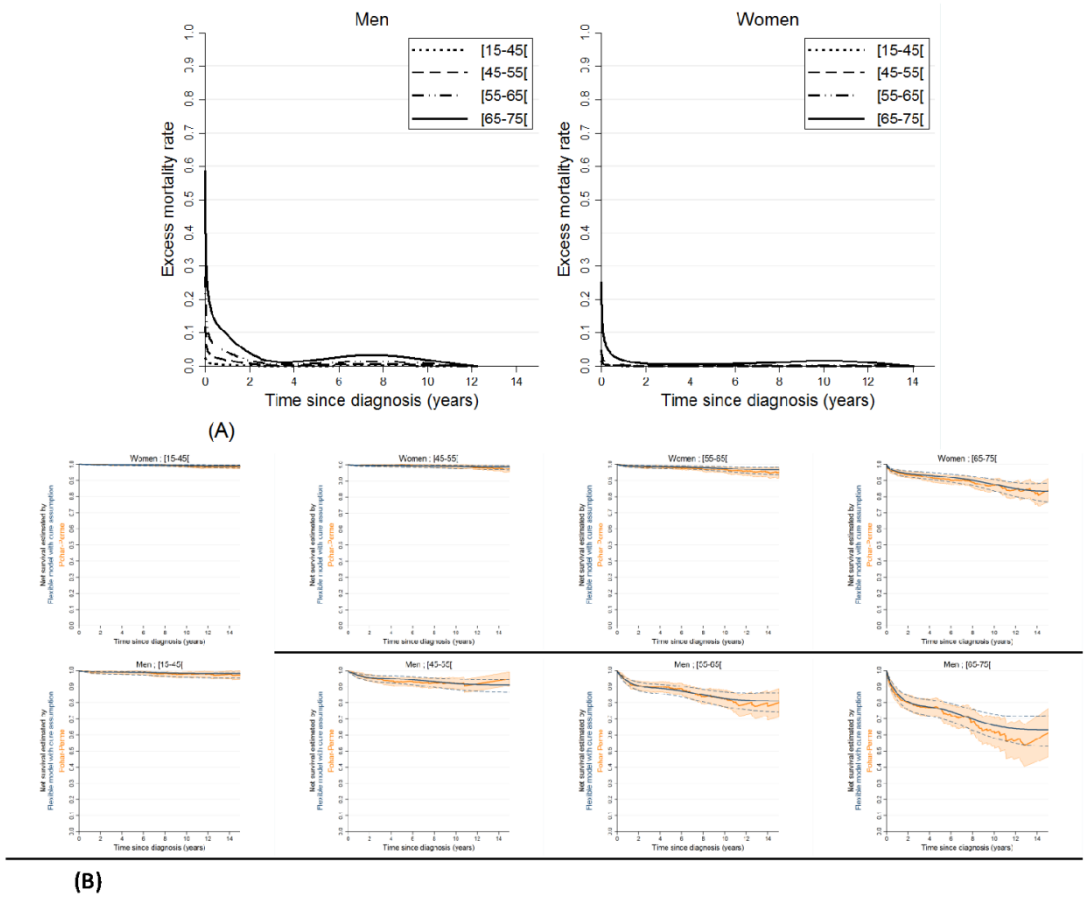


Figure S4 - Thyroid cancer. Panels A: Excess mortality rates as estimated by a flexible model without cure assumption. Panels B: Net survivals as estimated by a flexible model with cure assumption and Pohar-Perme approach. The curves concern four age categories of in patients diagnosed with thyroid cancer (France, FRANCIM data, 1995 to 2010).

Supplementary file 2 from "A new approach to estimate time-to-cure with application to real cancer data from the French network of cancer registries (FRANCIM)" by Boussari *et al.*

Figure V.7 - Courbes pour la vérification de l'hypothèse de guérison dans le cancer de la thyroïde. (A) Taux de mortalité en excès ; (B) Survie nette

Tableau V-1 : Délai de guérison TTC estimé pour différent seuil de la probabilité d'être guéris au cours du temps (90%, 95% et 99%)

Table S1: For each cancer site and age class (FRANCIM data), Time to cure estimated by TTC for different cut-offs (90%, 95% and 99%)

Cancer site	Sex	Age at diagnosis	Cure proportion P	Time-to-cure (TTC)		
				90% cut off	95% cut off	99% cut off
Colon and rectum	Men	[15-45[	0.58	6.38	9.34	12.84
		[45-55[	0.55	7.16	9.88	13.03
		[55-65[	0.52	7.73	10.27	13.18
		[65-75[	0.46	8.65	10.88	13.44
	Women	[15-45[	0.62	5.20	7.58	11.84
		[45-55[	0.59	5.64	8.08	12.09
		[55-65[	0.58	5.99	8.54	12.34
		[65-75[	0.53	6.65	9.19	12.65
Pancreas	Men	[15-55[	0.09	6.38	7.89	10.9
		[55-65[	0.05	7.32	8.78	11.59
		[65-75[	0.05	6.68	8.17	11.12
	Women	[15-55[	0.15	6.92	8.43	11.32
		[55-65[	0.08	7.52	8.97	11.74
		[65-75[	0.04	8.07	9.48	12.11
Breast	Women	[15-45[	0.72	9.46	12.13	14.35
		[45-55[	0.80	6.31	10.55	13.93
		[55-65[	0.78	7.43	11.2	14.08
		[65-75[	0.72	9.78	12.27	14.4
Thyroid	Men	[15-45[	0.97	0	0	2.86
		[45-55[	0.91	0	1.67	9.74
		[55-65[	0.81	2.30	7.65	11.21
		[65-75[	0.63	7.85	9.69	12.34
	Women	[15-45[	0.99	0	0	0
		[45-55[	0.98	0	0	0.59
		[55-65[	0.96	0	0	9.83
		[65-75[	0.83	4.86	9.88	12.91

Supplementary file 2 from "A new approach to estimate time-to-cure with application to real cancer data from the French network of cancer registries (FRANCIM)" by Boussari *et al.*

VI Estimation du délai de guérison en France dans le cancer

Ce chapitre est une application sur les données en base de population. Pour chaque localisation de cancer, j'ai vérifié l'hypothèse de guérison puis estimé la proportion de patients guéris et le délai de guérison lorsque la probabilité d'appartenir au groupe des patients guéris atteignait 95%.

VI.1 Population étudiée

Les données proviennent de la base de données FRANCIM incluant 30 localisations de tumeurs solides et 7 sous-localisations. Il s'agit d'une base de données dite de « basse résolution » car les seules informations fournies sont : la localisation du cancer (topologie et morphologie, d'après classification internationale des maladies pour l'oncologie, 3ème révision[69]), la date et le département de résidence au diagnostic, le sexe, la date de naissance, la date de dernière nouvelle et le statut vital associé du patient. Les patients inclus dans cette étude ont été diagnostiqué d'un cancer entre 15 et 74 ans sur la période 1995-2009. Les patients ont été suivis 15 ans ou jusqu'à la date de point au 30 juin 2013.

Afin de s'assurer un nombre d'évènement suffisant pour appliquer les modèles de guérison, les localisations de cancers présentant moins de 500 cas et moins de 200 décès ont été exclues. Au total, cette étude porte sur 22 localisations de cancer chez les hommes et 21 localisations chez les femmes, ce qui représente 393 095 cas de cancer. Les analyses ont été faites séparément pour chaque localisation de cancer et pour chaque sexe. L'effet de l'âge a été modélisé en catégories : [15-45[, [45-55[, [55-65[, et [65-75[ans. Les deux premières classes d'âge ont été regroupées pour onze localisations de cancer chez les hommes et sept chez les femmes. La classe d'âge de référence était celle dont l'effectif était le plus élevé.

VI.2 Méthodes

VI.2.1 Modélisation sans hypothèse de guérison

La survie nette a été estimée à partir d'un modèle paramétrique flexible de Nelson *et al.*[36, 38] (Section III.4.2.c). Les nœuds internes pour la RCS (spline cubique restreinte) du taux en excès cumulé de base ont été placés aux centiles 25, 50, 75 et 95 des temps de décès observés. Les nœuds externes ont été placés au premier centile des temps de décès observés et à 17 ans, afin que la queue de distribution ne soit pas influencée par les données observées (Tableau VI-1, modèle a.). Le modèle tient compte de l'effet non proportionnel (NP) de l'âge en catégories. La position des nœuds pour la RCS de l'effet NP de l'âge a été définie aux centiles 33 et 67 des temps de décès observés. En cas de non-convergence de ce modèle (due à un modèle trop complexe), le nombre et la position des nœuds ont été modifiés, jusqu'à obtenir un modèle stable (Tableau VI-1, modèles b. à e.). La non-proportionnalité de l'effet de l'âge était alors testée à partir du modèle retenu en utilisant le test du rapport de vraisemblance.

Tableau VI-1 – Positions des nœuds pour les splines cubiques restreintes (RCS) du modèle paramétrique flexible de survie nette.

	Sans hypothèse de guérison	
	Position des nœuds pour la RCS du taux de base	Position des nœuds pour la RCS de l'effet NP de l'âge
(a.)	**p25 p50 p75 p95**	**p33 p67**
(b.)	p25 p50 p75 p95	p50
(c.)	p33 p67 p95	p50
(d.)	p50 p95	p50
(e.)	p50	p50

VI.2.2 Modélisation avec hypothèse de guérison

La survie nette a été estimée par le modèle de guérison paramétrique flexible de non-mélange qui dérive du modèle décrit précédemment[56]. La stratégie de modélisation est la même que celle décrite précédemment. La seule différence est qu'un nœud interne sur la RCS du taux de base a été ajouté au centile 99 des temps de décès observés (Tableau VI-2).

Tableau VI-2 – Positions des nœuds pour les splines cubiques restreintes (RCS) du modèle de guérison paramétrique flexible de survie nette.

	Avec hypothèse de guérison	
	Position des nœuds pour la RCS du taux de base	Position des nœuds pour la RCS de l'effet NP de l'âge
(a.)	**p25 p50 p75 p95 p99**	**p33 p67**
(b.)	p25 p50 p75 p95 p99	p50
(c.)	p33 p67 p95 p99	p50
(d.)	p50 p95 p99	p50
(e.)	p50 p99	p50

VI.2.3 Vérification de l'hypothèse de guérison par sexe et par classes d'âge

A partir de l'observation graphique du taux en excès et de la survie nette estimée (Annexe A) à l'issue de l'étape 1, l'existence de guérison statistique était vérifiée selon les conditions présentées dans la section IV.3.

VI.2.4 Indicateurs de guérison estimés

Par sous-groupe de sexe et classe d'âge, pour les cancers vérifiant l'hypothèse de guérison, les indicateurs de guérisons estimés étaient : la proportion de guérison (π) et le délai de guérison (*TTC*) [68] (Section V.1.3).

La commande `stpm2` dans STATA™, version 14 (STATA corp., College Station, Texas) a été utilisée pour modéliser la survie nette sans puis avec l'option « `cure` ».

VI.3 Principaux résultats

L'ensemble des graphiques permettant la vérification de l'hypothèse de guérison (courbes du taux de mortalité en excès et courbes de survie nette estimées avec et sans l'hypothèse de guérison) ainsi que les courbes de la probabilité d'appartenir au groupe des guéris au cours du temps et la représentation graphique des proportions de guéris et des délais de guérison estimés sont présentées dans l'Annexe A (cette annexe correspond à

l'« appendix A » de l'article (Section VI.5). Pour deux localisations la proportion de patients guéris (π) est > 95% donc la guérison est atteinte dès le diagnostic (TTC=0) :

- Testicule chez les hommes âgés <55 ans : π=96% ;
- Thyroïde chez les patients âgés <45 ans : π=98% chez les hommes et 99% chez les femmes ; et chez les femmes âgées <65 ans : π=99% pour les 45-54 ans et π=97% pour les 55-64 ans.

Les résultats montrent $\pi \geq 80\%$ avec TTC ≤ 5 ans pour les localisations de cancer telles que :

- Mélanome de la peau chez les femmes âgées <54 ans et chez les hommes <64 ans : 4 ans < TTC < 5 ans ;
- Testicule chez les hommes âgés ≥55 ans : 1 ans < TTC < 3 ans ;
- Thyroïde chez les hommes âgés entre 45 et 54 ans : TTC=2 ans.

Pour les cancers du sein et de la prostate (les cancers les plus fréquents en France), le délai de guérison est estimé à plus de 10 ans bien que la proportion des patients guéris soit élevée (π>70%). Les localisations de mauvais pronostic telles que TTC>10 ans et π<10% sont l'œsophage et le poumon chez les hommes, le foie et le système nerveux central pour les deux sexes.

VI.4 Discussion

Le délai de guérison estimé à partir de la nouvelle définition est inférieur à 12 ans pour la plupart des localisations de cancer. Des disparités existent en fonction de l'âge au diagnostic : l'hypothèse de guérison est plus souvent rejetée chez les patients âgés, et avec l'augmentation de l'âge au diagnostic la proportion de patients guéris diminue et le délai de guérison augmente. Dans cette étude, les délais de guérison sont estimés à partir du diagnostic et nous n'avons pas tenu compte des traitements reçus ou non, cette information n'étant pas disponible. Or, le délai pour le droit à l'oubli est fixé à 10 ans à partir de la fin des traitements. Donc, les résultats montent que le droit à l'oubli pourrait donc être abaissé notamment pour les patients ayant un cancer dont le délai de guérison est inférieur à 10 ans suivant le diagnostic.

Les estimations fournies à partir de la nouvelle définition du délai de guérison sont concordantes avec l'étude italienne de Dal Maso *et al.*[67], sauf pour le cancer de la thyroïde, notamment pour les hommes de [55-65[ans et les femmes de [65-75[ans. Le taux de mortalité

en excès n'étant pas monotone décroissant au cours du temps (Annexe A), le délai de guérison, tel que défini dans l'étude italienne, ne peut donc pas être calculé pour ce groupe de patients (Section V.2).

La principale limite de tous les modèles de guérison est l'absence d'un test statistique permettant de tester objectivement l'hypothèse de guérison. Dans cette étude les critères pour l'acception de la guérison sont : taux de mortalité en excès ≤0,05 et adéquation des courbes de survie nette. Cependant, le seuil de 0,05 est un choix arbitraire et l'adéquation des courbes est une vérification graphique subjective. Par conséquent, différents critères auraient pu conduire à différentes sélections de localisation, de combinaisons de sexe et d'âge. De plus, le choix du seuil de $P(t) \geq 95\%$ pour estimer le délai *TTC* est également un choix arbitraire et les intervalles de confiance de *TTC* estimés sont très larges lorsque le nombre de cas incidents ou que la mortalité sont faibles.

Afin de pallier ces différentes limites un nouveau modèle de guérison a été développé au sein de l'équipe en intégrant le délai de guérison comme un paramètre à estimer : Section VII.1.

VI.5 Article

Cancer Epidemiology 60 (2019) 93–101

Contents lists available at ScienceDirect

Cancer Epidemiology

journal homepage: www.elsevier.com/locate/canep

Time-to-cure and cure proportion in solid cancers in France. A population based study

Gaëlle Romain[a,b,c,d], Olayidé Boussari[a,c,d,e], Nadine Bossard[f,g,h,i], Laurent Remontet[f,g,h,i], Anne-Marie Bouvier[a,b,c,d], Morgane Mounier[b,c,d,j], Jean Iwaz[f,g,h,i], Marc Colonna[k], Valérie Jooste[a,b,c,d,*], French Network of Cancer Registries (FRANCIM)[1]

[a] Registre Bourguignon des Cancers Digestifs, Dijon, France
[b] Centre Hospitalier Universitaire de Dijon Bourgogne, Dijon, France
[c] INSERM, LNC UMR1231, Dijon, France
[d] Université de Bourgogne Franche-Comté, Dijon, France
[e] Laboratoire d'Excellence LabEX LipSTIC, ANR-11-LABX-0021, Dijon, France
[f] Hospices Civils de Lyon, Service de Biostatistique-Bioinformatique, Lyon, France
[g] Université de Lyon, Lyon, France
[h] Université Lyon 1, Villeurbanne, France
[i] CNRS UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Équipe Biostatistique Santé, Pierre-Bénite, France
[j] Registre des Hémopathies Malignes de Côte d'Or, Dijon, France
[k] Registre du Cancer de l'Isère, Centre Hospitalier Universitaire Grenoble-Alpes, Grenoble, France

ARTICLE INFO

Keywords:
Statistical cure
Time-to-cure
Cure proportion
Cure models
Population based cancer registry
Survival
Epidemiology

ABSTRACT

Background: In cancer care, the cure proportion (P) and time-to-cure (TTC) are important indicators for practitioners, patients, and healthcare policy makers. The recent definition of TTC as the time at which the probability of belonging to the cured group reaches 95% was used for the first time.
Methods: The data stem from the common database of French cancer registries including 335,358 solid tumours diagnosed between 1995 and 2009 at 27 sites. P and TTC were estimated through a flexible parametric net survival cure model for each cancer site, sex, and age at diagnosis with acceptable assumption of cure (excess mortality rate ≤ 0.05).
Results: TTC was ≤ 5 years and P was $> 80\%$ for skin melanoma and thyroid and testis cancers. It was 0 for testis cancer in men < 55 and for thyroid cancer in men < 45 and women < 65. TTC was between 5 and 10 years for all digestive cancers except small intestine and all gynaecologic cancers except breast. It was ≥ 10 years in prostate, breast, and urinary tract. The range of P according to age and sex was 37–79% for urinary tract 72–88% for prostate and breast, 4–16% for pancreatic and 47–62% for colorectal cancer.
Conclusion: Time-to-cure was estimated for the first time from a large national database and individual probabilities of cure. It was 0 in the younger patients with testis or thyroid cancer and < 12 years in most cancer

* Corresponding author at: Registre Bourguignon des cancers digestifs, UFR des sciences de santé, BP 87900, F-21079, Dijon Cedex, France.
E-mail address: valerie.jooste@u-bourgogne.fr (V. Jooste).

[1] **French Network of Cancer Registries (FRANCIM):** Brice Amadeo (Registre des cancers de Gironde), Patrick Arveux (Registre des cancers du sein et des cancers gynécologiques de Côte-d'Or), Isabelle Baldi (Registre des tumeurs primitives du système nerveux de la Gironde), Simona Bara (Registre des cancers de la Manche), Anne- Marie Bouvier (Registre bourguignon des cancers digestifs), Véronique Bouvier (Registre des tumeurs digestives du Calvados), Jacqueline Clavel (Registre des hémopathies malignes de l'enfant), Marc Colonna (Registre du cancer de l'Isère), Gaëlle Coureau (Registre des cancers de Gironde), Anne Cowppli-Bony (Registre des tumeurs de Loire-Atlantique et Vendée), Tania Dalmeida (Registre général des cancers de Haute Vienne), Laetitia Daubisse-Marliac (Registre des cancers du Tarn), Gautier Defossez (Registre général des cancers de Poitou-Charentes), Patricia Delafosse (Registre du cancer de l'Isère), Jacqueline Deloumeaux (Registre général des cancers de Guadeloupe), Pascale Grosclaude (Registre des cancers du Tarn), Anne-Valérie Guizard (Registre général des tumeurs du Calvados), Clarisse Joachim (Registre général des cancers de Martinique), Brigitte Lacour (Registre National des Tumeurs Solides de l'Enfant), Bénédicte Lapôtre-Ledoux (Registre du cancer de la Somme), Emilie Marrer (Registre des cancers du Haut-Rhin), Marc Maynadié (Registre des hémopathies malignes de Côte d'Or), Florence Molinié (Registre des tumeurs de Loire-Atlantique et Vendée), Alain Monnereau (Registre des hémopathies malignes de la Gironde), Jean-Baptiste Nousbaum (Registre finistérien des tumeurs digestives), Juliette Plenet (Registre des Cancers de Guyane), Sandrine Plouvier (Registre général des cancers de Lille et de sa région),Camille Pouchieu (Registre des tumeurs primitives du système nerveux de la Gironde), Michel Robaszkiewicz (Registre finistérien des tumeurs digestives), Claire Schvartz (Registre des cancers thyroïdiens Marne-Ardennes), Brigitte Trétarre (Registre des tumeurs de l'Hérault), Xavier Troussard (Registre Régional des Hémopathies Malignes de Basse Normandie), Michel Velten (Registre des cancers du Bas-Rhin), Anne-Sophie Woronoff (Registre des tumeurs du Doubs et du Territoire de Belfort).

https://doi.org/10.1016/j.canep.2019.02.006
Received 3 September 2018; Received in revised form 21 January 2019; Accepted 7 February 2019

sites. These results should help improve access to credit and insurance for patients still alive past the estimated TTCs.

1. Introduction

Of the three million people living in France with a personal history of cancer [1], a large proportion will not die from cancer. Although clinical cure is difficult to evaluate, especially in large cohorts, statistical cure can be defined as the absence of death due to cancer. The proportion that will not die from cancer is therefore the cure proportion P (or cure fraction). P can be estimated using net survival cure models and cancer registry data. Up to now, cure models provided estimations of the cure proportion and of the whole net survival of fatal cases [2,3]. The time-to-cure is also of great interest for public health deciders, epidemiologists, physicians, and patients. Only one previous study has provided estimations of the time-to-cure using conditional net survival applied to Italian cancer registries data [4]. Boussari et al. have recently proposed a more intuitive definition of the time-to-cure, as the delay TTC after which the probability of belonging to the cured group (estimated from the cure proportion and the net survival) reaches a high pre-defined value [5].

In cancer studies, net survival is defined as the survival that would be observed if cancer were the only cause of death [6,7]. In large cohorts, the causes of death are not reliable; net survival is thus estimated using the excess mortality rate; i.e., the difference between the mortality rate observed in the cohort and the mortality rate expected in the general population that shares same socio-demographic characteristics [8,9]. When cure occurs, the excess mortality rate tends to zero; thus the net survival tends to the cure proportion [10]. The ratio between the cure proportion and the net survival is then the probability of belonging to the cured group [5,11].

As a follow-up of our previous study which proposed TTC as a definition of the time-to-cure and using net survival cure models, the aim of the present study was to estimate the cure proportion and the time-to-cure in France regarding solid cancer sites for which the assumption of statistical cure was deemed acceptable.

2. Patients and methods

2.1. Patients and main characteristics

The study concerned all solid tumours diagnosed January 1, 1995 to December 31, 2009 in patients aged 15–74 years (N = 335,358) and

Table 1
Results of checking the assumption of cure by age-group at diagnosis in men.

Cancer site	Number at diagnosis	Number at 5 years	Number at 10 years	Number deaths[a]	Percentage of loss to follow-up[a]	15–44 years	45–54 years	55–64 years	65–74 years
Bones	405	236	199	219	3.2	C	C	C	NC
Central nervous system	2925	614	479	2475	1.1	NC	C	NC	NC
Choroid[b]	281	194	161	123	1.1	–	–	–	–
Colon[c]	16029	9206	7769	8803	2.1	C	C	C	C
Colon and Rectum	27861	15959	13279	15538	2.0	C	C	C	C
Head and neck	14137	4829	3317	11297	1.0	NC	NC	NC	NC
Kidney	6183	4149	3580	2840	2.0	C	C	C	C
Lip[b]	400	327	276	167	4.5	–	–	–	–
Lung	30072	4620	3371	27052	0.8	C	C	NC	NC
Malignant neoplasm without specification of site[b]	132	33	30	103	4.3	–	–	–	–
Nasopharynx[c]	339	177	141	206	3.5	C	C	C	NC
Oral cavity[c]	3211	1305	908	2430	1.2	NC	NC	NC	NC
Rectum[c]	11832	6753	5510	6735	1.9	C	C	C	C
Skin melanoma	4571	3804	3508	1158	4.3	C	C	C	C
Stomach	5460	1415	1153	4397	1.8	C	C	C	C
Testis	3021	2864	2826	226	3.7	C	C	C	C
Thyroid	2060	1808	1708	381	2.6	C	C	C	C
Tissue sarcoma	1168	669	593	607	2.0	C	C	C	C
Tongue[c]	2624	913	658	2051	1.2	C	NC	NC	NC
						15-54 years		55-64 years	65-74 years
Biliary tract	1095	235	184	930	0.8	C		C	C
Bladder	7972	4343	3500	4823	2.1	C		C	C
Hypopharynx[c]	3531	960	597	3038	0.7	NC		NC	NC
Larynx	4347	2330	1800	2768	1.9	C		NC	NC
Liver	7165	1039	703	6498	0.9	C		NC	NC
Nasal cavity	562	284	220	360	1.4	C		C	C
Oropharynx[c]	3776	1322	913	2998	0.8	NC		NC	NC
Oesophagus	6047	839	530	5580	0.6	NC		NC	NC
Pancreas	4930	418	345	4607	1.1	C		C	C
Penis[b]	336	231	207	146	1.8	–		–	–
Pleural mesothelioma	592	32	15	579	0.7	NC		NC	NC
Prostate	52279	44567	39590	14736	2.8	C		C	C
Salivary glands[b]	368	220	190	192	2.2	–		–	–
Small intestine	710	359	296	435	1.7	C		C	NC
Urinary tract	722	324	268	472	1.9	C		C	C

C: cure; NC: no cure.
[a] At 15 years or on June 30, 2013.
[b] Not analysed because < 500 cases at diagnosis and < 200 deaths.
[c] Cancer sub-sites.

recorded by FRANCIM (the French network of cancer registries). FRANCIM data are checked for quality and completeness every four years by an independent audit committee (Comité d'Évaluation des Registres).

The considered cancer sites adopted the definitions of the International Classification of Diseases for Oncology, 3rd revision (ICD-O-3) [12]. To ensure sufficient numbers of events to fit the parametric flexible model that uses restricted cubic splines with potentially up to 20 parameters (see Methods section), cancer sites with < 500 cases and < 200 deaths were excluded from the analysis. These were choroid, lip, salivary gland cancers, malignant neoplasm without specification of site as well as penis cancer in men and bone, nasal cavity, pleural mesothelioma and urinary tract cancers in women. These exclusions left 27 sites and 7 subsites for analysis.

As in previous FRANCIM studies [13], age at diagnosis was considered in four groups: < 45, 45–54, 55–64, and 65–74 years. In 11 sites in men and 7 in women, the first two age groups had to be pooled because the number of cases was < 50. Patient vital status was followed over 15 years after diagnosis or up to June 30, 2013. The proportion of deceased patients was 49% and the proportion of patients lost to follow-up was 2.6% (Table 1 in men and Table 2 in women).

2.2. Methods

2.2.1. Estimation of net survival without assumption of cure

For each sex and Département (French administrative area), the expected mortality rates were derived from the general population mortality rates provided by the Institut National de la Statistique et des Études Économiques (INSEE). These rates were "observed" mortality rates (obtained simply by dividing the observed number of death by the corresponding number of person-years). These observed rates presented random (Poissonian) variation and were thus smoothed in order to obtain their expected values. This work was done by the Biostatistical unit of the Hospices Civils de Lyon, using *mgcv* package of R software.

The net survival was estimated using a flexible parametric model on the log cumulative excess hazard scale with restricted cubic spline function of log time [14,15]. In this model, the log cumulative excess hazard is forced to be linear beyond the boundary knots. The splines had four internal knots located at the 25th, 50th, 75th, and 95th percentiles of the observed death times and boundary knots located at the 1st percentile of the observed death times and at 17 years. A separate model was fitted for each cancer site and sex. Age was included in the model as a categorical variable. Its time-dependent effect was added for each category (splines with two internal knots located at the 33rd and 67th percentiles of the observed death times) and evaluated with a

Table 2
Results of checking the assumption of cure by age-group at diagnosis in women.

Cancer site	Number at diagnosis	Number at 5 years	Number at 10 years	Number deaths[a]	Percentage of loss to follow-up[a]	15–44 years	45–54 years	55–64 years	65–74 years
Bones[b]	323	214	197	131	3.1	–	–	–	–
Breast	71947	63289	58145	16125	3.4	C	C	C	C
Central nervous system	2148	510	435	1745	1.9	C	C	NC	NC
Cervix uteri	4407	3126	2918	1586	4.7	C	C	C	C
Choroid[b]	317	240	206	123	1.6	–	–	–	–
Colon[c]	11282	7129	6383	5182	3.8	C	C	C	C
Colon and Rectum	18510	11749	10454	8536	3.7	C	C	C	C
Corpus uteri	7315	5580	5118	2436	3.4	C	C	C	C
Head and neck	2397	1213	968	1503	1.9	C	NC	NC	NC
Kidney	2931	2188	1980	1051	3.8	C	C	C	C
Lip[b]	73	63	60	18	9.6	–	–	–	–
Lung	6785	1492	1223	5647	1.5	C	C	NC	NC
Malignant neoplasm without specification of site[b]	98	28	26	72	6.2	–	–	–	–
Nasopharynx[b,c]	109	67	61	51	4.6	–	–	–	–
Oral cavity[c]	708	413	323	416	2.3	NC	NC	NC	NC
Ovary	5753	2778	2256	3596	2.5	C	C	C	C
Rectum[c]	7228	4620	4071	3354	3.5	C	C	C	C
Skin melanoma	5466	5005	4794	783	6.2	C	C	C	C
Stomach	2264	779	672	1627	4.0	C	C	C	C
Thyroid	6702	6486	6342	448	4.2	C	C	C	C
Tissue sarcoma	849	528	473	397	4.7	C	C	C	C
Tongue[c]	605	310	262	360	2.8	C	NC	NC	NC
Vagina and Vulva	683	411	358	344	4.5	C	C	C	C
						15-54 years		55-64 years	65-74 years
Biliary tract	1084	224	194	899	2.1	C		C	C
Bladder	1233	655	571	707	3.6	C		C	C
Hypopharynx[b,c]	242	85	53	196	2.5	–		–	–
Larynx	419	250	217	224	2.4	C		NC	NC
Liver	1295	223	175	1125	1.8	C		C	C
Nasal cavity	165	82	69-	99	4.2	–		–	–
Oropharynx[b,c]	665	317	252	427	2.6	C		C	C
Oesophagus	862	169	126	745	1.6	NC		NC	NC
Pancreas	3239	357	286	2967	0.9	C		C	C
Pleural mesothelioma[b]	175	20	16	160	2.3	–		–	–
Salivary glands[b]	279	222	210	74	5.7	–		–	–
Small intestine	527	305	255	283	4.2	C		C	C
Urinary tract[b]	254	111	90	169	2.0	–		–	–

C: cure; NC: no cure.
[a] At 15 years or on June 30, 2013.
[b] Not analysed because < 500 cases at diagnosis and < 200 deaths.
[c] Cancer sub-sites.

likelihood ratio test with 0.05 as significance level. Net survival and excess mortality rate curves were drawn for each combination of cancer site, sex, and age group.

2.2.2. Estimation of net survival with assumption of cure

The net survival was estimated using the flexible parametric net survival cure model developed by Andersson et al. [16,17]. This model derives from the net survival model without assumption of cure [14]. The assumption of cure adds the constraint that the log cumulative excess hazard has to be constant (slope = 0) beyond the last boundary knot. The modelling strategy was identical to that without assumption of cure, except an additional internal knot located at the 99th percentile. The latter additional knot allowed an optimal estimation of net survival with the cure model.

2.2.3. Assumption of statistical cure

The assumption of statistical cure was assessed graphically by site, sex, and age group using the excess mortality rate modelled without assumption of cure. The estimated excess mortality rate was considered negligible in case of sustained value ≤0.05 and the net survival was modelled with assumption of cure. The adequacy of the estimates of net survival using flexible models with and without assumption of cure was evaluated by checking whether the curves overlap. In case of satisfactory adequacy, cure was considered acceptable and the cure proportion and time-to-cure were estimated.

2.2.4. Indicators of cure from cancer

Two cure indicators were estimated only for combinations of site, sex, and age for which the assumption of cure was accepted:

- The cure proportion (P) commonly provided in statistical cure studies [2–4,10] is the proportion of patients who will never die from the cancer under study. Its value ranges from 0 to 1.
- The time-to-cure was recently defined as TTC by Boussari et al. [5] using P(t), the individual probability for a patient with given characteristics (sex and age at diagnosis) of belonging to the cured group knowing that he/she is still alive at time t [11]:

$$P(t)=\frac{P}{S_n(t)}$$

with $S_n(t)$: the net survival estimated at t

Cure models assume that a given patient belongs to a given group (cured or uncured) since diagnosis. For $t = 0$, $S_n(t)=1$ thus $P(t) = P$. For $t \to +\infty$, $S_n(t) \to P$; thus $P(t) \to 1$. In other words, the probability P(t) that this patient belongs to the cured group is P at the time of diagnosis and increases as time goes by and the patient is still alive. We defined the TTC as TTC_95: i.e., the time t after diagnosis after which P(t) exceeds 95% [5]. The confidence intervals relative to P and TTC were calculated by the Delta method.

stpm2 command in STATA™, release 14 (STATA corp., College Station, Texas) was used to fit flexible parametric net survival models with and without cure option [17,18].

3. Results

Overall, 335 358 solid tumours from 22 cancer sites in men and 21 in women were included. The most frequent were breast (49% of women cancers), prostate (28% of men cancers), lung (16% in men and 5% in women) and colorectal cancer (15% in men and 13% in women).

The assumption of statistical cure for each cancer site, sex, and age-group combination was checked using the excess mortality rate and the net survival curves. The graphs used to check the assumption of statistical cure are presented in Appendix A and the results of the checks in Table 1 for men and Table 2 for women. The assumption of cure was rejected for: cancers of the oesophagus and the oral cavity in both sexes; larynx, lung, and tongue cancers in both sexes and age group 55–74; hypopharynx, pleural mesothelioma, and oropharynx cancers in men; bone, liver, and small intestine cancers in men aged 65–74; head and neck cancers in all men and in women aged 45–74; and central nervous system cancers in men aged 15–44 and men and women aged 55–74.

In cases with accepted assumption of cure, Tables 3 and 4 show, respectively, in men and women, P and TTC values (with their 95% CIs) and the lower left quadrant of Appendix A provides P(t). These results are described below according to three ranges of TTC: < 5 years, 5–10 years, and > 10 years. In Figs. 1–3, adapted from Verdecchia et al.

Table 3
Estimated values of the cure proportion (P) and time-to-cure (TTC) in men by cancer site and age at diagnosis.

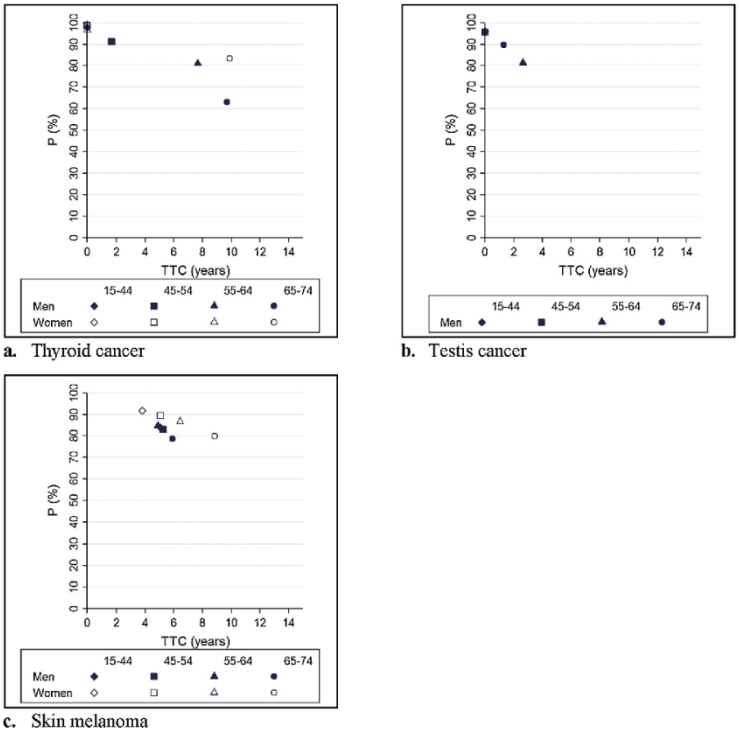
Cancer site	P (95% CI)	TTC (95% CI)	P (95% CI)	TTC (95% CI)	P (95% CI)	TTC (95% CI)	P (95% CI)	TTC (95% CI)
	15-44 years		45-54 years		55-64 years		65-74 years	
Bones	47 (39-54)	10.0 (0.6-19.5)	59 (40-73)	10.7 (0.0-22.6)	35 (19-50)	11.8 (0.0-26.0)	NC	NC
Central nervous system	NC	NC	15 (12-18)	10.7 (3.8-17.6)	NC	NC	NC	NC
Colon[a]	59 (55-64)	9.2 (4.3-14.1)	58 (55-60)	9.4 (6.6-12.3)	55 (53-57)	10.0 (8.1-11.9)	48 (46-50)	10.8 (9.3-12.3)
Colon and Rectum	58 (55-61)	9.3 (5.9-12.8)	55 (53-57)	9.9 (8.0-11.7)	53 (51-54)	10.3 (9.0-11.6)	47 (45-48)	10.9 (9.7-12.0)
Kidney	74 (69-79)	8.9 (3.0-14.8)	61 (57-64)	11.7 (9.4-14.0)	58 (55-62)	11.9 (9.9-13.9)	49 (45-53)	12.5 (10.5-14.6)
Lung	13 (11-15)	10.3 (5.2-15.4)	11 (10-12)	10.7 (8.2-13.1)	NC	NC	NC	NC
Nasopharynx[a]	55 (41-67)	11.0 (2.1-19.9)	41 (29-53)	11.8 (2.7-20.8)	32 (21-43)	12.2 (2.2-22.2)	NC	NC
Rectum[a]	57 (52-62)	9.3 (4.3-14.4)	53 (50-56)	10.3 (7.8-12.8)	50 (47-52)	10.6 (8.8-12.5)	45 (42-47)	11.1 (9.4-12.8)
Skin melanoma	84 (81-86)	5.1 (3.9-6.2)	83 (79-86)	5.2 (3.8-6.6)	85 (81-88)	4.9 (3.5-6.2)	79 (75-82)	5.9 (4.2-7.6)
Stomach	29 (24-34)	9.4 (2.2-16.5)	25 (22-28)	9.6 (4.8-14.4)	22 (20-24)	9.8 (6.0-13.6)	19 (17-21)	10.0 (6.7-13.3)
Testis	96 (95-97)	0	96 (92-98)	0	81 (71-88)	2.6 (0.0-7.2)	90 (72-96)	1.3 (0.0-3.4)
Thyroid	98 (95-99)	0	91 (87-94)	1.7 (0.1-3.3)	81 (74-86)	7.7 (4.7-10.6)	63 (53-72)	9.7 (5.3-14.1)
Tissue sarcoma	56 (50-62)	7.6 (2.9-12.4)	58 (49-66)	8.8 (2.7-14.9)	45 (38-53)	8.8 (1.8-15.8)	48 (39-56)	7.3 (0.4-14.2)
Tongue[a]	37 (30-44)	11.8 (6.5-17.1)	NC	NC	NC	NC	NC	NC
	15-54 years				55-64 years		65-74 years	
Biliary tracts	22 (16-30)	9.1 (0.0-22.4)			15 (11-21)	10.5 (0.1-20.9)	16 (13-20)	9.8 (1.0-18.6)
Bladder	56 (52-59)	11.6 (9.3-13.9)			49 (46-51)	12.0 (10.3-13.7)	37 (34-39)	12.6 (11.0-14.2)
Larynx	37 (34-40)	13.1 (11.1-15.1)			NC	NC	NC	NC
Liver	15 (12-17)	10.0 (4.5-15.4)			NC	NC	NC	NC
Nasal cavity	42 (33-51)	12.0 (6.5-17.4)			33 (24-41)	12.4 (6.1-18.7)	33 (25-41)	12.3 (6.2-18.5)
Pancreas	10 (8-12)	7.9 (0.0-15.8)			5 (4-7)	8.8 (0.7-16.9)	6 (4-7)	8.2 (0.8-15.6)
Prostate	82 (78-85)	11.5 (9.6-13.3)			88 (86-89)	10.0 (9.0-11.0)	80 (78-82)	12.3 (11.7-12.9)
Small intestine	50 (40-58)	11.3 (5.0-17.6)			42 (33-51)	11.6 (4.9-18.3)	NC	NC
Urinary tracts	48 (37-58)	8.1 (0.0-18.9)			43 (35-50)	8.5 (0.2-16.9)	36 (30-43)	9.0 (1.2-16.7)

NC: No Cure.
[a] Cancer subsite.

Table 4
Estimated values of the cure proportion (P) and time-to-cure (TTC) in women by cancer site and age at diagnosis.

Cancer site	P (95% CI)	TTC (95% CI)	P (95% CI)	TTC (95% CI)	P (95% CI)	TTC (95% CI)	P (95% CI)	TTC (95% CI)
	15-44 years		45-54 years		55-64 years		65-74 years	
Breast	72 (71-74)	12.1 (11.5-12.7)	80 (80-81)	10.6 (10.0-11.1)	79 (78-79)	11.2 (10.6-11.8)	73 (71-74)	12.3 (11.7-12.9)
Central nervous system	38 (33-42)	10.4 (6.2-14.5)	19 (15-23)	10.8 (3.3-18.3)	NC	NC	NC	NC
Cervix uteri	77 (75-79)	6.5 (4.3-8.7)	63 (60-67)	8.5 (4.6-12.5)	49 (45-53)	11.0 (7.6-14.5)	46 (42-51)	10.7 (6.3-15.2)
Colon[a]	62 (58-66)	7.4 (3.8-11.1)	62 (59-64)	6.8 (4.5-9.2)	58 (56-60)	8.1 (6.1-10.1)	55 (53-57)	8.6 (6.9-10.2)
Colon and Rectum	62 (59-65)	7.6 (4.8-10.4)	59 (57-61)	8.1 (6.2-10.0)	58 (57-60)	8.5 (7.1-10.0)	53 (52-55)	9.2 (7.9-10.5)
Corpus uteri	81 (75-86)	6.8 (0.3-13.3)	77 (73-79)	8.4 (5.0-11.8)	75 (73-77)	8.8 (6.5-11.1)	63 (60-65)	10.9 (8.8-13.0)
Head and Neck	40 (33-47)	12.3 (8.2-16.4)	NC	NC	NC	NC	NC	NC
Kidney	79 (73-84)	7.9 (1.1-14.6)	73 (68-78)	9.7 (5.0-14.5)	68 (63-72)	11.4 (8.7-14.1)	54 (49-58)	12.2 (9.7-14.7)
Lung	23 (20-27)	9.6 (4.0-15.2)	16 (14-18)	10.6 (6.7-14.5)	NC	NC	NC	NC
Ovary	62 (58-66)	8.4 (5.5-11.4)	43 (39-46)	10.3 (7.6-12.9)	32 (30-35)	10.4 (7.7-13.2)	22 (20-25)	11.1 (7.9-14.3)
Rectum[a]	62 (56-67)	8.5 (3.8-13.2)	57 (53-60)	9.3 (6.4-12.2)	59 (56-61)	9.1 (6.8-11.4)	50 (48-53)	10.1 (7.9-12.3)
Skin melanoma	92 (90-93)	3.8 (3.0-4.6)	90 (87-92)	5.1 (3.6-6.5)	87 (83-89)	6.4 (4.4-8.5)	80 (75-84)	8.8 (5.2-12.5)
Stomach	31 (25-38)	9.1 (0.0-18.9)	32 (26-37)	8.9 (0.7-17.2)	32 (27-36)	9.6 (3.7-15.5)	25 (22-28)	10.1 (5.5-14.7)
Thyroid	99 (98-100)	0	99 (97-100)	0	97 (94-98)	0	83 (76-89)	9.9 (7.5-12.3)
Tissue sarcoma	65 (58-71)	9.3 (3.0-15.6)	63 (54-71)	9.5 (1.9-17.1)	42 (34-49)	11.3 (4.3-18.2)	35 (27-42)	11.6 (4.6-18.7)
Tongue[a]	44 (31-56)	12.0 (4.5-19.5)	NC	NC	NC	NC	NC	NC
Vagina and Vulva	70 (57-80)	10.7 (3.4-17.9)	61 (50-71)	11.4 (5.5-17.4)	43 (34-52)	12.4 (7.0-17.7)	38 (31-45)	12.4 (8.2-16.6)
	15-54 years				55-64 years		65-74 years	
Biliary tracts	28 (21-35)	7.2 (0.0-17.0)			20 (16-25)	7.6 (0.0-16.6)	13 (11-16)	8.1 (0.4-15.9)
Bladder	46 (39-54)	11.8 (6.1-17.5)			44 (38-51)	11.9 (7.0-16.8)	40 (35-45)	12.1 (8.4-15.8)
Larynx	52 (41-61)	13.0 (8.4-17.7)			NC	NC	NC	NC
Liver	22 (17-27)	8.9 (0.8-17.1)			14 (10-17)	9.5 (0.4-18.5)	6 (4-7)	10.1 (0.3-20.0)
Oropharynx[a]	27 (21-32)	12.1 (6.6-17.5)			31 (24-39)	11.9 (5.7-18.0)	25 (17-34)	12.1 (3.5-20.7)
Pancreas	16 (13-19)	8.4 (1.6-15.2)			9 (7-11)	9.0 (2.0-15.9)	4 (3-5)	9.5 (2.3-16.7)
Small intestine	52 (42-61)	11.2 (5.6-16.7)			46 (36-55)	11.4 (5.4-17.5)	29 (21-37)	12.1 (5.3-18.9)

NC: No Cure.
[a] Cancer subsite.



a. Thyroid cancer b. Testis cancer

c. Skin melanoma

Fig. 1. Cure proportion (P) and time-to-cure (TTC) in cancer sites with TTC ≤ 5 years for some combinations of sex and age at diagnosis. a) Thyroïd cancer, b) Testis cancer and c) Skin melanoma.

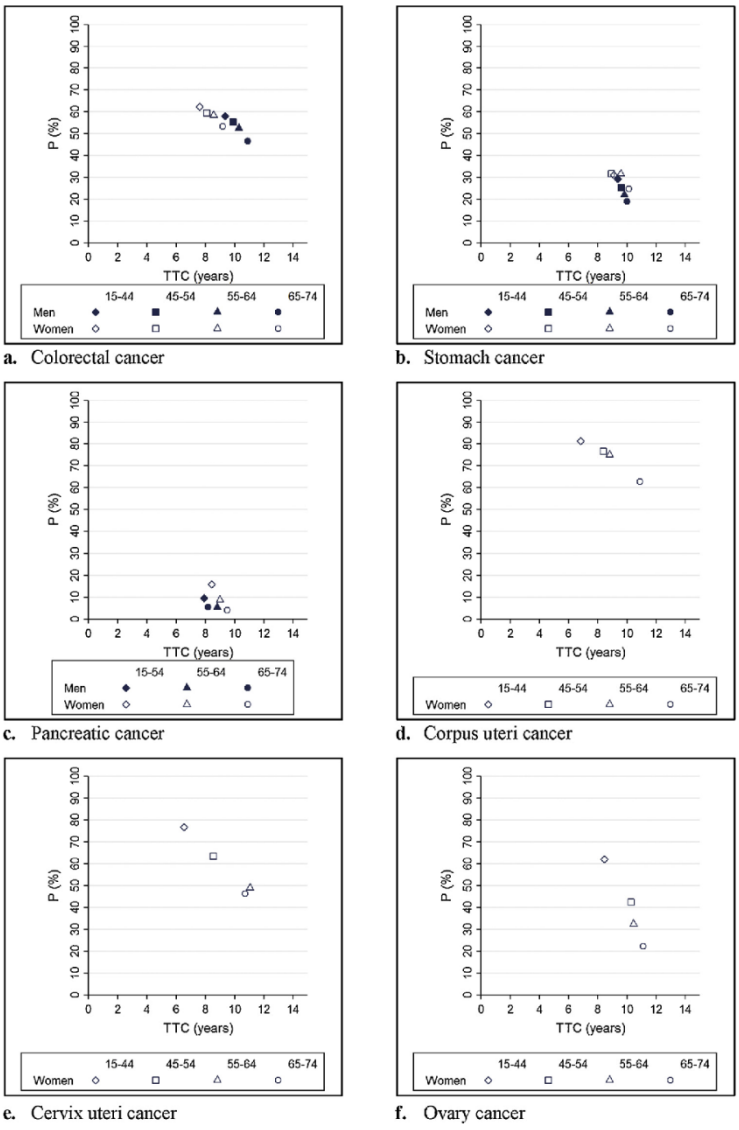


Fig. 2. Cure proportion (P) and time-to-cure (TTC). Examples of cancer sites with TTC between 5 and 10 years for some combinations of sex and age at diagnosis. a) Colorectal cancer, b) Stomach cancer, c) Pancreatic cancer, d) Corpus uteri cancer, e) Cervix uteri cancer and f) Ovary cancer.

[10], P is plotted against TTC to show changes of P and TTC values with sex and age and allow visual comparisons between cancer sites.

3.1. Time-to-cure < 5 years (Fig. 1)

Thyroid and testis cancers and skin melanoma had the shortest TTC. Cure was considered reached right after diagnosis (P > 95% thus TTC = 0) in thyroid cancer in men and women aged 15–44 and in women aged < 65 (Fig. 1a) and in testis cancer in men aged < 55 (Fig. 1b). TTC was < 2 years and P was > 80% in men aged 45–54 with thyroid cancer and men aged ≥55 with testis cancer. Cure from skin melanoma was reached within 5 years (Fig. 1c) in all patients aged < 55 and in men aged 55–64; P was around 90% in women and 85% in men. Though P was > 60%, cure was reached after 6–10 years in all patients aged 65–74 and men aged 55–64 with thyroid cancer and in all patients aged 65–74 and women aged 55–64 with skin melanoma.

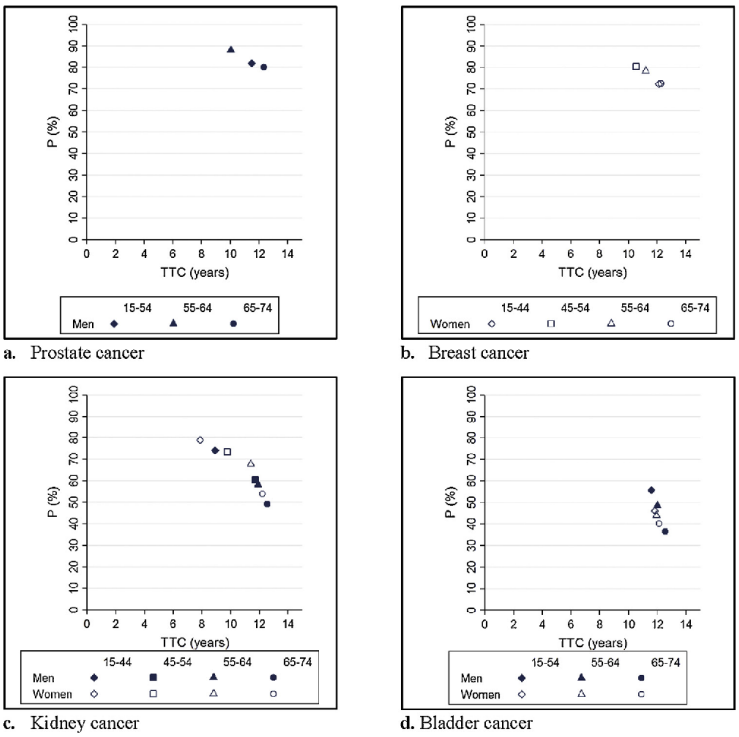


Fig. 3. Cure proportion (P) and time-to-cure (TTC). Examples of cancer sites with TTC > 10 years for most combinations of sex and age at diagnosis. a) Prostate cancer, b) Breast cancer, c) Kidney cancer and d) Bladder cancer.

3.2. Time-to-cure 5–10 years (Fig. 2)

In all digestive cancers (except small intestine) and all gynaecologic cancers (except breast), cure was reached within 5–10 years.

Among digestive cancers, P was the highest in colorectal cancer (Fig. 2a). P was slightly higher in women vs. men and decreased with age from 62 to 55% in women and from 59 to 48% in men. The TTC increased with age and cure was reached earlier in women vs. men. The TTC ranged from 6.8 to 9.2 years in women and from 9.3 to 10.8 years in men. P was the lowest in pancreas cancer (Fig. 2c). Under 45 years, P was 10% in men and 16% in women and about 5% in older patients. The TTC ranged from 8 to 9 years and did not change with sex or age. For stomach cancer (Fig. 2b), P decreased with age from 30 to 19% in men and from 30 to 25% in women. The TTC ranged from 9 to 10 years whatever the sex and age.

Among gynaecologic cancers, P was overall higher for corpus uteri than for cervix uteri or ovary (Fig. 2d). It decreased with increasing age from 81 to 63%. The corresponding figures for cervix uteri were 77% and 46% (Fig. 2e). For corpus and cervix uteri, the TTC was similar and increased with age from 7 to 11 years. For ovary, P decreased with increasing age from 62 to 22% (Fig. 2f). Cure was reached at 8.4 years in the youngest and 11 years in the oldest.

3.3. Time-to-cure > 10 years (Fig. 3)

Cure was reached past 10 years in urinary tract (kidney and bladder), breast and prostate cancers. For prostate and breast cancers (Fig. 3a and b), TTC ranged between 10 and 12 years though P was very high. P varied with age between 80 and 88% for prostate cancer and between 72 and 80% for breast cancer.

P was higher for kidney than for bladder cancer (Fig. 3c and d). For kidney, P decreased with increasing age from 74 to 49% in men and from 79 to 54% in women. For bladder, P ranged from 56 to 37% in men and from 46 to 40% in women. For kidney cancers diagnosed after age 45, the TTC ranged from 12 to 13 years in men and from 10 to 12 years in women. For bladder, cure was reached within 12–13 years whatever the sex and age.

4. Discussion

Among the 335,358 cancer cases registered 1989–2009 in patients aged 15–74 years, the TTC was < 12 years for most cancer sites.

With cure proportions > 95%, cure occurred right after diagnosis in men aged < 55 with testis cancer and in men < 45 and women < 65 with thyroid cancer. For skin melanoma, testis cancers and thyroid cancers diagnosed before age 65, the cure proportions were > 80% and cure was reached within 5 years. For breast and prostate cancers (the most frequent in France), cure occurred after 10 years despite very high cure proportions (72–88%). For colorectal cancer, cure was reached between 5 and 10 years.

Previous studies have reported net survival advantages in women vs. men [19]. Here, we show higher cure proportions in women than in men in nearly all cancer sites, except tissue sarcoma, biliary tract, and bladder cancers. Regarding TTC, sex disparity was not uniform; e.g., the TTC was shorter in women regarding colorectal or kidney cancer but longer regarding skin melanoma or tissue sarcoma. In both sexes and

most cancer sites, the TTC increased and the cure proportion decreased with increasing age at diagnosis. Plotting TTC vs. P depicts the proportion of patients to whom the TTC applies. For instance, when P is very low (melanoma or cancer of the lung or the pancreas), the TTC has less importance than when P is very high.

Regarding age disparities, the rejection of the cure assumption in the older age groups in several cancer sites was in agreement with net survival studies on the same French registry data [13] (absence of plateau in the corresponding sites/age groups).

To be excluded, a cancer site had to meet both conditions of < 500 incident cases and < 200 deaths. This allowed keeping in the analysis cancers with good prognosis and reasonable numbers of incident cases. Among the cancer sites excluded because the number of cases was < 500 and the number of deaths was < 200, only penile cancer had a low excess mortality rate. However, in France, penile cancer is rare: its crude incidence rate is 1.14 per 100,000 person-years [20].

This study is the first to provide cure proportions and TTC for cancers in France. Its strengths are: i) the exhaustiveness of the study due to a population-based recruitment; ii) the use of flexible parametric cure models that have shown advantages over other approaches [21]; iii) almost exact overlap of model-estimated net survivals with and without cure when the assumption of cure was accepted; and iv) the use of a recent definition of TTC [5]. Indeed, for the first time, the TTC was estimated from the individual probability of being cured [5]. This intuitive definition of TTC indicates the time after which patients can be reasonably confident in belonging to the cured group. From this time, the observed mortality rate equals the mortality of the general population. Although statistically cured patients would not die from cancer, they are still subject to death from other causes with the same probability as the general population.

Only two previous studies have estimated the time-to-cure from the net survival of the uncured [22] and from the 10-year conditional net survival (survival for additional 10 years) [4], but with limitations [5]: the former can be over-influenced by the early excess mortality because it depends only on the survival of the uncured. For a given long-term excess mortality, a high early excess mortality brings the estimate of this time-to-cure closer to the date of diagnosis than a lower early excess mortality. The latter assumes a monotonously increasing function for the conditional net survival and requires long-term follow-ups. Only one study has been published with estimations of the time-to-cure [4] using the conditional net survival. Our results are similar to those obtained by Dal Maso et al. [4], particularly regarding sex-related and age-related differences. However, there are two discrepancies i) when P > 95%, TTC values here are 0 vs. 1–2 year in men < 55 years with testicular cancer and 1–4 years in women < 65 years with thyroid cancer in the Italian study; ii) TTC of 10 years here vs. 5 years in women aged 65–75 with thyroid cancer in the Italian study (probably because of lack of monotony: in our data, the 10-year conditional net survival decreased between 5 and 8 years then increased to reach 95% only 10 years after diagnosis). The similarity of the results between our study and the Italian study are in agreement with the 5-year survivals found in EUROCARE studies [23]. Applied to other countries with different health policies (e.g., cancer prevention or screening) or with different risk factors due to different lifestyles, this method might show different TTCs for same cancer sites.

One limitation of the present study is the lack of information on well-recognised prognostic factors such as cancer stage. The main limitation of all cure models is the absence of implemented statistical tests. This imposes a graphical checking of the assumption of cure that requires criteria. Our criteria (excess mortality ≤ 0.05 and adequacy of net survival curves) may be questioned: different criteria would have led to different selections of cancer site, sex, and age combinations; furthermore, the threshold used to define TTC was arbitrarily chosen. Moreover, the post estimation of the TTC might have yielded very wide confidence intervals in case of low cancer incidence or lethal cancer. We are currently developing a cure model that includes time-to-cure (defined as the time at which the excess mortality rate becomes null) as a parameter to estimate. This approach will allow testing the assumption of cure and will provide a more accurate estimation of the time at which cure is reached because it will not use an approximation based on a predefined threshold.

Although statistical cure is not directly related to clinical cure, the cure proportion and the TTC are important indicators for policy making, medical practice, and patient information. In France, cancer patients cannot access to credit or have to pay high extra insurance premiums; however, since 2015, cancer declaration is no more mandatory ≥ 10 years after the end of cancer treatment (also called: “right to be forgotten”) [24]. The present results suggest that the durations of extra premiums should be shortened, not only after testis cancer, thyroid cancer, and skin melanoma where statistical cure occurs shortly after diagnosis, but and also after other cancers whose cure is reached ≤ 10 years after diagnosis.

Authorship contribution

- Substantial contributions to conception and design: GR, OB, NB, LR, MC, VJ.
- Acquisition of data: MC, VJ.
- Analysis and interpretation of data: GR, OB, VJ.
- Drafting the article or revising it critically for important intellectual content: all authors.
- Final approval of the version to be published: all authors.

Funding

This work was supported by the Institut National du Cancer, France [grant INCa 2014-087].

Declarations of interest

None.

Acknowledgements

The authors wish to thank the French network of cancer registries (FRANCIM) for providing the study data. They also thank Stéphanie Normand (Registre Bourguignon des Cancers Digestifs, France) for her technical assistance.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.canep.2019.02.006.

References

[1] M. Colonna, N. Mitton, N. Bossard, A. Belot, P. Grosclaude, Total and partial cancer prevalence in the adult French population in 2008, BMC Cancer 15 (2015) 153.
[2] S. Francisci, R. Capocaccia, E. Grande, M. Santaquilani, A. Simonetti, C. Allemani, G. Gatta, M. Sant, G. Zigon, F. Bray, M. Janssen-Heijnen, The cure of cancer: a European perspective, Eur. J. Cancer (Oxford, England: 1990) 45 (6) (2009) 1067–1079.
[3] M. Cvancarova, B. Aagnes, S.D. Fossa, P.C. Lambert, B. Moller, F. Bray, Proportion cured models applied to 23 cancer sites in Norway, Int. J. Cancer 132 (7) (2013) 1700–1710.
[4] L. Dal Maso, S. Guzzinati, C. Buzzoni, R. Capocaccia, D. Serraino, A. Caldarella, A.P. Dei Tos, F. Falcini, M. Autelitano, G. Masanotti, S. Ferretti, F. Tisano, U. Tirelli, E. Crocetti, R. De Angelis, Long-term survival, prevalence, and cure of cancer: a population-based estimation for 818 902 Italian patients and 26 cancer types, Ann. Oncol. 25 (11) (2014) 2251–2260.
[5] O. Boussari, G. Romain, L. Remontet, N. Bossard, M. Mounier, A.M. Bouvier, C. Binquet, M. Colonna, V. Jooste, A new approach to estimate time-to-cure from cancer registries data, Cancer Epidemiol. 53 (2018) 72–80.
[6] P.W. Dickman, A. Sloggett, M. Hills, T. Hakulinen, Regression models for relative survival, Stat. Med. 23 (1) (2004) 51–64.
[7] K.A. Cronin, E.J. Feuer, Cumulative cause-specific mortality for cancer patients in

the presence of other causes: a crude analogue of relative survival, Stat. Med. 19 (13) (2000) 1729–1740.
[8] J. Esteve, E. Benhamou, M. Croasdale, L. Raymond, Relative survival and the estimation of net survival: elements for further discussion, Stat. Med. 9 (5) (1990) 529–538.
[9] T. Hakulinen, L. Tenkanen, Regression analysis of relative survival rates, Appl. Stat. 36 (1987) 309–317.
[10] A. Verdecchia, R. De Angelis, R. Capocaccia, M. Sant, A. Micheli, G. Gatta, F. Berrino, The cure for colon cancer: results from the EUROCARE study, Int. J. Cancer 77 (3) (1998) 322–329.
[11] R. Sposto, Cure model analysis in cancer: an application to data from the Children's Cancer Group, Stat. Med. 21 (2) (2002) 293–312.
[12] A. Fritz, C. Percy, A. Jack, International Classification of Diseases for Oncology, 3rd ed., World Health Organization, Geneva, 2000.
[13] A. Cowppli-Bony, Z. Uhry, L. Remontet, N. Voirin, A.V. Guizard, B. Tretarre, A.M. Bouvier, M. Colonna, N. Bossard, A.S. Woronoff, P. Grosclaude, Survival of solid cancer patients in France, 1989–2013: a population-based study, Eur. J. Cancer Prev. 26 (6) (2017) 461–468.
[14] P. Lambert, P. Royston, Further development of flexible parametric models for survival analysis, Stata J. 9 (2009) 265–290.
[15] C.P. Nelson, P.C. Lambert, I.B. Squire, D.R. Jones, Flexible parametric models for relative survival, with application in coronary heart disease, Stat. Med. 26 (30) (2007) 5486–5498.
[16] T.M. Andersson, P.W. Dickman, S. Eloranta, P.C. Lambert, Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models, BMC Med. Res. Methodol. 11 (2011) 96.
[17] T.M.L. Andersson, P.C. Lambert, Fitting and modeling cure in population-based cancer studies within the framework of flexible parametric survival models, Stata J. 12 (2012) 623–628.
[18] P.C. Lambert, P. Royston, Further development of flexible parametric models for survival analysis, Stata J. 9 (2) (2009) 265–290.
[19] M.L. Janssen-Heijnen, A. Gondos, F. Bray, T. Hakulinen, D.H. Brewster, H. Brenner, J.W. Coebergh, Clinical relevance of conditional survival of cancer patients in europe: age-specific analyses of 13 cancers, J. Clin. Oncol. 28 (15) (2010) 2520–2528.
[20] L. Daubisse-Marliac, M. Colonna, B. Tretarre, G. Defossez, F. Molinie, K. Jehannin-Ligier, E. Marrer, P. Grosclaude, Long-term trends in incidence and survival of penile cancer in France, Cancer Epidemiol. 50 (Pt A) (2017) 125–131.
[21] X.Q. Yu, R. De Angelis, T.M. Andersson, P.C. Lambert, D.L. O'Connell, P.W. Dickman, Estimating the proportion cured of cancer: some practical advice for users, Cancer Epidemiol. 37 (6) (2013) 836–842.
[22] M. Chauvenet, C. Lepage, V. Jooste, V. Cottet, J. Faivre, A.M. Bouvier, Prevalence of patients with colorectal cancer requiring follow-up or active treatment, Eur. J. Cancer (Oxford, England: 1990) 45 (8) (2009) 1460–1465.
[23] R. De Angelis, M. Sant, M.P. Coleman, S. Francisci, P. Baili, D. Pierannunzio, A. Trama, O. Visser, H. Brenner, E. Ardanaz, M. Bielska-Lasota, G. Engholm, A. Nennecke, S. Siesling, F. Berrino, R. Capocaccia, Cancer survival in Europe 1999–2007 by country and age: results of EUROCARE–5-a population-based study, Lancet Oncol. 15 (1) (2014) 23–34.
[24] F. Binder-Foucard, N. Bossard, P. Delafosse, A. Belot, A.S. Woronoff, L. Remontet, Cancer incidence and mortality in France over the 1980–2012 period: solid tumors, Rev. Epidemiol. Sante Publ. 62 (2) (2014) 95–108.

VII Comparaison de trois modèles paramétriques de guérison

Tous les modèles de guérisons présentés précédemment permettent d'estimer la proportion de patients guéris mais aucun ne permet d'estimer directement le délai de guérison (le délai de guérison est une post-estimation) ni de vérifier statistiquement l'hypothèse de guérison. Il était nécessaire de répondre objectivement à la question de l'existence d'une guérison statistique et d'améliorer l'estimation du délai de guérison.

La première partie de ce chapitre présentera le nouveau modèle de guérison proposé par l'équipe. Ce modèle permet d'évaluer statistiquement l'hypothèse de guérison et d'estimer simultanément le délai de guérison. La deuxième partie, correspondant au second objectif de cette thèse, présentera les résultats de la comparaison que j'ai effectuée des performances de ce modèle à celles des deux autres modèles de guérison existants dans des conditions contrôlées. Les trois modèles seront appliqués à des données réelles. La comparaison des trois modèles de guérison a fait l'objet d'un article qui sera soumis à Biometrical Journal pour le numéro spécial « ISCB » (International Society for Clinical Biostatistics) (Annexe C).

VII.1 Nouveau modèle de guérison de Boussari *et al.*

L'objectif a été d'introduire le délai de guérison comme un paramètre à estimer dans ce modèle. Le délai de guérison a été défini comme le temps, noté τ, à partir duquel le taux instantané de mortalité en excès devient strictement nul. Ce modèle a été appelé modèle TNEH pour « Time-to-Null-Excess-Hazard ». La fonction du taux instantané de mortalité en excès (λ_{exc}) de ce modèle doit donc répondre à différents critères :

- Tendre vers 0 pour $0 \leq t \leq \tau$;
- Être égal à 0 pour $t > \tau$;
- Être une fonction suffisamment souple pour s'adapter à la forme du taux en excès qui peut être très variable selon la localisation de cancer.

Le taux de mortalité en excès du modèle TNEH s'écrit comme suit :

$$\lambda_{exc}(t) = \begin{cases} \left(\frac{t}{\tau}\right)^{\alpha-1}\left(1-\frac{t}{\tau}\right)^{\beta-1} & ,pour\ 0 \le t \le \tau \\ 0 & ,pour\ t > \tau \end{cases} \quad \text{(VII.1)}$$

Avec : $\tau > 0$, le temps écoulé jusqu'à ce que la mortalité en excès atteigne 0 (délai de guérison) ; $\beta > 1$ et $\alpha > 0$, les paramètres d'échelle. Les paramètres τ et α peuvent dépendre de covariables qui ne sont pas nécessairement les mêmes.

Cette écriture du taux de mortalité en excès a la structure d'une fonction de densité d'une loi Beta, ce qui permet d'exprimer le taux de mortalité en excès cumulé par la fonction de répartition d'une loi Beta.

Si Y, une variable aléatoire, suit une loi Beta de paramètres α et β ($B(\alpha,\beta)$) alors, la fonction de répartition est :

$$F_{Be}(y) = \frac{B_x(\alpha,\beta)}{B(\alpha,\beta)} = \begin{cases} \int_0^y \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}dx & ,pour\ 0 \le x \le 1 \\ 1 & ,pour\ x > 1 \end{cases} \quad \text{(VII.2)}$$

Cette fonction est définie sur [0, 1], où $B_x(\alpha,\beta)$ est la fonction Beta incomplète et $B(\alpha,\beta)$, la valeur de la fonction $B(\alpha,\beta)$ en 1. $B(\alpha,\beta)$ est la valeur maximale de $B_x(\alpha,\beta)$. Diviser $B_x(\alpha,\beta)$ par $B(\alpha,\beta)$ permet de régulariser $F_{Be}(y)$, ainsi $F_{Be}(y)$ varie entre 0 et 1.

A partir de l'équation (VII.1), le taux de mortalité en excès cumulé peut s'écrire tel que :

$$\Lambda_{exc}(t) = \int_0^t \lambda_{exc}(u)du = \begin{cases} \int_0^t \left(\frac{u}{\tau}\right)^{\alpha-1}\left(1-\frac{u}{\tau}\right)^{\beta-1}du & ,pour\ 0 \le u \le \tau \\ Cst & ,pour\ u > \tau \end{cases} \quad \text{(VII.3)}$$

Son domaine de définition est [0, τ]. En faisant un changement de variable tel que : $x = \frac{u}{\tau}$ son domaine de définition devient [0, 1] :

$$\Lambda_{exc}(t) = \begin{cases} \int_0^{t/\tau} x^{\alpha-1}(1-x)^{\beta-1}\tau dx & ,pour\ 0 \leq t \leq \tau \\ \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\tau dx & ,pour\ t > \tau \end{cases} \quad \text{(VII.4)}$$

Et en multipliant $\Lambda_{exc}(t)$ par $\frac{B(\alpha,\beta)}{B(\alpha,\beta)}$ on retrouve la fonction de répartition d'une loi Beta (équation (VII.2)) :

$$\Lambda_{exc}(t) = \tau B(\alpha,\beta) \begin{cases} \int_0^{t/\tau} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} dx & ,pour\ 0 \leq t \leq \tau \\ 1 & ,pour\ x > 1 \end{cases} \quad \text{(VII.5)}$$

$$\Lambda_{exc}(t) = \tau B(\alpha,\beta) F_{Be}\left(\frac{t}{\tau}, \alpha, \beta\right)$$

Où $F_{Be}\left(\frac{t}{\tau}, \alpha, \beta\right)$ est la fonction de réparation de la loi Beta, B(α, β), de paramètres α et β.

La survie nette peut donc s'écrit comme :

$$S_n(t) = exp[-\Lambda_{exc}(t)] = exp\left[-\tau B(\alpha,\beta) F_{Be}\left(\frac{t}{\tau}, \alpha, \beta\right)\right] \quad \text{(VII.6)}$$

La proportion de guéris (π) correspond à la survie nette lorsque $t = \tau$, soit :

$$\pi = exp[-\Lambda_{exc}(\tau)] = exp\left[-\tau B(\alpha,\beta) F_{Be}\left(\frac{\tau}{\tau}, \alpha, \beta\right)\right] \quad \text{(VII.7)}$$

Sachant que : $F_{Be}\left(\frac{\tau}{\tau}, \alpha, \beta\right) = 1$ pour tout t > τ, alors l'équation (VII.7) équivaut à :

$$\pi = exp[-\tau B(\alpha,\beta)] \quad \text{(VII.8)}$$

Les paramètres du modèle TNEH sont estimés en utilisant la méthode de maximisation de la vraisemblance (L) (Section 0). La log-vraisemblance s'écrit alors comme suit :

$$ln\, L = \sum_{i=1}^{N} \delta_i\, ln\left[\lambda_{pop}(t_{pop,i}; z_{pop,i}) + \lambda_{exc}(t_i; z_i)\right] + ln[S_n(t_i; z_i)] \quad \text{(VII.9)}$$

La méthode d'optimisation de la procédure d'estimation des paramètres utilise une contrainte de boîtes (L-BFGS-B) de Byrd *et al.*[70]. Les écart-types et les intervalles de

confiances des paramètres du modèle ainsi que ceux de la proportion de guéris et de la survie nette sont estimés en utilisant la Delta-méthode.

L'utilisation de la loi Beta pour l'écriture du nouveau modèle TNEH présente deux avantages : avoir une écriture analytique du taux de mortalité en excès cumulé et permettre une implémentation dans les logiciels plus simple car les fonctions de la loi Beta sont déjà implémentées.

L'évaluation des performances du modèle TNEH pour sa validation modèle-a été faite sur 1000 échantillons simulés de taille 250, 500, 1000 et 2000 en fixant les paramètres du modèle TNEH avec les paramètres α et τ dépendant de l'âge au diagnostic centré-réduit. Les paramètres du modèle TNEH ont ensuite été estimés à partir des données simulées. Les performances du modèle ont alors pu être étudiées en comparant les paramètres estimés aux paramètres fixés selon différents indicateurs.

Ce nouveau modèle ainsi que l'étude de ses performances sur données simulées ont fait l'objet d'un article actuellement en révision à la revue « Biometrics » (Annexe D et abstract ci-dessous) et a fait l'objet d'un poster présenté le 17 juillet 2019 lors du congrès de l'International Society for Clinical Biostatistics (ISCB) :

Biometrics XX, 1–22 DOI: XX
Month YEAR

Modeling excess hazard with time–to–cure as a parameter

Olayidé Boussari[1,2,*], Laurent Bordes[3], Gaëlle Romain[1,2], Marc Colonna[4], Nadine Bossard[5,6], Laurent Remontet[5,6], and Valérie Jooste[1,2]

[1]UMR 1231, EPICAD team, INSERM, Université Bourgogne–Franche–Comté, Dijon, F–21000, France

[2]Registre Bourguignon des Cancers Digestifs, Dijon-Bourgogne University Hospital, Dijon, F–21000, France

[3]UMR 5142, LMAP–IPRA, CNRS, E2S UPPA, Université Pau & Pays Adour, Pau, F–64000, France

[4]Registre du Cancer de l'Isère, Grenoble University Hospital, Grenoble, F–38000, France

[5]Department of Biostatistics and Bioinformatics, Hospices Civils de Lyon, Lyon, F–69003, France

[6]UMR 5558, LBBE, Biostatistics Health Group, CNRS, University Lyon 1, Lyon, F–69100, France

**email:* olayide.boussari@u-bourgogne.fr

Summary: Cure models have been widely developed to estimate the cure fraction when some subjects never experience the event of interest. However these models were rarely focused on the estimation of the time–to–cure i.e. the delay elapsed between the diagnosis and "the time from which cure is reached", an important indicator, for instance to address the question of access to insurance or loans for subjects with personal history of cancer. We propose a new excess hazard regression model that includes the time–to–cure as a covariate dependent parameter to be estimated. The model is written similarly to a Beta probability distribution function and is shown to be a particular case of the non-mixture cure models. Parameters are estimated through a maximum likelihood approach and simulation studies demonstrate good performance of the model. Illustrative applications to two cancer data sets are provided and some limitations as well as possible extensions of the model are discussed. The proposed model offers a simple and comprehensive way to estimate more accurately the time–to–cure.

Key words: Cancer; Cure model; Cure time; Net survival; Right to be forgotten.

This paper has been submitted for consideration for publication in *Biometrics*

VII.2 Méthodes pour la simulation des données

VII.2.1 Objectif

L'étude de simulation précédente a permis de valider le modèle TNEH. L'objectif, est maintenant de comparer, à partir de données simulées, ses performances avec celles des deux autres modèles de guérison existants : le modèle de guérison paramétrique flexible de Andersson *et al.*[56] et le modèle de guérison de mélange[51]. Les performances des trois modèles ont été comparées sur la survie nette et la proportion de patients guéris en considérant trois scénarios afin d'étudier les performances des modèles dans différentes conditions. Chaque scénario a été simulé à partir :

- d'un modèle de guérison de mélange avec une distribution d Weibull. Les taux de mortalité en excès obtenu à partir des données simulées tendent vers zéro au cours du temps mais sans jamais être égaux à zéro. Ces simulations permettent alors de comparer les performances du modèle TNEH lorsque le taux de mortalité en excès est très proche de zéro sans jamais l'atteindre.
- du modèle TNEH. Ces simulations ont pour objectif de vérifier le comportement des modèles de guérison actuels lorsque le taux de mortalité en excès devient nul.

VII.2.2 Simulation des données de survie

Des temps jusqu'au décès ont été simulés suivant trois scénario inspirés de données réelles : scénario de bon, moyen et mauvais pronostic. Pour chaque scénario, 1000 échantillons de taille : $N = 2000$ ont été simulés avec le modèle TNEH et un modèle de mélange suivant une loi de Weibull. Une analyse de sensibilité a été faite sur les données simulées de moyen pronostic pour des tailles d'échantillons de $N = 1000$ et $N = 500$.

Par la suite, nous noterons $\mathcal{U}[a, b]$ la loi uniforme sur [a, b] et $\mathcal{B}(p)$ la loi de Bernoulli de paramètres p.

VII.2.2.a Simulation de l'âge au diagnostic

Selon la taille de l'échantillon, *N* âges au diagnostic sont générés $A(a_1, \dots, a_n)$ suivant une distribution uniforme sur les intervalles : [15-44], [45-59] et [60-74] pour le scénario de bon pronostic et [15-59], [60-69] et [70-74] pour les scénarios de moyen et mauvais pronostic.

Les proportions d'âges provenant de ces trois intervalles sont respectivement :

- Pour les scénarios de bon et moyen pronostic : 30%, 40% et 30% ;
- Pour le scénario de mauvais : 35%, 40% et 25%.

VII.2.2.b Simulation du temps jusqu'à la censure

La censure peut être administrative ou parce que le patient est perdu de vue (Section III.1). Le temps de suivi maximum (t_{max}) est fixé à 15 ans à partir du diagnostic. On définit, C_1 le temps de censure correspondant à un mélange d'une proportion $p_1$de patients en vie à la fin du suivi mais suivi moins de 15 ans, une proportion $1 - p_1$ de patients en vie et suivis plus de 15 ans ; et C_2 le temps de censure correspondant à un mélange d'une proportion p_2 de patients perdus de vue, une proportion $1 - p_2$ de patients en vie et suivis plus de 15 ans. *N* temps de $C_1(c_{1_1}, \dots, c_{1_n})$ et *N* temps de $C_2(c_{2_1}, \dots, c_{2_n})$ sont générés tels que :

$$C_j = P_j \times U_j + \big(1 - P_j\big)t_{max}\ , pour\ j \in \{1, 2\} \tag{VII.10}$$

Où $U_j \sim \mathcal{U}[0,15]$ et $P_j = \mathcal{B}(p_j)$.
Pour chaque patients *i*, le temps jusqu'à la censure, noté $C(c_1, \dots, c_n)$, est donc le temps minimum entre C_1 et C_2 *:*

$$c_i = min\big(c_{1_i}, c_{2_i}\big)\ , pour\ 1 \leq i \leq n \tag{VII.11}$$

La proportion p_1 de censure administrative est fixée 0,25 pour le scénario de bon pronostic, 0,6 pour le scénario de moyen pronostic et 0,1 pour le scénario de mauvais pronostic. La p_2 de patients perdus de vue est respectivement fixés à 0,02 pour les trois scénarios.

VII.2.2.c Simulation du temps jusqu'au décès

Dans le cadre de la survie nette, le temps de décès observé correspond soit au temps de décès dû au cancer, noté $T_e(t_{e_1}, \dots, t_{e_n})$, soit au temps de décès dû à d'autres causes, noté $T_{pop}(t_{pop_1}, \dots, t_{pop_n})$. Dans la réalité, ne connaissant pas la cause du décès, il faut donc simuler les temps à partir du diagnostic jusqu'au décès par cancer et « autres causes ».

Temps de décès « autres causes »

N temps de décès « autres causes » sont générés à partir d'une loi de Weibull dont les paramètres prennent des valeurs permettant de reproduire un taux de mortalité similaire à celui des tables de mortalité fourni par l'INSEE. Dans les tables de mortalité, l'information disponible se réfère à T_{birth}, le temps de survie à partir de la naissance jusqu'au décès « autres causes ». La fonction de survie associée S_{birth} est, la survie à partir de la naissance. La survie à partir du diagnostic dans la population générale correspond donc à la survie à partir de la naissance sachant que le patient a survécu jusqu'au diagnostic :

$$S_{birth}(t_{birth}|a) = Prob(T_{birth} > t_{birth}|T_{birth} > a) = \frac{S_{birth}(t_{birth})}{S_{birth}(a)} \quad \text{(VII.12)}$$

Où : *a* étant l'âge au diagnostic, alors $S_{birth}(a)$ et $S_{birth}(t_{birth})$ sont, respectivement, la survie attendue de la population générale au moment du diagnostic et au moment du décès « autres causes ».

Afin de reproduire un taux de mortalité similaire à celui des tables de mortalité, on suppose que la variable aléatoire T_{birth} suit une loi de Weibull de paramètre k=11 et m=88. Donc la survie attendue dans la population générale est :

$$S_{birth}(t_{birth}|a) = \frac{exp\left[-\left(\frac{t_{birth}}{k}\right)^m\right]}{S_{birth}(a)} \quad \text{(VII.13)}$$

N valeurs de X_1 ($x_{1_1}, \dots, x_{1_n}$) sont générées, telles que $X_1 \sim \mathcal{U}[0,1]$, avec :

$$X_1 = S_{birth}(t_{birth}|a) = \frac{exp\left[-\left(\frac{t_{birth}}{k}\right)^m\right]}{S_{birth}(a)} \quad \text{(VII.14)}$$

D'où le temps à partir de la naissance jusqu'au décès :

$$t_{birth} = k\{-ln[X_1 S_{birth}(a)]\}^{1/m} \quad \text{(VII.15)}$$

Avec : $S_{birth}(a) = \exp\left[-\left(\frac{a}{k}\right)^m\right] / S_{birth}(a)$

A partir de l'équation (VII.15), le temps de survie « autres causes » à partir du diagnostic est :

$$\begin{aligned} t_{pop_i} &= t_{birth_i} - a_i \\ t_{pop_i} &= k\{-ln[x_{1_i} S_{birth}(a_i)]\}^{1/m} - a_i \end{aligned}, pour\ 1 \le i \le n \quad \text{(VII.16)}$$

Temps de décès dû au cancer

N temps de survie à partir du diagnostic jusqu'au décès dû au cancer sont générés ainsi que la survie nette associée à partir (a) d'un modèle de guérison de mélange et (b) du modèle de guérison TNEH.

(a) A partir d'un modèle guérison de mélange

Sous l'hypothèse de guérison, la survie nette peut être décrite par un modèle de mélange utilisant une distribution de Weibull pour la survie des non-guéris (Section IV.2.1) :

$$S_n(t_e) = \pi + (1-\pi)S_u(t_e) \tag{VII.17}$$

Où π est la proportion de guéris ; et $S_u(t_e)$ est la fonction de survie nette des « non-guéris ». On suppose que la variable aléatoire T_e suit une loi de Weibull de paramètre μ_e et ε_e et on génère *N* valeurs de X_2 ($x_{2_1}, \ldots, x_{2_n}$), telles que $X_2 \sim \mathcal{U}[0,1]$, avec :

$$\begin{aligned} X_2 &= S_n(t_e) = \pi + (1-\pi)S_u(t_e) \\ X_2 &= \pi + (1-\pi)exp[-k_e t^{m_e}] \end{aligned} \tag{VII.18}$$

A partir de l'équation (VII.18), le temps jusqu'au décès dû au cancer à partir du diagnostic (pour $1 \leq i \leq n$) est :

$$t_{e_i} = \begin{cases} \left\{ \dfrac{-ln\left[\dfrac{x_{2_i} - \pi(z_i;\omega)}{1-\pi(z_i;\omega)}\right]}{k_e(z_i;\eta)} \right\}^{1/m_e(z_i;\gamma)} & , si\ x_{2_i} \geq \pi(z_i;\omega) \\ +\infty & , si\ x_{2_i} < \pi(z_i;\omega) \end{cases} \tag{VII.19}$$

Avec : les paramètres k_e, m_e et π dépendent de la covariable z_i correspondant à l'âge au diagnostic en classes et associés respectivement aux vecteurs de paramètres η, γ et ω.

(b) A partir d'un modèle de guérison TNEH

N valeurs de X_2 ($x_{2_1}, \dots, x_{2_n}$) sont générées, telles que $X_2 \sim \mathcal{U}[0,1]$. A partir de l'écriture de la fonction de survie nette dans le modèle TNEH (VII.6), on pose :

$$X_2 = S_n(t_e) = exp\left[-\tau B(\alpha,\beta) F_{Be}\left(\frac{t_e}{\tau}, \alpha, \beta\right)\right] \quad \text{(VII.20)}$$

Où $F_{Be}\left(\frac{t}{\tau}, \alpha, \beta\right)$ est la fonction de réparation de la loi Beta, B(α, β), de paramètres α et β.
On peut donc calculer le temps jusqu'au décès dû au cancer à partir du diagnostic (pour $1 \leq i \leq n$) :

$$t_{e_i} = \begin{cases} \tau(z_i;\eta) F_{Be}^{-1}\left(\frac{ln(x_{2_i})}{ln[-\tau(z_i;\eta)B(\alpha(z_i;\gamma),\beta)]}\,;\alpha(z_i;\gamma);\beta\right) & ,si\ x_{2_i} \geq \pi(z_i;\omega) \\ +\infty & ,si\ x_{2_i} < \pi(z_i;\omega) \end{cases} \quad \text{(VII.21)}$$

Avec : les paramètres α et τ dépendent de la covariable z_i correspondant à l'âge au diagnostic en classes et associés respectivement aux vecteurs de paramètres η et γ.

VII.2.2.d Temps de suivi

Pour un patient *i*, le temps suivi observé, noté $T(t_i, \dots, t_n)$, est le minimum entre le temps jusqu'à la censure et le temps de décès observé ; lui-même est défini comme le temps le minimum entre le temps de décès dû au cancer et le temps de décès dû à d'autres causes. Donc, à partir du temps de censure (équation (VII.11)) et des deux temps de décès simulés ((VII.16) et (VII.18) ou (VII.21)), on peut associer à chaque patient *i*, le couple de variables $\{t_i, \delta_i\}$, où :

$$\begin{aligned} t_i &= min\left[c_i, min\left(t_{e_i}, t_{pop_i}\right)\right], pour\ 1 \leq i \leq n \\ \delta_i &= \begin{cases} 1 & ,si\ min\left(t_{e_i}, t_{pop_i}\right) \leq c_i \\ 0 & ,si\ min\left(t_{e_i}, t_{pop_i}\right) > c_i \end{cases}, pour\ 1 \leq i \leq n \end{aligned} \quad \text{(VII.22)}$$

VII.2.2.e Description des scénarios simulés

Pour simuler les données, les paramètres du modèle de guérison de mélange et du modèle TNEH ont été fixés de manières à ce que les proportions de guéris et les courbes de survie nette et du taux de mortalité en excès soient très proches.Afin que la seule différence

entre les types de simulations porte uniquement sur le délai pour que le taux de mortalité en excès atteigne zéro.

Scénario de bon pronostic

Le scénario de bon pronostic est inspiré des localisations telles que le cancer du testicule ou du mélanome de la peau. Pour les classes d'âge <45, [45-59] et ≥60 ans, la proportion de guéris du modèle de mélange est fixée, respectivement, à 80%, 85% et 90% ; et le délai τ du modèle TNEH est fixé, respectivement, à 4, 4,5 et 6 ans.

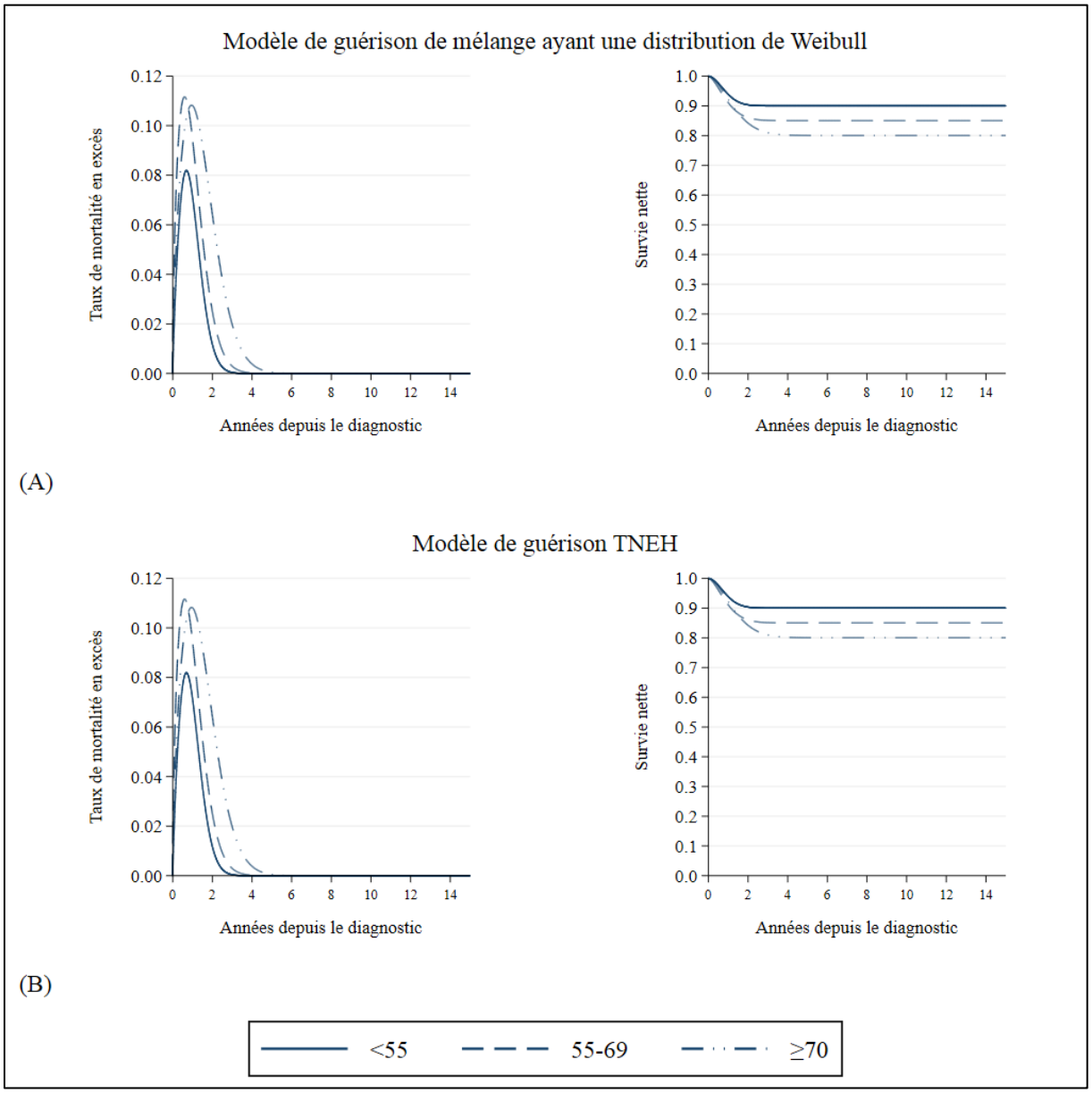


Figure VII.1 – Scénario de bon pronostic : Survie nette et taux de mortalité théoriques calculés à partir des modèles de guérison : (A) de mélange avec une distribution de Weibull ; (B) TNEH.

Scénario de moyen pronostic

Le scénario de moyen pronostic est inspiré du cancer colon-rectum. Pour les classes d'âge <60, [60-69] et ≥70 ans, la proportion de guéris du modèle de mélange est fixée, respectivement, à 65%, 60% et 50% ; et le délai τ du modèle TNEH est fixé, respectivement, à 7, 7,5 et 9,5 ans.

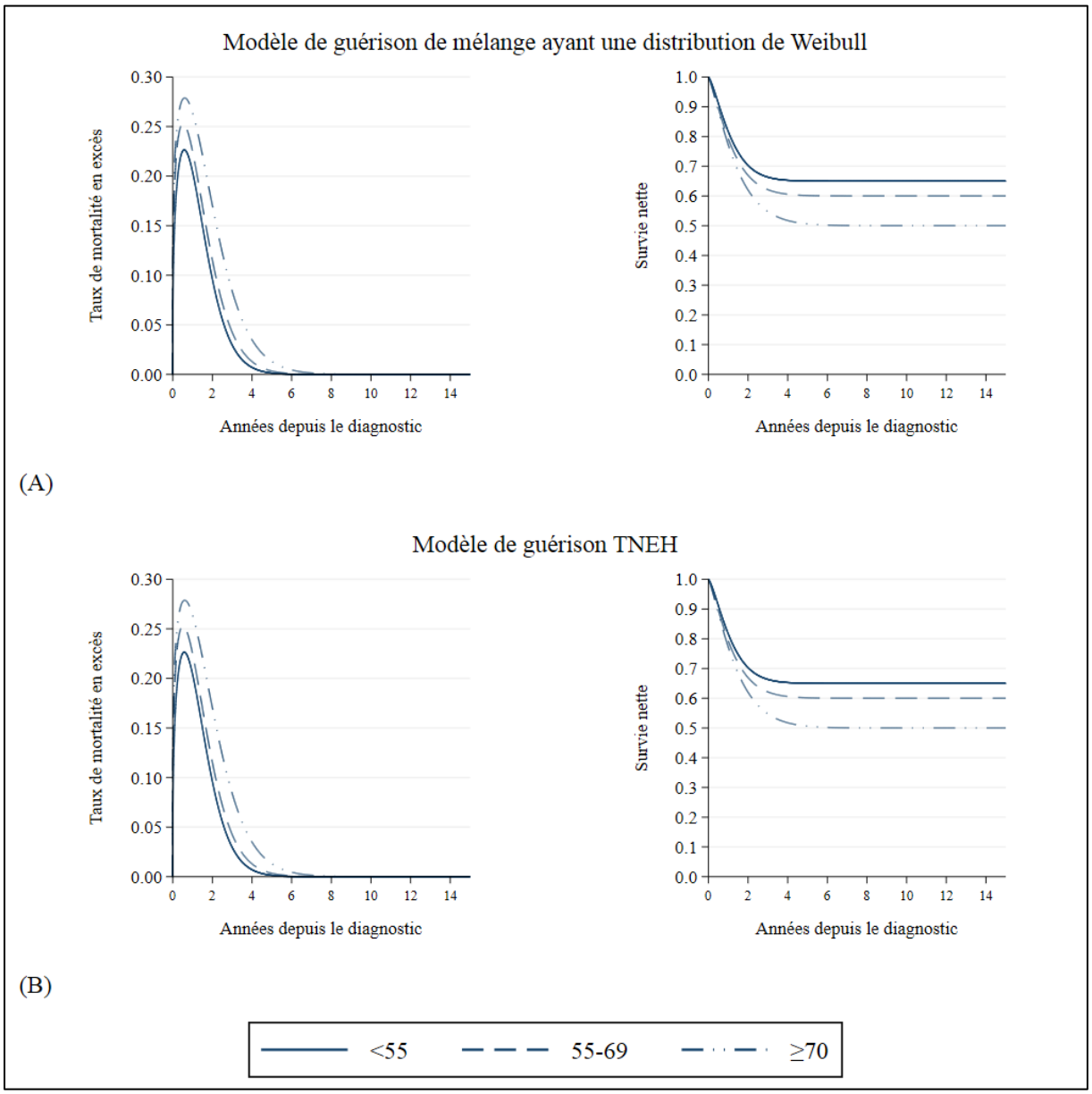


Figure VII.2 - Scénario de moyen pronostic : Survie nette et taux de mortalité théoriques calculés à partir des modèles de guérison : (A) de mélange avec une distribution de Weibull ; (B) TNEH.

Scénario de mauvais pronostic

Le scénario de mauvais pronostic est inspiré du cancer pancréas. Pour les classes d'âge <60, [60-69] et ≥70 ans, la proportion de guéris du modèle de mélange est fixée, respectivement, à 20%, 15% et 10% ; et le délai τ du modèle TNEH est fixé, respectivement, à 9, 10 et 11 ans.

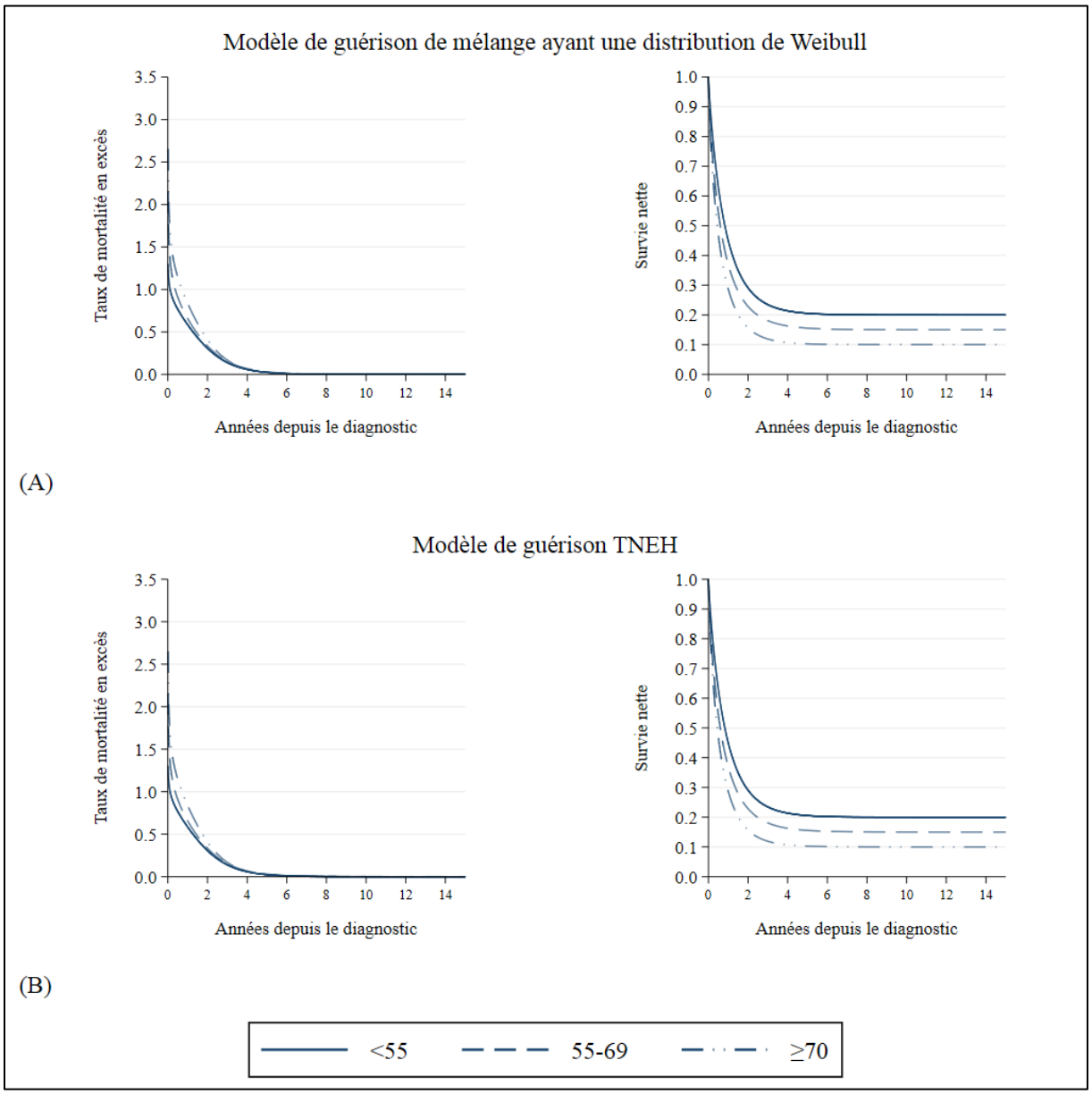


Figure VII.3 - Scénario de mauvais pronostic : Survie nette et taux de mortalité théoriques calculés à partir des modèles de guérison : (A) de mélange avec une distribution de Weibull ; (B) TNEH.

VII.2.3 Estimation de la survie nette et de la proportion de guéris sur les données simulées

Trois modèles de guérison ont été appliqués aux données simulées pour estimer la survie nette ($S_n(t)$) et la proportion de guéris (π) :

- Modèle de guérison de mélange [51, 54], noté MGM : pour les trois scénarios, la survie des « non-guéris » est modélisée par une distribution de Weibull. Les paramètres du modèle (la proportion de guéris et les paramètres de la distribution de Weibull) sont estimés pour chaque catégorie d'âge.
- Modèle de guérison paramétrique flexible[56], noté MGF : en suivant la même stratégie de modélisation sous l'hypothèse de guérison que précédemment (Section VI.2)
- Modèle TNEH : pour les trois scénarios, les paramètres du modèle (la proportion de guéris et les paramètres de la distribution Beta) sont estimés pour chaque catégorie d'âge.

La première catégorie d'âge est définie comme référence pour chaque modèle et chaque scénario.

VII.2.4 Indicateurs de performances pour la comparaison des modèles

Les performances des trois modèles présentés précédemment ont été comparées sur la survie nette estimée à 3, 5 et 10 ans et sur la proportion de patients guéris. Les valeurs théoriques de $\pi, S_n(3), S_n(5), S_n(10)$ sont obtenues à partir des paramètres fixés pour les simulations (des modèles de guérison de mélange et TNEH). Quatre indicateurs de performances ont été utilisés :

- Le biais : différence entre la valeur théorique et la moyenne des estimations. Un biais proche de 0 indique l'estimation est proche de la varie valeur, donc que le modèle étudié estime de manière non biaisée π, $S_n(3)$, $S_n(5)$ ou $S_n(10)$;
- L'erreur-type empirique : écart-type des 1000 estimations. Il donne une indication sur l'homogénéité des estimations, plus l'erreur-type empirique est faible plus la précision des estimations fournies par le modèle est élevée ;
- L'erreur-type moyenne : moyenne des 1000 erreur-types. L'estimation de l'erreur-type est considérée comme non biaisée lorsque cette dernière est proche

de l'erreur-type empirique. Si l'erreur-type moyenne est inférieure à l'erreur-type empirique alors le modèle sous-estime la véritable variabilité dans l'estimation ;

- Le taux de couverture : pourcentage d'échantillons pour lesquels la valeur théorique est comprise dans l'intervalle de confiance de l'estimation. Un bon estimateur doit avoir un taux de couverture proche de la valeur nominale de 95%.

VII.3 Résultats sur données simulées

Les données ont été simulées de façon à ce que les valeurs théoriques de survie nette et proportion de guéris soient similaires avec le modèle de guérison de mélange et le modèle TNEH mais contrairement au premier cas (modèle de mélange), dans le second cas, le taux de mortalité en excès atteint zéro à un temps fixé. L'ensemble des paramètres fixés des deux modèles pour chaque scénario sont présentés dans l'Annexe B. Dans cette partie, nous présenterons donc uniquement les résultats des simulations obtenues à partir du modèle de guérison de mélange. Les tableaux de résultats des simulations sur de 2000 échantillons à partir du modèle TNEH sont fournis en Annexe E.

VII.3.1 Scénario de bon pronostic

Les résultats des indicateurs de performances sont présentés dans le Tableau VII-1. Le biais de la survie nette et de la proportion de guéris estimées par le modèle de guérison de mélange (MGM) et le modèle TNEH varie entre -0.002 et 0.001 pour les trois classes d'âge. Le biais des estimations de $S_n(t)$ et π obtenu à partir du modèle de guérison flexible (MGF) est plus élevé mais reste acceptable, il varie entre -0,011 et 0,014. L'erreur-type empirique des estimations obtenues par le modèle TNEH est très proche de celle des estimations obtenues par les deux autres modèles de guérison. De plus, pour toutes les classes d'âge, on remarque que l'erreur-type empirique est faible, entre 0,012 et 0,018, et est très proche de l'erreur-type moyenne de $S_n(t)$ et π estimées par les trois modèles. Enfin le taux de couverture des estimations du MGM et du modèle TNEH sont d'environ 95%, alors que pour le MGF, le taux de couverture est entre 78,8% et 95,5%.

Tableau VII-1 - Scénario de bon pronostic : Résultats des 1000 échantillons de taille N=2000 simulés à partir du modèle guérison de mélange avec une distribution de Weibull

	Biais			Erreur-type empirique			Erreur-type moyenne			Taux de couverture		
	MGM	MGF	TNEH	MGM	MGF	TNEH	MGM	MGF	TNEH	MGM	MGF	TNEH
<45 ans												
$S_n(3) = 0,900$	<0,001	0,014	0,001	0,012	0,011	0,012	-0,012	0,011	0,012	95,9%	78,8%	95,2%
$S_n(5) = 0,900$	<0,001	0,003	0,001	0,012	0,012	0,012	-0,012	0,012	0,012	96,0%	95,5%	95,1%
$S_n(10) = 0,900$	<0,001	-0,003	0,001	0,012	0,013	0,012	-0,012	0,013	0,012	96,0%	93,7%	95,1%
$\pi = 0,900$	<0,001	-0,003	0,001	0,012	0,013	0,012	0,012	0,013	0,012	96,0%	93,8%	95,1%
[45-59] ans												
$S_n(3) = 0,851$	<0,001	0,014	-0,002	0,013	0,012	0,013	-0,013	0,012	0,013	94,9%	79,7%	94,5%
$S_n(5) = 0, 850$	-0,001	-0,001	-0,001	0,013	0,013	0,013	-0,013	0,013	0,013	94,7%	94,3%	94,4%
$S_n(10) = 0,850$	-0,001	-0,010	-0,001	0,013	0,013	0,013	-0,013	0,014	0,013	94,7%	86,7%	94,4%
$\pi = 0,850$	-0,001	-0,011	-0,001	0,013	0,014	0,013	0,013	0,014	0,013	94,7%	85,5%	94,4%
≥60 ans												
$S_n(3) = 0,810$	-0,001	0,012	<0,001	0,016	0,015	0,016	-0,017	0,016	0,017	96,1%	90,9%	95,8%
$S_n(5) = 0,800$	-0,001	0,001	<0,001	0,018	0,017	0,018	-0,019	0,018	0,019	95,7%	96,0%	95,6%
$S_n(10) = 0,800$	-0,002	-0,010	<0,001	0,018	0,018	0,018	-0,019	0,019	0,019	95,8%	91,4%	95,7%
$\pi = 0,800$	-0,002	-0,011	<0,001	0,018	0,018	0,018	0,019	0,019	0,019	95,9%	90,7%	95,7%

VII.3.2 Scénario de moyen pronostic

Les résultats des indicateurs de performances sont présentés dans le Tableau VII-2. Pour le modèle de guérison de mélange, le biais est proche de zéro. Alors que le modèle de guérison flexible fournit des estimations avec un biais globalement plus élevé mais acceptable : il est inférieur à 3%. Pour le modèle TNEH, seules les estimations de la survie nette à 3 ans présentent un biais plus élevé : entre à 0,012 et 0,014. L'erreur-type empirique est faible pour l'ensemble des estimations de survie nette et de proportion de guéris obtenu par les 3 modèles. De plus l'erreur-type empirique est très proche de l'erreur-type moyenne. Pour le modèle de guérison de mélange, le taux de couverture est très proche de 95%. On observe d'avantage de variations pour le modèle de guérison flexible et TNEH. Le taux de couverture varie respectivement pour les 2 modèles, entre 73,6% et 96,1%, et entre 85,2% et 94,8%.

Tableau VII-2 - Scénario de moyen pronostic : Résultats des 1000 échantillons de taille N=2000 simulés à partir du modèle guérison de mélange avec une distribution de Weibull

	Biais			Erreur-type empirique			Erreur-type moyenne			Taux de couverture		
	MGM	MGF	TNEH	MGM	MGM	MGF	MGM	MGF	TNEH	MGM	MGF	TNEH
<60 ans												
$S_n(3) = 0,663$	0,001	0,026	0,017	0,019	0,018	0,018	-0,020	0,018	0,019	95,4%	73,6%	85,2%
$S_n(5) = 0,651$	0,001	0,006	0,010	0,020	0,019	0,020	-0,020	0,020	0,020	95,2%	95,1%	92,3%
$S_n(10) = 0,650$	0,001	-0,010	0,007	0,020	0,020	0,020	-0,020	0,020	0,021	95,2%	90,7%	93,5%
$\pi = 0,650$	0,001	-0,012	0,007	0,020	0,020	0,020	0,020	0,020	0,021	95,2%	89,6%	93,5%
[60-69] ans												
$S_n(3) = 0,621$	<0,001	0,016	0,012	0,017	0,017	0,016	-0,018	0,017	0,017	95,7%	85,3%	88,9%
$S_n(5) = 0,601$	<0,001	<0,001	0,002	0,019	0,018	0,019	-0,019	0,018	0,019	94,8%	95,3%	94,8%
$S_n(10) = 0,600$	<0,001	-0,017	-0,002	0,019	0,018	0,020	-0,019	0,019	0,020	94,6%	83,7%	94,3%
$\pi = 0,600$	<0,001	-0,019	-0,002	0,019	0,018	0,020	0,019	0,019	0,020	94,8%	81,5%	94,3%
≥70 ans												
$S_n(3) = 0,548$	-0,001	0,018	0,014	0,021	0,020	0,020	-0,022	0,021	0,020	95,3%	89,5%	90,8%
$S_n(5) = 0,506$	-0,002	0,019	0,003	0,024	0,021	0,025	-0,024	0,022	0,023	96,0%	89,3%	93,4%
$S_n(10) = 0,500$	-0,002	0,005	-0,008	0,025	0,022	0,028	-0,025	0,023	0,027	95,5%	95,8%	92,0%
$\pi = 0,500$	-0,002	0,003	-0,008	0,025	0,022	0,028	0,025	0,023	0,027	95,9%	96,1%	92,1%

VII.3.3 Scénario de mauvais pronostic

Les résultats des indicateurs de performances sont présentés dans le Tableau VII-3. Avec le modèle de guérison de mélange, le biais de l'estimation de la survie nette et la proportion de guéris est proche de zéro. Avec le modèle TNEH, seule la survie nette estimé à 3 ans chez les patients de moins de 60 ans présente un biais de 0,012. Avec le modèle de guérison flexible, les patients de moins de 60 ans présentent le biais est le plus élevé (entre 0,016 et 0,026). Pour les trois modèles de guérison l'erreur-type empirique de $S_n(t)$ et π est faible et est proche de l'erreur-type moyenne. Le taux de couverture des estimations du modèle de guérison de mélange est proche de 95%. Cependant, pour le modèle de guérison flexible et TNEH le taux de couverture est, respectivement, entre 59.9% et 94.8%, et entre 79.8% et 93.4%.

Tableau VII-3 - Scénario de mauvais pronostic : Résultats des 1000 échantillons de taille N=2000 simulés à partir du modèle guérison de mélange avec une distribution de Weibull

	Biais			Erreur-type empirique			Erreur-type moyenne			Taux de couverture		
	MGM	MGF	TNEH	MGM	MGF	TNEH	MGM	MGF	TNEH	MGM	MGF	TNEH
<60 ans												
$S_n(3) = 0,235$	<0,001	0,026	0,012	0,016	0,015	0,015	-0,016	0,015	0,015	94,5%	61,6%	88,1%
$S_n(5) = 0,205$	<0,001	0,026	0,010	0,016	0,015	0,017	-0,016	0,015	0,016	94,1%	59,9%	89,0%
$S_n(10) = 0,200$	<0,001	0,017	0,007	0,017	0,014	0,018	-0,016	0,015	0,017	93,4%	79,9%	90,4%
$\pi = 0,200$	<0,001	0,016	0,007	0,017	0,014	0,018	0,016	0,015	0,017	93,1%	82,2%	90,4%
[60-69] ans												
$S_n(3) = 0,181$	<0,001	0,011	<0,001	0,013	0,012	0,013	-0,014	0,013	0,013	94,6%	89,3%	93,4%
$S_n(5) = 0,155$	<0,001	0,010	<0,001	0,013	0,012	0,014	-0,014	0,012	0,013	94,8%	89,4%	92,8%
$S_n(10) = 0,150$	<0,001	0,003	-0,001	0,014	0,012	0,015	-0,014	0,012	0,014	95,3%	94,7%	92,9%
$\pi = 0,150$	<0,001	0,002	-0,001	0,014	0,012	0,015	0,014	0,012	0,014	95,3%	94,8%	93,0%
≥70 ans												
$S_n(3) = 0,119$	<0,001	0,009	-0,001	0,015	0,013	0,015	-0,015	0,013	0,014	95,1%	88,6%	92,1%
$S_n(5) = 0,103$	<0,001	0,004	-0,009	0,015	0,012	0,017	-0,015	0,012	0,015	95,0%	93,5%	85,5%
$S_n(10) = 0,100$	-0,001	-0,003	-0,013	0,016	0,012	0,018	-0,015	0,012	0,016	95,5%	93,9%	79,8%
$\pi = 0,100$	-0,001	-0,004	-0,013	0,016	0,012	0,018	0,016	0,012	0,016	95,1%	93,3%	79,9%

VII.3.4 Analyse de sensibilité de la taille des échantillons

En comparant les résultats des Tableau VII-4 et Tableau VII-5 pour une taille d'échantillon de 1000 et 500 au Tableau VII-2 (taille d'échantillon de 2000), on observe que le biais de l'estimation de la survie nette et de la proportion de guéris augmente lorsque la taille de l'échantillon diminue. Cela est vrai pour les trois modèles de guérison. De plus, l'erreur-type des estimations obtenues par le modèle TNEH et les deux autres modèles existants, est divisé par deux lorsque la taille d'échantillon est multipliée par quatre. Cela signifie que le taux de convergence asymptotique est atteint tel que cela a été démontré dans l'article précédent (Annexe D).

Tableau VII-4 - Scénario de moyen pronostic : Résultats des 1000 échantillons de taille N=1000 simulés à partir du modèle guérison de mélange avec une distribution de Weibull

	Biais			Erreur-type empirique			Erreur-type moyenne			Taux de couverture		
	MGM	MGF	TNEH	MGM	MGM	MGF	TNEH	MGM	MGM	MGF	TNEH	MGM
<60 ans												
$S_n(3) = 0,663$	0,000	0,025	0,017	0,027	0,026	0,026	-0,028	0,026	0,026	95,4%	86,0%	89,9%
$S_n(5) = 0,651$	0,000	0,005	0,010	0,028	0,028	0,028	-0,028	0,028	0,028	95,3%	94,4%	92,9%
$S_n(10) = 0,650$	0,000	-0,011	0,008	0,028	0,029	0,028	-0,028	0,029	0,029	95,2%	92,2%	93,9%
$\pi = 0,650$	0,000	-0,013	0,008	0,028	0,029	0,028	0,028	0,029	0,029	95,2%	91,6%	93,9%
[60-69] ans												
$S_n(3) = 0,621$	0,000	0,016	0,013	0,024	0,023	0,023	-0,025	0,024	0,024	95,4%	91,7%	92,9%
$S_n(5) = 0,601$	0,000	0,000	0,003	0,026	0,025	0,025	-0,026	0,026	0,027	95,2%	95,1%	95,5%
$S_n(10) = 0,600$	-0,001	-0,018	-0,002	0,026	0,026	0,027	-0,027	0,026	0,028	95,2%	89,4%	95,9%
$\pi = 0,600$	-0,001	-0,019	-0,002	0,026	0,026	0,027	0,027	0,026	0,028	95,1%	88,3%	95,9%
≥70 ans												
$S_n(3) = 0,548$	0,000	0,019	0,015	0,031	0,029	0,029	-0,031	0,029	0,029	94,6%	92,1%	91,6%
$S_n(5) = 0,506$	0,000	0,020	0,004	0,034	0,031	0,034	-0,033	0,031	0,033	94,1%	91,9%	94,7%
$S_n(10) = 0,500$	-0,001	0,006	-0,008	0,036	0,032	0,038	-0,035	0,032	0,038	94,3%	95,0%	94,7%
$\pi = 0,500$	-0,001	0,004	-0,008	0,036	0,032	0,038	0,035	0,032	0,038	94,1%	95,1%	94,7%

Tableau VII-5 - Scénario de moyen pronostic : Résultats des 1000 échantillons de taille N=500 simulés à partir du modèle guérison de mélange avec une distribution de Weibull

	Biais			Erreur-type empirique			Erreur-type moyenne			Taux de couverture		
	MGM	MGF	TNEH	MGM	MGM	MGF	TNEH	MGM	MGM	MGF	TNEH	MGM
<60 ans												
$S_n(3) = 0,663$	-0,002	0,024	0,016	0,038	0,036	0,035	-0,039	0,036	0,037	96,0%	91,6%	92,9%
$S_n(5) = 0,651$	-0,002	0,004	0,008	0,039	0,039	0,038	-0,040	0,039	0,040	96,2%	96,0%	96,0%
$S_n(10) = 0,650$	-0,002	-0,013	0,005	0,039	0,040	0,038	-0,040	0,040	0,042	96,0%	93,5%	96,7%
$\pi = 0,650$	-0,002	-0,014	0,005	0,039	0,040	0,038	0,040	0,041	0,042	95,7%	93,3%	96,7%
[60-69] ans												
$S_n(3) = 0,621$	-0,001	0,015	0,012	0,035	0,033	0,032	-0,035	0,034	0,034	95,6%	94,7%	94,9%
$S_n(5) = 0,601$	-0,002	-0,001	0,002	0,036	0,035	0,036	-0,037	0,036	0,038	95,2%	96,3%	96,7%
$S_n(10) = 0,600$	-0,003	-0,019	-0,002	0,037	0,036	0,037	-0,038	0,037	0,040	95,5%	92,7%	96,9%
$\pi = 0,600$	-0,003	-0,020	-0,003	0,037	0,036	0,037	0,038	0,037	0,040	96,2%	91,6%	96,9%
≥70 ans												
$S_n(3) = 0,548$	-0,001	0,018	0,014	0,043	0,040	0,040	-0,043	0,042	0,041	95,2%	95,4%	93,2%
$S_n(5) = 0,506$	0,000	0,019	0,004	0,047	0,043	0,047	-0,047	0,044	0,047	95,3%	94,9%	94,5%
$S_n(10) = 0,500$	-0,002	0,004	-0,008	0,050	0,044	0,052	-0,049	0,045	0,053	95,7%	95,8%	95,2%
$\pi = 0,500$	-0,002	0,003	-0,008	0,050	0,044	0,053	0,049	0,045	0,054	95,6%	95,8%	95,4%

VII.4 Application aux données réelles

VII.4.1 Population étudiée et Méthode

Les données utilisées pour appliquer les trois modèles de guérison sont les mêmes que celles de l'étude précédente (Section VI.1). Parmi les localisations de cancer disponibles nous avons choisi trois localisations : cancer du testicule (bon pronostic), du côlon (moyen pronostic) et du pancréas (mauvais pronostic). L'âge au diagnostic a été inclus en catégories dans les trois modèles de guérison : <55 ans et ≥55 ans pour le cancer du testicule et, <55, 55-64 et ≥65 ans pour les cancers du côlon et du pancréas. L'effet non proportionnel de l'âge a également été testé et inclus dans les trois modèles. La classe d'âge de référence était celle dont l'effectif était le plus élevé.

Pour l'application aux données réelles nous avons également comparé graphiquement les estimations de la survie nette obtenues par les trois modèles de guérison aux estimations fournies par l'estimateur de Pohar-Perme.

VII.4.2 Résultats sur données réelles

VII.4.2.a Cancer du testicule

Le taux instantané de mortalité en excès est faible dès le diagnostic pour les patients <55 ans et il devient vite très proche de zéro (Figure VII.4.A). Pour cette classe d'âge, les courbes de survie nette estimée par les trois modèles de guérison et l'estimateur de Pohar-Perme se superposent (Figure VII.4.B). La proportion de guéris est estimée entre 96,0% et 96,7% selon le modèle (Tableau VII-6). Pour les patients de 55 ans et plus au moment du diagnostic, le taux de mortalité en excès est plus élevé les premières années suivant le diagnostic, puis il diminue pour devenir très proche de zéro avec les trois modèles de guérison. Cependant, alors que les courbes du taux de mortalité en excès estimé par les modèles de guérison de mélange et TNEH sont relativement proches entre elles et sont proches de zéro à partir de 5 ans, celle du taux estimé par le modèle de guérison flexible se rapproche de zéro beaucoup plus tardivement (Figure VII.4.A). Ce phénomène s'observe également avec les courbes de la survie nette (Figure VII.4.B). Les courbes de survie nette estimée à partir du modèle de guérison de mélange et TNEH se superposent pour atteindre

une proportion de guéris estimée, respectivement, à 86,6% et 86,8% (Tableau VII-6), alors que la proportion de guéris estimée par le modèle de guérison flexible est de 84,8%.

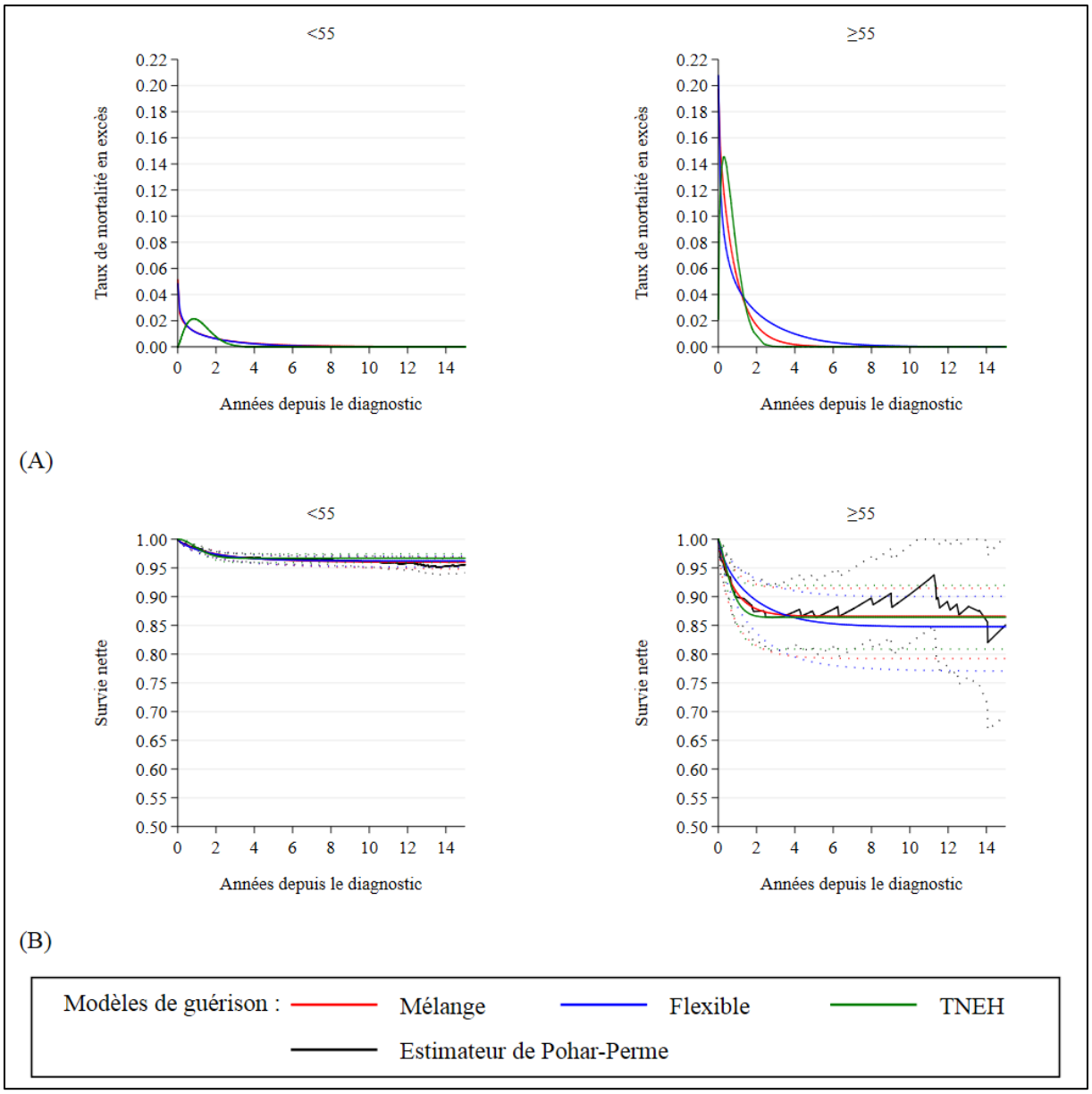


Figure VII.4 – Courbes du taux de mortalité en excès (A) et de survie nette (B) estimées par classe d'âge pour le cancer du testicule, à partir du modèle de guérison de mélange, flexible et TNEH.

Tableau VII-6 – Estimation de survie nette à 3, 5 et 10 ans et de la proportion de patients guéris (IC 95%) par classe d'âge pour le cancer du testicule, à partir du modèle de guérison de mélange (MGM), flexible (MGF) et TNEH.

Age au diagnostic		MGM		MGF		TNEH	
< 55	$S_n(3)$	97,0%	(96,2-97,6)	96,9%	(96,2-97,5)	96,7%	(96,0-97,5)
	$S_n(5)$	96,4%	(95,6-97,1)	96,5%	(95,6-97,1)	96,7%	(95,9-97,5)
	$S_n(10)$	96,1%	(95,0-96,9)	96,2%	(95,2-97,0)	96,7%	(95,9-97,5)
	π	96,0%	(96,0-94,7)	96,2%	(95,1-97,1)	96,7%	(95,9-97,5)
≥ 55	$S_n(3)$	87,0%	(80,2-91,6)	87,5%	(81,1-91,8)	86,4%	(80,9-92,0)
	$S_n(5)$	86,6%	(79,4-91,5)	85,7%	(78,6-90,6)	86,4%	(80,9-92,0)
	$S_n(10)$	86,6%	(79,2-91,5)	84,8%	(77,2-90,0)	86,4%	(80,9-92,0)
	π	86,6%	(79,2-91,5)	84,8%	(77,1-90,1)	86,4%	(80,9-92,0)

VII.4.2.b Cancer du côlon

Pour les deux premières classes d'âge les courbes du taux de mortalités en excès estimé par les trois modèles de guérison sont très proches (Figure VII.5.A). Seul le taux de mortalité chez les 65 ans et plus semble atteindre zéro plus rapidement avec le modèle TNEH qu'avec les deux autres (Figure VII.5.B). En effet la courbe de survie nette estimée par le modèle TNEH atteint un plateau plus rapidement. Cependant, la survie nette estimée par les modèles de guérison de mélange et flexible est comparable à la survie nette fournie par l'estimateur de Pohar-Perme. Chez les patients de 65 ans et plus, la proportion de guéris estimée par les modèles de guérison existant est d'environ 50% alors que le modèle TNEH estime une proportion de guéris de 56,7% (Tableau VII-7). Le délai de guérison estimé par le modèle TNEH est supérieur à 40 ans, autrement dit, d'après la définition du délai de guérison de ce modèle, la guérison statistique dans le côlon n'est pas atteinte.

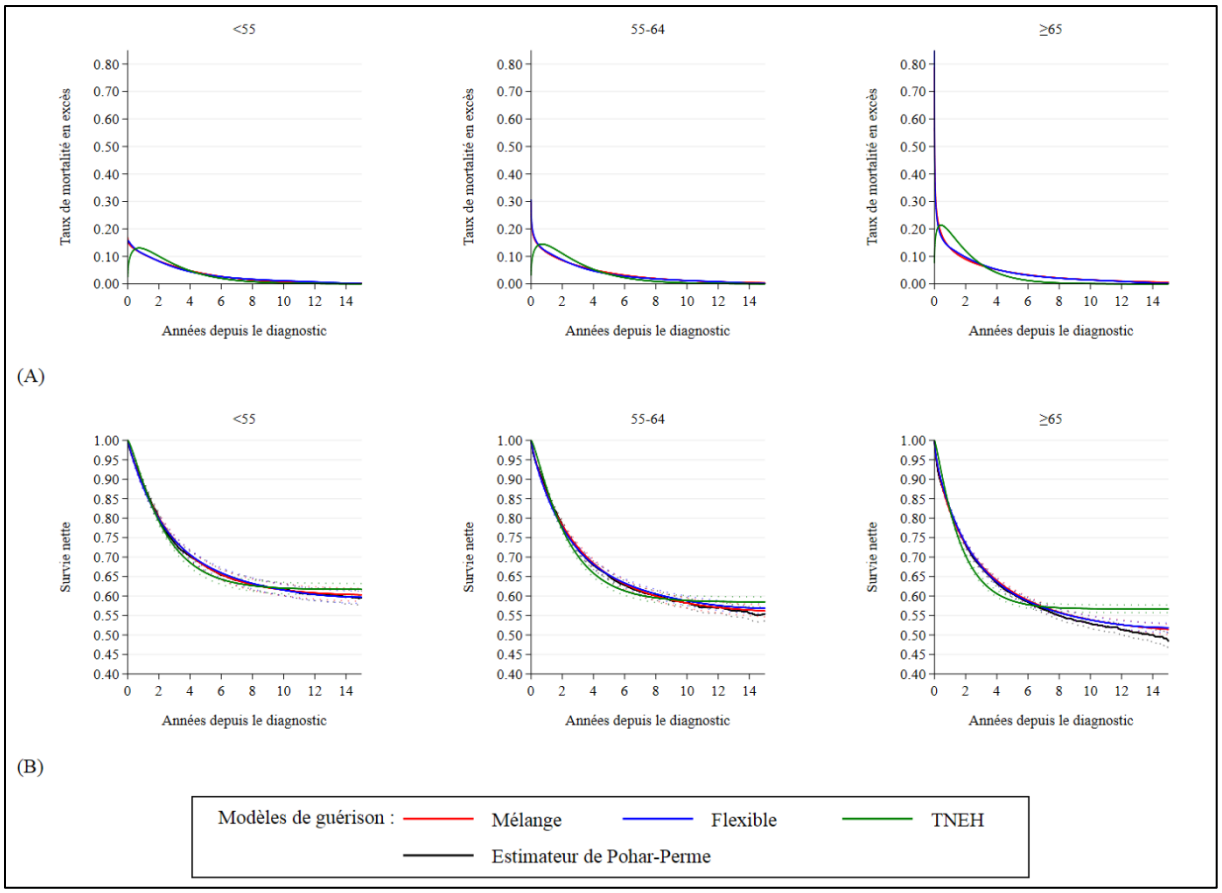


Figure VII.5 – Courbes du taux de mortalité en excès (A) et de survie nette (B) estimées par classe d'âge pour le cancer du côlon, à partir du modèle de guérison de mélange, flexible et TNEH.

Tableau VII-7 - Estimation de survie nette à 3, 5 et 10 ans et de la proportion de patients guéris (IC 95%) par classe d'âge pour le cancer du côlon, à partir du modèle de guérison de mélange (MGM), flexible (MGF) et TNEH.

Age au diagnostic		MGM		MGF		TNEH	
< 55	$S_n(3)$	74,5%	(73,3-75,6)	74,4%	(73,3-75,5)	72,8%	(71,6-73,9)
	$S_n(5)$	67,6%	(66,3-68,9)	68,0%	(66,7-69,2)	66,%	(64,7-67,3)
	$S_n(10)$	61,6%	(60,0-63,0)	61,5%	(60,0-63,0)	62,%	(60,6-63,5)
	π	60,0%	(60,0-58,2)	59,7%	(58,1-61,2)	61,7%	(60,2-63,2)
55-64	$S_n(3)$	72,6%	(71,6-73,5)	72,0%	(71,1-72,9)	70,3%	(69,4-71,3)
	$S_n(5)$	65,4%	(64,3-66,5)	65,4%	(64,3-66,4)	63,1%	(62,0-64,2)
	$S_n(10)$	58,2%	(56,9-59,5)	58,8%	(57,5-60,0)	58,9%	(57,6-60,1)
	π	55,3%	(55,3-53,5)	56,8%	(55,5-58,2)	58,5%	(57,2-59,8)
≥ 65	$S_n(3)$	67,9%	(67,1-68,7)	67,7%	(66,9-68,4)	64,%	(63,2-64,9)
	$S_n(5)$	61,1%	(60,2-62,0)	60,8%	(59,9-61,6)	58,8%	(57,9-59,7)
	$S_n(10)$	53,9%	(52,8-55,0)	53,9%	(52,8-54,9)	56,8%	(55,8-57,8)
	π	49,6%	(47,9-51,4)	51,8%	(50,6-53,0)	56,7%	(55,7-57,7)

VII.4.2.c Cancer du pancréas

Pour les trois classes d'âge, le taux de mortalité en excès est élevé près du diagnostic puis diminue lentement. Pour les trois modèles, le taux de mortalité en excès est proche de zéro à partir de 8 ans (Figure VII.6.A). Les courbes de survie nette estimée par les trois modèles de guérison sont également superposées pour les trois classes d'âge et elles sont comparables à l'estimateur de Pohar-Perme (Figure VII.6.B). La proportion de guéris varie entre 11,7% et 13,3% chez les moins de 55 ans, entre 5,7% et 7,5% chez les 55-64 ans et entre 4,6% et 5,8% chez les 65 ans et plus (Tableau VII-8). Cependant, tout comme pour le cancer du côlon, le délai de guérison estimé par le modèle TNEH est supérieur à 40 ans, autrement dit, d'après la définition du délai de guérison de ce modèle, la guérison statistique dans le pancréas n'est pas atteinte.

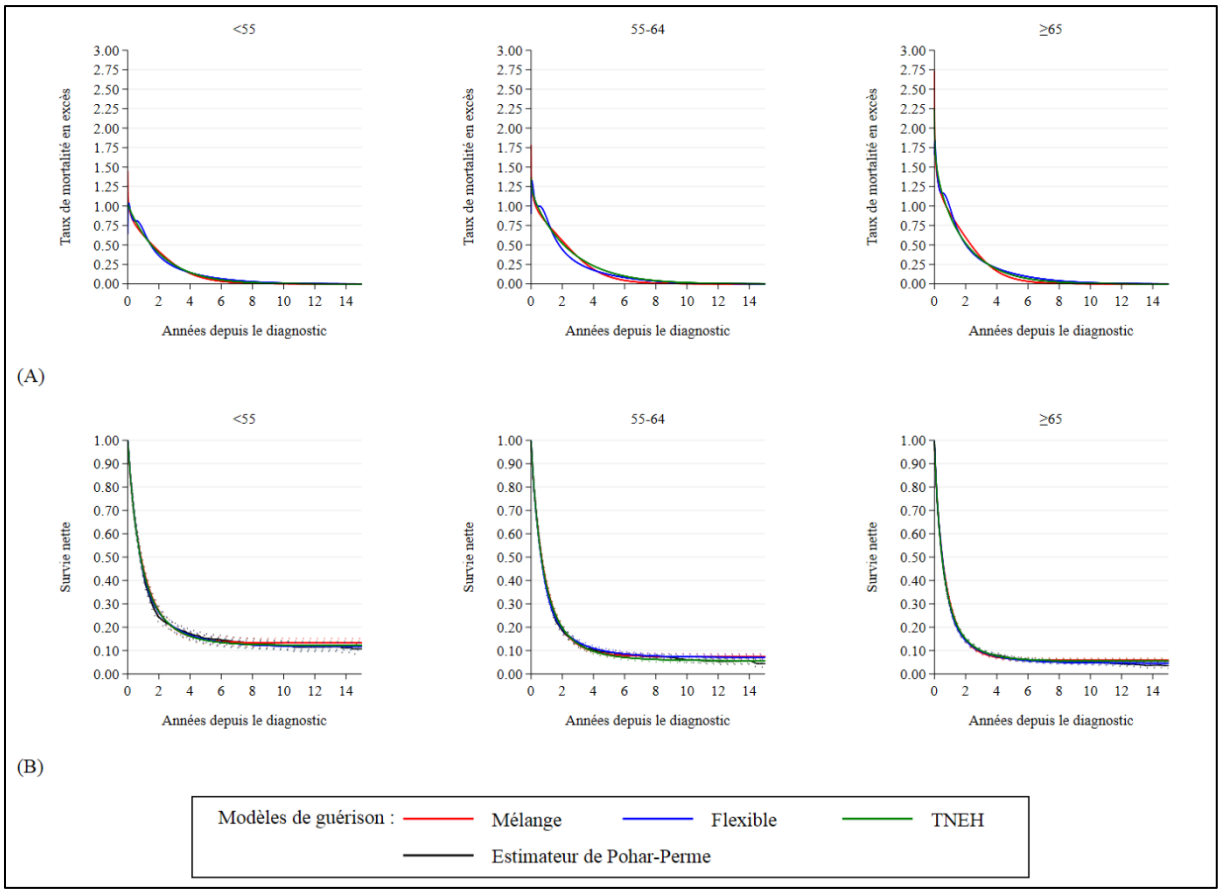


Figure VII.6 - Courbes du taux de mortalité en excès (A) et de survie nette (B) estimées par classe d'âge pour le cancer du pancréas, à partir du modèle de guérison de mélange, flexible et TNEH.

Tableau VII-8 – Estimation de survie nette à 3, 5 et 10 ans et de la proportion de patients guéris (IC 95%) par classe d'âge pour le cancer du pancréas, à partir du modèle de guérison de mélange (MGM), flexible (MGF) et TNEH.

Age au diagnostic		MGM		MGF		TNEH	
< 55	$S_n(3)$	19,4%	(17,6-21,4)	19,9%	(18,1-21,7)	19,5%	(17,8-21,3)
	$S_n(5)$	14,6%	(12,9-16,4)	14,7%	(13,1-16,4)	14,4%	(12,7-16,1)
	$S_n(10)$	13,3%	(11,5-15,2)	11,9%	(10,5-13,5)	12,4%	(10,6-14,1)
	π	13,3%	(13,3-11,5)	11,7%	(10,2-13,3)	12,3%	(10,5-14,1)
55-64	$S_n(3)$	12,6%	(11,4-13,8)	13,7%	(12,5-14,9)	12,9%	(11,8-14,0)
	$S_n(5)$	8,4%	(7,4-9,6)	9,5%	(8,5-10,5)	8,0%	(7,0-9,0)
	$S_n(10)$	7,5%	(6,4-8,7)	7,3%	(6,4-8,3)	5,9%	(4,8-6,9)
	π	7,5%	(7,5-6,4)	7,2%	(6,3-8,2)	5,7%	(4,6-6,7)
≥ 65	$S_n(3)$	9,1%	(8,3-10,0)	9,7%	(8,8-10,5)	10,0%	(9,2-10,7)
	$S_n(5)$	6,4%	(5,6-7,2)	6,4%	(5,7-7,1)	6,7%	(6,0-7,5)
	$S_n(10)$	5,9%	(5,1-6,7)	4,8%	(4,2-5,4)	5,5%	(4,7-6,3)
	π	5,8%	(5,1-6,7)	4,6%	(4,0-5,3)	5,5%	(4,6-6,3)

VII.5 Discussion

L'étude des performances des modèles montre que le modèle TNEH fournit des estimations de survie nette et proportion de guéris non biaisées avec un taux de couverture proche de 95% lorsque la situation contrôlée est favorable (le taux de mortalité en excès devient bien nul). Les estimations du modèle de guérison flexible présentent un biais plus élevé et un taux de couverture plus petit. Cependant ce biais étant inférieur est à 3% et le taux de couverture est supérieur à 75%. Sur données réelles de bon (cancer du testicule) et mauvais (cancer du pancréas) pronostic, les estimations de survie nette et de proportion de guéris sont proches de celles des deux autres modèles de guérison. Pour le cancer du testicule le délai de guérison estimé par le modèle TNEH est à 3,5 ans pour patients <55 ans, 4,0 ans pour les 55-64 ans et 5,8 ans pour les ≥65 ans ; alors que pour le cancer du pancréas et du côlon, le délai de guérison n'est pas estimé dans un délai cliniquement raisonnable (respectivement, à plus de 40 ans et plus de 25 ans). Or, le taux de mortalité en excès est très faible à 10 ans pour toutes les classes d'âge. Pour ces deux localisations le taux de mortalité

diminue progressivement au cours du temps, il n'y a pas de rupture de pente. Dans le cas de cancer du pancréas, les courbes de survie nette présentent une rupture de pente franche notamment car il s'agit d'une localisation très létale ($\pi < 15\%$) avec un taux de mortalité très élevé proche du diagnostic.

D'autres modèles de guérison pourraient être utilisés pour générer les temps de décès dus au cancer. Un modèle paramétrique de guérison flexible permettrait de simuler des données pour lesquelles le taux de mortalité en excès aurait une forme plus complexe au cours du temps, tel qu'il peut être observé sur données réelles (par exemple un « rebond » sur le taux de mortalité en excès pour le cancer de la thyroïde, Section VI.3) [68]. Une autre étude de simulation pourrait inclure la simulation de l'année du diagnostic, ce qui permettrait d'utiliser les tables de mortalité de la population générale fournies par l'INSEE.

Notre étude met en évidence différentes limites du modèle TNEH par rapport aux autres modèles. Les conditions d'application du modèle sont strictes, il ne peut être appliqué que lorsque le taux de mortalité en excès est suffisamment proche de zéro. Dans le cas contraire, le délai de guérison estimé par le modèle sera très grand et les estimations de survie nette et de proportion de guéris ne seront pas interprétables. Les résultats présenté dans l'Annexe E, les modèles de guérison de mélange et flexible fournissent de bonne estimations de la survie nette et de la proportion de guéris lorsque le taux de mortalité en excès atteint zéro (dans la période de suivi). Le modèle TNEH est cependant le seul modèle permettant actuellement d'estimer directement le délai de guérison et donc de déterminer si la guérison est atteinte dans un délai cliniquement raisonnable ou non.

VIII Discussion générale et Perspectives

La première étape de ce travail a été de vérifier l'hypothèse de guérison, c'est-à-dire que la probabilité de décéder du cancer devienne nulle. Dans la réalité, le taux de mortalité dû au cancer atteint rarement zéro donc la survie nette atteint rarement un plateau franc. Nous avons choisi de déterminer un seuil à partir duquel la probabilité de décéder du cancer devenait négligeable. En fixant ce seuil à 0,05, nous avons pu estimer le délai de guérison pour 27 localisations de cancers en France à partir de la probabilité d'appartenir au groupe des patients guéris. Nous avons mis en évidence que, pour la plupart des cancers, la guérison statistique et atteinte dans les 12 ans suivant le diagnostic.

Afin de d'estimer directement le délai de guérison, le nouveau modèle TNEH a été développé au sein de l'équipe. Il permet de déterminer objectivement si la guérison est atteinte ou non. L'étude de simulation a permis de montrer que le modèle TNEH permet une bonne estimation de la survie nette, de la proportion de guéris et du délai de guérison lorsque le taux de mortalité en excès est suffisamment proche de zéro. Cependant, sur données réelles, en dehors du cancer du testicule, ce critère d'application est trop contraignant.

L'ensemble des résultats de cette thèse suggèrent qu'il persiste un sur-risque de décéder chez les patients atteints de cancer, même si la probabilité de décéder directement ou indirectement du cancer peut être considérée comme négligeable. Capocaccia *et al.*[71] ont montré qu'au bout d'un certain temps suivant le diagnostic de cancer, le taux de mortalité observé des patients atteints de cancers se rapproche du taux de mortalité observé dans la population générale sans jamais le rejoindre. Autrement dit, le taux de mortalité en excès ne devient pas strictement nul. L'hypothèse habituelle est que le taux de mortalité « autres causes » peut être approximé par le taux de mortalité observé dans la population générale. Phillips *et al.*[72] font l'hypothèse que les individus atteints de certains cancers ont des taux de mortalité « autres causes » différents de ceux de la population générale. Dans ce sens, Botta *et al.*[73] ont développé un nouveau modèle dans lequel le taux instantané de mortalité observé dans la population de patient s'écrit : $\lambda_{obs}(t) = {\lambda_{exc}}^{c}(t) + r.\lambda_{pop}(t)$, où ${\lambda_{exc}}^{c}(t)$ est le taux instantané de mortalité en excès corrigé et $r.\lambda_{pop}(t)$ le taux instantané de mortalité observé dans la population générale corrigé par un facteur r, correspondant à un sur-risque de décès « autres causes » chez les patients par rapport à la population générale.

Le modèle TNEH pourrait être modifié selon la méthode de Botta *et al.*, ce qui permettrait de corriger le taux de mortalité en excès en supprimant le sur-risque. Il serait alors possible que le taux de mortalité en excès s'annule pour d'autres localisations que le testicule et que le délai de guérison puisse ainsi être estimé par le modèle TNEH.

A court terme, je vais appliquer la méthode de Botta *et al.*[73] aux données du réseau FRANCIM afin de modéliser le taux instantané de mortalité en excès corrigé et fournir des estimations corrigées des délais de guérison.

A moyen terme, je vais poursuivre ces travaux. A titre d'exemple :

Je vais implémenter le modèle TNEH dans le logiciel STATA.

Il sera intéressant, pour certaines localisations d'intérêt clinique ou épidémiologique par exemple le côlon-rectum ou le sein, de modéliser plus finement l'effet de l'âge.

Le recours aux données dites « hautes résolutions » permettra d'étudier différents facteurs pronostiques tels que le stade. Nous porterons un intérêt particulier aux patients ayant pu recevoir un traitement à visée curative, qui sont davantage concernés par le doit à l'oubli.

Références

1. Bray, F., J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, and A. Jemal, *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.

2. Defossez, G., S. Le Guyader Peyrou, Z. Uhry, P. Grosclaude, M. Colonna, E. Dantony, P. Delafosse, F. Molinié, A.S. Woronoff, A.M.R. Bouvier, L., N. Bossard, and A. Monnereau, *Synthèse - Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018*. SaintMaurice : Santé publique France ed. 2019. 20.

3. Dickman, P.W., A. Sloggett, M. Hills, and T. Hakulinen, *Regression models for relative survival.* Stat Med, 2004. **23**(1): p. 51-64.

4. Dumas, A., F. De Vathaire, and G. Vassal, *Access to loan-related insurance for French cancer survivors.* Lancet Oncol, 2016. **17**(10): p. 1354-1356.

5. Parkin, D.M., *The evolution of the population-based cancer registry.* Nat Rev Cancer, 2006. **6**(8): p. 603-12.

6. Bray, F., M. Colombet, L. Mery, M. Piñeros, A. Znaor, R. Zanetti, and J. Ferlay, *Cancer Incidence in Five Continents, Volume XI*. Vol. XI. 2017, Lyon: International Agency for Research on Cancer.

7. Kalbfleisch, J. and R. Prentice, *The statistical analysis of failure time data. 2nd ed*. Vol. 77. 2002.

8. Hakulinen, T. and L. Tenkanen, *Regression Analysis of Relative Survival Rates.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1987. **36**(3): p. 309-317.

9. Estève, J., E. Benhamou, M. Croasdale, and L. Raymond, *Relative survival and the estimation of net survival: Elements for further discussion.* Statistics in Medicine, 1990. **9**(5): p. 529-538.

10. Schaffar, R., B. Rachet, A. Belot, and L.M. Woods, *Estimation of net survival for cancer patients: Relative survival setting more robust to some assumption violations than cause-specific setting, a sensitivity analysis on empirical data.* Eur J Cancer, 2017. **72**: p. 78-83.

11. Kaplan, E.L. and P. Meier, *Nonparametric Estimation from Incomplete Observations.* Journal of the American Statistical Association, 1958. **53**(282): p. 457-481.

12. Rich, J.T., J.G. Neely, R.C. Paniello, C.C. Voelker, B. Nussenbaum, and E.W. Wang, *A practical guide to understanding Kaplan-Meier curves.* Otolaryngol Head Neck Surg, 2010. **143**(3): p. 331-6.

13. Cox, D.R., *Regression Models and Life-Tables.* Journal of the Royal Statistical Society. Series B (Methodological), 1972. **34**(2): p. 187-220.

14. Fine, J.P. and R.J. Gray, *A Proportional Hazards Model for the Subdistribution of a Competing Risk.* Journal of the American Statistical Association, 1999. **94**(446): p. 496-509.

15. Skyrud, K.D., F. Bray, and B. Moller, *A comparison of relative and cause-specific survival by cancer site, age and time since diagnosis.* Int J Cancer, 2014. **135**(1): p. 196-203.

16. Percy, C., E. Stanek, 3rd, and L. Gloeckler, *Accuracy of cancer death certificates and its effect on cancer mortality statistics.* American journal of public health, 1981. **71**(3): p. 242-250.

17. Pinheiro, P.S., C.R. Morris, L. Liu, T.J. Bungum, and S.F. Altekruse, *The impact of follow-up type and missed deaths on population-based cancer survival studies for Hispanics and Asians.* J Natl Cancer Inst Monogr, 2014. **2014**(49): p. 210-7.

18. Talback, M. and P.W. Dickman, *Estimating expected survival probabilities for relative survival analysis--exploring the impact of including cancer patient mortality in the calculations.* Eur J Cancer, 2011. **47**(17): p. 2626-32.

19. Andersen, P.K. and M. Vaeth, *Simple parametric and nonparametric models for excess and relative mortality.* Biometrics, 1989. **45**(2): p. 523-35.

20. Berkson, J. and R.P. Gage, *Calculation of survival rates for cancer.* Proc Staff Meet Mayo Clin, 1950. **25**(11): p. 270-86.

21. Hill, C., A. Laplanche, and A. Rezvani, *Comparison of the mortality of a cohort with the mortality of a reference population in a prognostic study.* Stat Med, 1985. **4**(3): p. 295-302.

22. Breslow, N.E., J.H. Lubin, P. Marek, and B. Langholz, *Multiplicative Models and Cohort Analysis.* Journal of the American Statistical Association, 1983. **78**(381): p. 1-12.

23. Bolard, P., C. Quantin, J. Esteve, J. Faivre, and M. Abrahamowicz, *Modelling time-dependent hazard ratios in relative survival: application to colon cancer.* J Clin Epidemiol, 2001. **54**(10): p. 986-96.

24. Buckley, J.D., *Additive and multiplicative models for relative survival rates.* Biometrics, 1984. **40**(1): p. 51-62.

25. Danieli, C., L. Remontet, N. Bossard, L. Roche, and A. Belot, *Estimating net survival: the importance of allowing for informative censoring.* Stat Med, 2012. **31**(8): p. 775-86.

26. Perme, M.P., J. Stare, and J. Esteve, *On estimation in relative survival.* Biometrics, 2012. **68**(1): p. 113-20.

27. Robins, J.M. and D.M. Finkelstein, *Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests.* Biometrics, 2000. **56**(3): p. 779-88.

28. Lambert, P.C. and P. Royston, *Further Development of Flexible Parametric Models for Survival Analysis.* The Stata Journal, 2009. **9**(2): p. 265-290.

29. Grzebyk, M. and I. Urmès, *STNS: Stata module for estimation of net survival*. 2012, Statistical Software Components, Boston College Department of Economics.

30. Ederer, F., L.M. Axtell, and S.J. Cutler, *The relative survival rate: a statistical methodology.* Natl Cancer Inst Monogr, 1961. **6**: p. 101-21.

31. Ederer, F. and H. Heise, *The effect of eliminating deaths from cancer on general population survival rates.* Methodological Note, 1959. **11**(23).

32. Hakulinen, T., *Cancer survival corrected for heterogeneity in patient withdrawal.* Biometrics, 1982. **38**(4): p. 933-42.

33. Giorgi, R., M. Abrahamowicz, C. Quantin, P. Bolard, J. Esteve, J. Gouvernet, and J. Faivre, *A relative survival regression model using B-spline functions to model non-proportional hazards.* Stat Med, 2003. **22**(17): p. 2767-84.

34. Remontet, L., N. Bossard, A. Belot, J. Estève, and F. the French network of cancer registries, *An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies.* Statistics in Medicine, 2007. **26**(10): p. 2214-2228.

35. Mahboubi, A., M. Abrahamowicz, R. Giorgi, C. Binquet, C. Bonithon-Kopp, and C. Quantin, *Flexible modeling of the effects of continuous prognostic factors in relative survival.* Statistics in Medicine, 2011. **30**(12): p. 1351-1365.

36. Nelson, C.P., P.C. Lambert, I.B. Squire, and D.R. Jones, *Flexible parametric models for relative survival, with application in coronary heart disease.* Stat Med, 2007. **26**(30): p. 5486-98.

37. Royston, P., *Flexible Parametric Alternatives to the Cox Model, and more.* The Stata Journal, 2001. **1**(1): p. 1-28.

38. Royston, P. and M.K. Parmar, *Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects.* Stat Med, 2002. **21**(15): p. 2175-97.

39. Durrleman, S. and R. Simon, *Flexible regression models with cubic splines.* Stat Med, 1989. **8**(5): p. 551-61.

40. Smith, P.L., *Splines As a Useful and Convenient Statistical Tool.* The American Statistician, 1979. **33**(2): p. 57-62.

41. Lambert, P., *STPM2: Stata module to estimate flexible parametric survival models*. 2010, Boston College Department of Economics.

42. Boag, J.W., *Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy.* Journal of the Royal Statistical Society. Series B (Methodological), 1949. **11**(1): p. 15-53.

43. Berkson, J. and R.P. Gage, *Survival Curve for Cancer Patients Following Treatment.* Journal of the American Statistical Association, 1952. **47**(259): p. 501-515.

44. Li, C.S. and J.M. Taylor, *A semi-parametric accelerated failure time cure model.* Stat Med, 2002. **21**(21): p. 3235-47.

45. Maller, R.A. and S. Zhou, *Testing for the presence of immune or cured individuals in censored survival data.* Biometrics, 1995. **51**(4): p. 1197-205.

46. Peng, Y. and K.B. Dear, *A nonparametric mixture model for cure rate estimation.* Biometrics, 2000. **56**(1): p. 237-43.

47. Sy, J.P. and J.M. Taylor, *Estimation in a Cox proportional hazards cure model.* Biometrics, 2000. **56**(1): p. 227-36.

48. *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. Series in Mathematical Biology and Medicine. Vol. Volume 1. 1996, WORLD SCIENTIFIC. 288.

49. Tsodikov, A.D., J.G. Ibrahim, and A.Y. Yakovlev, *Estimating Cure Rates From Survival Data: An Alternative to Two-Component Mixture Models.* J Am Stat Assoc, 2003. **98**(464): p. 1063-1078.

50. Chen, M.-H., J.G. Ibrahim, and D. Sinha, *A New Bayesian Model for Survival Data with a Surviving Fraction.* Journal of the American Statistical Association, 1999. **94**(447): p. 909-919.

51. De Angelis, R., R. Capocaccia, T. Hakulinen, B. Soderman, and A. Verdecchia, *Mixture models for cancer survival analysis: application to population-based data with covariates.* Stat Med, 1999. **18**(4): p. 441-54.

52. Verdecchia, A., R. De Angelis, R. Capocaccia, M. Sant, A. Micheli, G. Gatta, and F. Berrino, *The cure for colon cancer: results from the EUROCARE study.* Int J Cancer, 1998. **77**(3): p. 322-9.

53. Lambert, P.C., J.R. Thompson, C.L. Weston, and P.W. Dickman, *Estimating and modeling the cure fraction in population-based cancer survival analysis.* Biostatistics, 2007. **8**(3): p. 576-94.

54. Lambert, P.C., *Modeling of the Cure Fraction in Survival Studies.* The Stata Journal, 2007. **7**(3): p. 351-375.

55. Lambert, P.C., P.W. Dickman, P. Osterlund, T. Andersson, R. Sankila, and B. Glimelius, *Temporal trends in the proportion cured for cancer of the colon and rectum: a population-based study using data from the Finnish Cancer Registry.* Int J Cancer, 2007. **121**(9): p. 2052-9.

56. Andersson, T.M., P.W. Dickman, S. Eloranta, and P.C. Lambert, *Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models.* BMC Med Res Methodol, 2011. **11**: p. 96.

57. Andersson, T.M., *Quantifying cancer patient survival; extensions and applications of cure models and life expectancy estimation*, in *Department of Medical Epidemiology and Biostatistics*. 2013, Karolinska Institutet, Stockholm, Sweden. p. 66.

58. Yu, X.Q., R. De Angelis, T.M. Andersson, P.C. Lambert, D.L. O'Connell, and P.W. Dickman, *Estimating the proportion cured of cancer: some practical advice for users.* Cancer Epidemiol, 2013. **37**(6): p. 836-42.

59. Andersson, T.M. and P.C. Lambert, *Fitting and Modeling Cure in Population-Based Cancer Studies within the Framework of Flexible Parametric Survival Models.* Stata Journal, 2012. **12**: p. 623-638.

60. Yu, B., *A minimum version of log-rank test for testing the existence of cancer cure using relative survival data.* Biom J, 2012. **54**(1): p. 45-60.

61. Peng, Y., K.B. Dear, and K.C. Carriere, *Testing for the presence of cured patients: a simulation study.* Stat Med, 2001. **20**(12): p. 1783-96.

62. Cvancarova, M., B. Aagnes, S.D. Fosså, P.C. Lambert, B. Møller, and F. Bray, *Proportion cured models applied to 23 cancer sites in Norway.* International journal of cancer, 2013. **132**(7): p. 1700-1710.

63. Francisci, S., R. Capocaccia, E. Grande, M. Santaquilani, A. Simonetti, C. Allemani, G. Gatta, M. Sant, G. Zigon, F. Bray, M. Janssen-Heijnen, and E.W. Group, *The cure of cancer: a European perspective.* European journal of cancer (Oxford, England : 1990), 2009. **45**(6): p. 1067-1079.

64. Smith, L., A.W. Glaser, S.E. Kinsey, D.C. Greenwood, L. Chilton, A.V. Moorman, and R.G. Feltbower, *Long-term survival after childhood acute lymphoblastic leukaemia: population-based trends in cure and relapse by clinical characteristics.* Br J Haematol, 2018. **182**(6): p. 851-858.

65. Sposto, R., *Cure model analysis in cancer: an application to data from the Children's Cancer Group.* Stat Med, 2002. **21**(2): p. 293-312.

66. Chauvenet, M., C. Lepage, V. Jooste, V. Cottet, J. Faivre, and A.M. Bouvier, *Prevalence of patients with colorectal cancer requiring follow-up or active treatment.* Eur J Cancer, 2009. **45**(8): p. 1460-5.

67. Dal Maso, L., S. Guzzinati, C. Buzzoni, R. Capocaccia, D. Serraino, A. Caldarella, A.P. Dei Tos, F. Falcini, M. Autelitano, G. Masanotti, S. Ferretti, F. Tisano, U. Tirelli, E. Crocetti, and R. De Angelis, *Long-term survival, prevalence, and cure of cancer: a population-based estimation for 818 902 Italian patients and 26 cancer types.* Ann Oncol, 2014. **25**(11): p. 2251-60.

68. Boussari, O., G. Romain, L. Remontet, N. Bossard, M. Mounier, A.-M. Bouvier, C. Binquet, M. Colonna, and V. Jooste, *A new approach to estimate time-to-cure from cancer registries data.* Cancer Epidemiology, 2018. **53**: p. 72-80.

69. Fritz, A., C. Percy, A. Jack, K. Shanmugaratnam, L. Sobin, D.M. Parkin, and S. Whelan, *International Classification of Diseases for Oncology*. 3rd edition ed. 2000, Geneva: World Health Organization.

70. Byrd, R., P. Lu, J. Nocedal, and C. Zhu, *A Limited Memory Algorithm for Bound Constrained Optimization.* SIAM Journal on Scientific Computing, 2003. **16**.

71. Capocaccia, R., G. Gatta, and L. Dal Maso, *Life expectancy of colon, breast, and testicular cancer patients: an analysis of US-SEER population-based data.* Ann Oncol, 2015. **26**(6): p. 1263-8.

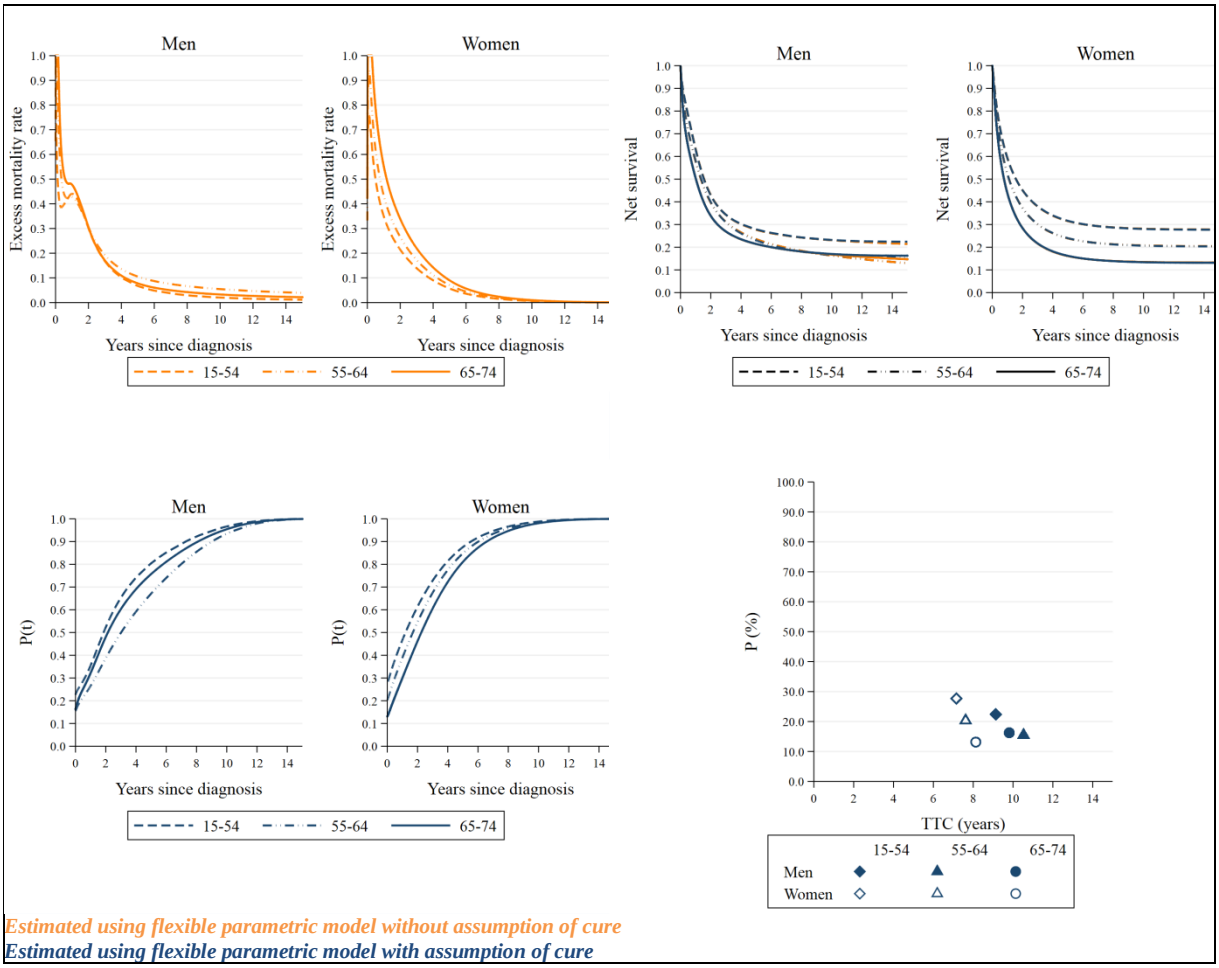
72. Phillips, N., A. Coldman, and M.L. McBride, *Estimating cancer prevalence using mixture models for cancer survival.* Stat Med, 2002. **21**(9): p. 1257-70.

73. Botta, L., G. Gatta, A. Trama, and R. Capocaccia, *Excess risk of dying of other causes of cured cancer patients.* Tumori, 2019. **105**(3): p. 199-204.

ANNEXE A : Vérification graphique de l'hypothèse de guérison, courbes de la probabilité d'être guéris au cours du temps et estimation de la proportion de guéris et du délai de guérison dans le cancer en France

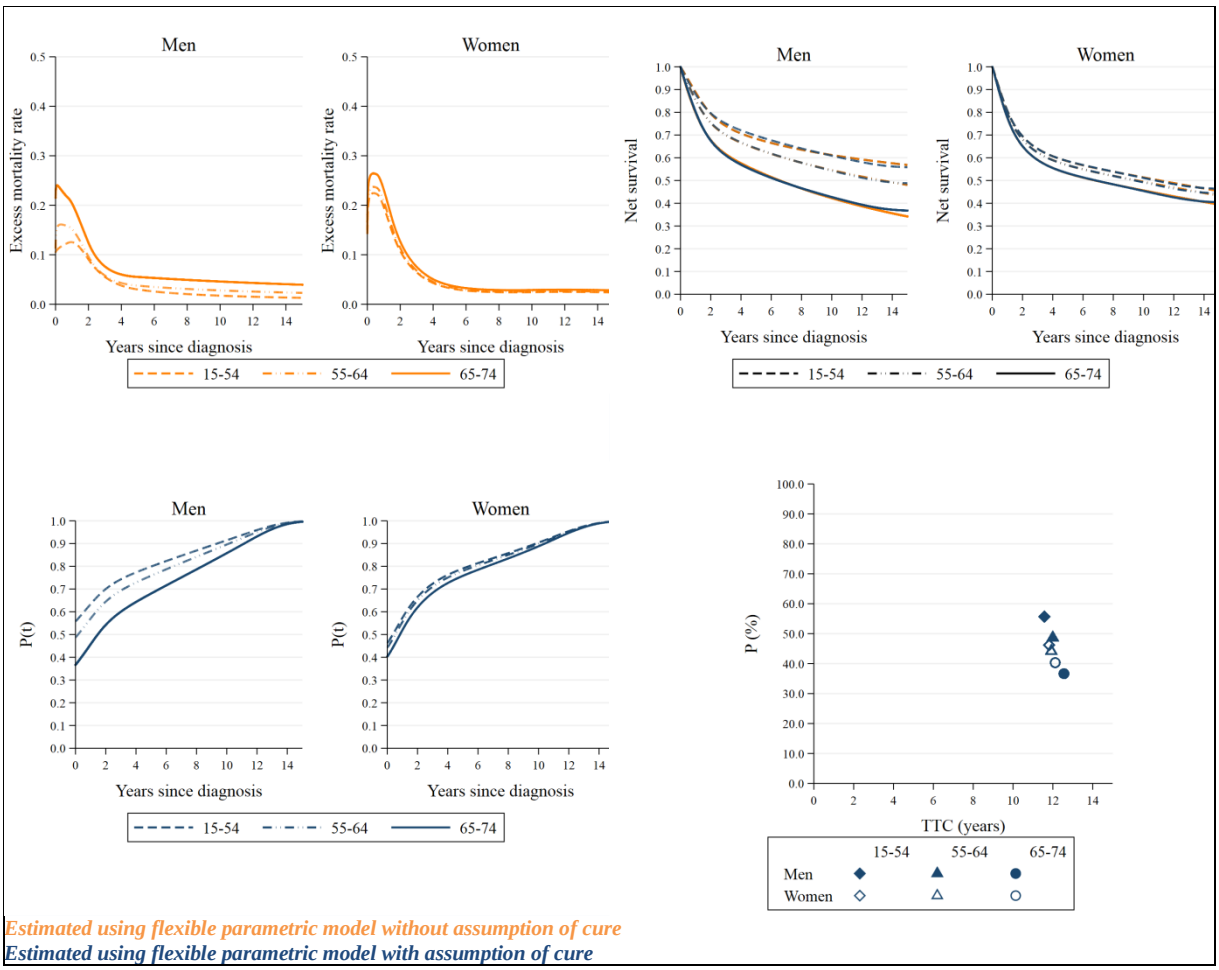
Supplementary file 1.
Excess mortality rate (**top left**) and net survival (**top right**) were estimated through a **flexible model without assumption of cure** and are provided, according to time since diagnosis up to 15 years, for all combinations of cancer site, sex and age at diagnosis.
For combinations of cancer site, sex and age at diagnosis for which the assumption of cure was accepted, net survival, patient's probability of belonging to the cured group (P(t)), cure proportion (P) and time-to-cure (TTC) were estimated through a **flexible model with assumption of cure**. Net survival (**top right**) and P(t) (**bottom left**) are provided according to time since diagnosis up to 15 years. P is plotted against TTC (**bottom right**) in order to describe changes in these parameters according to sex and age for each site (adapted from Verdecchia *et al.*[52]).

1. Biliary tracts



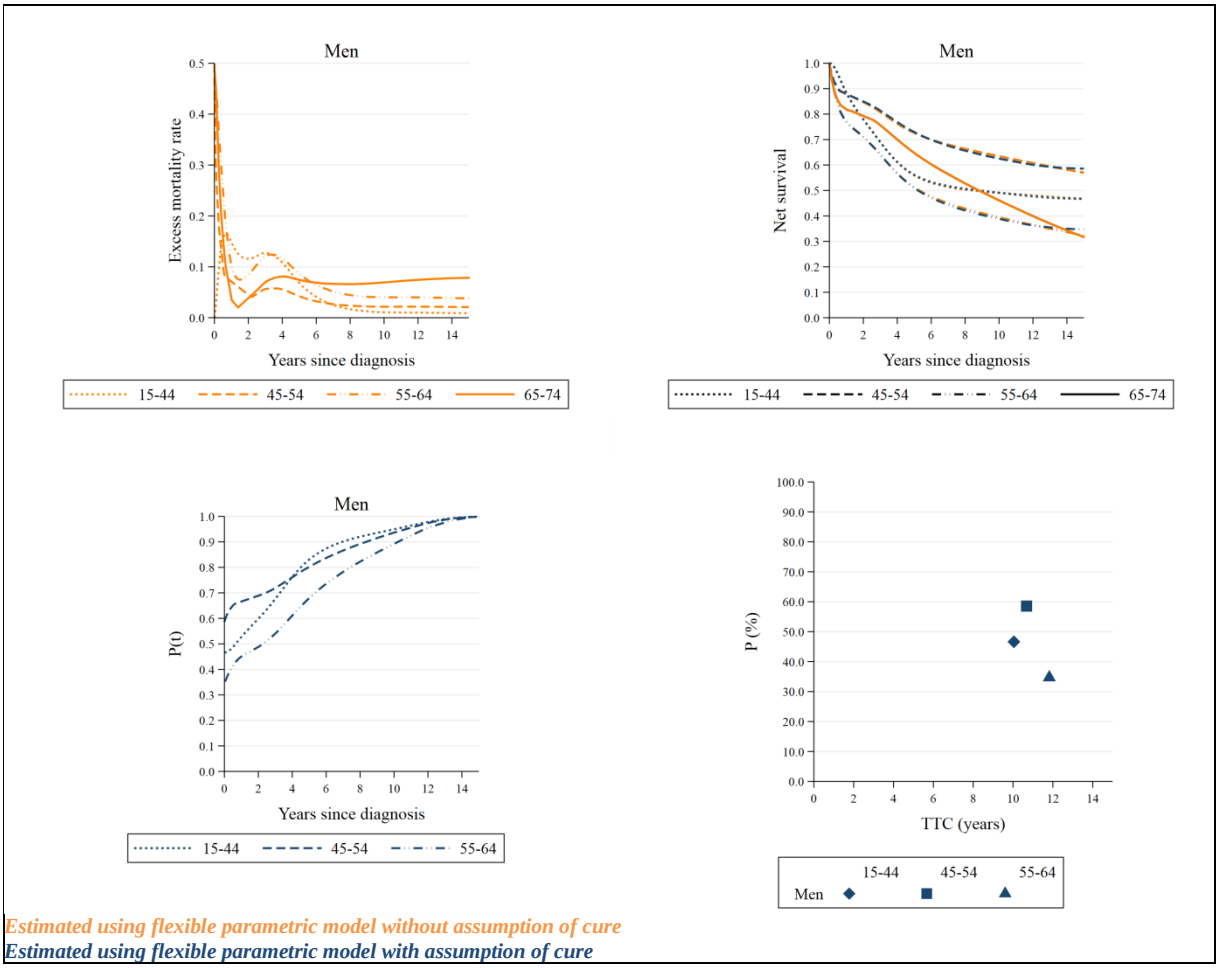
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

2. Bladder



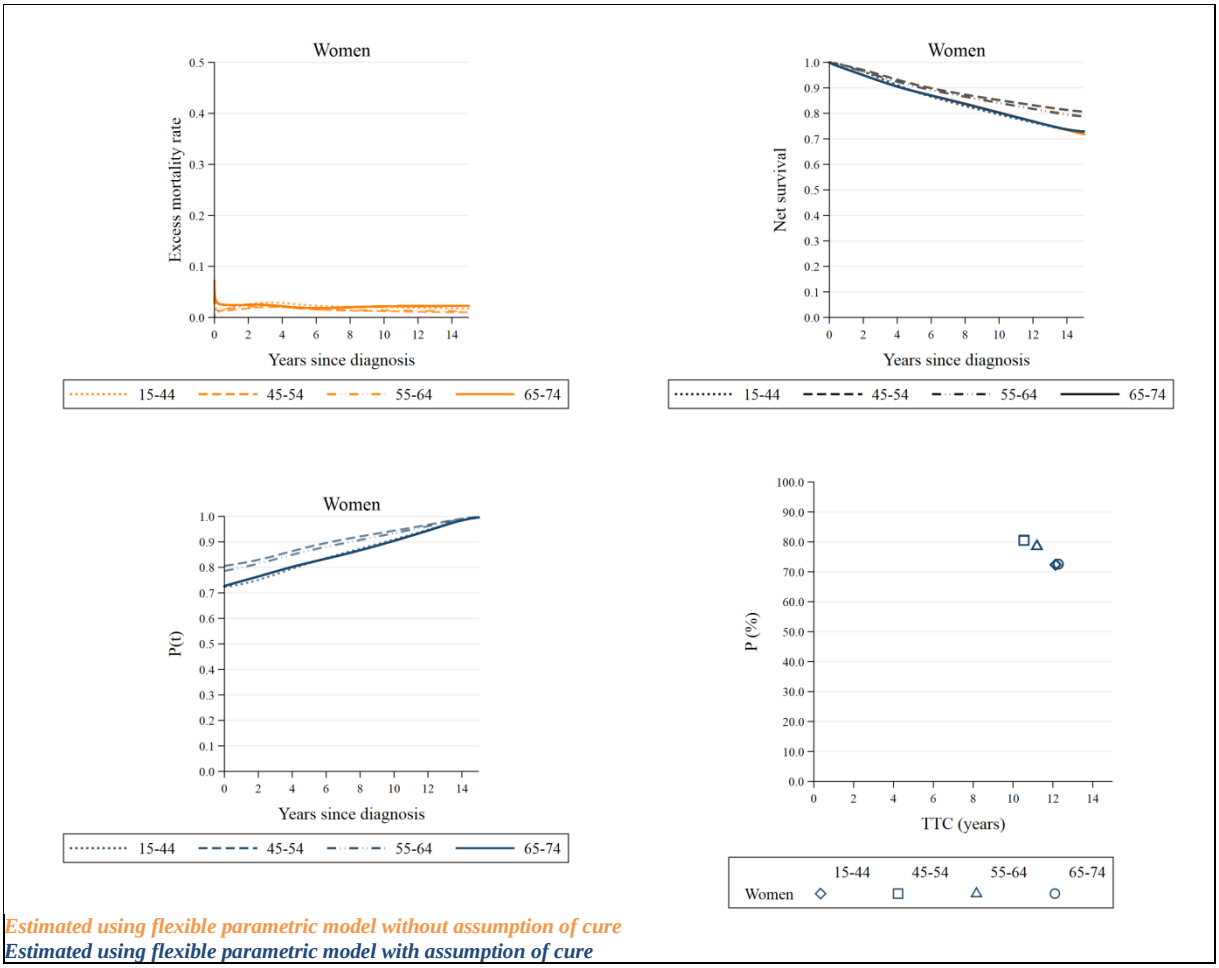
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

3. Bones



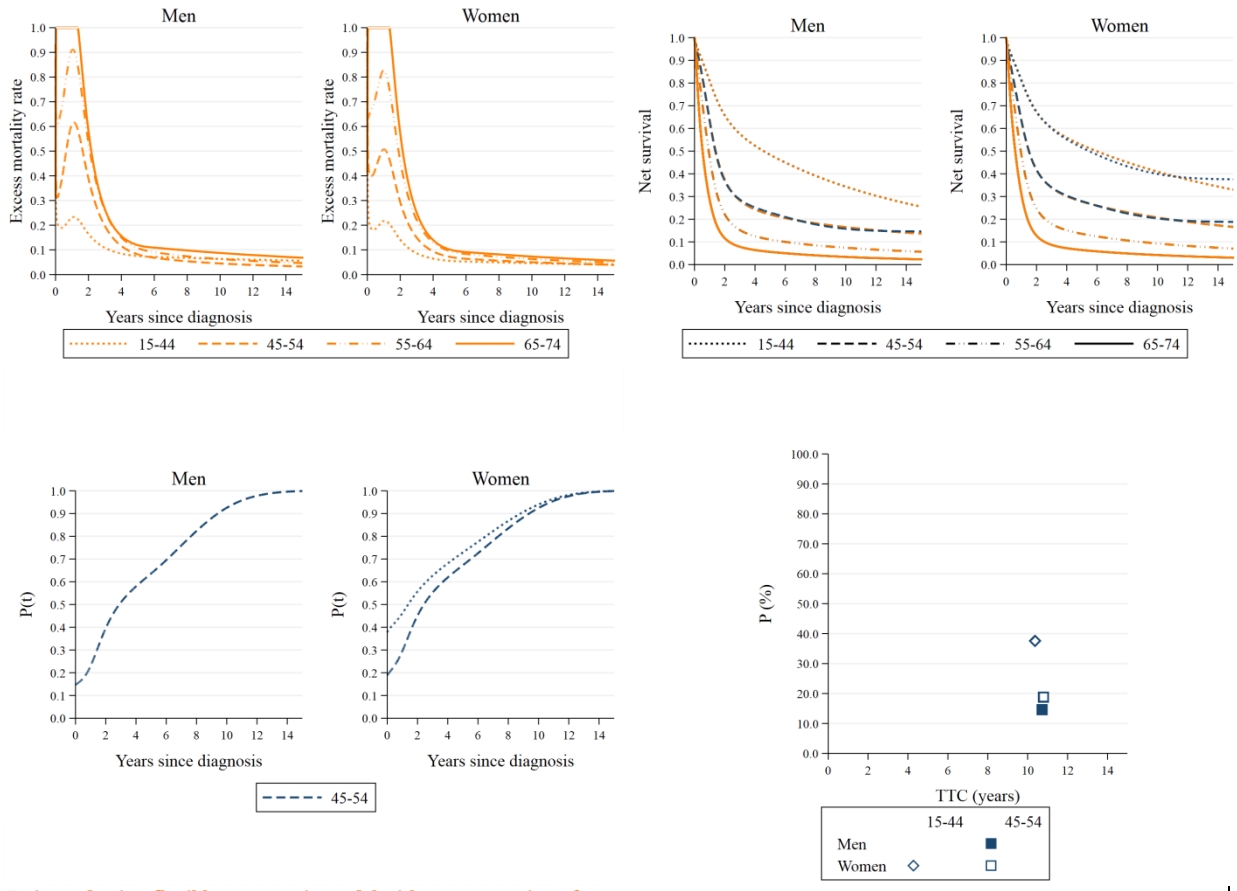
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

4. Breast



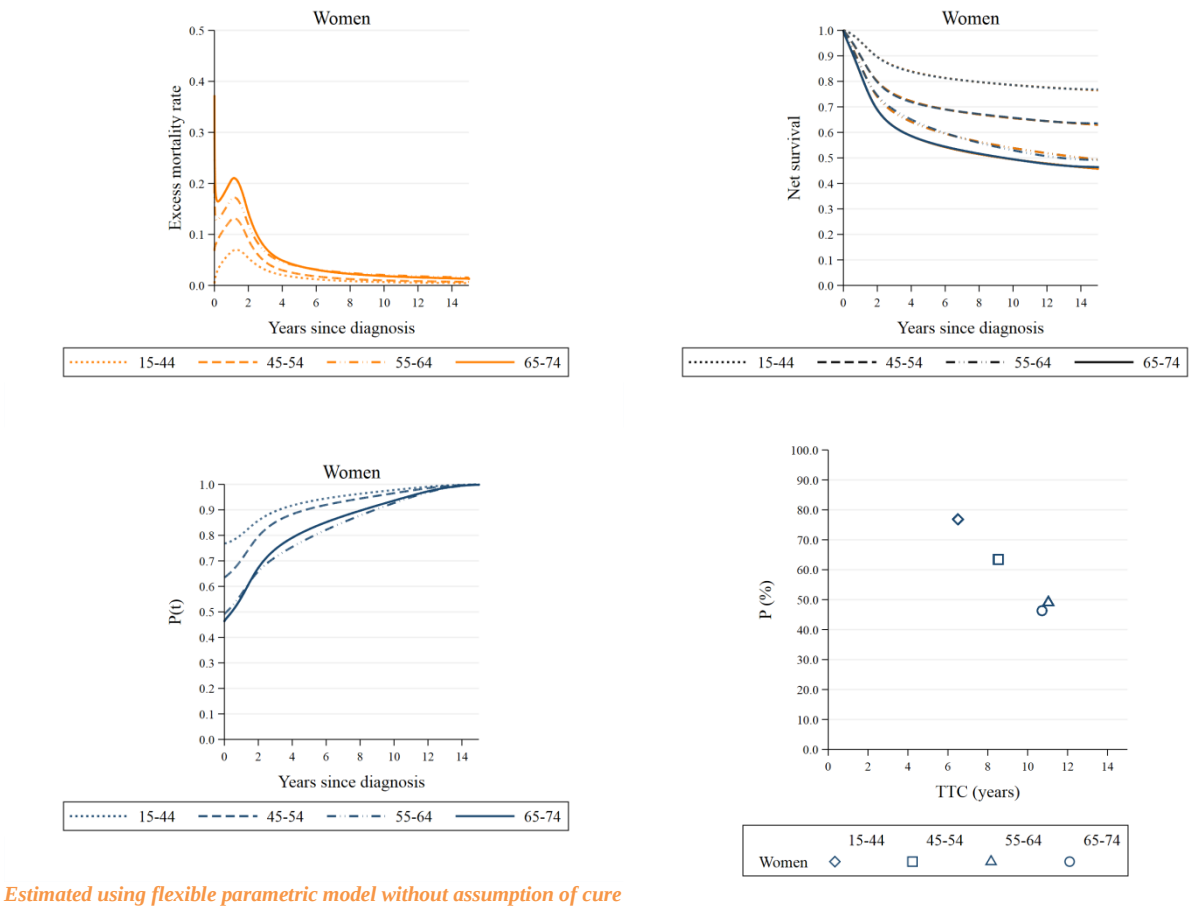
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

5. Central nervous system



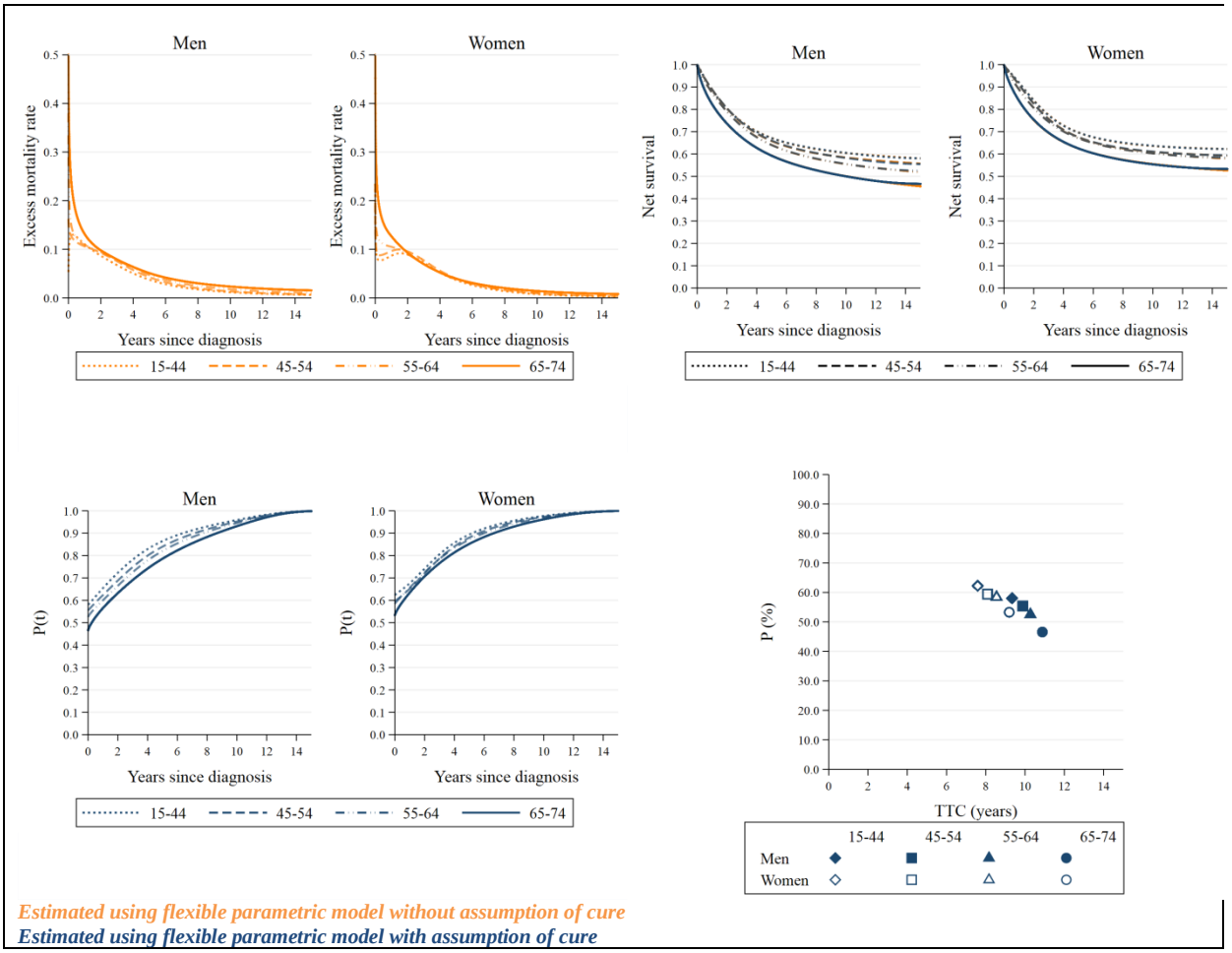
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

6. Cervix uteri



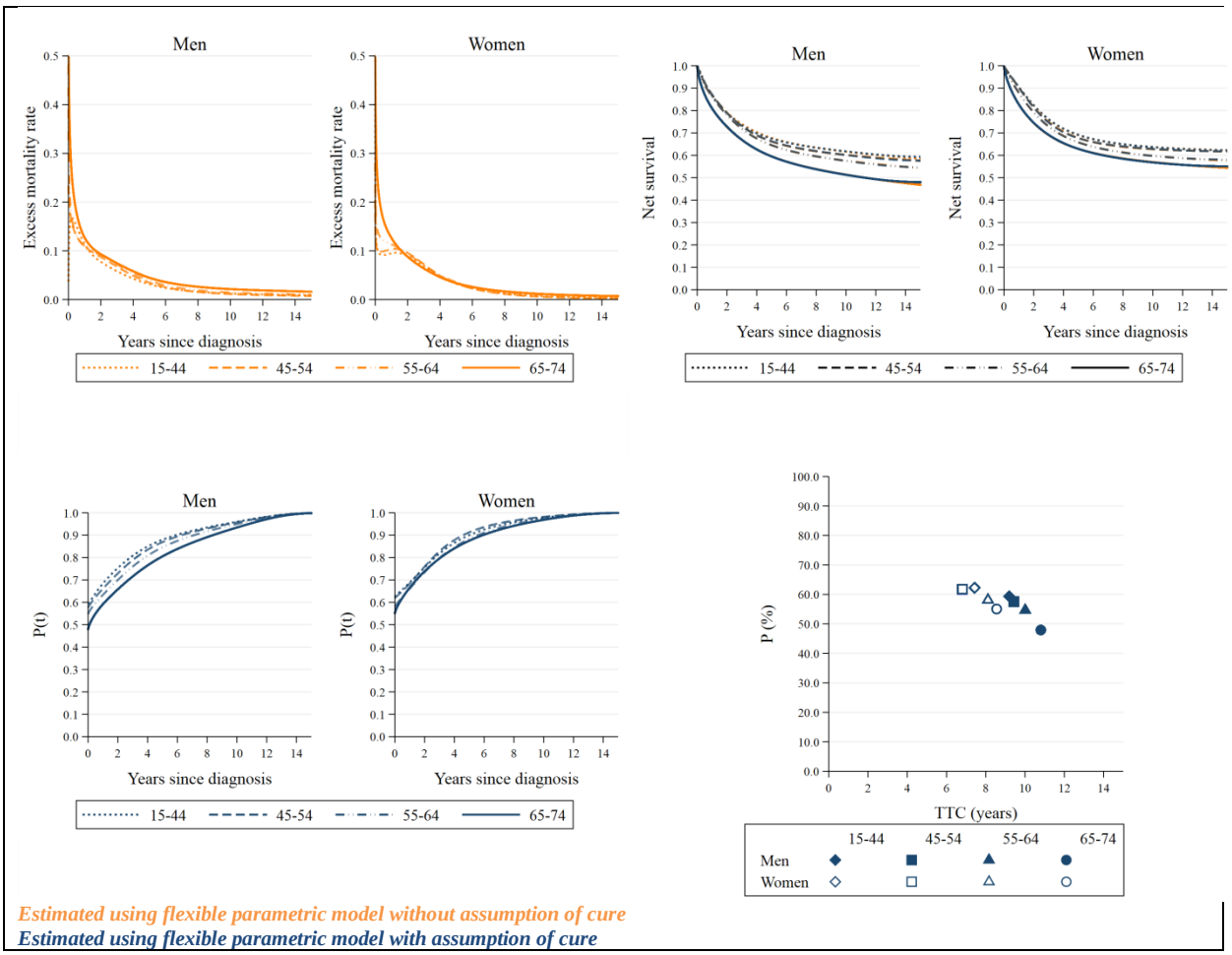
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

7. Colon and Rectum



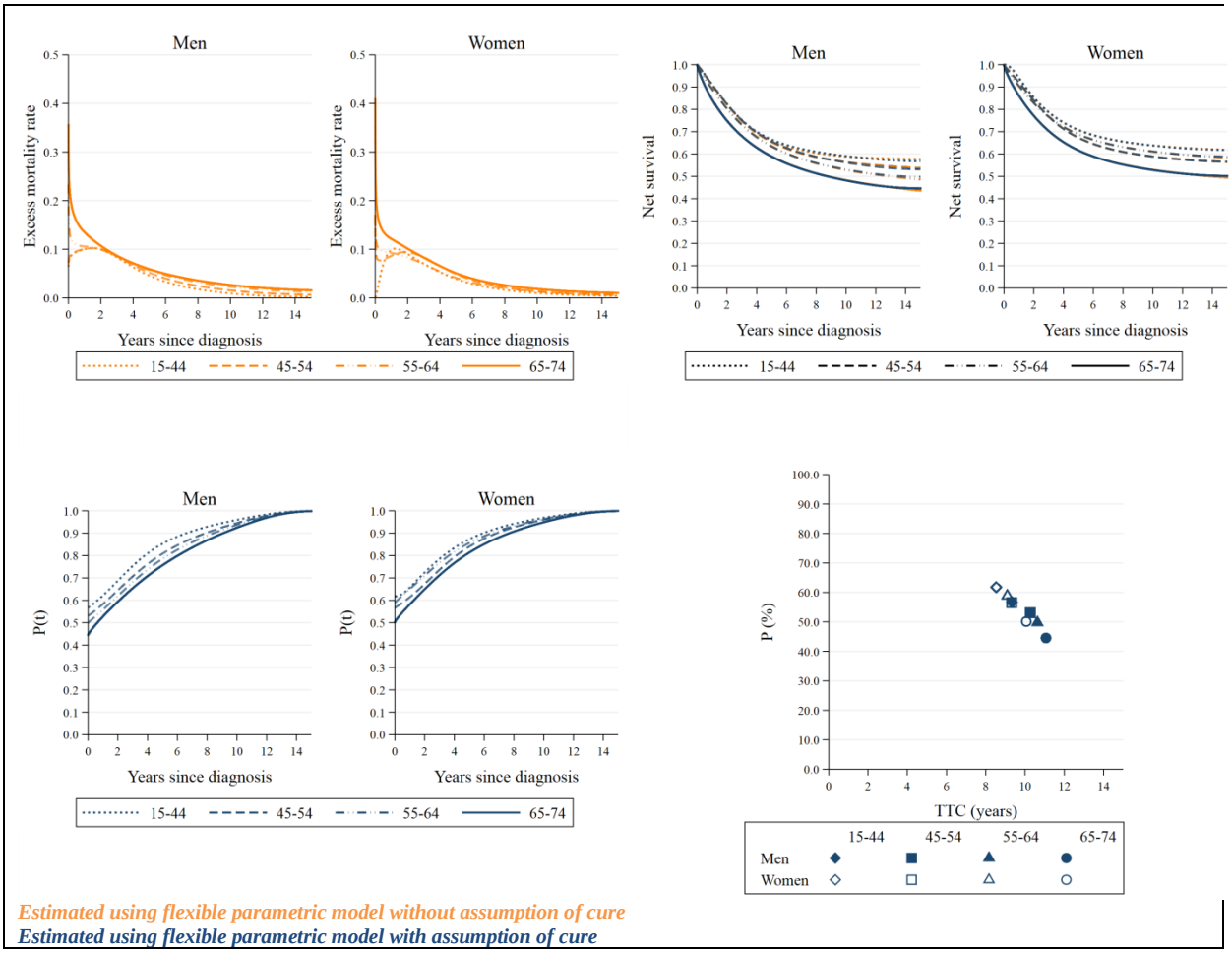
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

a. Colon



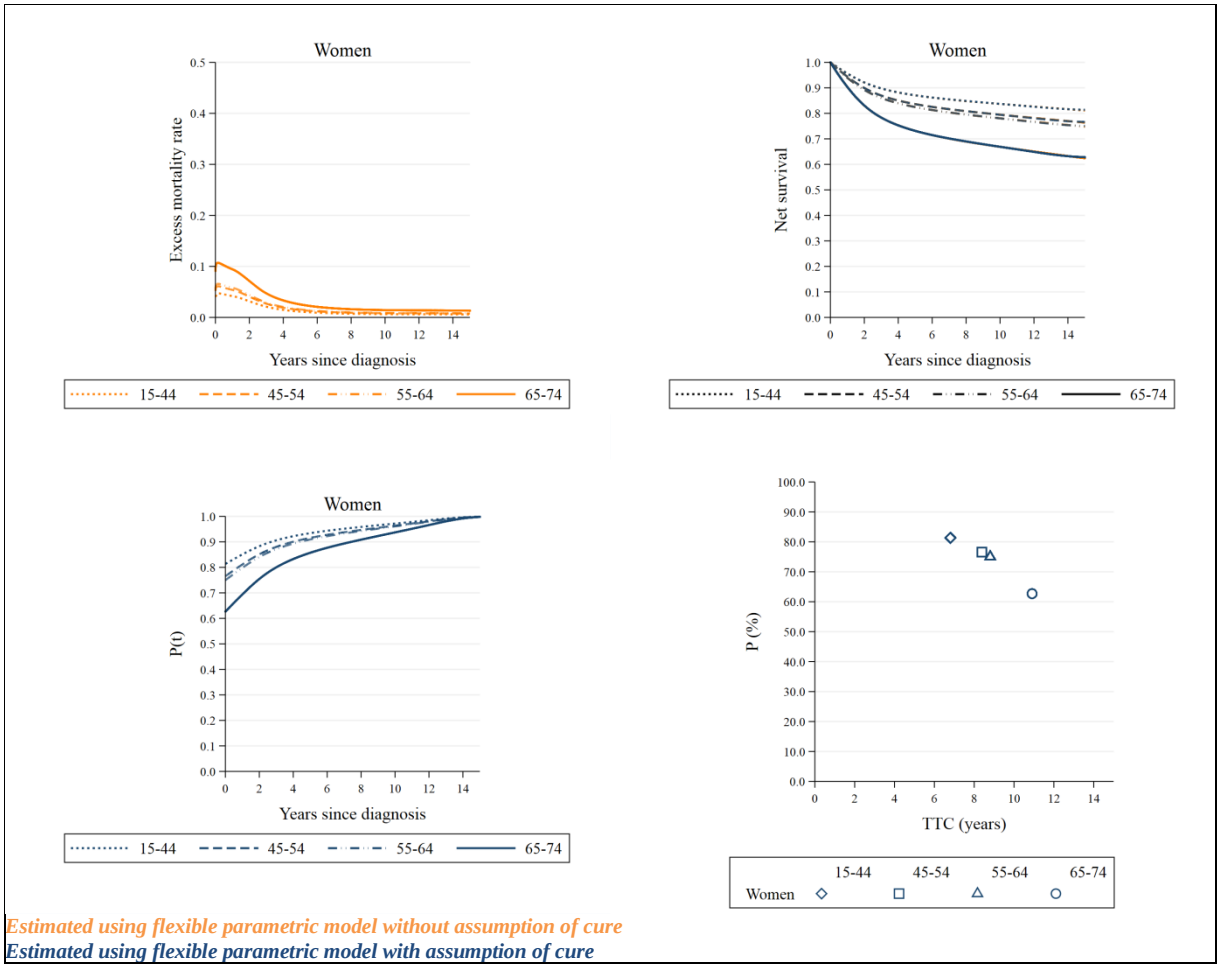
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

b. Rectum



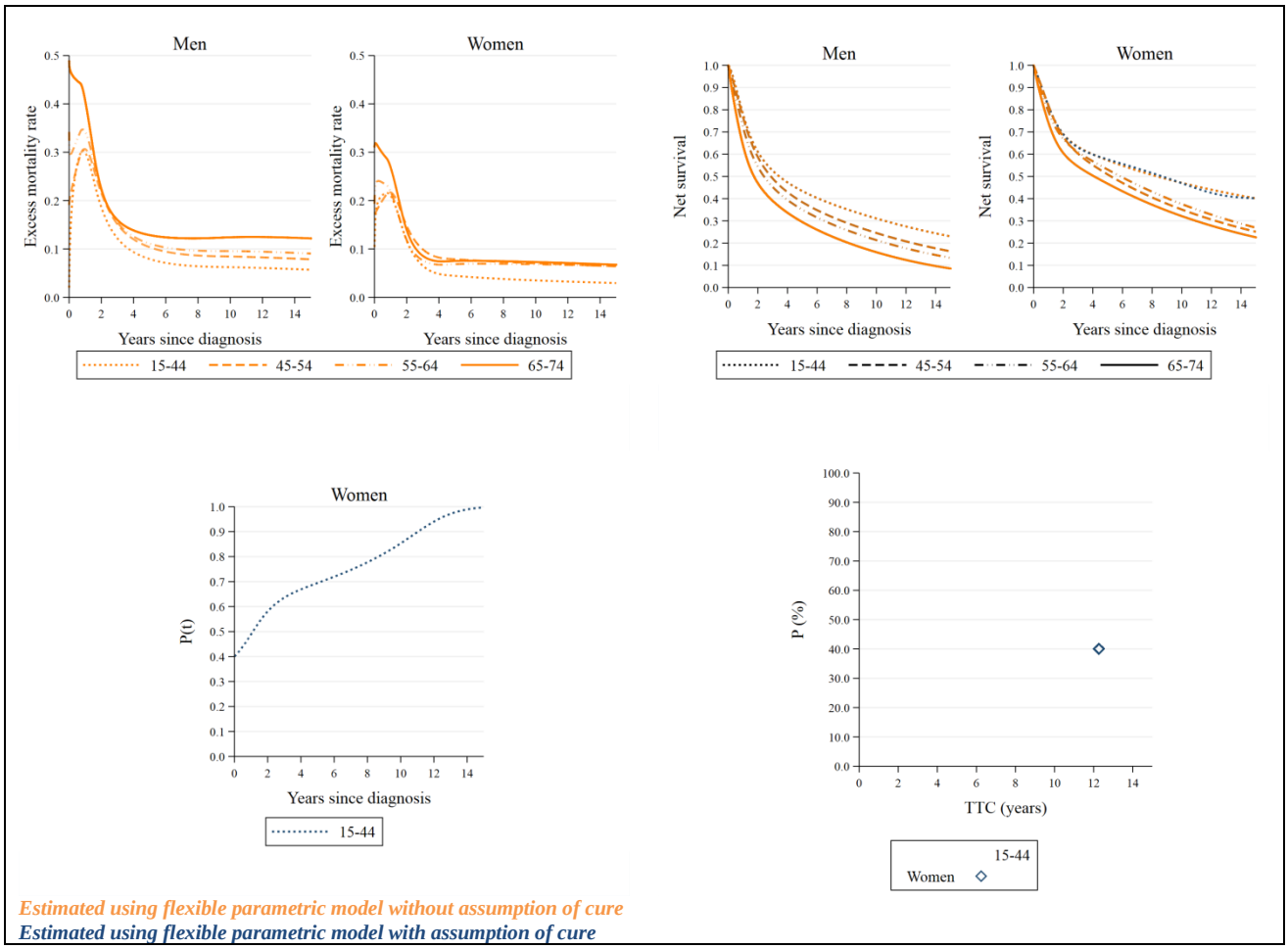
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

8. Corpus uteri



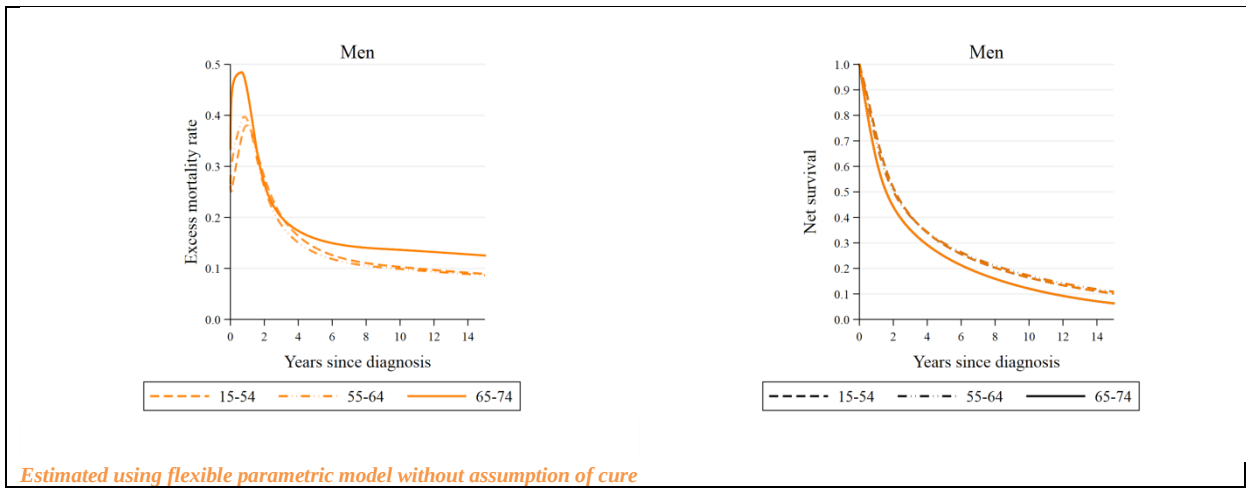
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

9. Head and neck



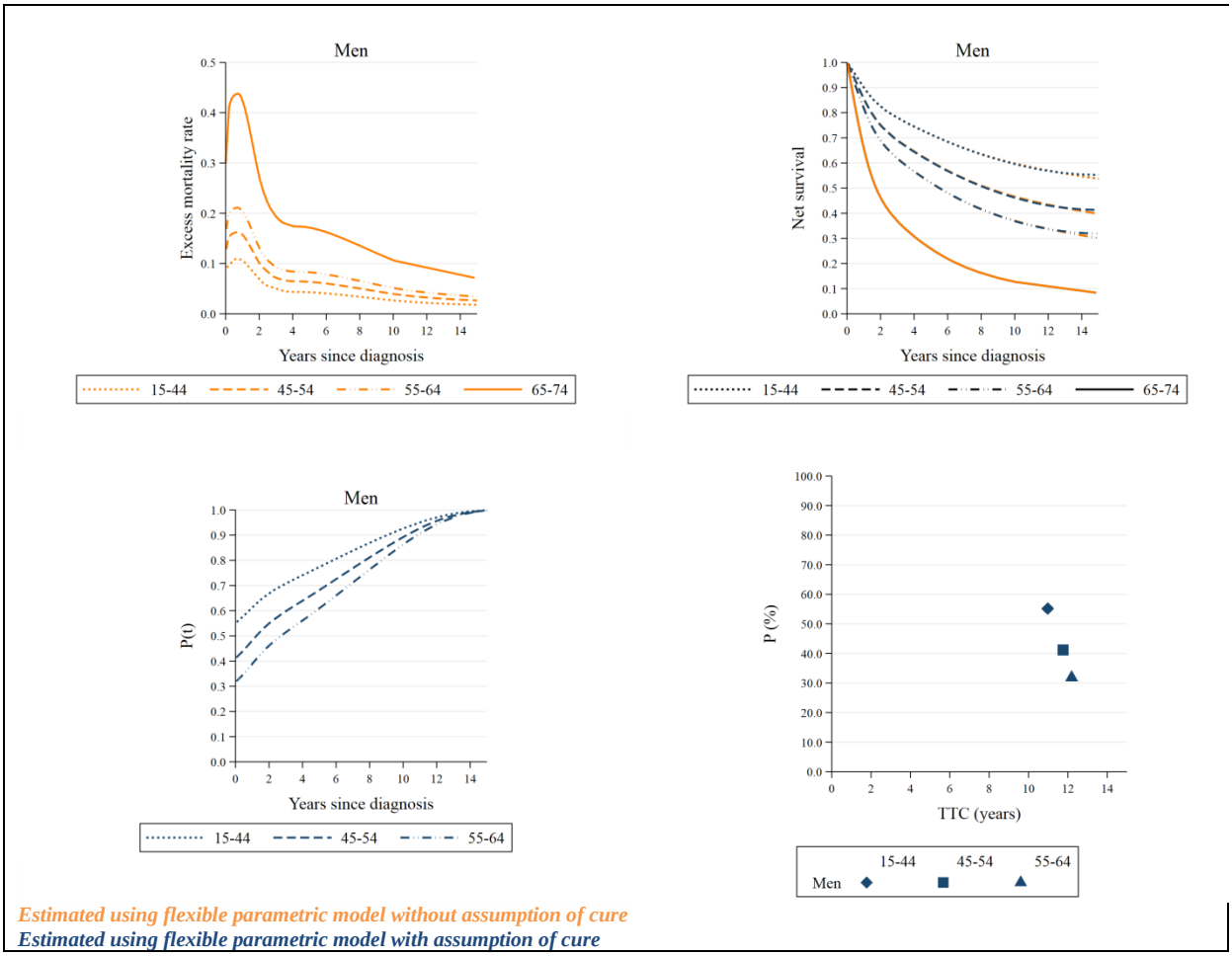
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

a. Hypopharynx



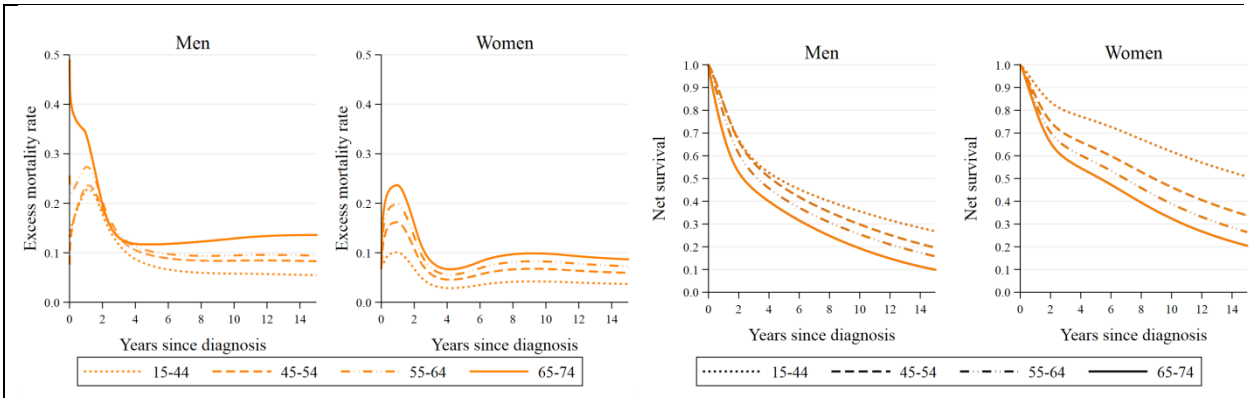
Estimated using flexible parametric model without assumption of cure

b. *Nasopharynx*



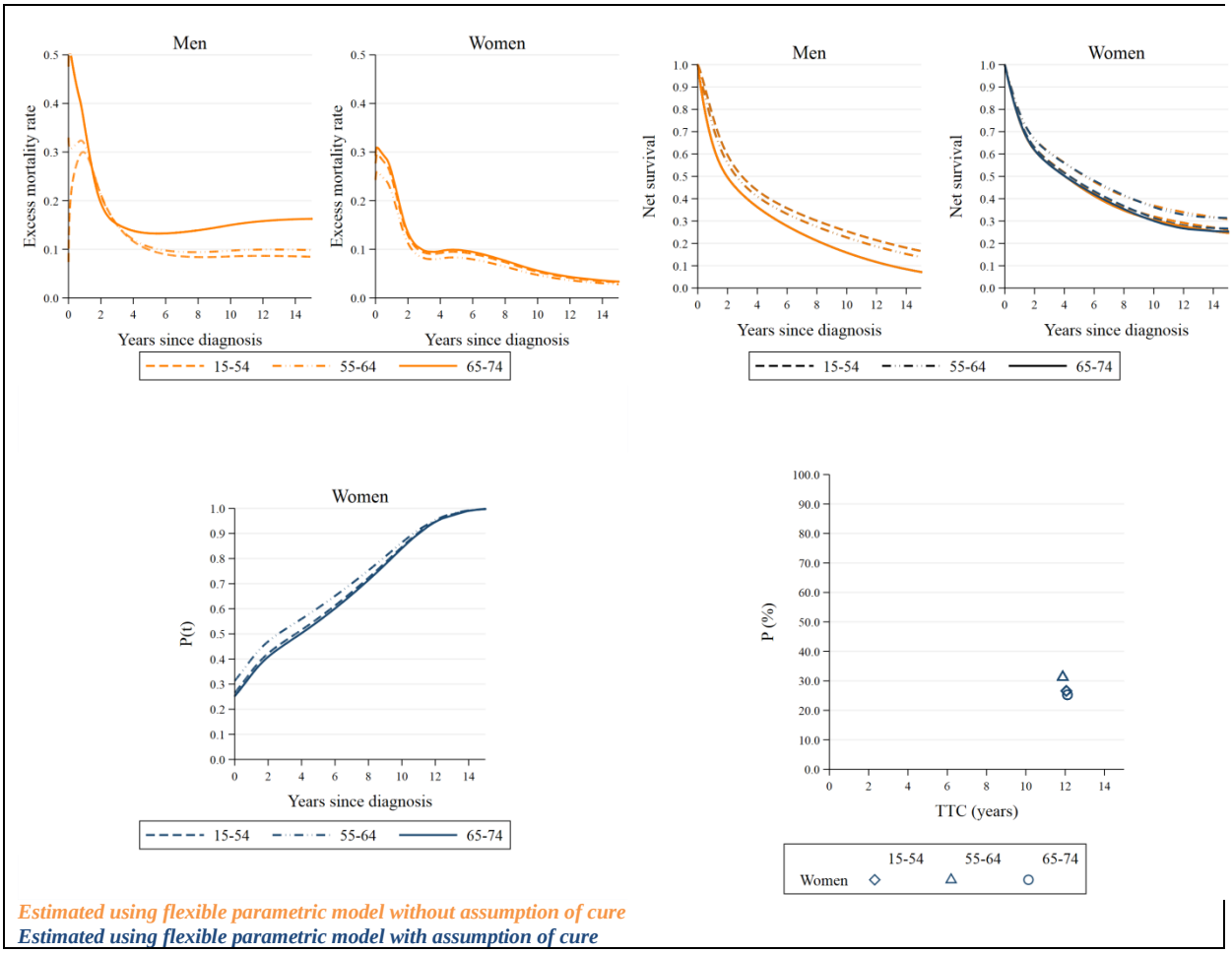
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

c. *Oral cavity*



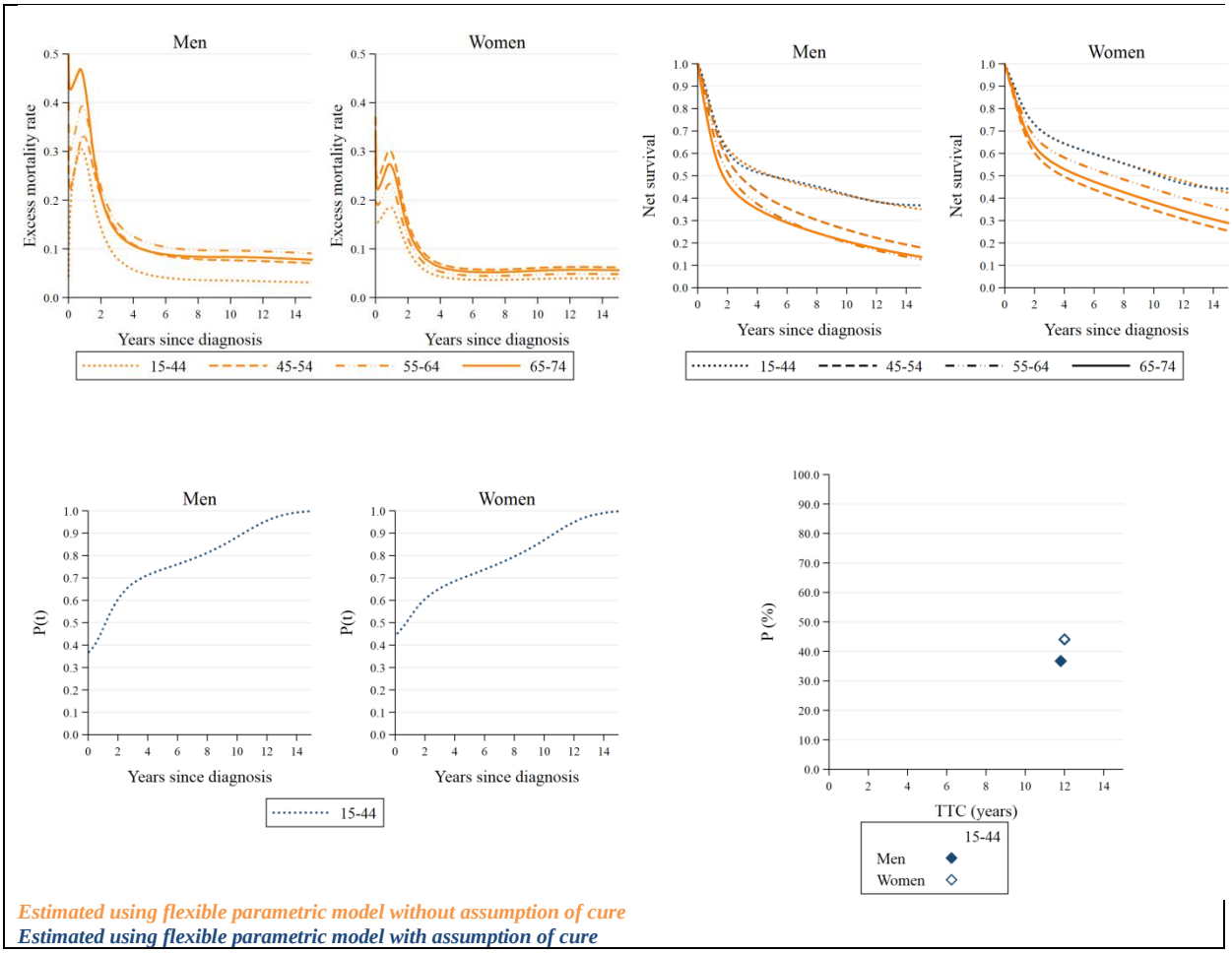
Estimated using flexible parametric model without assumption of cure

d. Oropharynx



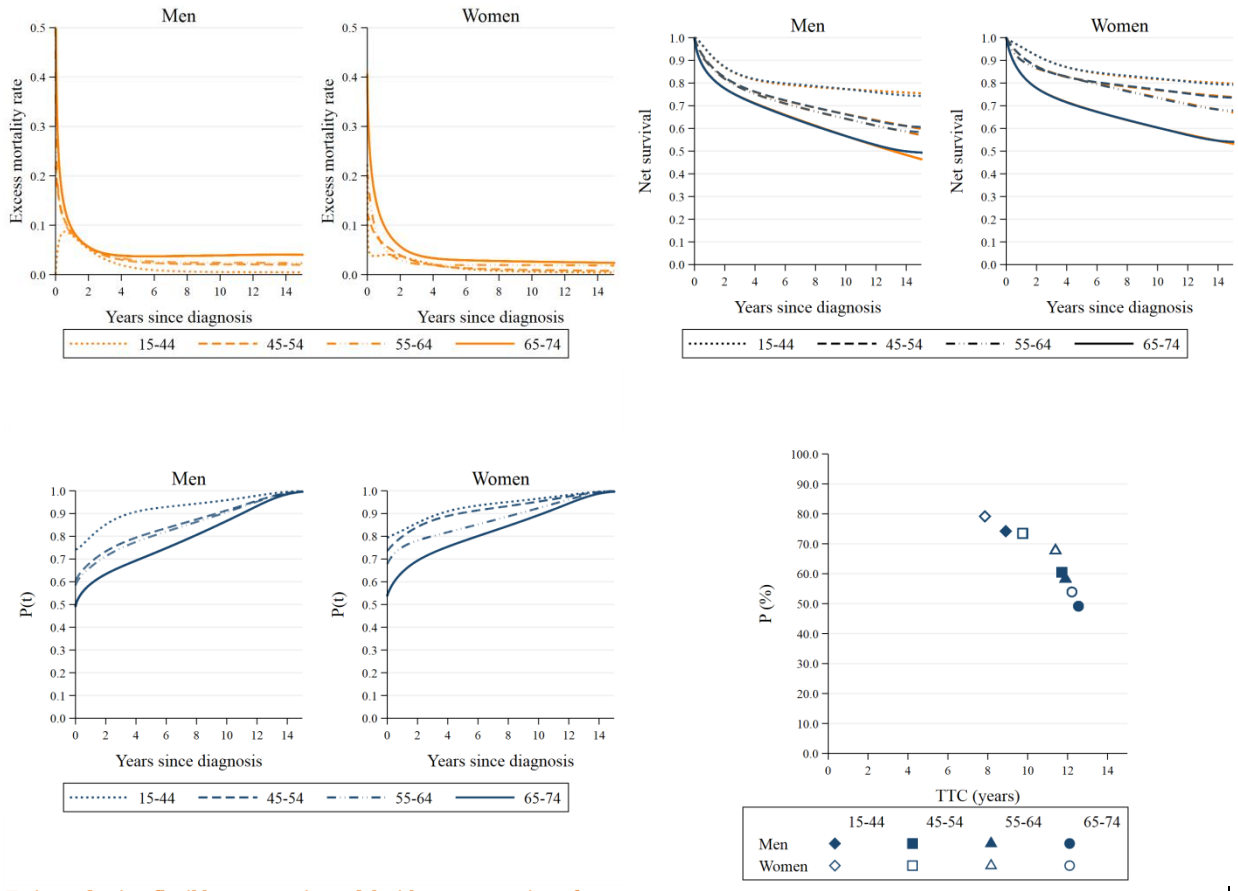
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

e. Tongue



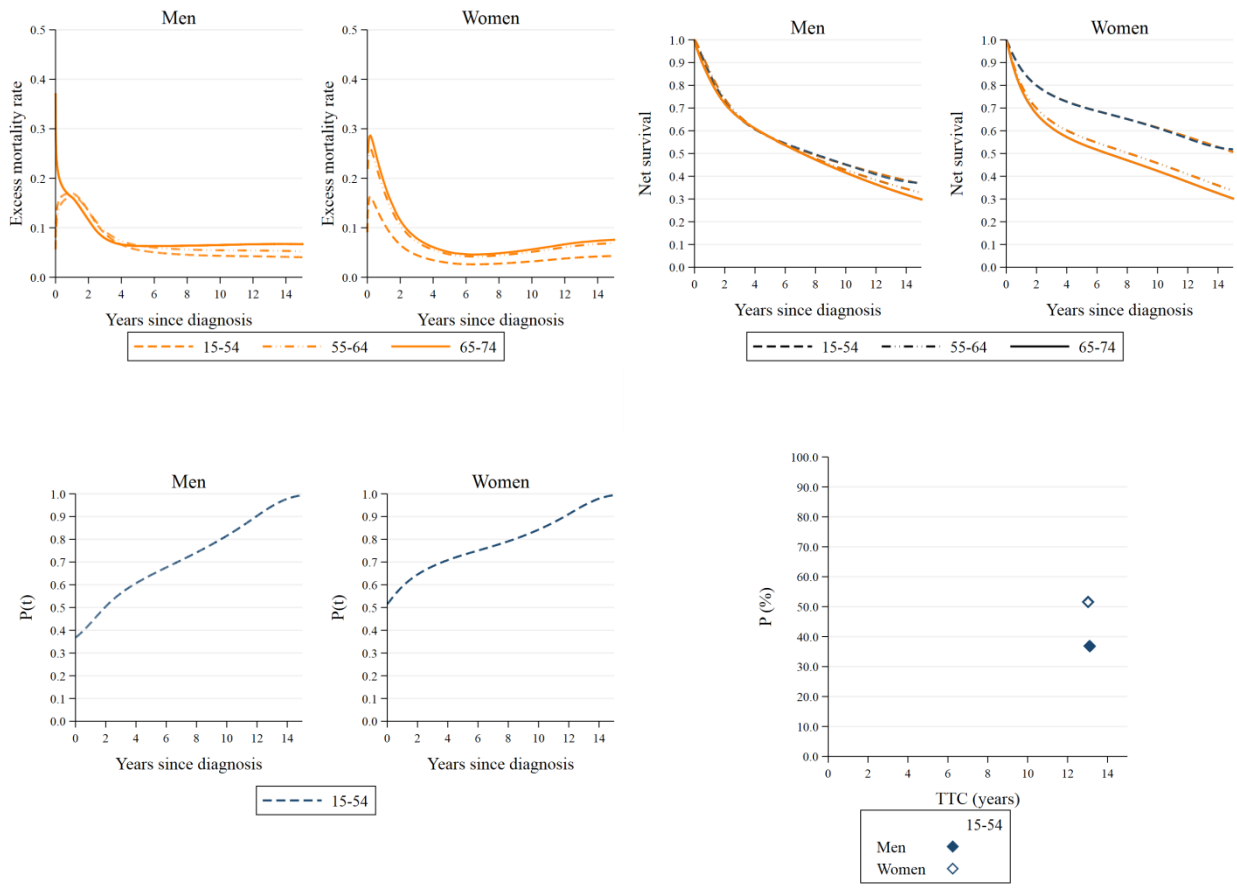
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

10. Kidney



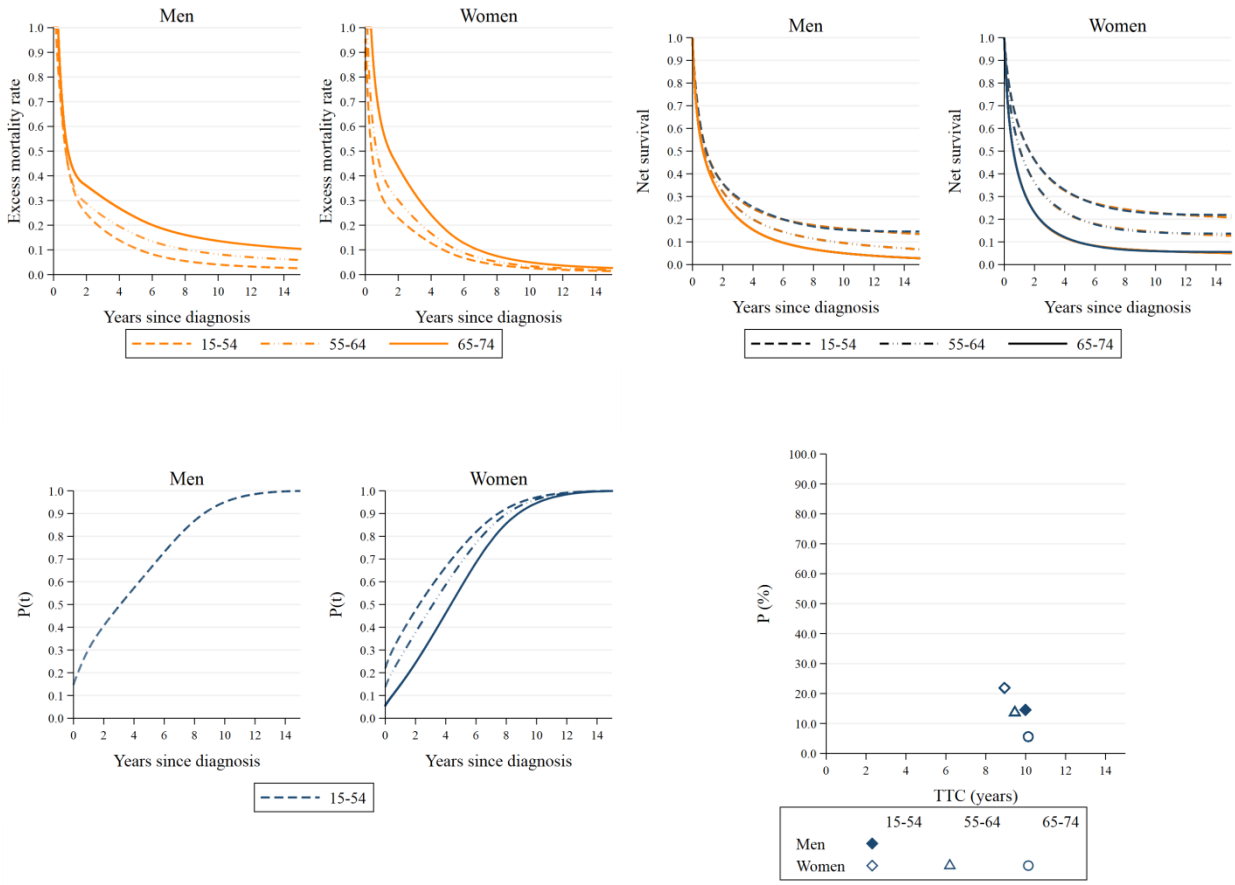
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

11. Larynx



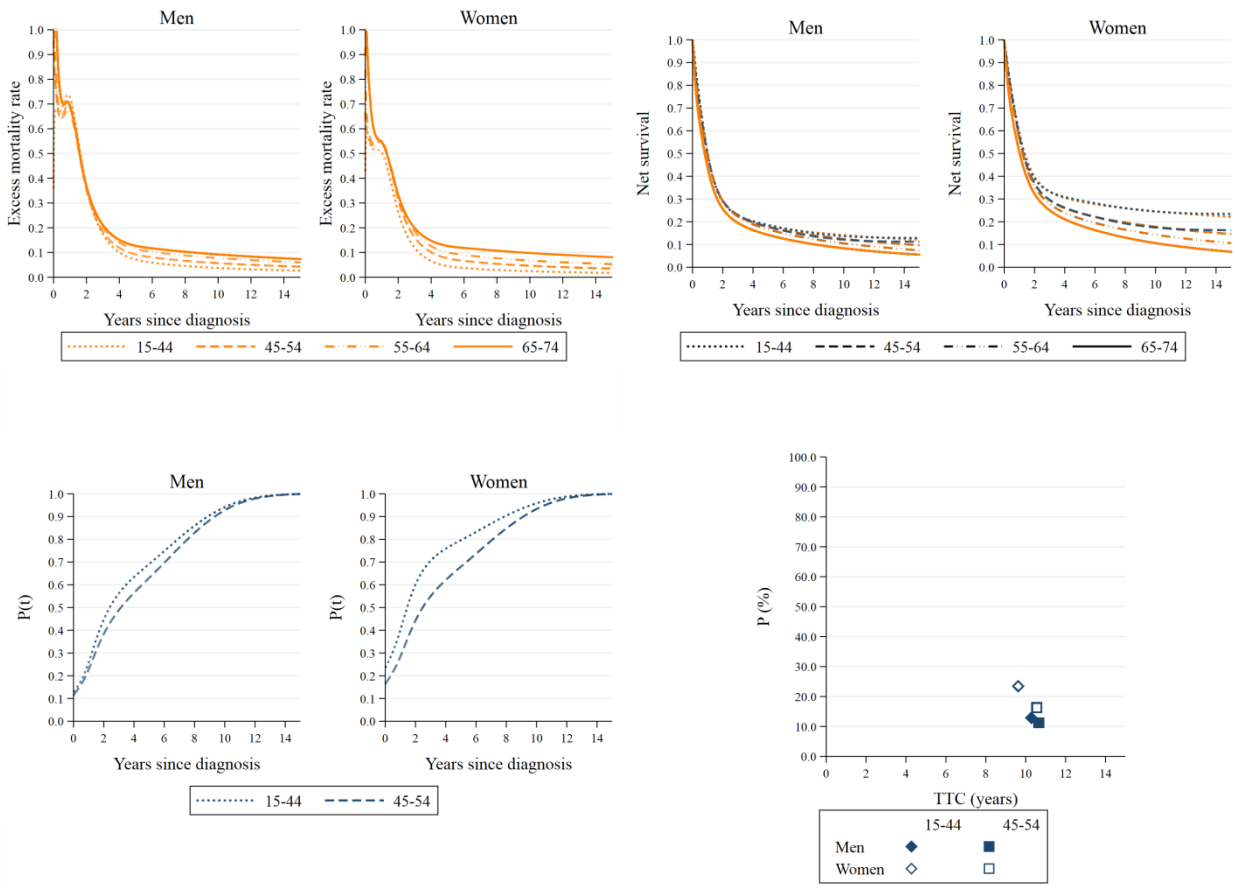
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

12. Liver



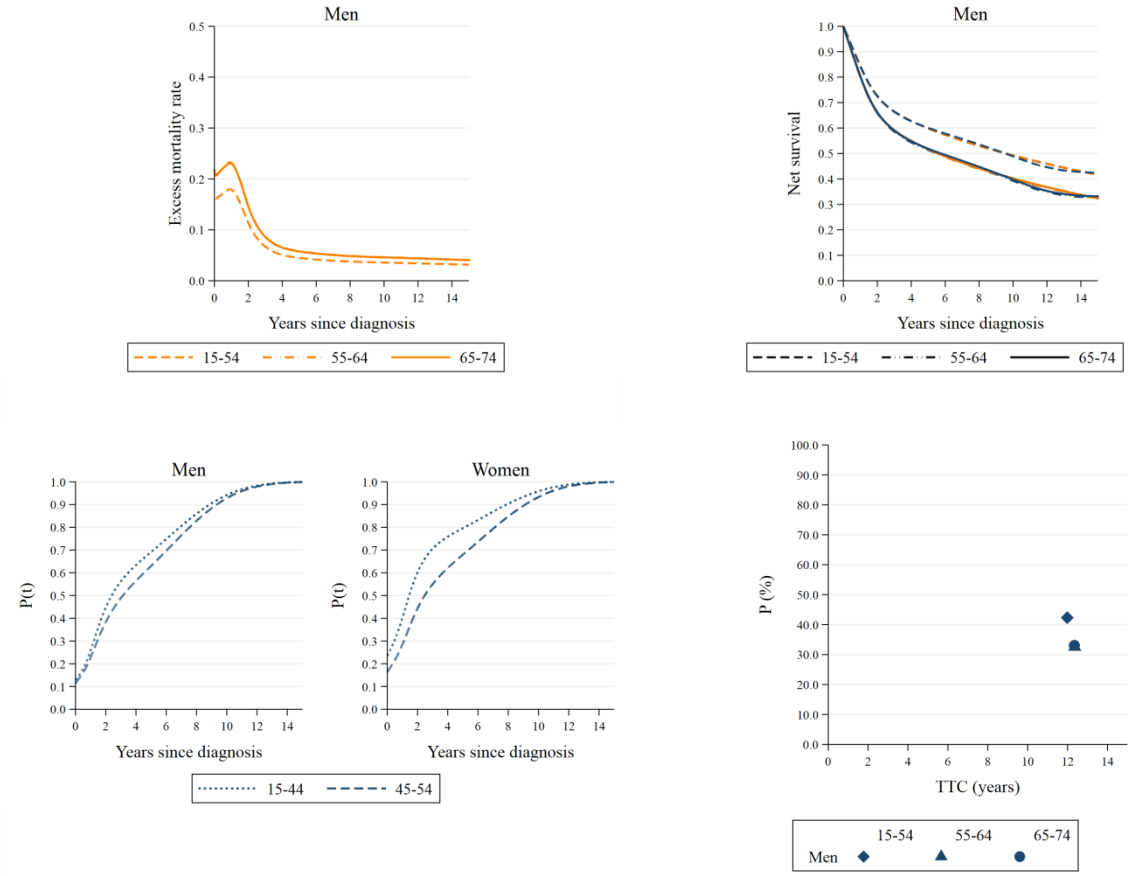
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

13. Lung



Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

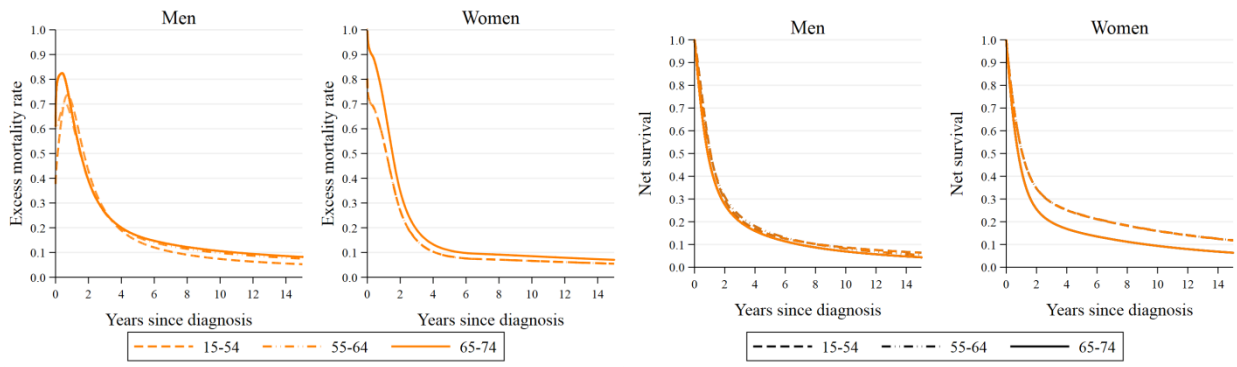
14. Nasal cavity



Estimated using flexible parametric model without assumption of cure

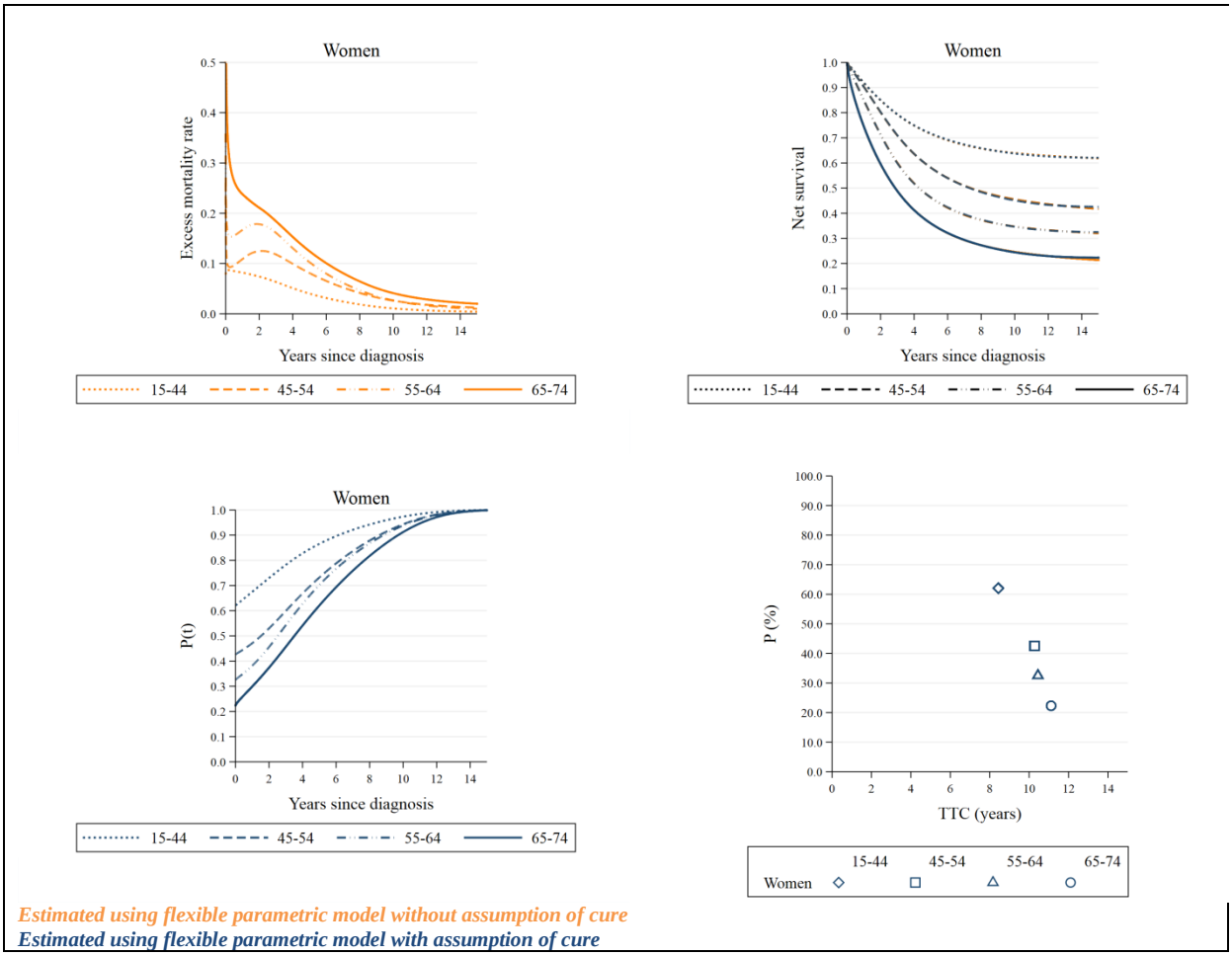
Estimated using flexible parametric model with assumption of cure

15. Oesophagus



Estimated using flexible parametric model without assumption of cure

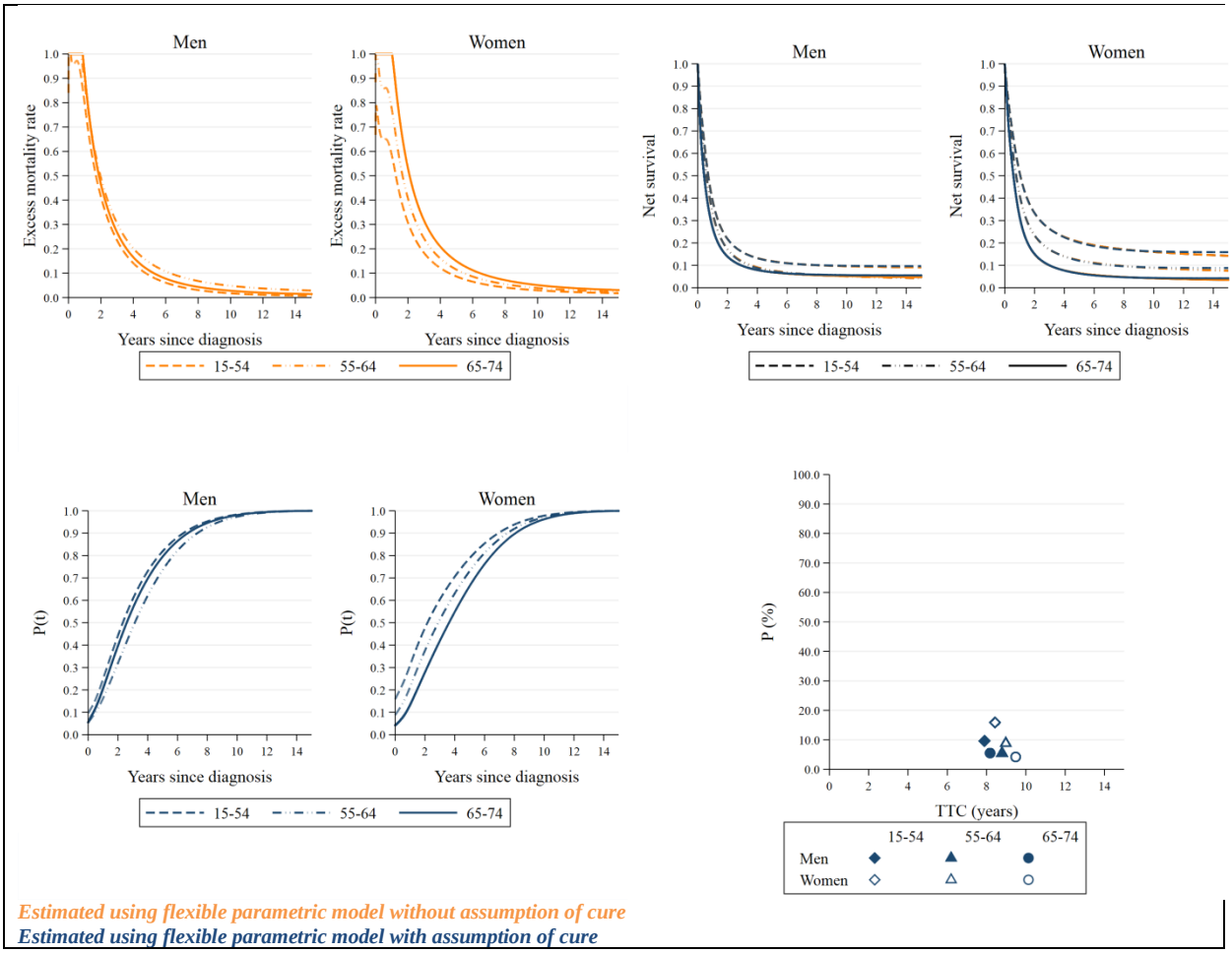
16. Ovary



Estimated using flexible parametric model without assumption of cure

Estimated using flexible parametric model with assumption of cure

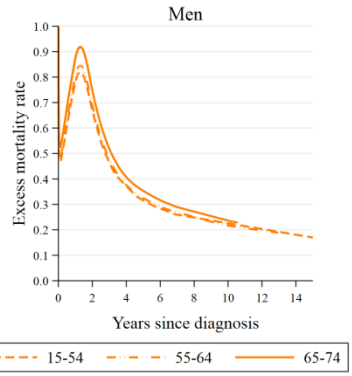
17. Pancreas

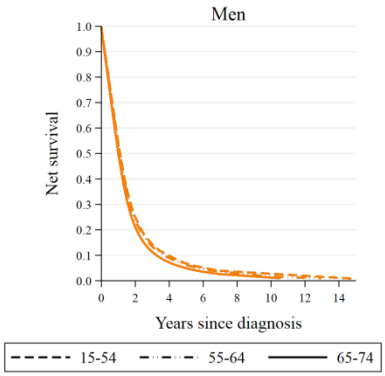


Estimated using flexible parametric model without assumption of cure

Estimated using flexible parametric model with assumption of cure

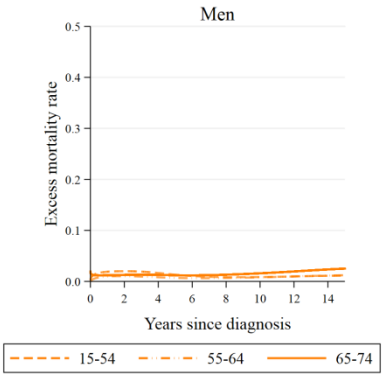
18. Pleura mesothelioma

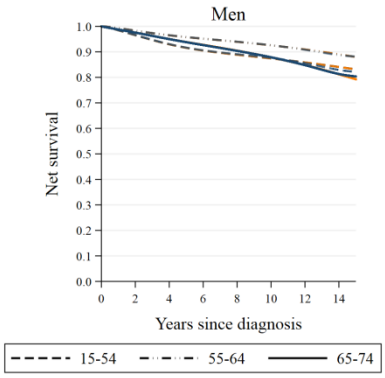


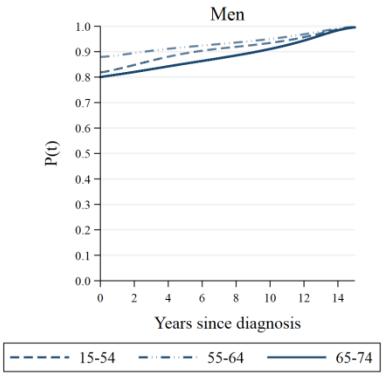


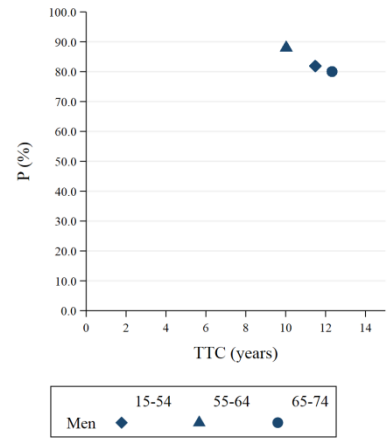
Estimated using flexible parametric model without assumption of cure

19. Prostate



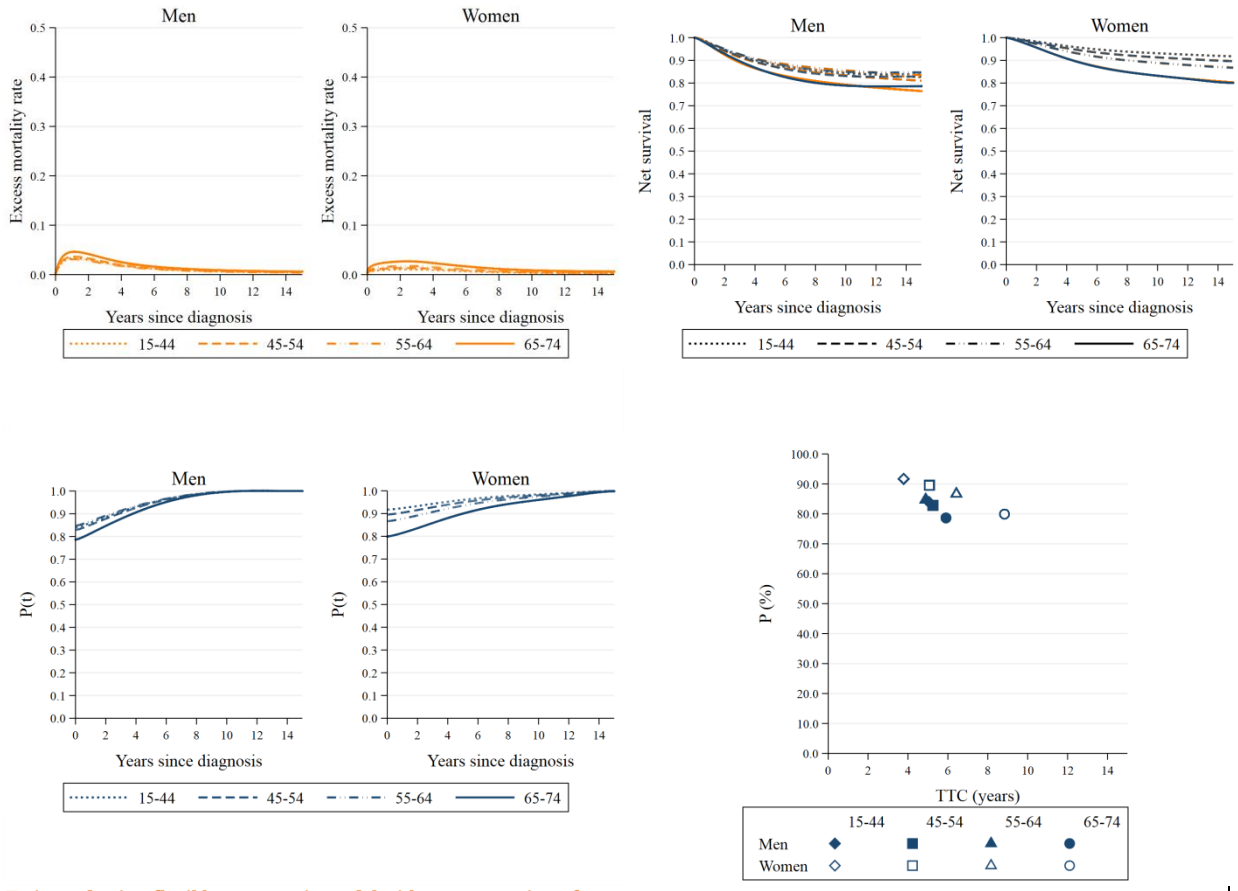






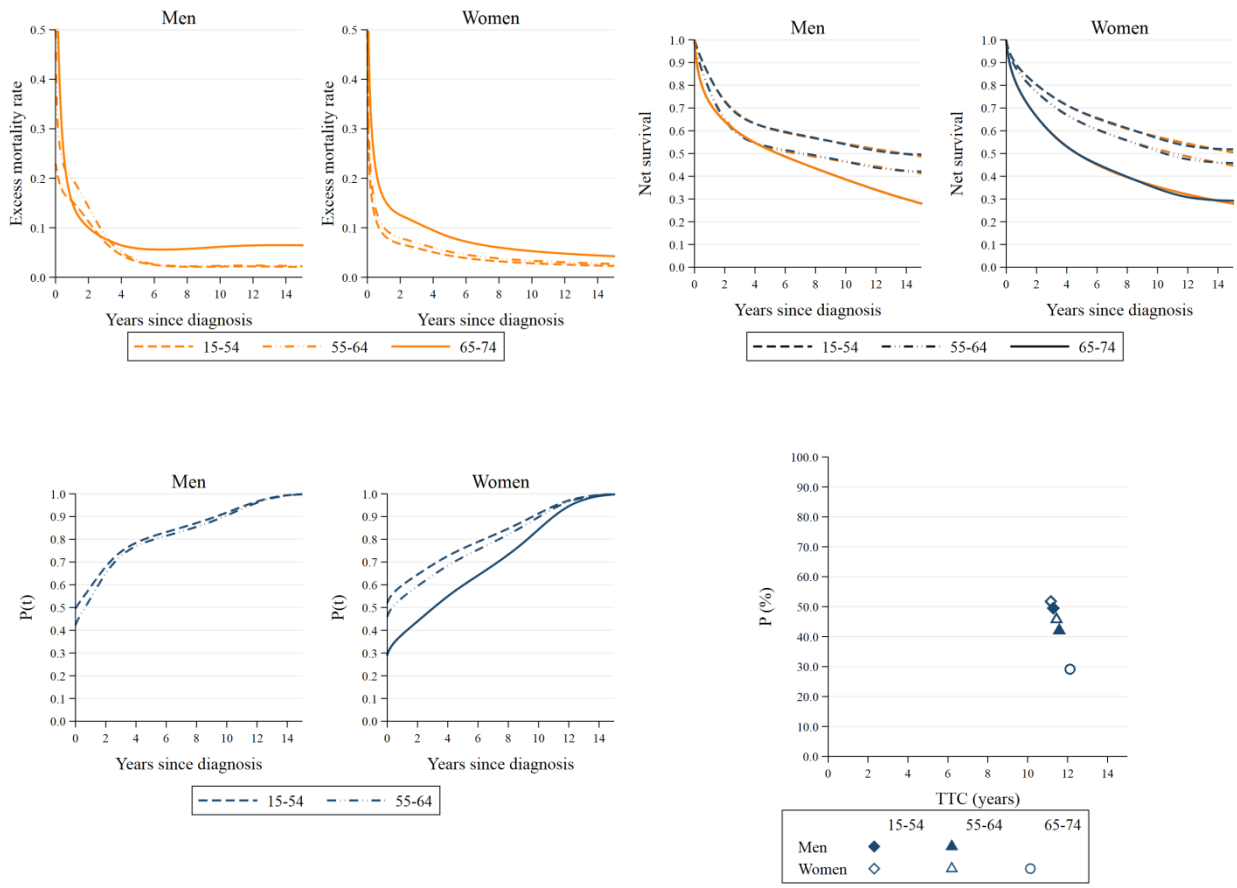
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

20. Skin melanoma



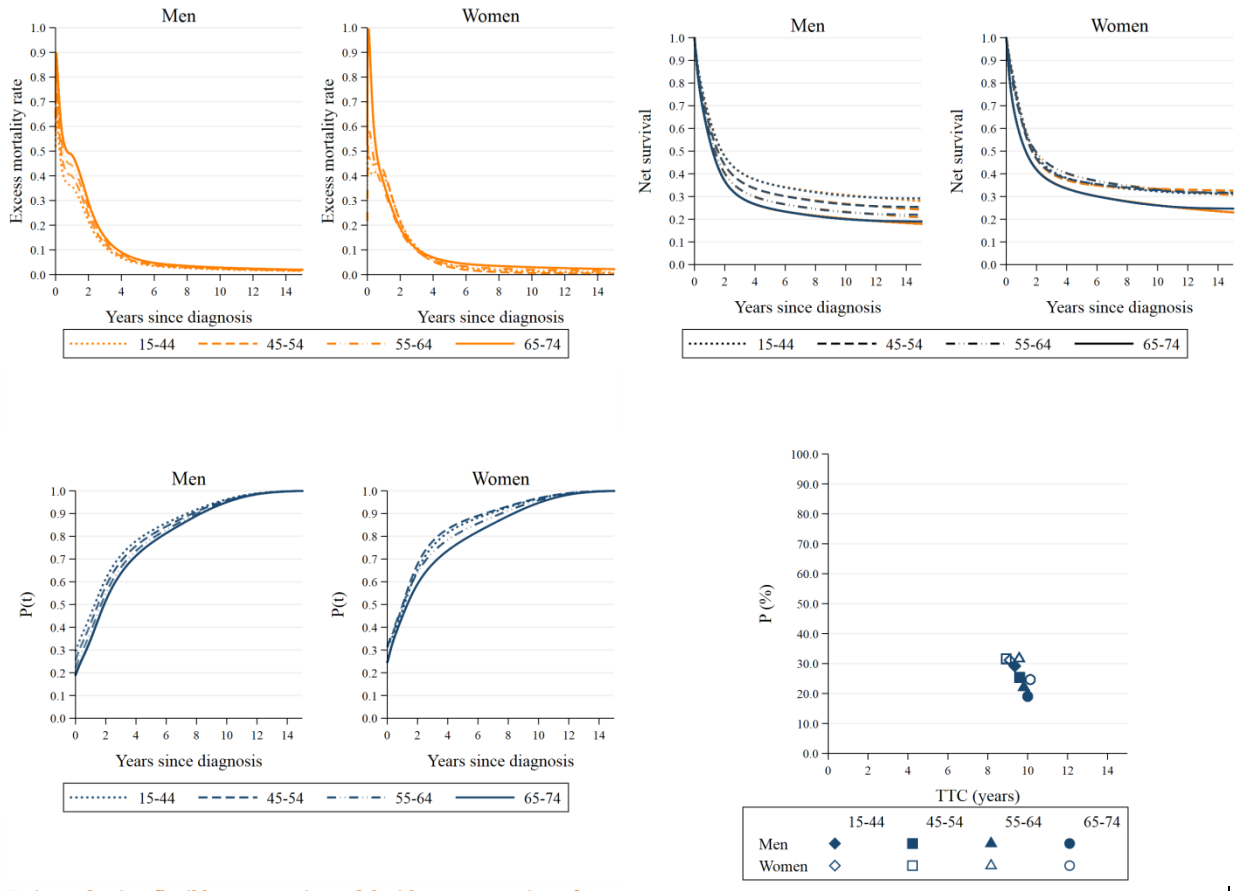
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

21. Small intestine



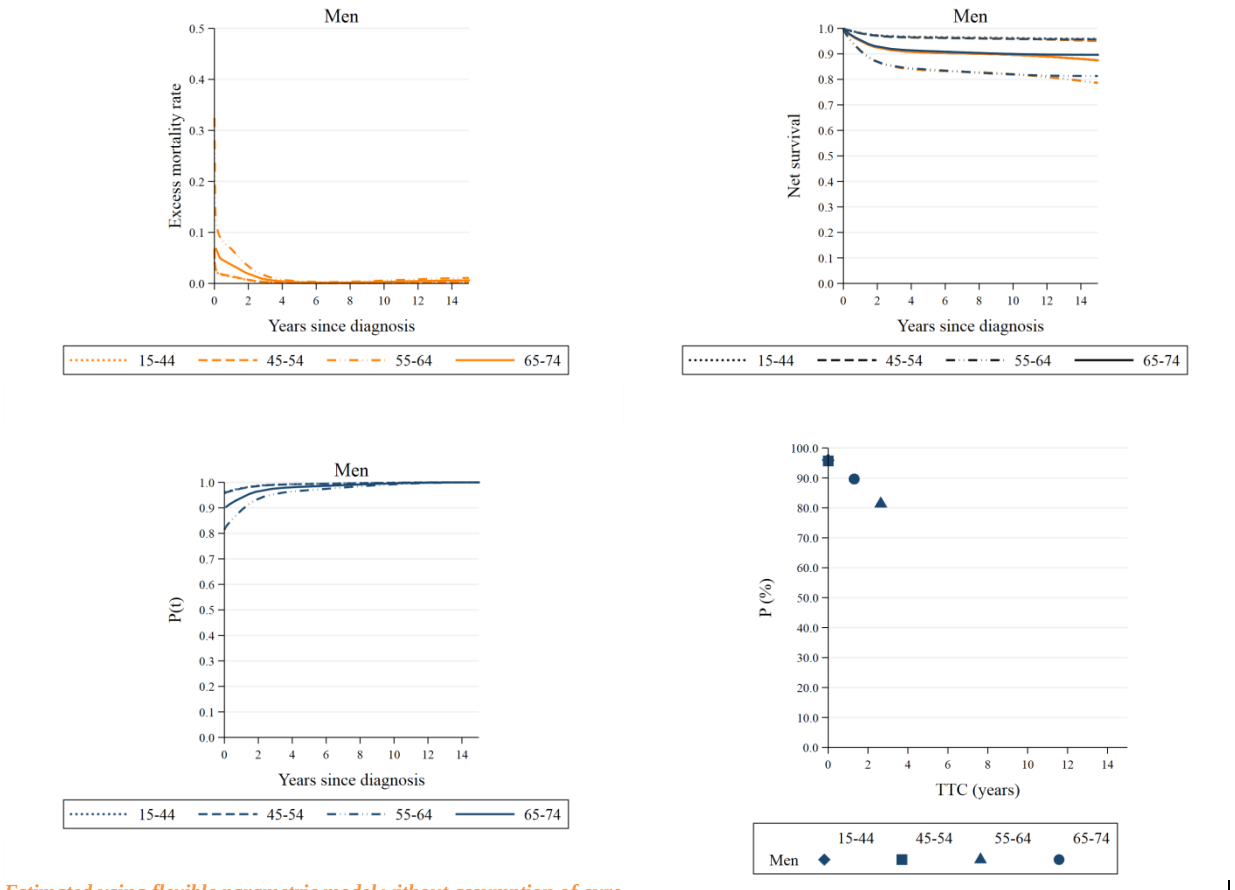
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

22. Stomach



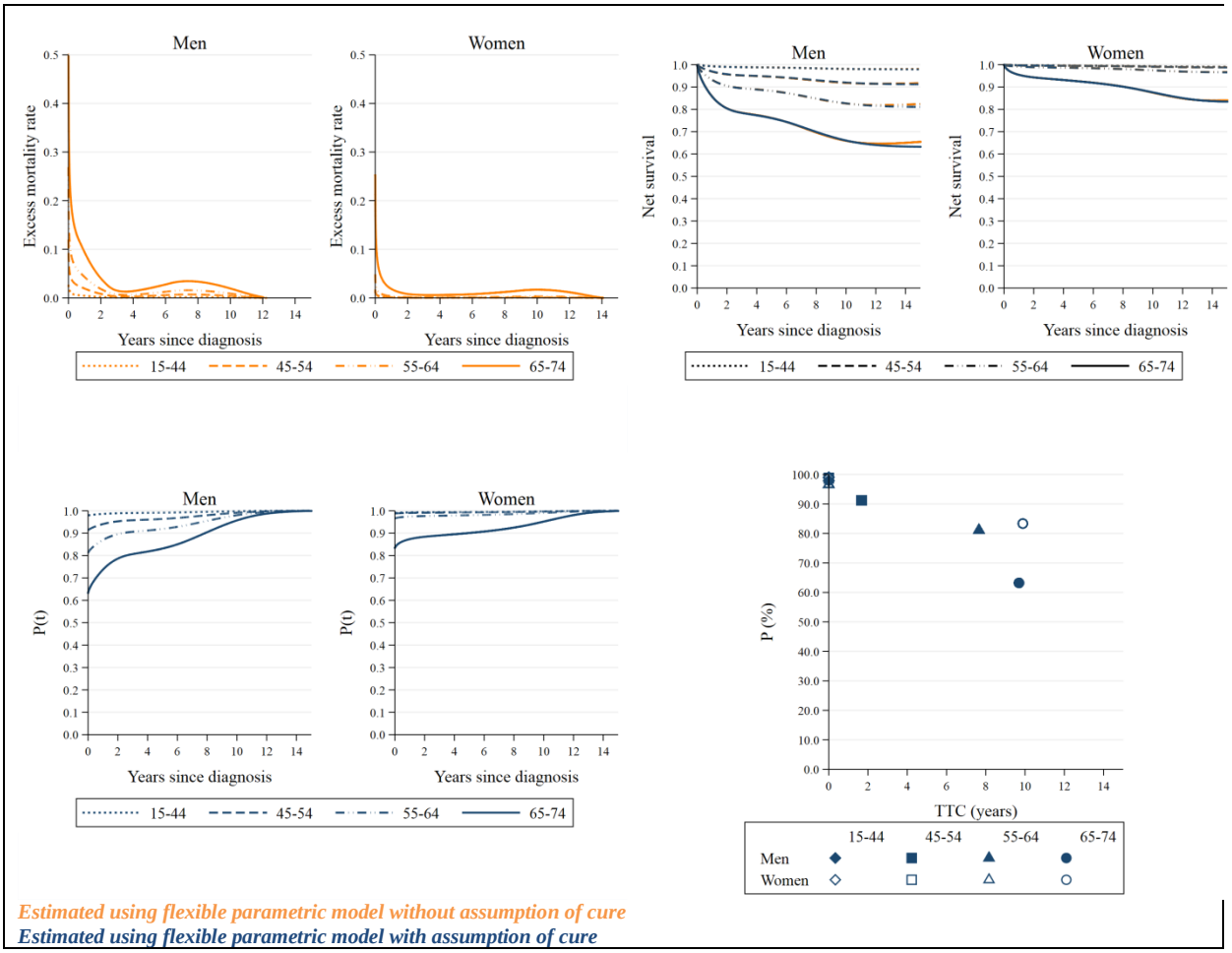
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

23. Testis



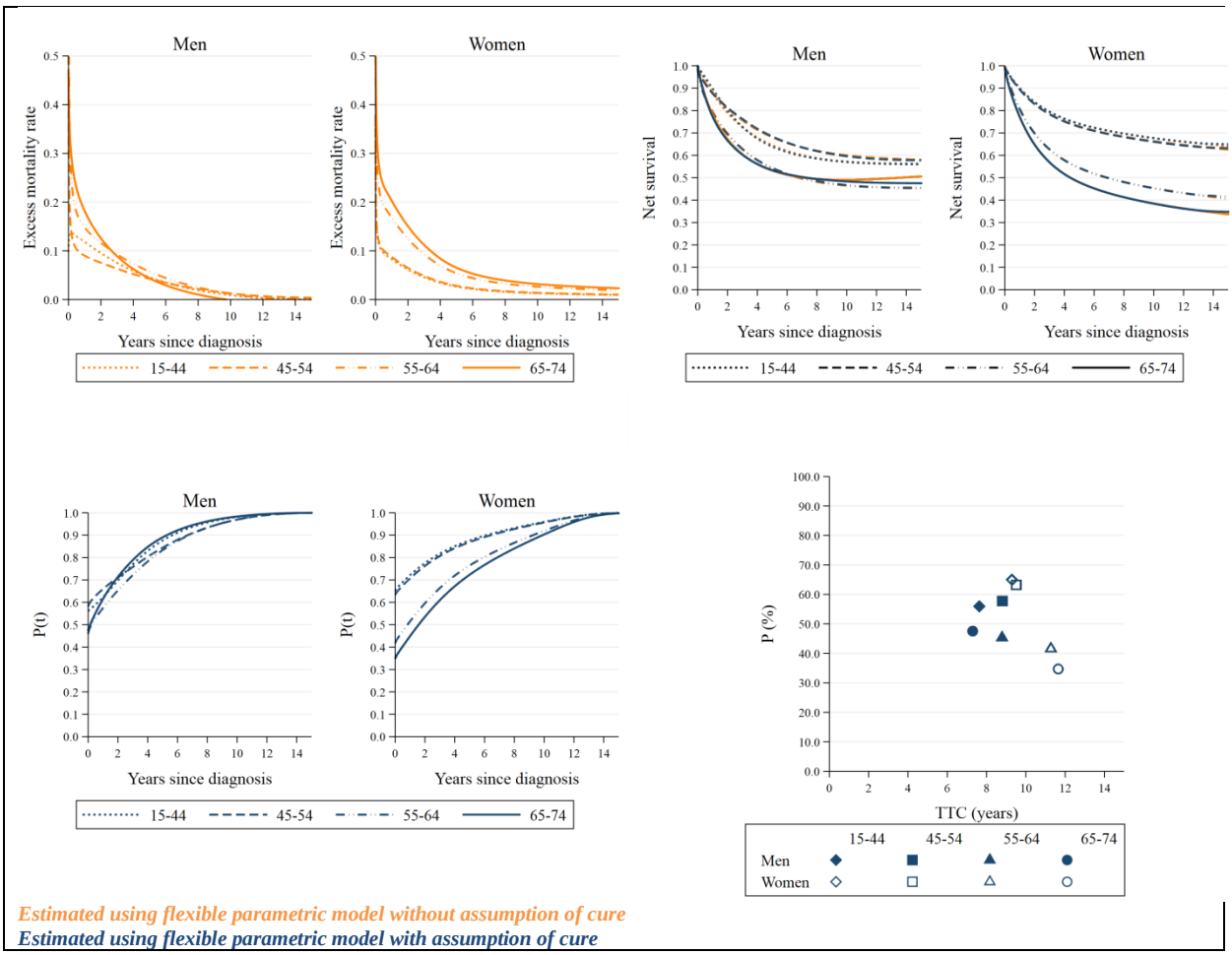
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

24. Thyroid



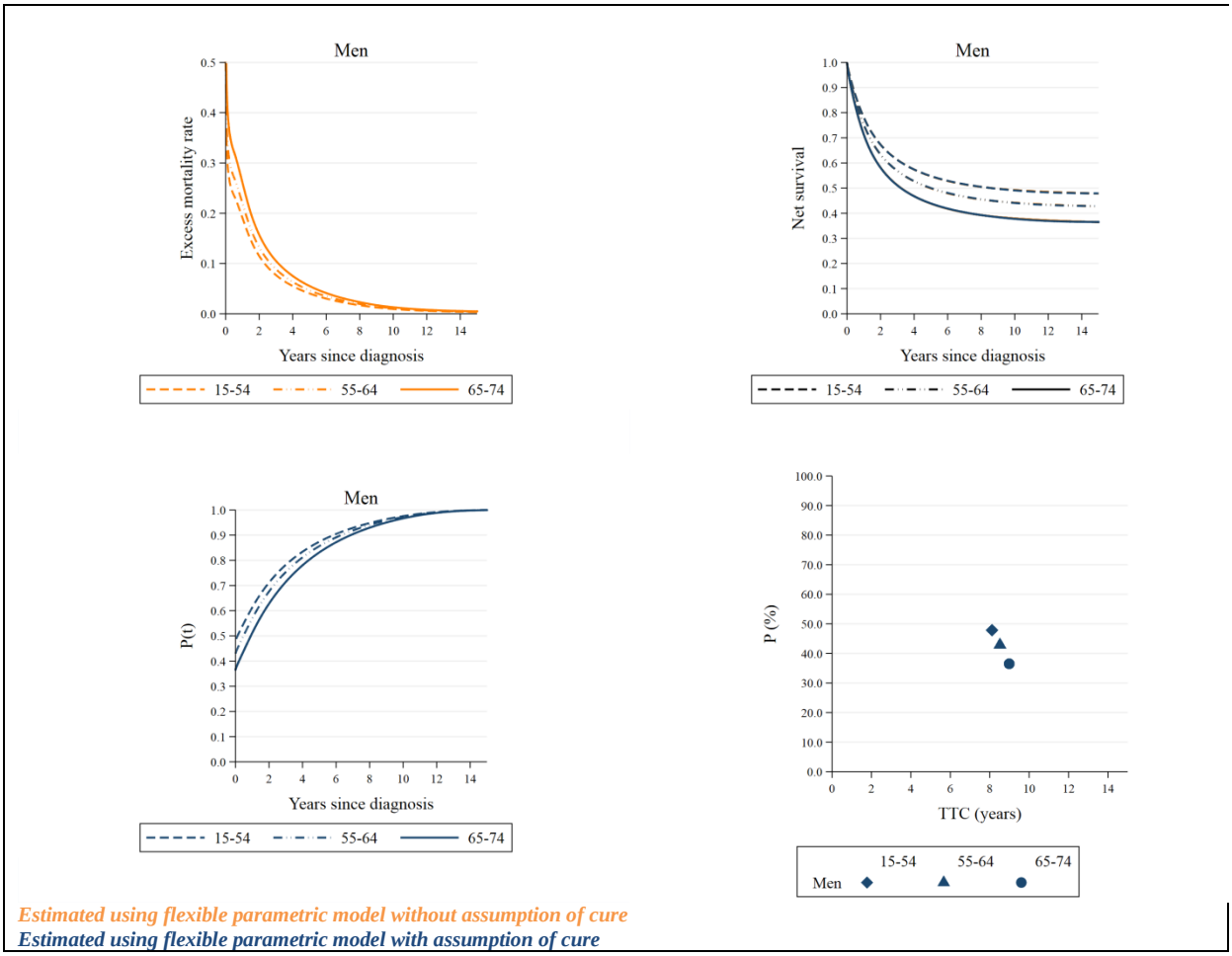
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

25. Tissue sarcoma



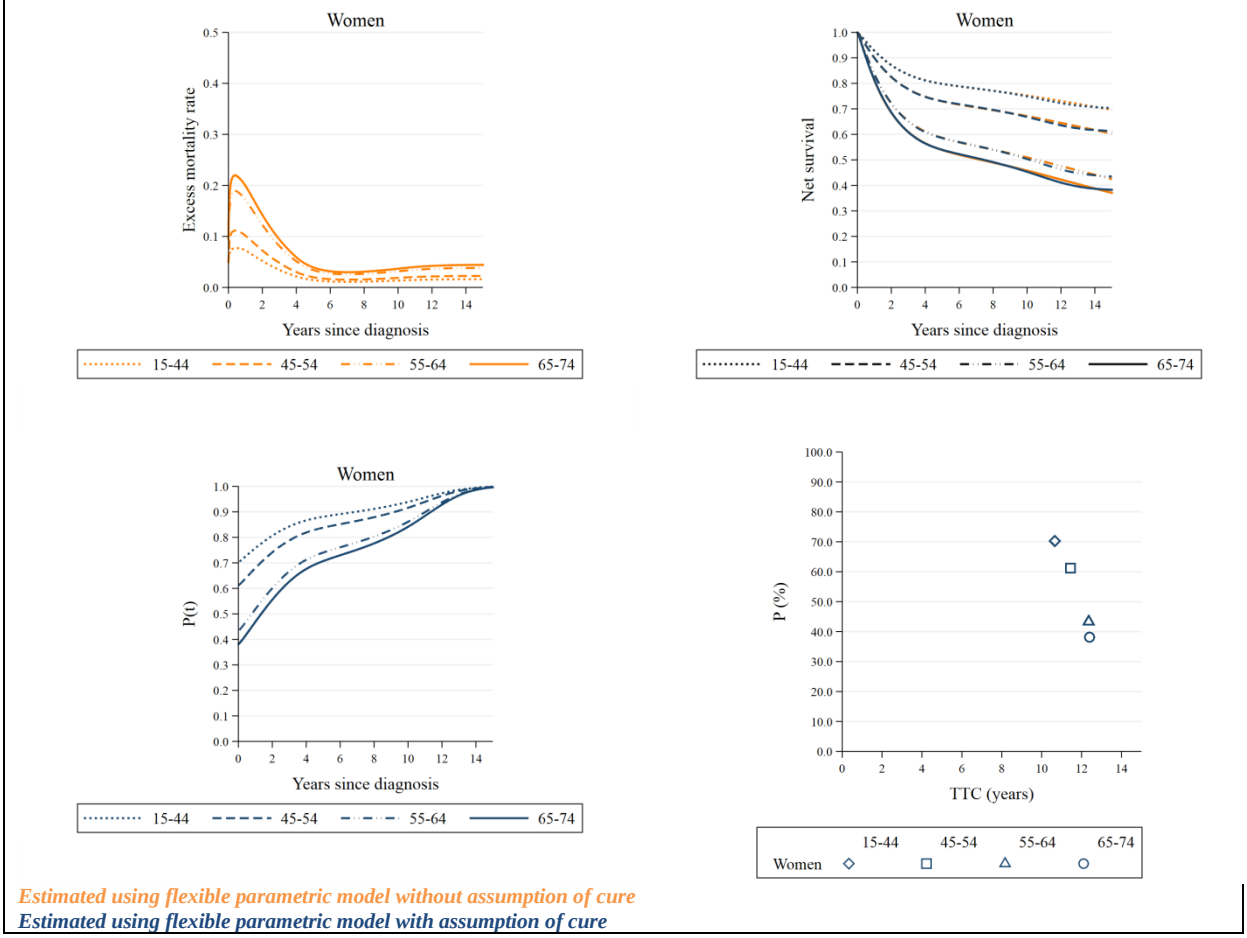
Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

26. Urinary tracts



Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

27. Vagina and Vulva



Estimated using flexible parametric model without assumption of cure
Estimated using flexible parametric model with assumption of cure

ANNEXE B : Paramètres du modèles de guérison de mélange et TNEH fixés

Tableau B-1 - Paramètres fixé pour simuler les données à partir du modèle de guérison de mélange avec une distribution de Weibull.

Scénario	$\pi = \pi_0 + \pi_1 a_1 + \pi_2 a_2$			$k_e = \exp(k_{e,0} + k_{e,1} a_1 + k_{e,2} a_2)$			$m_e = \exp(m_{e,0} + m_{e,1} a_1 + m_{e,2} a_2)$		
	π_0	π_1	π_2	$k_{e,0}$	$k_{e,1}$	$k_{e,2}$	$m_{e,0}$	$m_{e,1}$	$m_{e,2}$
Bon	0,90	-0,05	-0,10	-0,05	-0,05	-0,65	0,60	-0,15	-0,10
Moyen	0,65	-0,05	-0,15	-0,30	-0,02	-0,20	0,30	-0,06	-0,09
Mauvais	0,20	-0,05	-0,10	0,15	0,15	0,30	-0,10	-0,10	-0,10

Tableau B-2 - Paramètres fixé pour simuler les données à partir du modèle TNEH.

Scénario	$\alpha = \alpha_0 + \alpha_1 a_1 + \alpha_2 a_2$			β	$\tau = \tau_0 + \tau_1 a_1 + \tau_2 a_2$		
	α_0	α_1	α_2		τ_0	τ_1	τ_2
Bon	1,90	-0,20	-0,15	6,10	4,00	0,50	2,50
Moyen	1,40	-0,05	-0,08	6,55	7,00	0,50	2,50
Mauvais	0,95	-0,02	-0,03	6,40	9,00	1,00	3,00

Remarque : a_1 et a_2 correspondent, respectivement, à la 2ème et 3ème classe d'âge. La 1ère classe d'âge étant la classe de référence.

ANNEXE C : Article soumis au numéro spécial « ISCB » de Biometrical Journal

Biometrical Journal

Comparing performances of three cure models including a new model with time-to-cure as a parameter

Journal:	*Biometrical Journal*
Manuscript ID	bimj.201900361
Wiley - Manuscript type:	Research Paper
Date Submitted by the Author:	28-Nov-2019
Complete List of Authors:	Romain, Gaelle; Registre Bourguignon des Cancers Digestifs Boussari, Olayide; Fédération Francophone de Cancérologie Digestive Colonna, Marc; Registre du Cancer de l'Isere Jooste, Valerie; Registre Bourguignon des Cancers Digestifs
Keywords:	CANCER, CURE MODEL, NET SURVIVAL, SIMULATION STUDY, TIME-TO-CURE

Biometrical Journal **XX** (XXX) XX, ZZZZZZ / **DOI** XXX

Comparing performances of three cure models including a new model with time-to-cure as a parameter

Gaelle Romain[1,2], Olayide Boussari[1,3], Marc Colonna[4,5] and Valerie Jooste[*,1,2,5]

[1] Registre Bourguignon des Cancers Digestifs, UMR 1231, INSERM, Université Bourgogne-Franche-Comté, EPICAD team, 7 Boulevard Jeanne d'Arc, 21000 Dijon, France
[2] Dijon-Bourgogne University Hospital, 14 Rue Paul Gaffarel, 21000 Dijon, France
[3] Methodology department, Fédération Francophone de Cancérologie Digestive, 7 Boulevard Jeanne d'Arc, 21000 Dijon, France
[4] Grenoble University Hospital, Registre du Cancer de l'Isère, Boulevard de la Chantourne, 38700 La Tronche, France
[5] French Network of Cancer Registries (FRANCIM), 37 Allées Jules Guesde, 31000 Toulouse, France

Received zzz, revised zzz, accepted zzz

In population-based cancer studies, net survival (S_n) is modelled through the excess mortality rate (h_{exc}). For many cancer sites, a proportion of patients will not die from the studied cancer, representing the cure fraction (π). Cure models allow estimating S_n through the h_{exc} considering statistical cure: the asymptotic value of Sn is π. A new cure model allows a direct estimation of the time-to-cure by including the Time-to-Null-Excess-Hazard (TNEH) as a covariate dependent parameter to be estimated. The performances of the TNEH cure model were compared to that of mixture and flexible non-mixture cure models in both simulation study and application to real datasets. Three scenarios were simulated using Weibull mixture cure model to mimic Sn from poor, medium and good prognosis cancers corresponding to different values of π and shapes of h_{exc}. The performances of the three models, each including age group as covariate, were evaluated on π and Sn. These performances were satisfactory in the three scenarios in which h_{exc} reached zero. In applications on real data provided by the French cancer database, Sn curves and π from the three models were identical for testicular cancer. For colon and pancreatic cancers, where h_{exc} became low but did not reach zero, π and S_n estimated by TNEH differed from estimations provided by the other two cure models.

Key words: CANCER; CURE MODEL; NET SURVIVAL; SIMULATION STUDY; TIME-TO-CURE

Supporting Information for this article is available from the author

*Corresponding author: e-mail: vjooste@u-bourgogne.fr, Phone: +33-380-393-325, Fax: +33-380-668-251

1 Introduction

Statistical cure can be defined as the absence of death due to cancer. In cancer studies, net survival ($S_n(t)$) at time t since diagnosis is defined as the survival that would be observed if cancer were the only cause of death (Cronin & Feuer, 2000; Hakulinen & Tenkanen, 1987). In population-based cohort studies, $S_n(t)$ is estimated through the excess mortality rate ($h_{exc}(t)$) in the population of patients. The observed mortality ($h_{obs}(t)$) from the patients is the sum of the excess mortality and the expected mortality ($h_{pop}(t + a)$) in the general population sharing same demographic characteristics as the population of patients: $h_{obs}(t) = h_{exc}(t) + h_{pop}(t + a)$ with a, the age at diagnosis (Estève, Benhamou, Croasdale, & Raymond, 1990; Hakulinen & Tenkanen, 1987).

When the excess mortality rate becomes null ($h_{exc}(t) = 0$), statistical cure is reached and the mortality rate observed in the population of patients equals the expected mortality in the general population: $h_{obs}(t) = h_{pop}(t + a)$. Patients still alive at this time are considered "statistically cured". This is a statistical definition of cure at the population level; it does not imply that all patients are clinically cured. Graphically, S_n reaches a plateau at this time, its value corresponding to the cure fraction (π), which is the fraction of patients that would not die from their cancer (Boag, 1949; Verdecchia *et al.*, 1998). π is a widely used cure indicator and can be estimated from mixture or non-mixture cure models (Andersson, Dickman, Eloranta, & Lambert, 2011; Cvancarova *et al.*, 2013; Francisci *et al.*, 2009; Lambert, Thompson, Weston, & Dickman, 2007; Verdecchia *et al.*, 1998). The time from diagnosis to statistical cure, called time-to-cure, is also an interesting cure indicator for public health decision makers, epidemiologists, physicians, and patients. To date, only the model proposed by Boussari *et al.* allows direct estimation of the time-to-cure by including it as parameter to be estimated (Boussari *et al.*, July, 2019, Under review). Moreover, it objectively answers the question of the existence of a statistical cure. Until then, all cure models based the assessment of the cure assumption on a graphical observation of a plateau in S_n (Andersson & Lambert, 2012; De Angelis, Capocaccia, Hakulinen, Soderman, & Verdecchia, 1999).

The aim of this work was to compare the performances of the new cure model proposed by Boussari *et al.* to those of two other cure models: the mixture cure model proposed by De Angelis *et al.* (De Angelis et al., 1999) and the flexible parametric cure model proposed by Andersson *et al.* (Andersson et al., 2011). The considered three cure models were presented in Section 2 and compared on the net survival and the cure fraction estimations in a simulation study (Section 3); then on real datasets in Section 4.

2 Cure models

2.1 Mixture cure models

The concept of mixture cure models is to consider two sub-populations: a fraction of cured patients (π) who will not die from their cancer and whose net survival distribution $S_c(t)$ always equals 1 and a fraction of uncured patients ($1 - \pi$) that will die from their cancer following a specific proper survival distribution $S_u(t)$. In mixture cure model, the net survival is $S_n(t) = \pi S_c(t) + (1-\pi)S_u(t)$, thus $S_n(t) = \pi + (1-\pi)S_u(t)$.

We used the mixture cure model with a Weibull distribution modelling the net survival of uncured patients, similarly to that developed by De Angelis *et al.* (De Angelis et al., 1999). The net survival was:

$$S_n(t|z; \Theta) = \pi(z; \eta) + [1 - \pi(z; \eta)]\exp\left[-\lambda(z; \nu)t^{\gamma(z; \phi)}\right] \quad (1)$$

Where $\pi(\mathbf{z}; \boldsymbol{\eta})$ was the cure fraction, $\lambda(\mathbf{z}; \boldsymbol{\nu})$ and $\gamma(\mathbf{z}; \boldsymbol{\phi})$ the parameters of the Weibull distribution, depending on covariate z through the vector of parameters, respectively, η, ν and ϕ. Hence, $\boldsymbol{\Theta} = (\boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\phi})$ was the vector of parameters to be estimated. In this work, only age was included in the model as categorical variable.

The mixture cure model developed by De Angelis *et al.* (De Angelis et al., 1999)was implemented in STATA™,(STATA corp., College Station, Texas)by Lambert *et al.* (Lambert, 2007) through strsmix command.

2.2 Parametric flexible cure models

The flexible parametric cure model developed by Andersson *et al.* (Andersson et al., 2011) is a specific case of non-mixture cure models (Chen, Ibrahim, & Sinha, 1999; Lambert et al., 2007). This model fits the log cumulative excess hazard with restricted cubic spline (RCS) function of log time (Lambert & Royston, 2009; Nelson, Lambert, Squire, & Jones, 2007). The log cumulative excess hazard is forced to be linear beyond the two boundary knots of the RCS and to be constant beyond the last boundary knot. For K knots from k_1 to k_K, the RCS function is defined as:

$$s(x; \gamma) = \gamma_0 + \gamma_2 \upsilon_2(x) + \ldots + \gamma_{K-1} \upsilon_{K-1}(x) \quad (2)$$

Where $x = \ln(t)$ and the spline basis functions $\upsilon_j(x)$ are defined as:

$$\upsilon_j(x) = (x - k_{K-j})^3_+ - \phi_j(k_K - x)^3_+ - (1 - \phi_j)(k_1 - x)^3_+ \quad \text{for } j = 2,\ldots, K-1 \quad (3)$$

Where: $u_+ = u$ if $u > 0$ and else $u_+ = 0$; k_1 and k_K are the boundary knots and k_j is the j-th knots; $\phi_j = \frac{k_{K-j} - k_j}{k_K - k_1}$.

Thus the net survival written with the parametric flexible cure model is:

$$S_n(t) = \exp[-\exp(\gamma_{00})]^{\exp\left[\sum_{j=2}^{K-1} \gamma_{02}\upsilon_{0j}(x)\right]} \quad (4)$$

Covariable z was introduced in the model with a time-depend effect. Thus, the equation (4) became:

$$S_n(t|z, \Theta) = \exp[-\exp(\gamma_{00} + \beta z)]^{\exp\left[\sum_{i=2}^{K-1} s(x; \gamma_{0j}) + \sum_{i=1}^{D} s(x; \gamma_i)z_i\right]} \quad (5)$$

Where: the cure fraction was $\pi = \exp[-\exp(\gamma_{00} + \beta z)]$; $s(x; \boldsymbol{\gamma_{0j}})$ the spline function for the baseline log cumulative excess hazard as expressed in equation (2); D the number of time-dependent covariate effects; and $s(x; \boldsymbol{\gamma_i})$ the spline function for the i-th time-dependent covariate (z_i). Hence, $\Theta = (\boldsymbol{\gamma}, \boldsymbol{\beta})$ was the vector of parameters to be estimated.

In this work, age at diagnosis was introduced in the model as a categorical variable, noted z, with a time-depend effect, noted z_i. The RCS functions for the baseline log cumulative excess hazard were defined with five internal knots located at the 25th, 50th, 75th, 95th and 99th percentiles of the observed death times and boundary knots located at the 1st percentile of the observed death times and at 17 years. Age was included in the model as a categorical variable. Its time-dependent effect was added for each category (splines with two internal knots located at the 33rd and 67th percentiles of the observed death times) and evaluated with a likehood-ratio test with 0.05 as significance level.

The flexible parametric cure model was implemented in STATA™, (STATA corp., College Station, Texas) by Andersson *et al.* (Andersson & Lambert, 2012) through the stpm2 command with the cure option.

2.3 Time-to-Null-Excess-Hazard model (TNEH model)

The new model developed by Boussari *et al.* (Boussari *et al.*, July, 2019, Under review) aimed at estimating the time-to-cure. It was called Time-to-Null-Excess-Hazard model (TNEH model) because the time from diagnosis until the excess mortality rate becomes null was included in the model as a parameter τ to be estimated. (Boussari *et al.*, July, 2019, Under review). The net survival, including the age groups as covariate z, was written as:

$$S_n(t|z;\Theta) = \exp\left\{ - \tau(z;\eta)\, B[\alpha(z;\gamma),\beta] F_{Be}\left[\frac{t}{\tau(z;\eta)}, \alpha(z;\gamma), \beta\right]\right\} \quad (6)$$

Where $\tau(z; \boldsymbol{\eta}) > 0$ was the time to null excess mortality rate; B denoted the Beta function and F_{Be} [., $\alpha(z; \gamma)$, β] was the cumulative distribution function of a beta distribution with the shape parameters $\alpha(\mathbf{z}; \boldsymbol{\gamma}) > 0$ and $\beta > 1$. Hence, $\Theta = (\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})$ was the vector of parameters to be estimated. The net survival becomes constant from τ, then equals the cure fraction: $\pi = \exp\{- \tau(z; \eta)B[\alpha(z; \gamma), \beta]\}$.

We performed the estimations with the TNEH model in R software.

3 Simulation study

3.1 Simulation design and Scenarios

To compare net survival and the cure fraction estimated by each of the three cure models, times to death were simulated using three scenarios based on real data with good, medium and poor prognosis. The method to generate data is detailed in Appendix A.1. For each scenario, 1,000 samples of size N=2,000 were simulated with a mixture cure model with a Weibull distribution. The performances according to the sample size were also studied, for the medium prognosis scenario by simulating samples of sizes of N=1000 and N=500.

Age at diagnosis was generated to mimic age distribution of real data. It was uniformly distributed respectively on intervals [15-45), [45-60) and [60-75) for the scenario of good prognosis, with the proportions of age at diagnosis coming from these three intervals being, respectively, 30%, 40% and 30%. For the scenario of medium and poor prognosis, the age at diagnosis was uniformly distributed respectively on intervals [15-60), [60-70) and [70-75). The proportions of age at diagnosis coming from these three intervals are, respectively, 30%, 40% and 30% for the medium prognosis scenario and 35%, 40% and 25% for the poor prognostic scenario.

The censoring time was generated as the minimum between the administrative censoring time and the loss to follow-up time with a maximum follow-up time fixed at 15 years. The administrative censoring rate was fixed to 25%, 60% and 1%, respectively, for the good, medium and poor prognosis scenarios. The proportion of observations considered as lost to follow-up was fixed to 2% for the three scenarios.

The time of observed death was generated as the minimum between the time of death due to other causes since the diagnosis and the time of death due to cancer. The time of death due to other causes was generated from a Weibull distribution with the scale and shape parameters fixed at 88 and 11 respectively. These values correspond to the scale and shape parameters obtained by fitting a Weibull distribution to the life tables provided by the Institut National de la Statistique et des Études Économiques (INSEE). The time of death due to cancer was simulated from the mixture cure model with Weibull distribution as described in the Appendix A.1. The parameters of the mixture cure model were fixed to correspond to different values of π and shapes of h_{exc}.

The good prognosis scenario is based on cancer sites such as testicular cancer or skin melanoma. For age groups <45, [45 60) and ≥60 years, the cure fraction of the mixture cure model was set, respectively, at 80%, 85% and 90% (Figure 1). The medium prognosis scenario mimicked colon-rectum cancer. For age groups <60, [60-70) and ≥70 years, the cure fraction of the mixture model was set, respectively, at 65%, 60% and 50%, respectively (Figure 2). The bad prognosis scenario mimicked pancreatic cancer. For age groups <60, [60-70) and ≥70 years, the cure fraction of the mixture model was set, respectively, at 20%, 15% and 10% (Figure 3).

The mixture cure model with Weibull distribution, the flexible parametric and the TNEH cure models were fitted on simulated data to estimate the net survival and cure fraction by age groups.

Simulations and parameters estimation through mixture cure and flexible parametric cure models were performed in STATA™ version 15, (STATA corp., College Station, Texas). Parameters estimation through TNEH model was performed in R. Source code to reproduce the results

is available as Supporting Information on the journal's web page (http://onlinelibrary.wiley.com/doi/xxx/suppinfo).

3.2 Statistical indicators of performance

The performances of the three cure models presented above were assessed on the estimated 3, 5 and 10-year net survival and on the estimated cure fraction. The true values of π, $S_n(3)$, $S_n(5)$ and $S_n(10)$ were obtained from the parameters fixed to generate the data. The statistical indicators of performances were the bias (difference between the true value and the mean of estimates), the empirical standard error (standard deviation of the 1,000 estimates), the average standard error (mean of the 1,000 standard error estimates) and the coverage rate at 95% (percentage of the estimated 95% confidence intervals that include the true value).

3.3 Results

Good prognosis scenario

The performance indicators are presented in Table 1. The net survival ($S_n(t)$) and the cure fraction (π) estimated by the mixture cure model (MCM) and the TNEH model were unbiased (ranged from -0.002 to 0.001 for the three age groups). The biases of $S_n(t)$ and π estimated from the flexible cure model (FCM) were higher but remained acceptable (varying from -0.011 to 0.014). The empirical standard error of the TNEH model was very close to the estimates of the other two cure models. For all age groups, the empirical standard error was low, between 0.012 and 0.018, and was very close to the average standard error of $S_n(t)$ and π estimated by the three models.

Medium prognosis scenario

The results of the performance for the medium prognosis scenario are presented in Table 2. S_n and π estimated from the MCM were unbiased for the three age groups. Although the FCM provided estimates with a higher overall bias, this latter was acceptable being less than 3%. For the TNEH model, the bias was higher only for the 3-year S_n (between 0.012 and 0.014 according to age group). The empirical standard error was low for Sn and π estimated from the three cure models. The empirical standard error was very close to the average standard error. For the MCM, the coverage rate was very close to 95%. There were some variations for the FCM and TNEH models, the coverage rate varied between 73.6% and 96.1%, and between 85.2% and 94.8%, respectively.

Poor prognosis scenario

The results of the performances were presented in Table 3. S_n and π estimated from the mixture cure model were unbiased. With the TNEH model, only 3-year Sn estimated in patients aged < 60 years was biased (0.012). With the FCM, the bias was higher in patients aged <60 (between 0.016 and 0.026). For the three cure models, the empirical standard error of $S_n(t)$ and π were low and close to the average standard error. The coverage rate of the MCM estimates was close to 95%. However, for the FCM and TNEH the coverage rates were, respectively, between 59.9% and 94.8%, and between 79.8% and 93.4%.

Performances according to the sample size

We simulated 1,000 samples of size 1,000 and 500 using the same methodology used to simulate the data with medium prognosis. The tables of results are presented in Appendix A.2. As the sample size decreased down to 500, the bias and both empirical and average standard errors of estimated S_n and π increased, but remained acceptable, for the three cure models and all age groups.

4 Application to real data

4.1 Studied patients

Data emanated from the national database maintained by the French network of cancer registries FRANCIM. which includes cancer cases from 30 sites and 7 sub-sites defined as the International Classification of Diseases for Oncology, 3rd revision (ICDO-3) (Fritz *et al.*, 2000). Quality and completeness of this database is validated every four years by an independent audit committee (Comité d'Évaluation des Registres). Three cancer sites were chosen for application to real data: testicular, colon and pancreatic cancer for respectively good, medium and poor prognosis. Solid tumors diagnosed between January 1, 1995 and December 31, 2009 in patients aged [15-75) years were included. Age at diagnosis was considered in two age groups in testis cancer (<55 and ≥55) and in three age groups in colon and pancreatic cancer (<55, [55-65) and ≥65). Patient vital status was followed over 15 years after diagnosis or up to June 30, 2013.

4.2 Method

For each sex and Département (French administrative area), the expected mortality rates were derived from the general population mortality rates provided by the Institut National de la Statistique et des Études Économiques (INSEE). These rates were "observed" mortality rates (obtained simply by dividing the observed number of death by the corresponding number of person-years). These observed rates presented random (Poissonian) variation and were thus smoothed in order to obtain their expected values. This work was done by the Biostatistical unit of the Hospices Civils de Lyon, using `mgcv` package of `R` software.

The net survival S_n and the cure fraction π were estimated from the three cure models detailed in Section 2. The time-dependent effect of age at diagnosis was added and evaluated with the likehood ratio test with 0.05 as significance level. Net survival estimates by the cure models were also compared to the net survival provided by the Pohar-Perme estimator (Perme, Stare, & Esteve, 2012).

4.3 Results

Testicular cancer

The excess mortality rate was low at diagnosis for patients aged < 55years and became quickly close to zero (Figure 4.A). S_n curves predicted by the three cure models and estimated by the Pohar-Perme estimator overlapped (Figure 4.B). The estimated value of π varied between 96.0% and 96.7% depending on the model (Table 4). For patients aged ≥55years, the excess mortality rate was high in the first few years after diagnosis. The excess mortality rate curves estimated by the mixture cure model and TNEH model became close to zero around 5 years after diagnosis, whereas the estimation by the flexible cure model became close to zero later (Figure 4.A). The net survival curves (Figure 4.B) estimated by the mixture and the TNEH cure model overlapped and reached cure fraction estimated, respectively, at 86.6% and 86.8% (Table 4), whereas cure fraction estimated by the flexible cure model was 84.8%.

Colon cancer

The excess mortality rate estimated by the three cure models were similar (Figure 5.A) for patients aged <55 and 55-64 years but it was slightly different for patients aged ≥65 years. The S_n curves estimated by the three cure models and Pohar-Perme estimator overlapped for patients aged <55 and 55-64 years (Figure 5.B). According to the model, π varied between 60.0% and 61.7% in patients aged <55 and between 55.3% and 58.3% in patients aged 55-64 (Table 5). For patients age ≥65 years, S_n curve estimated by TNEH model was different from S_n curves estimated with the mixture or the

flexible cure model (Figure 5.B) and the TNEH model provided a higher, estimation of π (56.7%) than the mixture (49.6%) and the flexible cure model (51.8%, Table 5).

Pancreatic cancer

The curves of excess mortality rate estimated by the three cure models overlapped. Excess mortality rate was high near diagnosis and then decreased slowly to approach zero around eight years after diagnosis for the three age groups (Figure 6.A). Sn curves estimated by the three cure models and the Pohar-Perme estimator overlapped for the three age groups (Figure 6.B). π varied, depending on the cure model, between 11.7% and 13.3% in <55 years, between 5.7% and 7.5% in 55-64 years and between 4.6% and 5.8% in ≥ 65 years (Table 6).

5 Discussion

This work compared the net survival and the cure fraction estimated from two currently used cure models (mixture and parametric flexible cure models) to that provided by a new cure model (TNEH model). The performances of the TNEH model are satisfactory. In cases where the excess mortality rate became null, the TNEH model estimated unbiased S_n and π with a coverage rate close to 95%. Overall, the estimates of the flexible cure model had higher biases and smaller coverage rates. However, the bias was less than 3% and the coverage rate was greater than 75%. On real data, the time-to-cure estimated for testicular cancer by the TNEH model was, 3.5 years for patients aged <55 years and 5.8 years for patients aged ≥55 years; whereas for pancreatic and colon cancer, the estimated cure time was beyond the follow-up period, showing an absence of statistical cure within a clinically reasonable time. Although h_{exe} became very low10 years after diagnosis, it did not reach zero. It is worth noticing that from the epidemiological point of view, this finding has different impacts for pancreatic cancer with very few survivors and for colon cancer for which the number of 10-year survivors is high (Colonna *et al.*, 2018).

Further studies could involve new simulation designs, by using a flexible parametric cure model to generate time of death due to cancer. This would allow simulating excess mortality rate with complex shape over time, as it can be observed on real data (for example a "bounce" on excess mortality rate for the thyroid cancer) (Boussari *et al.*, 2018). Another simulation design could include the simulation of the year of diagnosis, which would allow using the real life table. Our study reveals different limitations of the TNEH model compared to the other two cure models. The conditions of application of the model are strict; this model can only be used when the excess mortality rate is sufficiently close to zero. If the conditions are not respected, the time-to-cure estimated will be very large and S_n and π will not be interpretable. Another limitation of the model is its relative lack of flexibility.

On real data, the long-term excess mortality rate reached zero for very few cancer sites. Previous studies estimated the time-to-cure when the conditional survival (Dal Maso *et al.*, 2014) or when the probability of belonging to the cures groups (Boussari *et al.*, 2018) was considered high enough. These conditions were reached held when the excess mortality rate was very low but this latter was not, necessarily, equal to zero at the time-to-cure. The statistical cure assumption was assessed graphically by comparing the net survival curves estimated from a cure model to the curves estimated from a classical net survival model (Dal Maso *et al.*, 2014; Romain *et al.*, 2019). The TNEH model is the only model that currently allows direct estimation of the Time-to-Null-Excess-Hazard, as opposed to the other cure models from which time-to-cure could be derived as a post-estimation via approximation methods. The TNEH model also allows determining if the excess mortality rate becomes null within a clinically reasonable time, therefore testing the cure assumption.

Acknowledgements The authors wish to thank the French network of cancer registries (FRANCIM) for providing the data for application on real data.

Conflict of Interest The authors have declared no conflict of interest.

Appendix

A.1 Data generation

In the following $\mathcal{U}[a; b]$ denotes the uniform distribution on [a; b] and $\mathcal{B}(p)$ denotes the Bernoulli distribution with parameter p. For each scenario, 1,000 data sets of size n=2,000 were generated.

The age at diagnosis A ($a_1, \ldots, a_n$) was generated from uniform distribution on interval [15-45), [45-60) and [60-75) for good prognosis scenario and [15-60), [60-70) and [70-75) medium and poor prognosis scenario. Covariate Z ($z_1, \ldots, z_n$) was age groups at diagnosis.

(i) The censoring time was generated as the minimum between the administrative censoring time, note C_1 ($c_{1_1}, \ldots, c_{1_n}$), and the lost to follow-up time, note C_2 ($c_{2_1}, \ldots, c_{2_n}$). Note that t_{max} was the maximum follow-up time. The censoring time C_1 was generated from two proportions: the proportion p_1, of patients alive at the end of the follow-up but followed less than t_{max}, and the proportion, $1\text{-}p_1$, of patients alive and followed more than t_{max}. The censoring time C_2 was also generated from two proportions: the proportion p_2, of patients lost to follow-up, and the proportion, $1\text{-}p_2$, of patients alive and followed more than t_{max}. C_1 et C_2 were generated as follow: $C_j = P_j \times U_j + (1 - P_j) \times t_{max}$, for $j \in \{1, 2\}$ with $U_j \sim \mathcal{U}[0, 15]$ and $P_j = \mathcal{B}(p_j)$. Thus, each individual censoring time C ($c_1, \ldots, c_n$) was: $c_i = \min(c_{1_i}, c_{2_i})$ for $1 \leq i \leq n$.

(ii) The time to death due to other causes (t_{pop}) was simulated from Weibull distribution with parameters allows to reproduce a mortality rate similar to those of the mortality tables provided by INSEE. The expected survival in the general population is the survival from birth knowing she/he is alive at diagnosis. It is expressed as a conditional survival:

$$S_{pop}(t_{pop_i} + a_i \mid a_i) = \frac{S_{pop}(t_{pop_i} + a_i)}{S_{pop}(a_i)} = \frac{\exp\left[- \left(\frac{t_{pop_i} + a_i}{\lambda} \right)^{\gamma} \right]}{S_{pop}(a_i)}, \text{ for } 1 \leq i \leq n \quad (1)$$

Where a_i is the age at diagnosis and λ and γ are the shape and scale parameters of the Weibull distribution. We draw n realization of U ($u_1, \ldots, u_n$) with $U \sim \mathcal{U}[0, 1]$ such as $S_{pop}(t_{pop_i} + a_i \mid a_i) = u_i$. Thus, for $1 \leq i \leq n$, the time to death due to other causes was:

$$t_{pop_i} = \lambda \{ - \log [u_i \times S_{pop}(a_i)] \}^{1/\gamma} - a_i, \text{ for } 1 \leq i \leq n \quad (2)$$

(iii) The time of death due to cancer (t_{exc}) was simulated from the mixture cure model with Weibull distribution for the three scenarios. We draw n realization of U^* ($u_1^*, \ldots, u_n^*$) with $U^* \sim \mathcal{U}[0, 1]$ such as the net survival was $S_n(t_{exc_i} | z_i; \Theta) = u_i^*$. From the mixture cure model with the survival of uncured expressed by a Weibull distribution:

$$t_{exc_i} = \begin{cases} \left\{ \dfrac{ - \log \left[\dfrac{u_i^* - \pi(z_i; \eta)}{1 - \pi(z_i; \eta)} \right] }{\lambda^*(z_i; \nu)} \right\}^{1/\gamma^*(z_i; \phi)} & \text{if } u_i^* \geq \pi(z_i; \eta) \\ +\infty & \text{if } u_i^* < \pi(z_i; \eta) \end{cases}, \text{ for } 1 \leq i \leq n \quad (3)$$

Where $\pi(\mathbf{z}; \boldsymbol{\eta})$ is the cure fraction, $\lambda^*(\mathbf{z}; \boldsymbol{\nu})$ and $\gamma^*(\mathbf{z}; \boldsymbol{\phi})$ are the parameters of the Weibull distribution, depending on covariates z through the vector of parameters, respectively, η, ν and ϕ. Hence, $\Theta = (\boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\phi})$ was the vector of parameters fixed to generated t_{exc} from the mixture cure model with a Weibull distribution.

(iv) The observed time to follow-up, T ($t_1, \ldots, t_n$) was calculated by taking the minimum between the censoring time and the observed time. This latter was defined as the minimum between the time to death due to cancer (t_{exc}) and the time to death due to other causes (t_{pop}). Thus, each subject i was associated to variables $\{t_i, \delta_i\}$, with δ_i the indicatrices of death :

$$t_i = \min[c_i, \min(t_{exc_i}, t_{pop_i})]$$

$$\delta_i = \begin{cases} 1 \text{ if } \min(t_{exc_i}, t_{pop_i}) \leq c_i \\ 0 \text{ if } \min(t_{exc_i}, t_{pop_i}) > c_i \end{cases} \quad \text{, for } 1 \leq i \leq n \tag{4}$$

A.2 Performances of the mixture, parametric flexible and TNEH cure model on 1,000 samples simulated from the mixture cure model of size 1,000 and 500.

Table Medium prognosis scenario: Performances of the mixture (MCM), parametric flexible (FCM) and the TNEH cure models (TNEH) over 1,000 samples simulated of size 1,000 from the mixture cure model with a Weibull distribution.

True value	Bias			Empirical standard error			Average standard error			Coverage rate		
	MCM	FCM	TNEH	MCM	FCM	TNEH	MCM	FCM	TNEH	MCM	FCM	TNEH
<60 years												
$S_n(3) = 0.663$	<0.001	0.025	0.017	0.027	0.026	0.026	-0.028	0.026	0.026	95.4%	86.0%	89.9%
$S_n(5) = 0.651$	<0.001	0.005	0.010	0.028	0.028	0.028	-0.028	0.028	0.028	95.3%	94.4%	92.9%
$S_n(10) = 0.650$	<0.001	-0.011	0.008	0.028	0.029	0.028	-0.028	0.029	0.029	95.2%	92.2%	93.9%
$\pi = 0.650$	<0.001	-0.013	0.008	0.028	0.029	0.028	0.028	0.029	0.029	95.2%	91.6%	93.9%
[60-70) years												
$S_n(3) = 0.621$	<0.001	0.016	0.013	0.024	0.023	0.023	-0.025	0.024	0.024	95.4%	91.7%	92.9%
$S_n(5) = 0.601$	<0.001	<0.001	0.003	0.026	0.025	0.025	-0.026	0.026	0.027	95.2%	95.1%	95.5%
$S_n(10) = 0.600$	-0.001	-0.018	-0.002	0.026	0.026	0.027	-0.027	0.026	0.028	95.2%	89.4%	95.9%
$\pi = 0.600$	-0.001	-0.019	-0.002	0.026	0.026	0.027	0.027	0.026	0.028	95.1%	88.3%	95.9%
≥70 years												
$S_n(3) = 0.548$	<0.001	0.019	0.015	0.031	0.029	0.029	-0.031	0.029	0.029	94.6%	92.1%	91.6%
$S_n(5) = 0.506$	<0.001	0.020	0.004	0.034	0.031	0.034	-0.033	0.031	0.033	94.1%	91.9%	94.7%
$S_n(10) = 0.500$	-0.001	0.006	-0.008	0.036	0.032	0.038	-0.035	0.032	0.038	94.3%	95.0%	94.7%
$\pi = 0.500$	-0.001	0.004	-0.008	0.036	0.032	0.038	0.035	0.032	0.038	94.1%	95.1%	94.7%

Table Medium prognosis scenario: Performances of the mixture (MCM), parametric flexible (FCM) and the TNEH cure models (TNEH) over 1,000 samples simulated of size 500 from the mixture cure model with a Weibull distribution.

True value	Bias			Empirical standard error			Average standard error			Coverage rate		
	MCM	FCM	TNEH	MCM	FCM	TNEH	MCM	FCM	TNEH	MCM	FCM	TNEH
<60 years												
$S_n(3) = 0.663$	-0.002	0.024	0.016	0.038	0.036	0.035	-0.039	0.036	0.037	96.0%	91.6%	92.9%
$S_n(5) = 0.651$	-0.002	0.004	0.008	0.039	0.039	0.038	-0.040	0.039	0.040	96.2%	96.0%	96.0%
$S_n(10) = 0.650$	-0.002	-0.013	0.005	0.039	0.040	0.038	-0.040	0.040	0.042	96.0%	93.5%	96.7%
$\pi = 0.650$	-0.002	-0.014	0.005	0.039	0.040	0.038	0.040	0.041	0.042	95.7%	93.3%	96.7%
[60-70) years												
$S_n(3) = 0.621$	-0.001	0.015	0.012	0.035	0.033	0.032	-0.035	0.034	0.034	95.6%	94.7%	94.9%
$S_n(5) = 0.601$	-0.002	-0.001	0.002	0.036	0.035	0.036	-0.037	0.036	0.038	95.2%	96.3%	96.7%
$S_n(10) = 0.600$	-0.003	-0.019	-0.002	0.037	0.036	0.037	-0.038	0.037	0.040	95.5%	92.7%	96.9%
$\pi = 0.600$	-0.003	-0.020	-0.003	0.037	0.036	0.037	0.038	0.037	0.040	96.2%	91.6%	96.9%
≥70 years												
$S_n(3) = 0.548$	-0.001	0.018	0.014	0.043	0.040	0.040	-0.043	0.042	0.041	95.2%	95.4%	93.2%
$S_n(5) = 0.506$	<0.001	0.019	0.004	0.047	0.043	0.047	-0.047	0.044	0.047	95.3%	94.9%	94.5%
$S_n(10) = 0.500$	-0.002	0.004	-0.008	0.050	0.044	0.052	-0.049	0.045	0.053	95.7%	95.8%	95.2%
$\pi = 0.500$	-0.002	0.003	-0.008	0.050	0.044	0.053	0.049	0.045	0.054	95.6%	95.8%	95.4%

References

Andersson, T. M., Dickman, P. W., Eloranta, S., & Lambert, P. C. (2011). Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC Med Res Methodol, 11*, 96. doi:10.1186/1471-2288-11-96

Andersson, T. M., & Lambert, P. C. (2012). Fitting and Modeling Cure in Population-Based Cancer Studies within the Framework of Flexible Parametric Survival Models. *Stata Journal, 12*, 623-638. doi:10.1177/1536867X1201200404

Boag, J. W. (1949). Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. *Journal of the Royal Statistical Society. Series B (Methodological), 11*(1), 15-53.

Boussari, O., Bordes, L., Romain, G., Colonna, M., Bossard, N., Remontet, L., & Jooste, V. (July, 2019). *Modeling excess hazard with time-to-cure as a parameter.* Paper presented at the 40th Annual Conference of the International Society for Clinical Biostatistics, Leuven, Belgium. Poster retrieved from https://kuleuvencongres.be/iscb40/images/iscb40-2019-e-versie.pdf

Boussari, O., Bordes, L., Romain, G., Colonna, M., Bossard, N., Remontet, L., & Jooste, V. (Under review). Modeling excess hazard with time-to-cure as a parameter. *Biometrics,* http://arxiv.org/abs/1911.12204.

Boussari, O., Romain, G., Remontet, L., Bossard, N., Mounier, M., Bouvier, A.-M., . . . Jooste, V. (2018). A new approach to estimate time-to-cure from cancer registries data. *Cancer Epidemiol, 53*, 72-80. doi:https://doi.org/10.1016/j.canep.2018.01.013

Chen, M.-H., Ibrahim, J. G., & Sinha, D. (1999). A New Bayesian Model for Survival Data with a Surviving Fraction. *Journal of the American Statistical Association, 94*(447), 909-919. doi:10.1080/01621459.1999.10474196

Colonna, M., Boussari, O., Cowppli-Bony, A., Delafosse, P., Romain, G., Grosclaude, P., & Jooste, V. (2018). Time trends and short term projections of cancer prevalence in France. *Cancer Epidemiol, 56*, 97-105. doi:10.1016/j.canep.2018.08.001

Cronin, K. A., & Feuer, E. J. (2000). Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. *Stat Med, 19*(13), 1729-1740. doi:10.1002/1097-0258(20000715)19:13<1729::aid-sim484>3.0.co;2-9

Cvancarova, M., Aagnes, B., Fosså, S. D., Lambert, P. C., Møller, B., & Bray, F. (2013). Proportion cured models applied to 23 cancer sites in Norway. *Int J Cancer, 132*(7), 1700-1710. doi:10.1002/ijc.27802

Dal Maso, L., Guzzinati, S., Buzzoni, C., Capocaccia, R., Serraino, D., Caldarella, A., . . . De Angelis, R. (2014). Long-term survival, prevalence, and cure of cancer: a population-based estimation for 818 902 Italian patients and 26 cancer types. *Ann Oncol, 25*(11), 2251-2260. doi:10.1093/annonc/mdu383

De Angelis, R., Capocaccia, R., Hakulinen, T., Soderman, B., & Verdecchia, A. (1999). Mixture models for cancer survival analysis: application to population-based data with covariates. *Stat Med, 18*(4), 441-454. doi:10.1002/(sici)1097-0258(19990228)18:4<441::aid-sim23>3.0.co;2-m

Estève, J., Benhamou, E., Croasdale, M., & Raymond, L. (1990). Relative survival and the estimation of net survival: Elements for further discussion. *Stat Med, 9*(5), 529-538. doi:10.1002/sim.4780090506

Francisci, S., Capocaccia, R., Grande, E., Santaquilani, M., Simonetti, A., Allemani, C., . . . Group, E. W. (2009). The cure of cancer: a European perspective. *Eur J Cancer, 45*(6), 1067-1079. doi:10.1016/j.ejca.2008.11.034

Fritz, A., Percy, C., Jack, A., Shanmugaratnam, K., Sobin, L., Parkin, D. M., & Whelan, S. (2000). *International Classification of Diseases for Oncology* (3rd edition ed.). Geneva: World Health Organization.

Hakulinen, T., & Tenkanen, L. (1987). Regression Analysis of Relative Survival Rates. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 36*(3), 309-317. doi:10.2307/2347789

Lambert, P. C. (2007). Modeling of the Cure Fraction in Survival Studies. *The Stata Journal, 7*(3), 351-375. doi:10.1177/1536867X0700700304

Lambert, P. C., & Royston, P. (2009). Further Development of Flexible Parametric Models for Survival Analysis. *The Stata Journal, 9*(2), 265-290. doi:10.1177/1536867X0900900206

Lambert, P. C., Thompson, J. R., Weston, C. L., & Dickman, P. W. (2007). Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics, 8*(3), 576-594. doi:10.1093/biostatistics/kxl030

Nelson, C. P., Lambert, P. C., Squire, I. B., & Jones, D. R. (2007). Flexible parametric models for relative survival, with application in coronary heart disease. *Stat Med, 26*(30), 5486-5498. doi:10.1002/sim.3064

Perme, M. P., Stare, J., & Esteve, J. (2012). On estimation in relative survival. *Biometrics, 68*(1), 113-120. doi:10.1111/j.1541-0420.2011.01640.x

Romain, G., Boussari, O., Bossard, N., Remontet, L., Bouvier, A. M., Mounier, M., . . . Jooste, V. (2019). Time-to-cure and cure proportion in solid cancers in France. A population based study. *Cancer Epidemiol, 60*, 93-101. doi:10.1016/j.canep.2019.02.006

Verdecchia, A., De Angelis, R., Capocaccia, R., Sant, M., Micheli, A., Gatta, G., & Berrino, F. (1998). The cure for colon cancer: results from the EUROCARE study. *Int J Cancer, 77*(3), 322-329. doi:10.1002/(sici)1097-0215(19980729)77:3<322::aid-ijc2>3.0.co;2-q

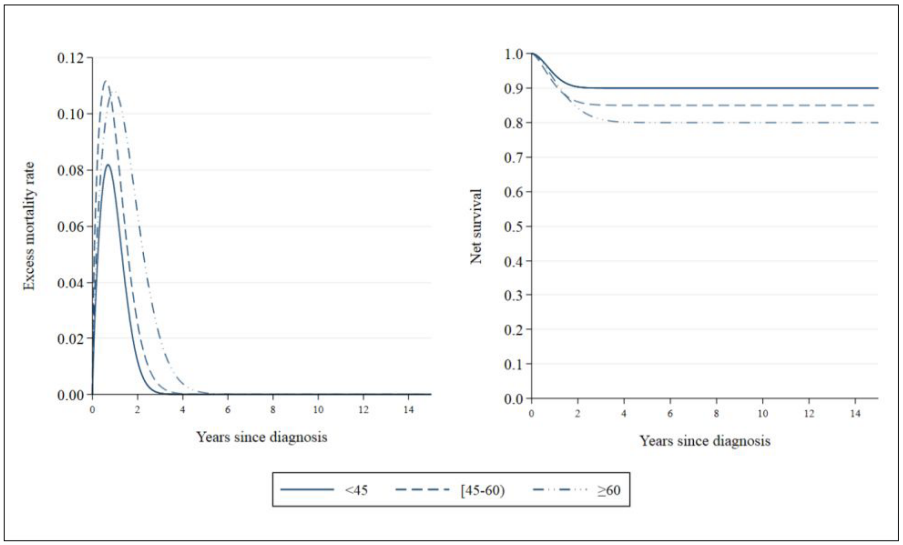


Figure 1 Good prognosis scenario: True net survival and excess mortality rates calculated from mixture cure models with a distribution of Weibull.

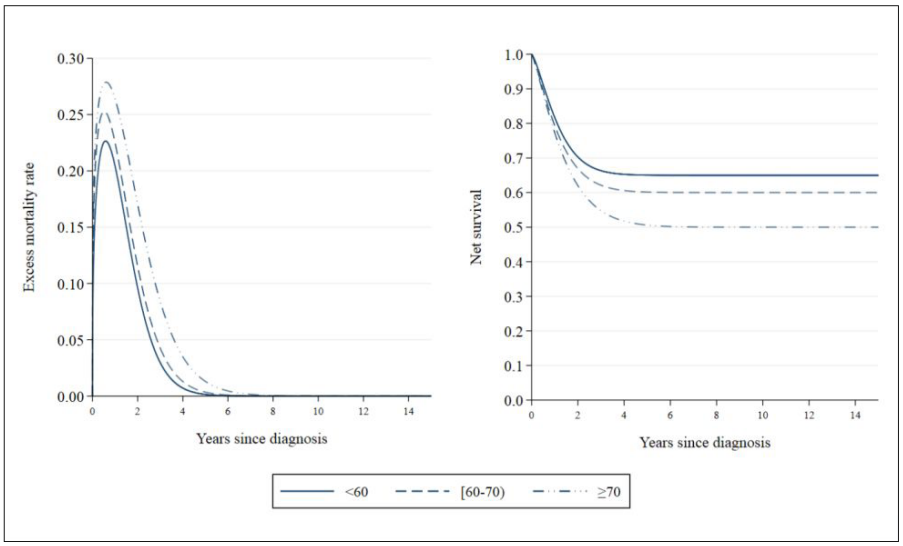


Figure 2 Medium prognosis scenario: True net survival and excess mortality rates calculated from mixture cure models with a distribution of Weibull.

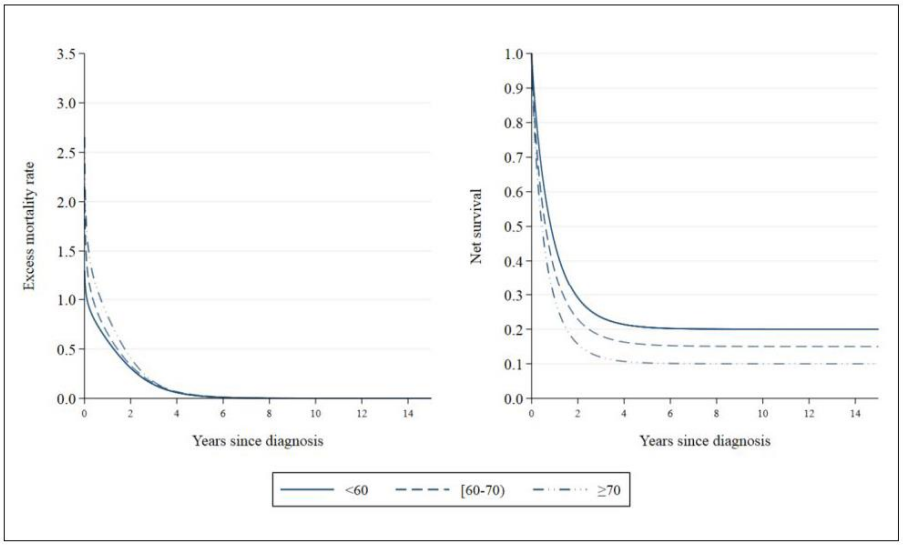


Figure 3 Poor prognosis scenario: True net survival and excess mortality rates calculated from mixture cure models with a distribution of Weibull.

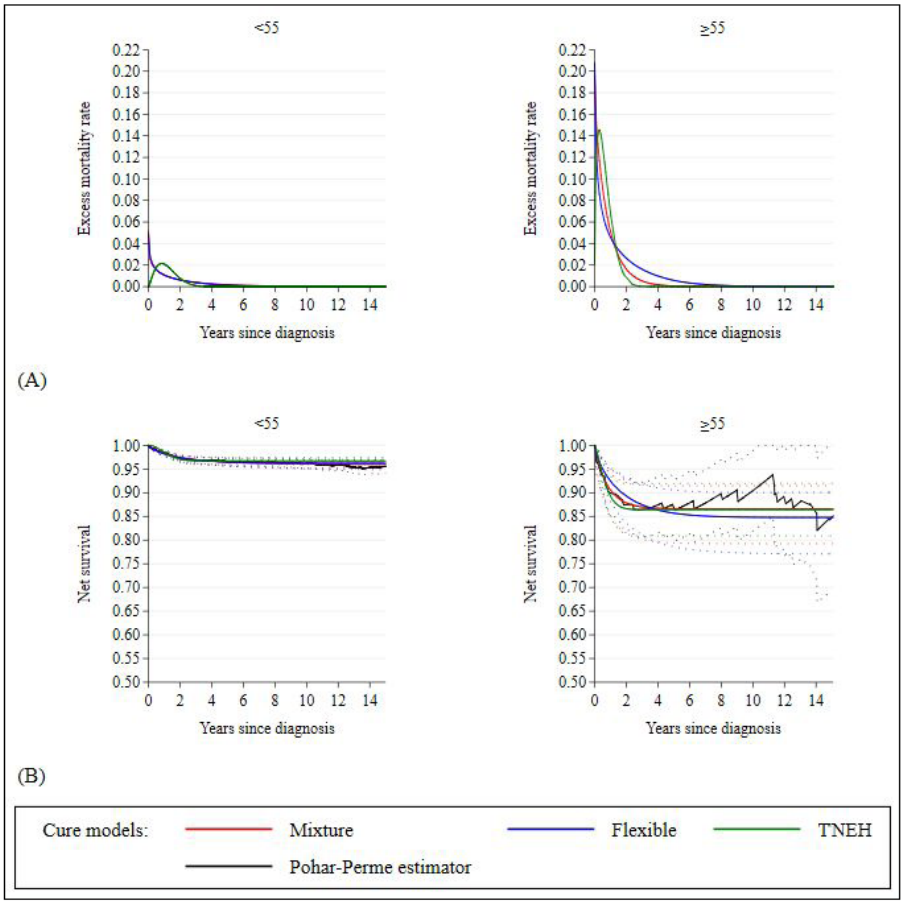


Figure 4 Testicular cancer: excess mortality rate **(A)** and net survivals with confidence intervals at 95% **(B)** estimated for each age group by the mixture cure model (Mixture), the parametric flexible cure model (Flexible) and the TNEH model (TNEH), and the Pohar-Perme estimator of net survival with confidence intervals at 95%.

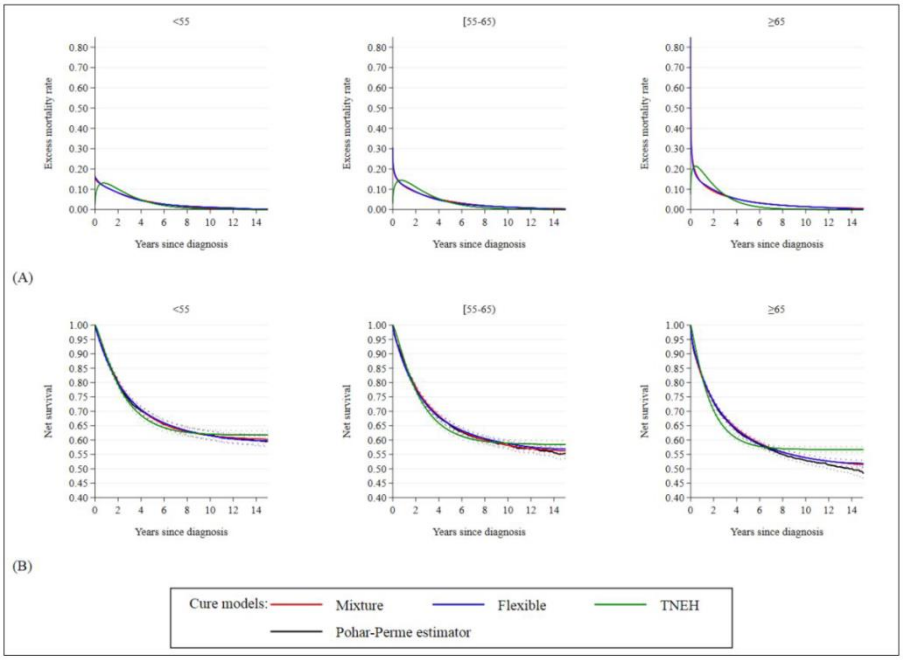


Figure 5 Colon cancer: excess mortality rate **(A)** and net survivals with confidence intervals at 95% **(B)** estimated for each age group by the mixture cure model (Mixture), the parametric flexible cure model (Flexible) and the TNEH model (TNEH), and the Pohar-Perme estimator of net survival with confidence intervals at 95%.

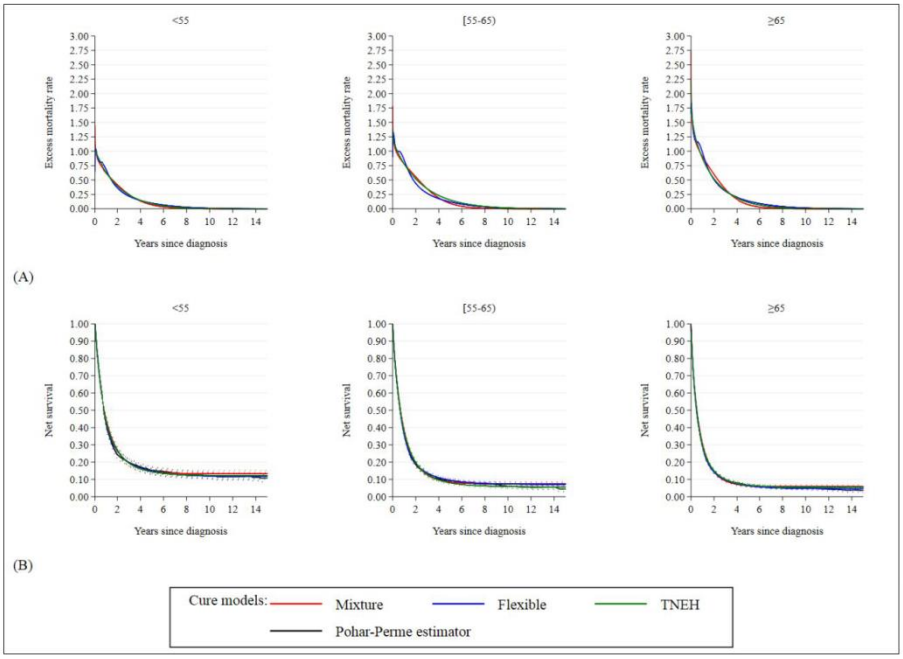


Figure 6 Pancreatic cancer: excess mortality rate **(A)** and net survivals with confidence intervals at 95% **(B)** estimated for each age group by the mixture cure model (Mixture), the parametric flexible cure model (Flexible) and the TNEH model (TNEH), and the Pohar-Perme estimator of net survival with confidence intervals at 95%.

Table 1 Good prognosis scenario: Performances of the mixture (MCM), parametric flexible (FCM) and the Time-to-Null-Excess-Hazard (TNEH) cure models over 1,000 samples of size 2,000 simulated with the mixture cure model with a Weibull distribution.

True values	Bias			Empirical standard error			Average standard error			Coverage rate		
	MCM	FCM	TNEH	MCM	FCM	TNEH	MCM	FCM	TNEH	MCM	FCM	TNEH
<45 years												
$S_n(3) = 0.900$	<0.001	0.014	0.001	0.012	0.011	0.012	-0.012	0.011	0.012	95.9%	78.8%	95.2%
$S_n(5) = 0.900$	<0.001	0.003	0.001	0.012	0.012	0.012	-0.012	0.012	0.012	96.0%	95.5%	95.1%
$S_n(10) = 0.900$	<0.001	-0.003	0.001	0.012	0.013	0.012	-0.012	0.013	0.012	96.0%	93.7%	95.1%
$\pi = 0.900$	<0.001	-0.003	0.001	0.012	0.013	0.012	0.012	0.013	0.012	96.0%	93.8%	95.1%
[45-60) years												
$S_n(3) = 0.851$	<0.001	0.014	-0.002	0.013	0.012	0.013	-0.013	0.012	0.013	94.9%	79.7%	94.5%
$S_n(5) = 0.850$	-0.001	-0.001	-0.001	0.013	0.013	0.013	-0.013	0.013	0.013	94.7%	94.3%	94.4%
$S_n(10) = 0.850$	-0.001	-0.010	-0.001	0.013	0.013	0.013	-0.013	0.014	0.013	94.7%	86.7%	94.4%
$\pi = 0.850$	-0.001	-0.011	-0.001	0.013	0.014	0.013	0.013	0.014	0.013	94.7%	85.5%	94.4%
≥60 years												
$S_n(3) = 0.810$	-0.001	0.012	<0.001	0.016	0.015	0.016	-0.017	0.016	0.017	96.1%	90.9%	95.8%
$S_n(5) = 0.800$	-0.001	0.001	<0.001	0.018	0.017	0.018	-0.019	0.018	0.019	95.7%	96.0%	95.6%
$S_n(10) = 0.800$	-0.002	-0.010	<0.001	0.018	0.018	0.018	-0.019	0.019	0.019	95.8%	91.4%	95.7%
$\pi = 0.800$	-0.002	-0.011	<0.001	0.018	0.018	0.018	0.019	0.019	0.019	95.9%	90.7%	95.7%

Table 2 Medium prognosis scenario: Performances of the mixture (MCM), parametric flexible (FCM) and the Time-to-Null-Excess-Hazard (TNEH) cure models over 1,000 samples of size 2,000 simulated with the mixture cure model with a Weibull distribution.

True values	Bias			Empirical standard error			Average standard error			Coverage rate		
	MCM	FCM	TNEH	MCM	FCM	TNEH	MCM	FCM	TNEH	MCM	FCM	TNEH
<60 years												
$S_n(3) = 0.663$	0.001	0.026	0.017	0.019	0.018	0.018	-0.020	0.018	0.019	95.4%	73.6%	85.2%
$S_n(5) = 0.651$	0.001	0.006	0.010	0.020	0.019	0.020	-0.020	0.020	0.020	95.2%	95.1%	92.3%
$S_n(10) = 0.650$	0.001	-0.010	0.007	0.020	0.020	0.020	-0.020	0.020	0.021	95.2%	90.7%	93.5%
$\pi = 0.650$	0.001	-0.012	0.007	0.020	0.020	0.020	0.020	0.020	0.021	95.2%	89.6%	93.5%
[60-70) years												
$S_n(3) = 0.621$	<0.001	0.016	0.012	0.017	0.017	0.016	-0.018	0.017	0.017	95.7%	85.3%	88.9%
$S_n(5) = 0.601$	<0.001	<0.001	0.002	0.019	0.018	0.019	-0.019	0.018	0.019	94.8%	95.3%	94.8%
$S_n(10) = 0.600$	<0.001	-0.017	-0.002	0.019	0.018	0.020	-0.019	0.019	0.020	94.6%	83.7%	94.3%
$\pi = 0.600$	<0.001	-0.019	-0.002	0.019	0.018	0.020	0.019	0.019	0.020	94.8%	81.5%	94.3%
≥70 years												
$S_n(3) = 0.548$	-0.001	0.018	0.014	0.021	0.020	0.020	-0.022	0.021	0.020	95.3%	89.5%	90.8%
$S_n(5) = 0.506$	-0.002	0.019	0.003	0.024	0.021	0.025	-0.024	0.022	0.023	96.0%	89.3%	93.4%
$S_n(10) = 0.500$	-0.002	0.005	-0.008	0.025	0.022	0.028	-0.025	0.023	0.027	95.5%	95.8%	92.0%
$\pi = 0.500$	-0.002	0.003	-0.008	0.025	0.022	0.028	0.025	0.023	0.027	95.9%	96.1%	92.1%

Table 3 Poor prognosis scenario: Performances of the mixture (MCM), parametric flexible (FCM) and the Time-to-Null-Excess-Hazard (TNEH) cure models over 1,000 samples of size 2,000 simulated with the mixture cure model with a Weibull distribution.

True values	Bias			Empirical standard error			Average standard error			Coverage rate		
	MCM	FCM	TNEH	MCM	MCM	FCM	TNEH	MCM	MCM	FCM	TNEH	MCM
<60 years												
$S_n(3) = 0.235$	<0.001	0.026	0.012	0.016	0.015	0.015	-0.016	0.015	0.015	94.5%	61.6%	88.1%
$S_n(5) = 0.205$	<0.001	0.026	0.010	0.016	0.015	0.017	-0.016	0.015	0.016	94.1%	59.9%	89.0%
$S_n(10) = 0.200$	<0.001	0.017	0.007	0.017	0.014	0.018	-0.016	0.015	0.017	93.4%	79.9%	90.4%
$\pi = 0.200$	<0.001	0.016	0.007	0.017	0.014	0.018	0.016	0.015	0.017	93.1%	82.2%	90.4%
[60-70) years												
$S_n(3) = 0.181$	<0.001	0.011	<0.001	0.013	0.012	0.013	-0.014	0.013	0.013	94.6%	89.3%	93.4%
$S_n(5) = 0.155$	<0.001	0.010	<0.001	0.013	0.012	0.014	-0.014	0.012	0.013	94.8%	89.4%	92.8%
$S_n(10) = 0.150$	<0.001	0.003	-0.001	0.014	0.012	0.015	-0.014	0.012	0.014	95.3%	94.7%	92.9%
$\pi = 0.150$	<0.001	0.002	-0.001	0.014	0.012	0.015	0.014	0.012	0.014	95.3%	94.8%	93.0%
≥70 years												
$S_n(3) = 0.119$	<0.001	0.009	-0.001	0.015	0.013	0.015	-0.015	0.013	0.014	95.1%	88.6%	92.1%
$S_n(5) = 0.103$	<0.001	0.004	-0.009	0.015	0.012	0.017	-0.015	0.012	0.015	95.0%	93.5%	85.5%
$S_n(10) = 0.100$	-0.001	-0.003	-0.013	0.016	0.012	0.018	-0.015	0.012	0.016	95.5%	93.9%	79.8%
$\pi = 0.100$	-0.001	-0.004	-0.013	0.016	0.012	0.018	0.016	0.012	0.016	95.1%	93.3%	79.9%

Table 4 Testicular cancer: Net survival at 3, 5 and 10 years after diagnosis and cure fraction (with confidence interval at 95%) estimated for each age group by the mixture cure model (Mixture), the parametric flexible cure model (Flexible) and the TNEH model.

Age at diagnosis		Mixture		Flexible		TNEH	
< 55	$S_n(3)$	97.0%	(96.2-97.6)	96.9%	(96.2-97.5)	96.7%	(96.0-97.5)
	$S_n(5)$	96.4%	(95.6-97.1)	96.5%	(95.6-97.1)	96.7%	(95.9-97.5)
	$S_n(10)$	96.1%	(95.0-96.9)	96.2%	(95.2-97.0)	96.7%	(95.9-97.5)
	π	96.0%	(96.0-94.7)	96.2%	(95.1-97.1)	96.7%	(95.9-97.5)
≥ 55	$S_n(3)$	87.0%	(80.2-91.6)	87.5%	(81.1-91.8)	86.4%	(80.9-92.0)
	$S_n(5)$	86.6%	(79.4-91.5)	85.7%	(78.6-90.6)	86.4%	(80.9-92.0)
	$S_n(10)$	86.6%	(79.2-91.5)	84.8%	(77.2-90.0)	86.4%	(80.9-92.0)
	π	86.6%	(79.2-91.5)	84.8%	(77.1-90.1)	86.4%	(80.9-92.0)

Table 5 Colon cancer: Net survival at 3, 5 and 10 years after diagnosis and cure fraction (with confidence interval at 95%) estimated for each age group by the mixture cure model, the parametric flexible cure model and the TNEH model.

Age at diagnosis		Mixture		Flexible		TNEH	
< 55	$S_n(3)$	74.5%	(73.3-75.6)	74.4%	(73.3-75.5)	72.8%	(71.6-73.9)
	$S_n(5)$	67.6%	(66.3-68.9)	68.0%	(66.7-69.2)	66.%	(64.7-67.3)
	$S_n(10)$	61.6%	(60.0-63.0)	61.5%	(60.0-63.0)	62.%	(60.6-63.5)
	π	60.0%	(60.0-58.2)	59.7%	(58.1-61.2)	61.7%	(60.2-63.2)
[55-65)	$S_n(3)$	72.6%	(71.6-73.5)	72.0%	(71.1-72.9)	70.3%	(69.4-71.3)
	$S_n(5)$	65.4%	(64.3-66.5)	65.4%	(64.3-66.4)	63.1%	(62.0-64.2)
	$S_n(10)$	58.2%	(56.9-59.5)	58.8%	(57.5-60.0)	58.9%	(57.6-60.1)
	π	55.3%	(55.3-53.5)	56.8%	(55.5-58.2)	58.5%	(57.2-59.8)
≥ 65	$S_n(3)$	67.9%	(67.1-68.7)	67.7%	(66.9-68.4)	64.%	(63.2-64.9)
	$S_n(5)$	61.1%	(60.2-62.0)	60.8%	(59.9-61.6)	58.8%	(57.9-59.7)
	$S_n(10)$	53.9%	(52.8-55.0)	53.9%	(52.8-54.9)	56.8%	(55.8-57.8)
	π	49.6%	(47.9-51.4)	51.8%	(50.6-53.0)	56.7%	(55.7-57.7)

Table 6 Pancreatic cancer: Net survival at 3, 5 and 10 years after diagnosis and cure fraction (with confidence interval at 95%) estimated for each age group by the mixture cure model, the parametric flexible cure model and the TNEH model.

Age at diagnosis		Mixture		Flexible		TNEH	
< 55	$S_n(3)$	19.4%	(17.6-21.4)	19.9%	(18.1-21.7)	19.5%	(17.8-21.3)
	$S_n(5)$	14.6%	(12.9-16.4)	14.7%	(13.1-16.4)	14.4%	(12.7-16.1)
	$S_n(10)$	13.3%	(11.5-15.2)	11.9%	(10.5-13.5)	12.4%	(10.6-14.1)
	π	13.3%	(13.3-11.5)	11.7%	(10.2-13.3)	12.3%	(10.5-14.1)
[55-65)	$S_n(3)$	12.6%	(11.4-13.8)	13.7%	(12.5-14.9)	12.9%	(11.8-14.0)
	$S_n(5)$	8.4%	(7.4-9.6)	9.5%	(8.5-10.5)	8.0%	(7.0-9.0)
	$S_n(10)$	7.5%	(6.4-8.7)	7.3%	(6.4-8.3)	5.9%	(4.8-6.9)
	π	7.5%	(7.5-6.4)	7.2%	(6.3-8.2)	5.7%	(4.6-6.7)
≥ 65	$S_n(3)$	9.1%	(8.3-10.0)	9.7%	(8.8-10.5)	10.0%	(9.2-10.7)
	$S_n(5)$	6.4%	(5.6-7.2)	6.4%	(5.7-7.1)	6.7%	(6.0-7.5)
	$S_n(10)$	5.9%	(5.1-6.7)	4.8%	(4.2-5.4)	5.5%	(4.7-6.3)
	π	5.8%	(5.1-6.7)	4.6%	(4.0-5.3)	5.5%	(4.6-6.3)

ANNEXE D : ARTICLE RÉVISÉ « MODELING EXCESS HAZARD WITH TIME-TO-CURE AS A PARAMETER », SOUMIS À BIOMETRICS

Biometrics XX, 1–22 DOI: XX
Month YEAR

Modeling excess hazard with time–to–cure as a parameter

Olayidé Boussari[1,2,*], Laurent Bordes[3], Gaëlle Romain[1,2], Marc Colonna[4], Nadine Bossard[5,6], Laurent Remontet[5,6], and Valérie Jooste[1,2]

[1]UMR 1231, EPICAD team, INSERM, Université Bourgogne–Franche–Comté, Dijon, F–21000, France

[2]Registre Bourguignon des Cancers Digestifs, Dijon-Bourgogne University Hospital, Dijon, F–21000, France

[3]UMR 5142, LMAP–IPRA, CNRS, E2S UPPA, Université Pau & Pays Adour, Pau, F–64000, France

[4]Registre du Cancer de l'Isère, Grenoble University Hospital, Grenoble, F–38000, France

[5]Department of Biostatistics and Bioinformatics, Hospices Civils de Lyon, Lyon, F–69003, France

[6]UMR 5558, LBBE, Biostatistics Health Group, CNRS, University Lyon 1, Lyon, F–69100, France

**email:* olayide.boussari@u-bourgogne.fr

Summary: Cure models have been widely developed to estimate the cure fraction when some subjects never experience the event of interest. However these models were rarely focused on the estimation of the time–to–cure i.e. the delay elapsed between the diagnosis and "the time from which cure is reached", an important indicator, for instance to address the question of access to insurance or loans for subjects with personal history of cancer. We propose a new excess hazard regression model that includes the time–to–cure as a covariate dependent parameter to be estimated. The model is written similarly to a Beta probability distribution function and is shown to be a particular case of the non-mixture cure models. Parameters are estimated through a maximum likelihood approach and simulation studies demonstrate good performance of the model. Illustrative applications to two cancer data sets are provided and some limitations as well as possible extensions of the model are discussed. The proposed model offers a simple and comprehensive way to estimate more accurately the time–to–cure.

Key words: Cancer; Cure model; Cure time; Net survival; Right to be forgotten.

This paper has been submitted for consideration for publication in *Biometrics*

1. Background

Since their first formulation by Boag (1949), Mixture Cure Models (MCM) have been widely developed to deal with survival data including a fraction of subjects who never experience the event of interest ("cured subjects"). Various ways have been considered by authors (see e.g. Kuk and Chen (1992); Li and Taylor (2002); Zhang and Peng (2009)) to model the baseline of both the survival function of the "uncured subjects" and the cure fraction, as well as the covariates effects on these two quantities, and these led to an extensive development of the MCM. A large review of MCM can be found in Maller and Zhou (1996) or Klein et al. (2016). In Yakovlev, Tsodikov, and Asselain (1996) a new family of cure models was introduced, the bounded cumulative hazard cure models also known as non-mixture cure models (NMCM). Suitable reviews and interpretations of the NMCM were proposed in Chen, Ibrahim, and Sinha (1999); Tsodikov, Ibrahim, and Yakovlev (2003); and Cooner et al. (2007). Other cure models can be found in the litterature (e.g. Yin and Ibrahim (2005b); Gu, Sinha, and Banerjee (2011)) and some approaches based on the Box-Cox transformation have been developed to unify different types of cure models (Yin and Ibrahim (2005a); Zeng, Yin, and Ibrahim (2006); Taylor and Liu (2007)). For practical use and interpretation of covariates effects, each type of cure model has both advantages and disadvantages and the issue of cure model selection was addressed for example in Peng and Xu (2012).

Since the late 1990s, cure models have been extended to the framework of the net survival (survival that would be observed if no death could occur from other causes than the disease of interest) with applications emphasized on cancer data (see Verdecchia et al. (1998); Yu et al. (2005); Lambert et al. (2006); Andersson et al. (2011) among others).

Let us recall briefly hereafter the basic concept of cure models within the net survival framework. We consider a population of patients suffering from a disease (say cancer). For a given patient, let denote A the age at diagnosis, X_1 (X_2 respectively) the latent variable

corresponding to the time elapsed between the diagnosis and the death due to cancer (other causes respectively), C the right censoring time, $\Delta = \mathbb{1}\{X < C\}$ the censoring indicator where $X = \min(X_1, X_2)$ and $\boldsymbol{Z}$ a vector of covariates in $\mathbb{R}^d$. We assume that conditionally on $\boldsymbol{Z}$, X and C are independent. Then an observation is a quadruple $(T, \Delta, A, \boldsymbol{Z})$, where $T = \min(X, C)$ is the observed time since diagnosis then $T + A$ is the observed time since birth. As it is well known that the age at diagnosis is one of the covariates that influenced the risk of dying from cancer, A is often also included in the vector of covariates $\boldsymbol{Z}$. When cause of death is not available, the most used methodology to estimate the net survival is to assume that conditionally on $(A, \boldsymbol{Z}) = (a, \boldsymbol{z})$, the observed hazard λ_{obs} of T equals the sum of λ_{pop} the known background mortality hazard in the general population (provided by life tables from national statistics) and λ_{exc} the excess hazard due to cancer:

$$\lambda_{\text{obs}}(t|\boldsymbol{z}) = \lambda_{\text{pop}}(t + a|\boldsymbol{z}) + \lambda_{\text{exc}}(t|\boldsymbol{z}). \tag{1}$$

From (1) the link between the survival functions is given by:

$$S_{\text{obs}}(t|\boldsymbol{z}) = S_{\text{pop}}(t + a|\boldsymbol{z}) \times S_{\text{net}}(t|\boldsymbol{z}), \tag{2}$$

where S_{obs} (S_{pop} respectively) represents the observed (the background respectively) survival distribution function, and S_{net} corresponds to the net survival distribution function, linked to λ_{exc} through

$$S_{\text{net}}(t|\boldsymbol{z}) = \exp\Big\{-\Lambda_{\text{exc}}(t|\boldsymbol{z})\Big\} = \exp\left\{-\int_0^t \lambda_{\text{exc}}(x|\boldsymbol{z})\mathrm{d}x\right\}, \tag{3}$$

with Λ_{exc} denoting the cumulative excess hazard function.

In situations where it is assumed that a fraction of patients will not die from cancer (meaning that $\lim_{t\to+\infty} \Lambda_{\text{exc}}(t|\boldsymbol{z}) < +\infty$ or equivalently that $\lim_{t\to+\infty} S_{\text{net}}(t|\boldsymbol{z}) > 0$), the observed subjects can be partitioned into two groups (cured and uncured subjects). The net survival can then be expressed as a mixture cure model:

$$S_{\text{net}}(t|\boldsymbol{z}) = \pi(\boldsymbol{z})S_1(t|\boldsymbol{z}) + \{1 - \pi(\boldsymbol{z})\}S_2(t|\boldsymbol{z}) = \pi(\boldsymbol{z}) + \{1 - \pi(\boldsymbol{z})\}S_2(t|\boldsymbol{z}), \tag{4}$$

where $S_1(t|\boldsymbol{z}) \equiv 1$ and $S_2(t|\boldsymbol{z})$ are the net survival functions of cured and uncured patients respectively, the later being a proper survival distribution. The fraction of cured subjects is $\pi(\boldsymbol{z})$, it depends on covariates $\boldsymbol{z}$ and from (4), is equal to $\lim_{t\to+\infty} S_{\text{net}}(t|\boldsymbol{z})$.

Although cure models have been originally designed to estimate the fraction of cured subjects, they can be used to estimate another important epidemiological indicator: the time–to–cure i.e. the delay elapsed between the diagnosis and the "time from which cure is reached". Existing cure models do not allow direct estimation of the time–to–cure although this indicator seems crucial; as for instance it can be used to improve the estimation of the delay for the right to be forgotten for cancer survivors. Indeed the right to be forgotten provision is an important milestone in European policymaking. However, it is not universally accessible to cancer survivors across Europe nor does it address all their specific issues. Cancer survivors are often disadvantaged when applying for essential services such as loans, mortgages or child adoption (Youth Cancer Europe, 2018).

Different methods to estimate the time–to–cure after fitting a cure model have been proposed: Dal Maso et al. (2014) defined the time–to–cure as the delay elapsed between the diagnosis and the time from which the 5–year conditional net survival (defined as the ratio between the net survival at time $t + 5$ years and the net survival at time t) becomes greater than 0.95. In a recent paper Boussari et al. (2018) proposed to consider for a given patient i, the probability $p_i(t)$ of being cured at a given time t after diagnosis knowing that he/she was alive up to t (this probability is nothing but the ratio between the cure fraction and the net survival at time t); then the time–to–cure is estimated as the delay from which $p_i(t)$ reaches 0.95.

This work considers a natural definition of the time–to–cure, named hereafter *time–to–null–excess–hazard* (TNEH), as the delay elapsed between the diagnosis and the time from which the excess hazard becomes null, and proposes a new excess hazard model where the TNEH

is a covariate dependent parameter to be estimated.

We obtain from (3), (4) and the definitions of both the cure fraction and the TNEH:

$$\pi(\boldsymbol{z}) = \lim_{t\to+\infty} S_{\text{net}}(t|\boldsymbol{z}) = \exp\Big\{-\Lambda_{\text{exc}}(\tau(\boldsymbol{z})|\boldsymbol{z})\Big\} = \exp\left\{-\int_0^{\tau(\boldsymbol{z})} \lambda_{\text{exc}}(x|\boldsymbol{z})\mathrm{d}x\right\}, \qquad (5)$$

where $\tau(\boldsymbol{z})$ is the TNEH depending on the covariates $\boldsymbol{z}$ and $\lambda_{\text{exc}}(t|\boldsymbol{z})$=0 whenever $t > \tau(\boldsymbol{z})$. An illustrative plot of the three hazards functions (described in (1)) is given in Figure 1 for an individual diagnosed at 55 or 70 years, assuming that cure is reached and that TNEH depends on the age at diagnosis.

[Figure 1 about here.]

The rest of the paper is organized as follows. In the next Section the new model and its properties are presented as well as the parameters estimation procedure. We illustrate the performances of the estimators derived from our model through both a simulation study in Section 3 and applications to survival data from French cancer registries in Section 4. The last Section is devoted to some concluding remarks summarizing the paper and providing some future related researches.

2. New cure model including the TNEH as parameter

2.1 *Model specification*

Various parametric excess hazard functions have been explored to incorporate the TNEH as a parameter (τ). According to the definitions of both the TNEH and the excess hazard function, one of the conditions that must be held by any candidate excess hazard function is to be continuous on $[0, +\infty)$, positive on $[0, \tau)$ and null from τ. Besides, the function must have the ability to reproduce a large panel of excess hazard curves encountered, in particular

in cancer survival study. The new excess hazard model is written as follows:

$$\lambda_{\text{exc}}(t|\boldsymbol{z};\boldsymbol{\theta}) = \left\{\frac{t}{\tau(\boldsymbol{z};\boldsymbol{\eta})}\right\}^{\alpha(\boldsymbol{z};\boldsymbol{\gamma})-1}\left\{1-\frac{t}{\tau(\boldsymbol{z};\boldsymbol{\eta})}\right\}^{\beta-1}\mathbb{1}_{\{0\leqslant t\leqslant\tau(\boldsymbol{z};\boldsymbol{\eta})\}}, \tag{6}$$

where $\tau(\boldsymbol{z};\boldsymbol{\eta}) > 0$ is the TNEH depending on covariates $\boldsymbol{z}$ through the vector of parameters $\boldsymbol{\eta}$. Both $\beta > 1$ and $\alpha(\boldsymbol{z};\boldsymbol{\gamma}) > 0$ are shape parameters, the later depending on covariates $\boldsymbol{z}$ through the vector of parameters $\boldsymbol{\gamma}$. Hence $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \beta, \boldsymbol{\eta})$ is the vector of parameters to be estimated. Note that we constrained β to be larger than 1 in order to insure the nullity and the continuity of $\lambda_{\text{exc}}(t|\boldsymbol{z};\boldsymbol{\theta})$ at $\tau(\boldsymbol{z};\boldsymbol{\eta})$. The shape of the excess hazard function is therefore dependent of the value of $\alpha(\boldsymbol{z};\boldsymbol{\gamma})$, either it belongs to $(0,1)$ and the excess hazard function is non increasing on $[0,\tau(\boldsymbol{z};\boldsymbol{\eta})]$ with $\lambda_{\text{exc}}(t|\boldsymbol{z};\boldsymbol{\theta})$ tending to infinity as t tends to 0, or it is larger than 1 and $\lambda_{\text{exc}}(t|\boldsymbol{z};\boldsymbol{\theta})$ is N-shaped with a maximum located at $\{\alpha(\boldsymbol{z};\boldsymbol{\gamma})-1\}/\{\alpha(\boldsymbol{z};\boldsymbol{\gamma})+\beta-1\}$; thus because of the linear link between the covariates and the parameters interpreting the effect of covariates in the shape of excess hazard rate seems easy. Again, the covariates effects on the TNEH ($\tau(\boldsymbol{z};\boldsymbol{\eta})$) are easy to interpret because of the linear link. In the sequel we refer to the above specified model as the beta–TNEH model. An example of the beta–TNEH model, with the sex (sex) and age at the diagnosis of disease (age) as covariates can be expressed:

$$\lambda_{\text{exc}}(t|\boldsymbol{z};\boldsymbol{\theta}) = \left(\frac{t}{\eta_0+\eta_1\times age}\right)^{(\gamma_0+\gamma_1\times age+\gamma_2\times sex)-1}\left(1-\frac{t}{\eta_0+\eta_1\times age}\right)^{\beta-1}\mathbb{1}_{\{0\leqslant t\leqslant(\eta_0+\eta_1\times age)\}},$$

with $\boldsymbol{z} = (age, sex)$ and $\boldsymbol{\theta} = (\gamma_0, \gamma_1, \gamma_2, \beta, \eta_0, \eta_1)$.

As the equation (6) has the structure of a Beta probability density function, we obtain:

The cumulative excess hazard function

$$\begin{aligned}\Lambda_{\text{exc}}(t|\boldsymbol{z};\boldsymbol{\theta}) &= \begin{cases}\tau(\boldsymbol{z};\boldsymbol{\eta})\int_0^{\frac{t}{\tau(\boldsymbol{z};\boldsymbol{\eta})}} x^{\alpha(\boldsymbol{z};\boldsymbol{\gamma})-1}(1-x)^{\beta-1}\,\mathrm{d}x & \text{if } 0\leqslant t\leqslant\tau(\boldsymbol{z};\boldsymbol{\eta})\\ \tau(\boldsymbol{z};\boldsymbol{\eta})\int_0^{1} x^{\alpha(\boldsymbol{z};\boldsymbol{\gamma})-1}(1-x)^{\beta-1}\,\mathrm{d}x & \text{if } t>\tau(\boldsymbol{z};\boldsymbol{\eta})\end{cases}\\ &= \tau(\boldsymbol{z};\boldsymbol{\eta})\,\mathrm{B}\Big(\alpha(\boldsymbol{z};\boldsymbol{\gamma}),\beta\Big)\,\mathrm{F}_{\mathcal{B}e}\Big(\frac{t}{\tau(\boldsymbol{z};\boldsymbol{\eta})};\alpha(\boldsymbol{z};\boldsymbol{\gamma}),\beta\Big),\end{aligned}$$

where B denotes the beta function and $\mathrm{F}_{\mathcal{B}e}\Big(\cdot\,;\alpha(\boldsymbol{z};\boldsymbol{\gamma}),\beta\Big)$ is the cumulative distribution function of a beta distribution with parameters $\alpha(\boldsymbol{z};\boldsymbol{\gamma})$ and β.

The net survival function

$$\begin{aligned} S_{\text{net}}(t|\boldsymbol{z};\boldsymbol{\theta}) &= \exp\Big\{-\Lambda_{\text{exc}}(t|\boldsymbol{z};\boldsymbol{\theta})\Big\} \\ &= \exp\left\{-\tau(\boldsymbol{z};\boldsymbol{\eta})\ \mathrm{B}\Big(\alpha(\boldsymbol{z};\boldsymbol{\gamma}),\beta\Big)\mathrm{F}_{\mathcal{B}e}\Big(\frac{t}{\tau(\boldsymbol{z};\boldsymbol{\eta})};\alpha(\boldsymbol{z};\boldsymbol{\gamma}),\beta\Big)\right\}. \end{aligned}$$

The cure fraction

$$\begin{aligned} \pi(\boldsymbol{z};\boldsymbol{\theta}) &= \exp\Big\{-\Lambda_{\text{exc}}\Big(\tau(\boldsymbol{z};\boldsymbol{\eta})|\boldsymbol{z};\boldsymbol{\theta}\Big)\Big\} \\ &= \exp\Big\{-\tau(\boldsymbol{z};\boldsymbol{\eta})\ \mathrm{B}\Big(\alpha(\boldsymbol{z};\boldsymbol{\gamma}),\beta\Big)\Big\}, \end{aligned}$$

hence for covariates that influenced only to the TNEH, it is easy to derive a log-linear link with the cure rate function, otherwise interpretation of the effect of covariates on the cure rate may be more complex.

REMARK 1:

(i) From the above two latest results we obtain: $S_{\text{net}}(t|\boldsymbol{z};\boldsymbol{\theta}) = \Big\{\pi(\boldsymbol{z};\boldsymbol{\theta})\Big\}^{\mathrm{F}_{\mathcal{B}e}\left(\frac{t}{\tau(\boldsymbol{z};\boldsymbol{\eta})};\alpha(\boldsymbol{z};\boldsymbol{\gamma}),\beta\right)}$. We recognize here the form of the NMCM, then the beta–TNEH model can be seen as a special case of the non-mixture cure models family.

(ii) The issue of covariates incorporation in model (6) (especially in both the two shape parameters α and β) is discussed in Section 5.

(iii) A huge literature about mixture models identifiability exists. Recently general results about the identifiability of parameters of a cure model have been obtained by Hanin and Huang (2014). Because of the specificity of our model we provide in the appendix conditions under which a direct proof of parameters identifiability is obtained.

2.2 *Parameters estimation*

Maximum Likelihood Estimator (MLE): For a subject i we observe $(t_i, \delta_i, a_i, \boldsymbol{z}_i)$ a realization of $(T_i, \Delta_i, A_i, \boldsymbol{Z}_i)$. The contribution of the i^{th} subject to the log-likelihood is

$$\begin{aligned}\ell(\boldsymbol{\theta}|t_i, \delta_i, a_i, \boldsymbol{z}_i) &= \delta_i \log\left(\lambda_{\text{obs}}(t_i, a_i|\boldsymbol{z}_i, \boldsymbol{\theta})\right) - \Lambda_{\text{obs}}(t_i, a_i|\boldsymbol{z}_i, \boldsymbol{\theta}) \\ &\equiv \delta_i \log\left(\lambda_{\text{pop}}(t_i + a_i|\boldsymbol{z}_i) + \lambda_{\text{exc}}(t_i|\boldsymbol{z}_i, \boldsymbol{\theta})\right) - \Lambda_{\text{exc}}(t_i|\boldsymbol{z}_i, \boldsymbol{\theta}). \qquad (7)\end{aligned}$$

Hence for a sample of n subjects, the MLE satisfies

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}|t_i, \delta_i, a_i, \boldsymbol{z}_i).$$

Standard errors of the MLE: It is well known (see e.g. Newey and McFadden (1994)) that $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(0, I_0^{-1}\right)$ as n tends to infinity, where I_0 denotes the Fisher information matrix. Using standard martingale methods for counting processes (see Andersen et al. (2012), Section VI.1), I_0 is consistently estimated by $\widehat{I}$ defined by

$$\widehat{I} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i \frac{\partial \lambda_{\text{exc}}}{\partial \boldsymbol{\theta}}(t_i|\boldsymbol{z}_i, \widehat{\boldsymbol{\theta}})}{\lambda_{\text{pop}}(t_i + a_i|\boldsymbol{z}_i) + \lambda_{\text{exc}}(t_i|\boldsymbol{z}_i, \widehat{\boldsymbol{\theta}})}\right\}^{\otimes 2}, \qquad (8)$$

where for a column vector $\boldsymbol{v}$, $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^T$. Thus the standard error of the i-th component of $\widehat{\boldsymbol{\theta}}$ is estimated by the square root of the i-th diagonal entry of $n^{-1}\widehat{I}^{-1}$. Moreover the standard errors of other quantities related to $\widehat{\boldsymbol{\theta}}$ such as the cure fraction or the net survival could be derived easily using the delta method.

REMARK 2: In the estimation procedure we use the bound constrained optimization method (L-BFGS-B) of Byrd et al. (1995). This method takes the advantage of the BFGS algorithm which is shown to have good performance even for non-smooth optimization functions (Lewis and Overton, 2009), and uses simple bounds constraints. The optimization algorithm looks for a $\widehat{\boldsymbol{\theta}}$ belonging to a predefined set having the form $\prod_{j=1}^{k}[\theta_{j,min}\ ,\ \theta_{j,max}] \subset \mathbb{R}^k$ where $k \in \mathbb{N}$ denotes the number of the parameters. The initial value of $\boldsymbol{\theta}$ must belong to the predefined set. A help to set the boundaries is for instance, informations derived from the model constraints ($\tau > 0$, $\beta > 1$, $\alpha > 0$, see (6)). Note that if any of the estimates equals a

boundary, the predefined set of boundaries must be expanded, followed by a new estimation of the parameters. We provide in Web Appendix A, the sets of boundaries used in the estimations steps of the following numerical studies (Sections 3 and 4). The L-BFGS-B method is already implemented in the R software, package *stats*, function *optim*, (R Core Team, 2019); it requires very short time to run.

3. Simulation study

3.1 *Simulated examples*

For the data generation algorithm, one can refer to Web Appendix B.

In the following the model complexity was reduced, without loss of generality, by considering only the age at diagnosis as covariate. Hence we consider the following beta–TNEH model:

$$\lambda_{\text{exc}}(t|a;\boldsymbol{\theta}) = \left\{\frac{t}{\tau(a;\boldsymbol{\eta})}\right\}^{\alpha(a;\boldsymbol{\gamma})-1}\left\{1-\frac{t}{\tau(a;\boldsymbol{\eta})}\right\}^{\beta-1}\mathbb{1}_{\{0\leqslant t\leqslant\tau(a;\boldsymbol{\eta})\}}, \tag{9}$$

where $\tau(a;\boldsymbol{\eta}) = \eta_0+\eta_1\times a^*$ and $\alpha(a;\boldsymbol{\gamma}) = \gamma_0+\gamma_1\times a^*$ with a^* the age at diagnosis standardized using the mean and the standard deviation of its specified distribution. The vector of unknown parameters, to be estimated from a sample of size n, is $\boldsymbol{\theta} = (\gamma_0,\gamma_1,\beta,\eta_0,\eta_1)$. The population hazard (expected hazard) is assumed to follow a Weibull distribution, the scale and shape parameters being 75 and 11 respectively. We considered three different settings for the simulations: one illustrating a low excess hazard with a short TNEH and a high censoring rate, the second illustrating a high excess hazard with a moderate TNEH and a low censoring rate, the third illustrating a low excess hazard with a longer TNEH (the excess hazard becomes null very later and the net survival decreases slowly) and a moderate censoring rate. A graphical illustration of the discrepancy between the three settings is provided in Web Appendix C.

In the first setting, the vector of true parameters is $\boldsymbol{\theta}$ =(2.3, -0.1, 4.8, 5.5, 0.9). The age at diagnosis is uniformly distributed on intervals $[20,40)$, $[40,65)$ and $[65,80]$, and the proportions

of age at diagnosis coming from these three intervals are 0.36, 0.29 and 0.35 respectively. The maximum follow-up time (from diagnosis) is fixed to 15 years and the censoring rate is about 60%. The MLE performances for several sample sizes $n \in \{250, 500, 1000, 2000\}$ are reported in Table 1, Setting 1. The bias decreases as the sample size n increases and becomes very small for $n = 2000$. The discrepancies between the standard deviations of the n estimates (sd) and the empirical means of the standard deviation estimates ($\overline{se^*}$) are low, particularly when $n = 2000$. Furthermore, the standard deviations are reduced by half when the sample size n is quadrupled showing that the root–of–n asymptotic convergence rate is reached. For $n = 2000$ the coverage probabilities (cp) are close to 0.95 which is another indicator that the MLE behaves well for a sample of size 2000.

Table 1, Setting 2, summarizes the MLE performances from the second setting of simulation. Here the vector of true parameters is $\boldsymbol{\theta}$ =(1.25, -0.05, 3.5, 9, 0.3). The age at diagnosis is uniformly distributed on intervals $[20, 50)$, $[50, 70)$ and $[70, 80]$, and the proportions of age at diagnosis coming from these three intervals are 0.15, 0.60 and 0.25 respectively. The maximum follow-up time is fixed to fifteen years and the censoring rate is about 20%. The MLE performances from the third setting of simulation are summarized in Table 1, Setting 3, where the vector of true parameters is $\boldsymbol{\theta}$ =(3.01, -0.2, 2.98, 18, 1.2). The age at diagnosis is distributed as in the first setting, the maximum follow-up time is fixed to twenty–five years and the censoring rate is about 46%. Overall the MLE performances are slightly better in the first and second settings than in the third setting. The findings from the first simulation setting are consolidated showing that good estimates can be obtained with the MLE even with moderate sample sizes. Similar results regarding the MLE performances of the beta–TNEH model were obtained when a more extensive simulation setting was considered with λ_{exc} depending on three covariates: age at diagnosis (continuous), sex, stage of cancer (3 stages = I, II and III). See Web Appendix D for full details.

Moreover, we checked the beta–TNEH model performances when data were generated from distributions that do not fulfill the assumptions underlying the new model. We fit the beta–TNEH model to data simulated from another cure model, the Weibull mixture cure model (a mixture cure model where the survival time of uncured subjects follows a Weibull distribution) and assess performances of the beta–TNEH model by computing the bias, the root mean square error and the coverage probability for the cure fraction and for the net survival at time $t = 5$, 10 and 15 years. Results were provided in the appendix E. Overall, the beta–TNEH model gives good estimations only if the excess hazard from the Weibull mixture cure model becomes almost null within the follow-up interval.

We note that when data are generated from a distribution that do not allow cure fraction (i.e. the cure fraction is null), the beta–TNEH model does not fit the data and gives poor estimations. Indeed, a null cure fraction hypothesis is not compatible with the beta–TNEH model because it corresponds to a TNEH equal to infinity which is outside of the area of validity of the model.

[Table 1 about here.]

3.2 *Sensitivity to the initial value in the optimization procedure*

In this Section we investigate whether the MLE is robust with respect to the chosen initialization point for likelihood maximization algorithm. We then have to verify if the maximization algorithm converges to the same value $\widehat{\boldsymbol{\theta}}$ whatever the chosen initial value. We do this through a simulation study considering again the simulated example 1 with $B = 1000$ repetitions of a sample data of size $n = 2000$. We generate K initial values of $\boldsymbol{\theta}$ following a multiple uniform distribution on a given space defined by the bounds fixed for the optimization method (see Remark 2). For a given initial value $\boldsymbol{\theta}_k^{(0)}$, $k = 1, \ldots, K$, we compute the empirical mean $\widehat{\boldsymbol{\theta}}_k$ of the B estimates of $\boldsymbol{\theta}$, obtained from the B simulated samples data respectively. Then we are interested in the biases between the $\widehat{\boldsymbol{\theta}}_k$, $k = 1, \ldots, K$ and the true parameter.

Figure 2 shows for $K = 30$ initial values the boxplots obtained from the K estimates of biases between the empirical means and the true parameter. The ranges of the computed biases are 0.009, 0.001, 0.048, 0.055 and 0.018 for $\gamma_0 = 2.3$, $\gamma_1 = -0.1$, $\beta = 4.8$, $\eta_0 = 5.5$ and $\eta_1 = 0.9$ respectively (relative biases vary from 0.4% to 2%). These values are very low (near 0) which means that the estimates of $\boldsymbol{\theta}$ are almost identical whatever the chosen initial value. Thus the estimates are robust to the choice of the initial point of the optimization algorithm.

[Figure 2 about here.]

4. Illustrative examples on real data

Data were provided by the French network of cancer registries (FRANCIM). The analysis included all patients diagnosed from 1995 to 2010, aged 15 to 74 years at diagnosis and followed up to June 30, 2013. The follow-up time was censored at fifteen years. The variable "age at diagnosis" was categorized as in previous analyzes of the FRANCIM data (Cowppli-Bony et al., 2016). Thus we define $\tau(a;\boldsymbol{\eta}) = \eta_0 + \eta_1 \times \mathbb{1}_{A_1}(a) + \cdots + \eta_J \times \mathbb{1}_{A_J}(a)$ and $\alpha(a;\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 \times \mathbb{1}_{A_1}(a) + \cdots + \gamma_J \times \mathbb{1}_{A_J}(a)$ where A_j $(0 \leqslant j \leqslant J)$ are the age at diagnosis groups and, η_j and γ_j $(1 \leqslant j \leqslant J)$ denote respectively the deviations from the effects η_0 and γ_0 of age at diagnosis in the reference group A_0. In the examples A_0 was the group with the largest size. Model (9) was fitted on two data sets. Expected mortality rates were derived from the observed mortality rates in the general population available by sex, annual age, year of death, Department of residence and provided by the Institut National de la Statistique et des Etudes Economiques - France (see Web Appendix F for more details and plots of the expected mortality).

4.1 *Testicular cancer data*

We considered data of 2834 subjects diagnosed with testicular cancer, for which excess mortality is low. Death was observed for 182 subjects (6.4% of the cohort) and the observed

median survival time since diagnosis for fatal cases was 2.13 years.

Age at diagnosis was categorized into 4 groups $A_0 = [15,45)$, $A_1 = [45,55)$, $A_2 = [55,65)$ and $A_3 = [65,75)$, and the unknown parameters as well as their standard errors (in brackets) were estimated: $\widehat{\gamma_0} = 2.41(0.16)$, $\widehat{\gamma_1} = -0.13(0.18)$, $\widehat{\gamma_2} = -0.90(0.15)$, $\widehat{\gamma_3} = -0.84(0.28)$, $\widehat{\beta} = 8.40(2.06)$, $\widehat{\eta_0} = 5.47(1.49)$, $\widehat{\eta_1} = -1.07(0.98)$, $\widehat{\eta_2} = -0.30(0.92)$, $\widehat{\eta_3} = -2.94(1.64)$; with a log-likelihood equal to -1152.12, corresponding to an Akaike Information Criterion (AIC) equal to 2322.24.

According to the above estimations we pooled A_0 and A_1 as well as A_2 and A_3 and fitted another model with age at diagnosis categorized into 2 groups $A_0 = [15,55)$ and $A_1 = [55,75)$. The estimates were then: $\widehat{\gamma_0} = 2.39(0.15)$, $\widehat{\gamma_1} = -0.87(0.14)$, $\widehat{\beta} = 8.40(2.09)$, $\widehat{\eta_0} = 5.31(1.45)$ and $\widehat{\eta_1} = -0.91(0.81)$ and the log-likelihood $= -1153.87$, corresponding to an AIC $= 2317.74$. Based on the AIC values we selected the model with 2 age groups.

Figure 3A shows the resulting excess hazard functions while the net survivals from the model with 2 age groups are given in Figure 3 B and C. The excess hazard peaks around the first year after diagnosis whatever the age group. A significant discrepancy is observed between the two excess hazards during the two first years after diagnosis; they become almost identical from two years after diagnosis and close to zero a year later. For each of the age group, we plotted in the same panel (Figure 3 B and C), the net survival (with confidence bounds) estimation by the beta–TNEH model and the nonparametric estimation of the net survival (with confidence bounds) by the method of Perme, Stare, and Estève (2012), denoted hereafter PP model. The two curves match well enough whatever the age group showing that the beta–TNEH model provides a reasonable description of the data. Hence from the testicular cancer data, the cure fraction estimates followed by their 95% confidence intervals were 0.97 [0.93, 1.00] and 0.86 [0.72, 1.00] in the [15, 55) and the [55, 75) age at diagnosis groups respectively. The TNEH estimate in the [15, 55) age at diagnosis

group is $\widehat{\eta_0} = 5.31$ years with a 95% confidence interval equal to [2.45, 8.17]. In the [55, 75) age at diagnosis group, the TNEH estimate is $\widehat{\eta_0} + \widehat{\eta_1} = 4.40$ years with a 95% confidence interval equal to [1.91, 6.89].

When fitting a flexible parametric cure model (Andersson et al., 2011) to the testicular cancer data, the derived time–to–cure estimates from Dal Maso et al. (2014) approach with a cut-off at 0.95 were 0 year and 2.1 years in the [15, 55) and the [55, 75) age at diagnosis groups respectively.

[Figure 3 about here.]

4.2 *Women pancreatic cancer data*

We considered data of 3239 women diagnosed with pancreatic cancer, which is a cancer with a very high excess mortality. Death was observed for 91.6% of the subjects and half of the deaths were observed at least 0.6 years after diagnosis. The median and mean age at diagnosis were 66 years and 63.17 years respectively. We fitted a model with the age at diagnosis categorized into 3 groups A_0 = [65,75), A_1 = [55,65) and A_2 = [15,55).

In the estimation procedure (see Remark 2), the upper bound $\eta_{0,max}$ for the search of the baseline TNEH (i.e. TNEH for the reference age group A_0) was first fixed to 15 years (corresponding to the maximum follow-up time) and we observed that the estimate $\widehat{\eta_0}$ equals the upper bound. When varying $\eta_{0,max}$ up to 40 years, the estimation $\widehat{\eta_0}$ still reached the upper bound. According to the fixed value for $\eta_{0,max}$, the estimations of the other parameters varied but were always different from both their lower and upper bounds fixed for the estimation procedure. The parameter estimates and standard errors as well as the log-likelihood of the fitted model for three different values of $\eta_{0,max}$ (i.e. $\eta_{0,max} = 15, 18, 40$ years) are reported in Table 2. The likelihood increased with $\widehat{\eta_0}$, and for each parameter the standards errors estimates showed that the derived three confidence intervals overlapped (95% confidence intervals are $[9.73, 20.27]$, $[9.68, 26.32]$ and $[-13.37, 93.37]$ respectively). Note

that for $\widehat{\eta}_0 = 40$ years, we are far out of the follow-up interval (0 to 15 years) meaning that no data are available for consistent estimations of the parameters; this leads to large standard errors and consequently possible negative value for the lower bound of the 95% confidence interval, what must be interpreted with caution. For each age at diagnosis group, the 3 excess hazards corresponding to $\widehat{\eta}_0 = 15, 18$ and 40 years are plotted in Figure 4, panels A, B and C. Whatever the age group, the 3 curves were very similar showing that the changes in the parameter estimates had little impact on the corresponding excess hazards. The panels D, E and F of Figure 4 show for the 3 age groups, both the parametric and the nonparametric estimations of the survival function using the parametric beta–TNEH model and the nonparametric PP model. Overall whatever the age group, the two curves matched well, meaning that the beta–TNEH model provided a reliable estimation of the net survival function even if the TNEH was obviously underestimated. Note that the beta–TNEH model was able to provide satisfactory estimations because of the behavior of the excess hazard function which was high just after diagnosis, decreased rapidly and became very close to zero before the end of the follow-up. Although the net survival shows an apparent plateau (often translated as the existence of cure), this example shows a situation where either the TNEH had a non-finite value or the TNEH was too large, greater than a predefined threshold (for instance the maximum follow-up time) over which the estimated TNEH has no practical usefulness.

On the other hand, the approaches by both Dal Maso et al. (2014) and Boussari et al.(2018) provided approximations of the time–to–cure for the women pancreatic cancer data after fitting a flexible parametric cure model (Andersson et al., 2011). The time–to–cure estimates from the approach of Dal Maso et al. (2014) using a cut-off at 0.95, were 9 years in both the [15, 55) and the [55, 65) age at diagnosis groups and 10 years in the [65, 75) age at diagnosis

group; very close results were obtained from the approach proposed by Boussari et al. (2018) (see the related paper for more details).

[Table 2 about here.]

[Figure 4 about here.]

5. Concluding remarks

The time from which no more death occurs from a disease of interest (such as cancer) is a useful indicator in epidemiological studies, and can help to improve access to insurance and loans for people living with a personal history of cancer. In this paper we refer to this delay as the time–to–null–excess–hazard (TNEH). While sophisticated models have been proposed to estimate efficiently the fraction of cured patients, it seems that there is a lack of methods in the literature for the TNEH estimation.

We proposed a cure model based on a paradigm where the excess hazard function includes the TNEH as a parameter to be estimated. The proposed beta–TNEH model could be treated as a special case of the NMCM. The simulation study showed that the beta–TNEH presented good performances regarding the maximum likelihood estimation method. However we advise to do not use the new method when cure assumption is not reasonable because this will lead to poor estimations with meaningless results; the beta–TNEH model is really suitable for data showing an excess hazard which becomes null within the follow-up interval. Existing methods (for instance Dal Maso et al., 2014 and Boussari et al., 2018) can provide an approximation of the time–to–cure when the excess hazard becomes just relatively low and not necessarily null. But these methods are based on approximations requiring a choice of a cut-off what could easily be subject of criticism since the derived time–to–cure estimate could be very sensitive to the predefined value of the cut-off.

Two examples on real data sets were treated and for each of them, the net survival estimated

by the beta–TNEH model was very close to the estimation provided by the nonparametric PP model. With testicular cancer data, robust finite value of the TNEH was estimated. With women pancreatic cancer data, despite the fact that the excess hazard becomes almost zero around 10 years after diagnosis, the TNEH's baseline estimate equals the corresponding upper bound specified for the optimization procedure, even for a relatively large value of the upper bound (40 years). In such situations, one can conclude at least that the TNEH is greater than a predefined time T^* having practical usefulness. Hence treating the TNEH as a parameter would offer a way to test a cure hypothesis. Of course, by testing the hypotheses "TNEH $\leqslant T^*$" versus "TNEH $> T^*$", one could conclude at least whether there is evidence of cure before T^* or not.

Due to its similarity with a Beta probability distribution function, the new model could reproduce various excess hazard curves, offers a simple and comprehensible way to handle the TNEH as a model parameter and leads to satisfactory estimates as shown by the numerical studies. However, in some cases, it would not be flexible enough to capture some shapes for instance in situation where the excess hazard shows several local extrema before reaching zero. Thus the new paradigm should be adapted to more flexible cure models to account for a larger panel of hazard curves. For instance in the "flexible" NMCM model proposed by Andersson et al. (2011), the baseline cumulative excess hazard is modeled by a restricted cubic spline of the time since diagnosis. Instead of fixing arbitrary the last knot of the spline after the last observed event time as they did, one could consider this knot as a parameter corresponding to the TNEH. However, treating a spline knot as a model parameter to be estimated is a topic that would need more research.

For model (6) specification purpose we had considered the case where both the two shape parameters α and β were linked to covariates as well as the case where only β was linked to covariates. Simulations led to unsatisfactory parameter estimates for these cases (especially

when both α and β were linked to the same continuous covariates), probably due to numerical problems and/or identifiability issue. This kind of problem seemed inherent to cure models. Indeed, Li, Taylor and Sy (2001) stated that even if a cure model is formally identifiable, a possible near non-identifiability situation could occur as a flat or irregular likelihood surface for finite samples, with associated numerical problems. Farewell (1982) noted in parametric cure model a high correlation between the parameter estimates of the incidence part of the model and those of the latency part of the model, what we thought, could also lead to numerical problems. More generally, Hanin and Huang (2014) pointed that sharing of covariates between various components of cure models may prevent identifiability and then good parameter estimates.

Finally we would like to point out that in the beta–TNEH model, the TNEH is assumed to be deterministic, which means that the model estimates the same value of TNEH for subjects sharing the same characteristics (covariates). A way to improve our model would be to consider the TNEH as a random effect with a specified common distribution depending on covariates. This is an ongoing work.

Acknowledgements

The authors like to thank the French network of cancer registries (FRANCIM) for providing the two real data sets of cancers, Professor Debashis Ghosh (Co-Editor), as well as the Associate Editor and the three anonymous Referees for their constructive comments and helpful suggestions. This work was supported by the Institut National du Cancer (INCa) [grant number 2014–087], by the Fondation ARC pour la recherche sur le cancer [personal grant for author O.B. number PDF20151203665] and by the French Government (Agence Nationale de la Recherche, "Investissements d'Avenir", ANR–11-LABX–0021).

Supporting Information

Web Appendix A, referenced in Section 2, Web Appendix B, C, D and E referenced in Section 3 and Web Appendix F, referenced in Section 4 are available with this paper at the Biometrics website on Wiley Online Library.

References

Andersen, P. K., Borgan O., Gill, R. D., and Keiding, N. (2012). *Statistical Models Based on Counting Processes*. Springer Science & Business Media.

Andersson, T. M. L., Dickman, P. W., Eloranta, S., and Lambert, P. C. (2011). Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC Med Res Methodol* 11, 96.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J R Stat Soc Series B* 11, 15–53.

Boussari, O., Romain, G., Remontet, L., Bossard, N., Mounier, M., Bouvier, A.-M., Binquet, C., Colonna, M., and Jooste, V. (2018). A new approach to estimate time–to–cure from cancer registries data. *Cancer epidemiol* 53, 72–80.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J Sci Comput* 16, 1190–1208.

Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *J Am Stat Assoc* 94, 909–919.

Cooner, F., Banerjee, S., Carlin, B. P., and Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes. *J Am Stat Assoc* 102, 560–572.

Cowppli-Bony, A., Uhry, Z., Remontet, L., Guizard, A.-V., Voirin, N., Monnereau, A., et al. (2016). *Survie des Personnes Atteintes de Cancer en France Métropolitaine, 1989–2013. partie 1 – Tumeurs Solides*. Saint-maurice: Institut de veille sanitaire.

Dal Maso, L., Guzzinati, S., Buzzoni, C., Capocaccia, R., Serraino, D., Caldarella, A., et al. (2014). Long-term survival, prevalence, and cure of cancer: a population-based estimation for 818,902 italian patients and 26 cancer types. *Ann Oncol* 25, 2251–2260.

Farewell, V. T. (1982). The use of a mixture model for the analysis of survival data with long-term survivors. *Biometrics* 38, 1041–1046.

Gu, Y., Sinha, D., and Banerjee S. (2011). Analysis of cure rate survival data under proportional odds model. *Lifetime Data Anal* 17, 123–134.

Hanin, L. and Huang, L. S. (2014). Identifiability of cure models revisited. *J Multivar Anal* 130, 261–274.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., and Scheike T. H. (Eds.). (2016). *Handbook of Survival Analysis*. CRC Press.

Kuk, A. Y. C., and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* 79, 531–541.

Lambert, P. C., Thompson, J. R., Weston, C. L., and Dickman P. W. (2006). Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics* 8, 576–594.

Lehmann, E. L., and Casella, G. (1998). *Theory of point estimation*. 2nd Edition. New York: Springer-Verlag.

Lewis, A. S., and Overton, M. L. (2009). Nonsmooth optimization via bfgs. URL `https://cs.nyu.edu/overton/papers/pdffiles/bfgsinexactLS.pdf`.

Li, C.-S., and Taylor. J. M. G. (2002). Smoothing covariate effects in cure models. *Commun Stat Theory Methods* 31, 477–493.

Maller, R. A., and Zhou, X. (1996). *Survival Analysis with Long Term Survivors*. John Wiley and sons.

Newey, W. K., and McFadden, D. (1994). Large sample estimation and hypothesis testing.

Handbook of econometrics 4, 2111–2245.

Peng, Y. and Xu, J. (2012). An extended cure model and model selection. *Lifetime Data Anal* 18, 215–233.

Perme, M. P., Stare, J., and Estève, J. (2012). On estimation in relative survival. *Biometrics* 68, 113–120.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Taylor, J. M. G., and Liu, N. (2007). Statistical issues involved with extending standard models. In *Advances In Statistical Modeling And Inference: Essays in Honor of Kjell A. Doksum* 299–311. World Scientific.

Tsodikov, A. D., Ibrahim, J. G., and Yakovlev. A. Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *J Am Stat Assoc* 98, 1063–1078.

Verdecchia, A., De Angelis, R., Capocaccia, R., Sant, M., Micheli, A., Gatta, G., and Berrino F. (1998). The cure for colon cancer: results from the eurocare study. *Int J Cancer* 77, 322–329.

Yakovlev, A. Y., Tsodikov, A. D., and Asselain, B. (1996). *Stochastic models of tumor latency and their biostatistical applications* (Vol. 1). World Scientific.

Yin G., and Ibrahim, J. G. (2005a). Cure rate models: a unified approach. *Can J Stat* 33, 559–570.

Yin G., and Ibrahim, J. G. (2005b). A general class of bayesian survival models with zero and nonzero cure fractions. *Biometrics* 61, 403–412.

Youth Cancer Europe (2018). The right to be Forgotten for Cancer Survivors. URL `https://www.youthcancereurope.org/youth-cancer-news-events/2018/08/event-the-right-to-be-forgotten-for-cancer-survivors/` [accessed 24 August

2019]

Yu, B., Tiwari, R. C., Cronin, K. A., McDonald, C., and Feuer, E. J. (2005). Cansurv: a windows program for population-based cancer survival analysis. *Comput Methods Programs Biomed* 80, 195–203.

Zeng, D., Yin, G., and Ibrahim, J.G. (2006). Semiparametric transformation models for survival data with a cure fraction. *J Am Stat Assoc* 101, 670–684.

Zhang, J., and Peng, Y. (2009). Accelerated hazards mixture cure model. *Lifetime Data anal* 15, 455–467.

Received Month YEAR. *Revised Month* YEAR. *Accepted Month* YEAR.

Appendix

Model identifiability

Let us note $P_{\boldsymbol{\theta}}$ the probability distribution of an observation (T, Δ, Z) under the parametric model (6) where for simplicity we consider that the age at diagnosis A is a component of the covariate vector Z. According to Lehmann and Casella (1998) the model identifiability holds if $\boldsymbol{\theta} \mapsto P_{\boldsymbol{\theta}}$ is one–to–one on the parameter space Θ. Let us denote by $\mathscr{Z}$ the set of values taken by the covariates. It is straightforward to verify that identifiability of model (6) is equivalent to the identifiability of the class of functions

$$\mathscr{H} = \left\{(t, \boldsymbol{z}) \mapsto \lambda_{\text{exc}}(t|\boldsymbol{z}; \boldsymbol{\theta}) : [0, +\infty) \times \mathscr{Z} \to [0, +\infty); \boldsymbol{\theta} = (\boldsymbol{\gamma}, \beta, \boldsymbol{\eta}) \in \Theta_\alpha \times [0, +\infty) \times \Theta_\tau \equiv \Theta\right\}$$

that is $\lambda_{\text{exc}}(t|\boldsymbol{z}; \boldsymbol{\theta}) = \lambda_{\text{exc}}(t|\boldsymbol{z}; \boldsymbol{\theta}^*)$ for all $(t, \boldsymbol{z})$ in $[0, +\infty) \times \mathscr{Z}$ implies $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ for all $(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \in \Theta^2$. Let us introduce two additional classes of functions:

- $\mathscr{A} = \{\boldsymbol{z} \mapsto \alpha(\boldsymbol{z}; \boldsymbol{\gamma}) : \mathscr{Z} \to (0, +\infty); \boldsymbol{\gamma} \in \Theta_\alpha\}$;
- $\mathscr{T} = \{\boldsymbol{z} \mapsto \tau(\boldsymbol{z}; \boldsymbol{\eta}) : \mathscr{Z} \to (0, +\infty); \boldsymbol{\eta} \in \Theta_\tau\}$.

Proposition 1 *If the classes of functions $\mathscr{A}$ and $\mathscr{T}$ are identifiable, then model* (6) *is identifiable.*

Proof. We ever explained that it is sufficient to verify that $\mathscr{H}$ is identifiable. Let us consider $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \beta, \boldsymbol{\eta}) \in \Theta$ and $\boldsymbol{\theta}^* = (\boldsymbol{\gamma}^*, \beta^*, \boldsymbol{\eta}^*) \in \Theta$ such that $\lambda_{\text{exc}}(t|\boldsymbol{z}; \boldsymbol{\theta}) = \lambda_{\text{exc}}(t|\boldsymbol{z}; \boldsymbol{\theta}^*)$ for all $(t, \boldsymbol{z})$ in $[0, +\infty) \times \mathscr{Z}$. The supports of $\lambda_{\text{exc}}(\cdot|\boldsymbol{z}; \boldsymbol{\theta})$ and $\lambda_{\text{exc}}(\cdot|\boldsymbol{z}; \boldsymbol{\theta}^*)$ having to match for all $\boldsymbol{z} \in \mathscr{Z}$ we have $\tau(\boldsymbol{z}; \boldsymbol{\eta}) = \tau(\boldsymbol{z}; \boldsymbol{\eta}^*)$ for all $\boldsymbol{z} \in \mathscr{Z}$, thus $\boldsymbol{\eta} = \boldsymbol{\eta}^*$ since $\mathscr{T}$ is identifiable. As a consequence, for all $\boldsymbol{z} \in \mathscr{Z}$ and $t \in (0, \tau(\boldsymbol{z}; \boldsymbol{\eta}))$ we have

$$\left\{\frac{t}{\tau(\boldsymbol{z}; \boldsymbol{\eta})}\right\}^{\alpha(\boldsymbol{z}; \boldsymbol{\gamma})-1} \times \left\{1 - \frac{t}{\tau(\boldsymbol{z}; \boldsymbol{\eta})}\right\}^{\beta-1} = \left\{\frac{t}{\tau(\boldsymbol{z}; \boldsymbol{\eta})}\right\}^{\alpha(\boldsymbol{z}; \boldsymbol{\gamma}^*)-1} \times \left\{1 - \frac{t}{\tau(\boldsymbol{z}; \boldsymbol{\eta})}\right\}^{\beta^*-1}.$$

Taking the logarithm of the above equality and using the linear independence of functions $t \mapsto \log\left\{\frac{t}{\tau(\boldsymbol{z}; \boldsymbol{\eta})}\right\}$ and $t \mapsto \log\left\{1 - \frac{t}{\tau(\boldsymbol{z}; \boldsymbol{\eta})}\right\}$ we obtain $\beta = \beta^*$ and $\alpha(\boldsymbol{z}; \boldsymbol{\gamma}) = \alpha(\boldsymbol{z}; \boldsymbol{\gamma}^*)$ for all $\boldsymbol{z} \in \mathscr{Z}$. The later result with the identifiability of the class of functions $\mathscr{A}$ leads to $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$. This finishes the proof. □

It is easy to check that classes of functions $\mathscr{A}$ and $\mathscr{T}$ that we used for both the simulation study in Section 3 and the illustrations on real dataset in Section 4 are identifiable.

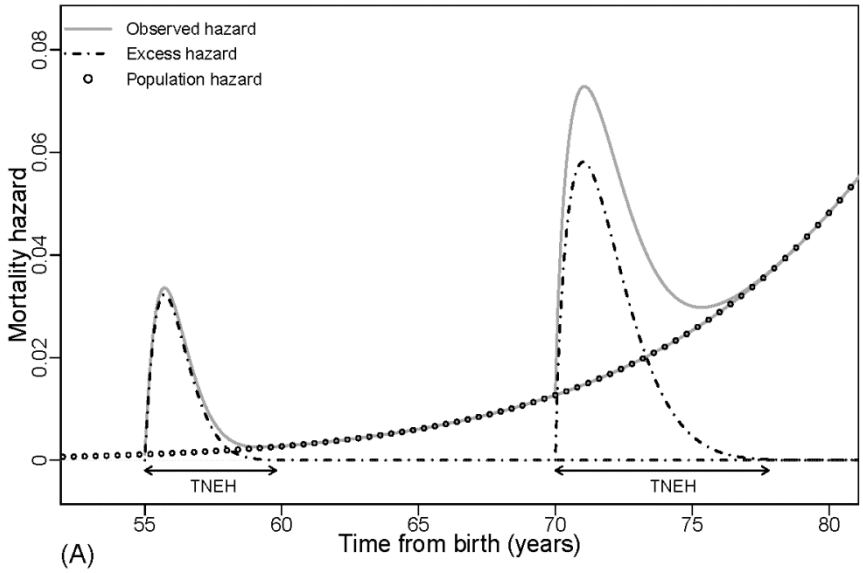


Figure 1. Illustrative plot of Observed, Excess and Population hazards functions for an individual diagnosed at 55, 70 years.

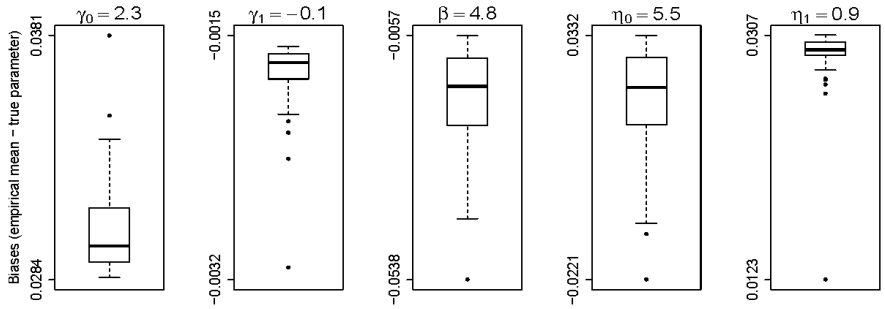


Figure 2. Distributions, based on 1000 simulated samples of size 2000, of the biases when using $K = 30$ initial values in the optimization procedure.

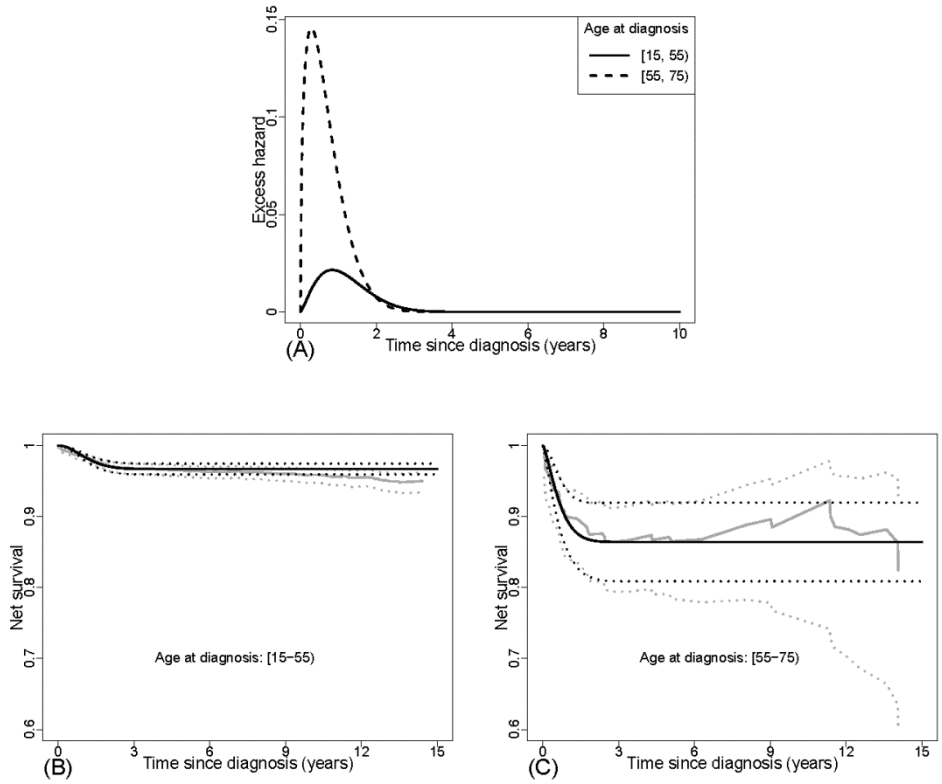


Figure 3. Testicular cancer: estimated Excess hazards function (A) by the beta–TNEH model, and estimated Net survivals functions with their estimated confidence bounds (B and C) from both the PP model (gray lines) and the beta–TNEH model (black lines).

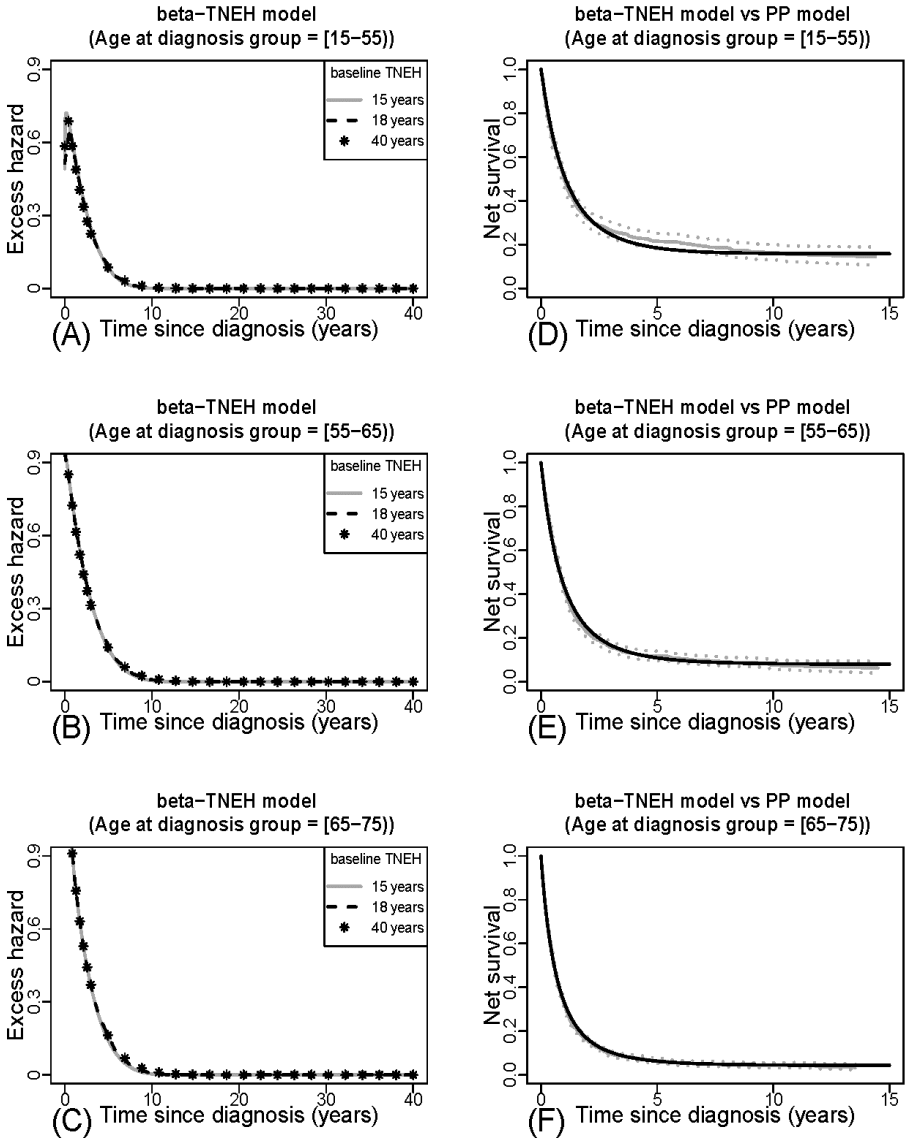


Figure 4. Women pancreatic cancer: on the left, the beta–TNEH model estimations of the excess hazards (panels A, B and C, one panel for each age at diagnosis group) when the TNEH's baseline estimate equals its upper bound fixed to 15 (solid gray line), 18 (dashed black lines) and 40 years (asterisk). On the right, the net survivals estimations for each age at diagnosis group (panels D, E and F, one panel for each age at diagnosis group) by the PP model with confidence bounds (solid gray and dotted gray lines) and by the beta–TNEH model (solid black lines) where the TNEH's baseline estimate equals 40 years.

Table 1

MLE performances for various sample sizes based on 1000 simulated samples: mean is the empirical mean, sd is the empirical standard deviation, $\overline{se^*}$ is the mean of the standard errors estimates and cp is the 95% coverage probability. The censoring rates are about 60% (Setting 1), 20% (Setting 2) and 46% (Setting 3).

Setting 1	n	*indicators*	$\gamma_0 = 2.3$	$\gamma_1 = -0.1$	$\beta = 4.8$	$\eta_0 = 5.5$	$\eta_1 = 0.9$
	250	*mean*	2.688	−0.122	4.249	5.084	1.031
		sd	0.651	0.325	1.941	2.406	1.264
		$\overline{se^*}$	0.618	0.329	2.674	3.143	1.332
		cp	0.965	0.988	0.722	0.730	0.858
	500	*mean*	2.457	−0.109	4.595	5.388	0.995
		sd	0.341	0.170	1.547	1.900	0.866
		$\overline{se^*}$	0.331	0.167	1.747	2.142	0.878
		cp	0.959	0.978	0.818	0.808	0.901
	1000	*mean*	2.375	−0.105	4.733	5.487	0.940
		sd	0.214	0.110	1.168	1.488	0.629
		$\overline{se^*}$	0.211	0.110	1.141	1.400	0.600
		cp	0.949	0.961	0.861	0.864	0.915
	2000	*mean*	2.333	−0.100	4.756	5.481	0.918
		sd	0.136	0.070	0.723	0.907	0.411
		$\overline{se^*}$	0.140	0.070	0.748	0.907	0.400
		cp	0.964	0.956	0.920	0.910	0.935
Setting 2	n	*indicators*	$\gamma_0 = 1.25$	$\gamma_1 = -0.05$	$\beta = 3.5$	$\eta_0 = 9$	$\eta_1 = 0.3$
	250	*mean*	1.260	−0.052	3.643	9.634	0.447
		sd	0.052	0.038	1.377	3.913	1.080
		$\overline{se^*}$	0.055	0.039	1.715	4.928	1.201
		cp	0.956	0.949	0.851	0.848	0.965
	500	*mean*	1.258	−0.052	3.466	9.027	0.320
		sd	0.035	0.025	0.831	2.308	0.664
		$\overline{se^*}$	0.037	0.027	0.914	2.557	0.704
		cp	0.952	0.959	0.870	0.859	0.961
	1000	*mean*	1.254	−0.051	3.517	9.110	0.331
		sd	0.026	0.019	0.601	1.716	0.468
		$\overline{se^*}$	0.026	0.019	0.620	1.730	0.474
		cp	0.947	0.940	0.905	0.901	0.955
	2000	*mean*	1.252	−0.050	3.487	8.992	0.322
		sd	0.018	0.013	0.419	1.160	0.319
		$\overline{se^*}$	0.018	0.013	0.415	1.145	0.319
		cp	0.953	0.946	0.926	0.921	0.957
Setting 3	n	*indicators*	$\gamma_0 = 3.01$	$\gamma_1 = -0.20$	$\beta = 2.98$	$\eta_0 = 18$	$\eta_1 = 1.2$
	250	*mean*	3.198	−0.196	2.843	17.350	1.125
		sd	0.390	0.259	0.454	3.513	2.360
		$\overline{se^*}$	0.359	0.249	0.434	3.300	2.338
		cp	0.952	0.952	0.855	0.829	0.909
	500	*mean*	3.081	−0.202	2.925	17.854	1.265
		sd	0.257	0.154	0.329	2.623	1.788
		$\overline{se^*}$	0.228	0.158	0.301	2.341	1.659
		cp	0.931	0.964	0.880	0.864	0.918
	1000	*mean*	3.046	−0.203	2.943	17.909	1.259
		sd	0.162	0.110	0.202	1.664	1.1490
		$\overline{se^*}$	0.155	0.108	0.204	1.602	1.138
		cp	0.944	0.946	0.921	0.927	0.939
	2000	*mean*	3.021	−0.205	2.967	18.005	1.268
		sd	0.109	0.076	0.148	1.177	0.792
		$\overline{se^*}$	0.106	0.075	0.142	1.135	0.804
		cp	0.949	0.945	0.923	0.926	0.950

Table 2
Parameter estimates with standard errors in brackets and log-likelihood (LL) when the upper bound of η_0 (the baseline TNEH) is fixed to 15, 18 and 40 years.

$\eta_{0,max}$	LL	$\widehat{\gamma_0}$	$\widehat{\gamma_1}$	$\widehat{\gamma_2}$	$\widehat{\beta}$	$\widehat{\eta_0}$	$\widehat{\eta_1}$	$\widehat{\eta_2}$
15	−3741.758	0.931	0.074	0.128	6.005	15.000	0.259	−1.959
		(0.007)	(0.013)	(0.018)	(1.093)	(2.690)	(0.962)	(1.039)
18	−3737.123	0.933	0.071	0.121	7.156	18.000	0.282	−2.608
		(0.007)	(0.013)	(0.018)	(1.712)	(4.243)	(1.208)	(1.390)
40	−3727.847	0.941	0.058	0.100	15.650	40.000	−0.535	−6.850
		(0.008)	(0.013)	(0.021)	(10.848)	(27.234)	(2.930)	(5.523)

Supporting Information for

Modeling excess hazard with time–to–cure as a parameter

by Boussari et al.

Web Appendix A: boundaries specified in the estimation procedure when using the L-BFGS-B method

$\boldsymbol{\theta}_{up}$ denotes the vector of upper bounds, $\boldsymbol{\theta}_{lo}$ the vector of lower bounds, $\boldsymbol{\theta}$ the vector of true parameters and $\widehat{\boldsymbol{\theta}}$ the vector of estimated parameters.

Simulations

Setting 1	$\boldsymbol{\theta}$	$\gamma_0 = 2.3$	$\gamma_1 = -0.1$	$\beta = 4.8$	$\eta_0 = 5.5$	$\eta_1 = 0.9$
	$\boldsymbol{\theta}_{lo}$	0.4	-3	1	1	-5
	$\boldsymbol{\theta}_{up}$	8	3	12	20	5
Setting 2	$\boldsymbol{\theta}$	$\gamma_0 = 1.25$	$\gamma_1 = -0.05$	$\beta = 3.5$	$\eta_0 = 9$	$\eta_1 = 0.3$
	$\boldsymbol{\theta}_{lo}$	0.4	-3	1	1	-5
	$\boldsymbol{\theta}_{up}$	8	3	12	20	5

Testicular cancer

$\widehat{\boldsymbol{\theta}}$	$\widehat{\gamma}_0 = 2.39$	$\widehat{\gamma}_1 = -0.87$	$\widehat{\beta} = 8.4$	$\widehat{\eta}_0 = 5.31$	$\widehat{\eta}_1 = -0.91$
$\boldsymbol{\theta}_{lo}$	0.4	-3	1	0.1	-5
$\boldsymbol{\theta}_{up}$	6.4	3	10	15	5

Women pancreatic cancer

$\widehat{\boldsymbol{\theta}}$	$\widehat{\gamma}_0 = 0.931$	$\widehat{\gamma}_1 = 0.074$	$\widehat{\gamma}_2 = 0.128$	$\widehat{\beta} = 6.005$	$\widehat{\eta}_0 = 15$	$\widehat{\eta}_1 = 0.259$	$\widehat{\eta}_2 = -1.959$
$\boldsymbol{\theta}_{lo}$	0.4	-5	-5	1	0.1	-10	-10
$\boldsymbol{\theta}_{up}$	10	5	5	40	15	10	10
$\widehat{\boldsymbol{\theta}}$	$\widehat{\gamma}_0 = 0.933$	$\widehat{\gamma}_1 = 0.071$	$\widehat{\gamma}_2 = 0.121$	$\widehat{\beta} = 7.156$	$\widehat{\eta}_0 = 18$	$\widehat{\eta}_1 = 0.282$	$\widehat{\eta}_2 = -2.608$
$\boldsymbol{\theta}_{lo}$	0.4	-5	-5	1	0.1	-10	-10
$\boldsymbol{\theta}_{up}$	10	5	5	40	18	10	10
$\widehat{\boldsymbol{\theta}}$	$\widehat{\gamma}_0 = 0.941$	$\widehat{\gamma}_1 = 0.058$	$\widehat{\gamma}_2 = 0.100$	$\widehat{\beta} = 15.650$	$\widehat{\eta}_0 = 40$	$\widehat{\eta}_1 = -0.535$	$\widehat{\eta}_2 = -6.850$
$\boldsymbol{\theta}_{lo}$	0.4	-5	-5	1	0.1	-10	-10
$\boldsymbol{\theta}_{up}$	10	5	5	40	40	10	10

Web Appendix B: data simulation algorithm

In the following $\mathcal{U}[a,b]$ denotes the uniform distribution on $[a,b]$ and $\mathcal{B}(p)$ denotes the Bernoulli distribution with parameter p.

(a) Generate n ($n \in \mathbb{N}$) realizations of age at diagnosis A $(a_1,\ldots,a_n)$ and covariates $\boldsymbol{Z}$ $(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n)$.

(b) Generate n realizations of the lifetime X_1 $(x_{11},\ldots,x_{1n})$ elapsed between the diagnosis and the death due to cancer: draw n realizations of U $(u_1,\ldots,u_n)$ with $U \sim \mathcal{U}[0,1]$, and set $x_{1i} = \begin{cases} \tau(\boldsymbol{z}_i;\boldsymbol{\eta})\mathrm{F}_{\mathcal{B}e}^{-1}\left(\frac{\log(u_i)}{\log\{\pi(\boldsymbol{z}_i;\boldsymbol{\theta})\}};\alpha(\boldsymbol{z}_i;\boldsymbol{\gamma}),\beta\right) & \text{if } u_i \geq \pi(\boldsymbol{z}_i;\boldsymbol{\theta}) \\ \infty & \text{if } u_i < \pi(\boldsymbol{z}_i;\boldsymbol{\theta}) \end{cases}$, for $1 \leq i \leq n$.

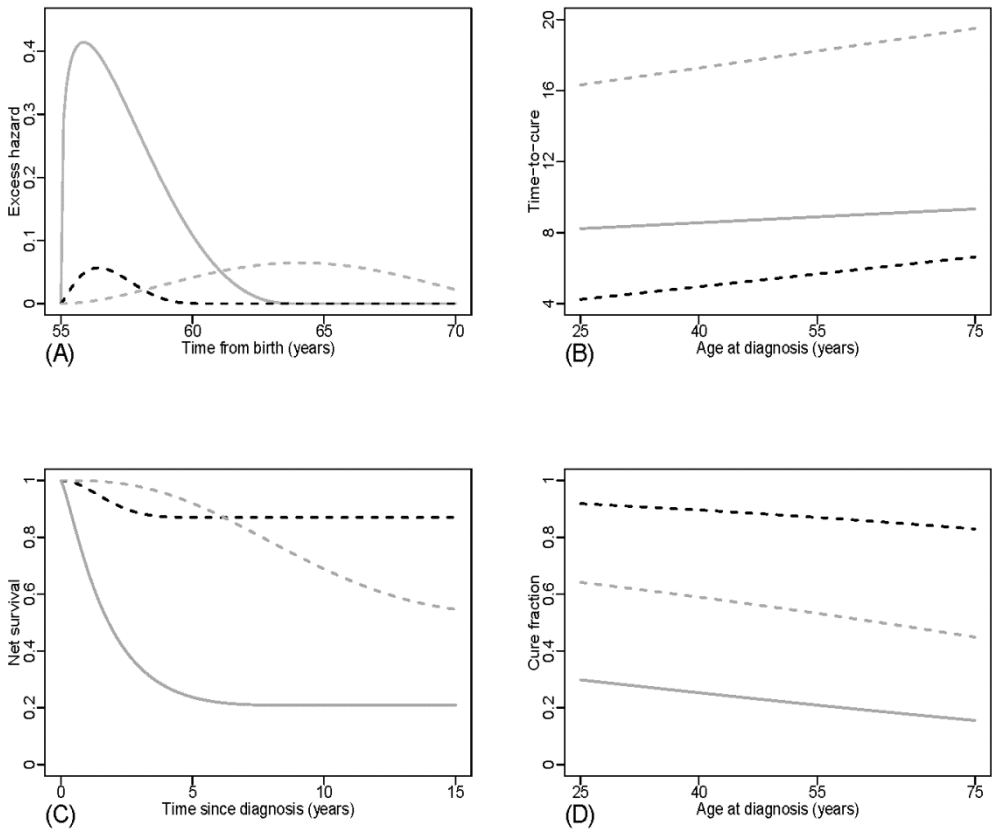
(c) Generate n realizations of the lifetime X_2 $(x_{21},\ldots,x_{2n})$ elapsed between the diagnosis and the death due to other causes: draw n realizations of U $(u_1^*,\ldots,u_n^*)$ with $U \sim \mathcal{U}[0,1]$, and set $x_{2i} = \Lambda_{pop}^{-1}\left[\log\left\{\frac{u_i^*}{S_{\text{pop}}(a_i)}\right\}\right] - a_i$, for $1 \leq i \leq n$.

(d) Generate n realizations of C_1 $(c_{11},\ldots,c_{1n})$ and n realizations of C_2 $(c_{21},\ldots,c_{2n})$ such that $C_j = P_j \times U_j + (1-P_j) \times t_{max}$ for $j \in \{1,2\}$ where t_{max} denotes a predefined maximum follow-up time, $P_j \sim \mathcal{B}(p_j)$ and $U_j \sim \mathcal{U}[0,t_{max}]$.

Note that C_1 describes censoring times corresponding to a mixture of a proportion p_1 of patients who are still alive at the endpoint of the follow-up and are followed less than t_{max}, and a proportion $1-p_1$ of patients followed more than t_{max} before the endpoint of the follow-up. C_2 describes censoring times corresponding to a mixture of a proportion p_2 of patients who are lost to follow-up, and a proportion $1-p_2$ of patients followed more than t_{max} before the endpoint of the follow-up. Considering that the censoring time C results from the competing times C_1 and C_2, the realizations $(c_1,\ldots,c_n)$ of the censoring time C are then equal to $c_i = \min(c_{1i},c_{2i})$ for $1 \leq i \leq n$.

(e) Finally the realizations of the survival time T $(t_1,\ldots,t_n)$ and the censoring indicator Δ $(\delta_1,\ldots,\delta_n)$ are obtained by setting $t_i = \min(x_{1i},x_{2i},c_i)$ and $\delta_i = \mathbb{1}_{\{\min(x_{1i},x_{2i})<c_i\}}$, for $1 \leq i \leq n$.

Web Appendix C: graphical illustration of the simulation settings



Excess hazards (A) and Net survivals (C) functions for a patient diagnosed at 55 years; Time–to–null–excess–hazard (B) and Cure fraction (D) according to age at diagnosis. The vector of true parameters are $\boldsymbol{\theta} = (2.3, -0.1, 4.8, 5.5, 0.9)$, $\boldsymbol{\theta} = (1.25, -0.05, 3.5, 9, 0.3)$ and $\boldsymbol{\theta} = (3.01, -0.2, 2.98, 18, 1.2)$ corresponding respectively to the simulation setting 1 (dashed black lines), the simulation setting 2 (grey lines) and the simulation setting 3 (dashed grey lines).

Web Appendix D: Performances of a multivariates beta–TNEH model

We consider the beta–TNEH model (equation (6) in the paper) where λ_{exc} depending on three covariates: age at diagnosis (continuous), sex, stage of cancer (3 stages = I, II and III) with α depending on age and stage and τ depending on age and sex: $\alpha(\boldsymbol{z}_1;\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 \times a^* + \gamma_2 \times \mathbb{1}_{\{\text{stage = II}\}} + \gamma_3 \times \mathbb{1}_{\{\text{stage = III}\}}$ and $\tau(\boldsymbol{z}_2;\boldsymbol{\eta}) = \eta_0 + \eta_1 \times a^* + \eta_2 \times \mathbb{1}_{\{\text{sex = man}\}}$; where a^* is the age at diagnosis standardized using the mean and the standard deviation of its specified distribution, $\boldsymbol{z}_1$ = (age, stage), $\boldsymbol{z}_2$ = (age, sex), $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ and $\boldsymbol{\eta} = (\eta_0, \eta_1, \eta_2)$. Hence $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\eta})$ is the vector of parameters to be estimated.

Here are some considerations for data generation: The vector of true parameters is $\boldsymbol{\theta}$ = (1.6, –0.02, –0.2, –0.5, 5.4, 7.5, 1.2, 1.5). The age at diagnosis is uniformly distributed on intervals $[20, 40)$, $[40, 65)$ and $[65, 80]$, and the proportions of age at diagnosis coming from these three intervals are 0.25, 0.25 and 0.50 respectively; 60% of cases are men; the proportions of stage I, II and III are 0.50, 0.30 and 0.20. The population hazard is assumed to follow a Weibull distribution, the scale and shape parameters being 75 and 11 respectively. The maximum follow-up time (from diagnosis) is fixed to fifteen years and the censoring rate is about 35%. We generate 1000 repetitions of each sample of size $n = \{250, 500, 1000, 2000\}$.

Table:

n	ind.	$\gamma_0 = 1.6$	$\gamma_1 = -0.02$	$\gamma_2 = -0.2$	$\gamma_3 = -0.5$	$\beta = 5.4$	$\eta_0 = 7.5$	$\eta_1 = 1.2$	$\eta_2 = 1.5$
250	*mean*	1.689	−0.023	−0.236	−0.572	4.967	7.081	1.296	1.490
	sd	0.191	0.056	0.175	0.192	2.044	3.201	1.209	1.498
	$\overline{se^*}$	0.190	0.056	0.161	0.185	2.965	4.621	1.433	1.866
	cp	0.9680	0.966	0.963	0.967	0.778	0.804	0.873	0.931
500	*mean*	1.647	−0.022	−0.221	−0.540	5.167	7.259	1.227	1.504
	sd	0.118	0.036	0.104	0.117	1.642	2.533	0.873	1.127
	$\overline{se^*}$	0.124	0.037	0.104	0.121	1.994	3.059	0.929	1.2409
	cp	0.964	0.964	0.955	0.955	0.838	0.818	0.887	0.947
1000	*mean*	1.622	−0.022	−0.209	−0.517	5.317	7.477	1.238	1.460
	sd	0.081	0.025	0.068	0.079	1.267	1.994	0.672	0.819
	$\overline{se^*}$	0.083	0.025	0.070	0.082	1.365	2.104	0.649	0.842
	cp	0.954	0.945	0.964	0.960	0.878	0.891	0.920	0.944
2000	*mean*	1.608	−0.020	−0.202	−0.506	5.394	7.534	1.241	1.496
	sd	0.056	0.017	0.048	0.055	0.896	1.386	0.446	0.573
	$\overline{se^*}$	0.057	0.017	0.048	0.056	0.930	1.428	0.445	0.579
	cp	0.947	0.955	0.955	0.950	0.922	0.924	0.935	0.956

Web Appendix E: Performances of the beta–TNEH model on data simulated from a Weibull mixture cure model

1. Times to death due to the disease are generated from a Weibull mixture cure model; i.e. the net survival is written as follows:

$$S_{\text{net}}(t;\mathbf{\Phi}) = \pi + (1-\pi)\exp\left\{-\left(\frac{t}{a}\right)^{b}\right\}$$

where t is the time since diagnosis, $\mathbf{\Phi} = (\pi, a, b)$ is the vector of parameters with π the cure fraction.

The population hazard (expected hazard) is assumed to follow a Weibull distribution, the scale and shape parameters being 75 and 11 respectively. The age at diagnosis is uniformly distributed on intervals $[15, 58)$, $[58, 65)$, $[65, 70)$ and $[70, 75]$, with the same proportion (0.25) of the observations coming from each interval. The maximum follow-up time (from diagnosis) is fixed to fifteen years and we generate 1000 repetitions of sample of size 2000. We consider two sets of parameters corresponding to two cases:

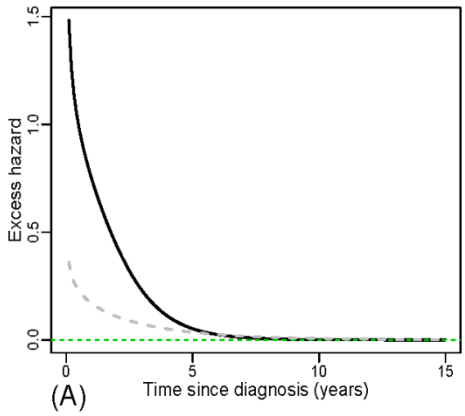
Case 1: $\mathbf{\Phi} = (0.10, 0.70, 0.80)$

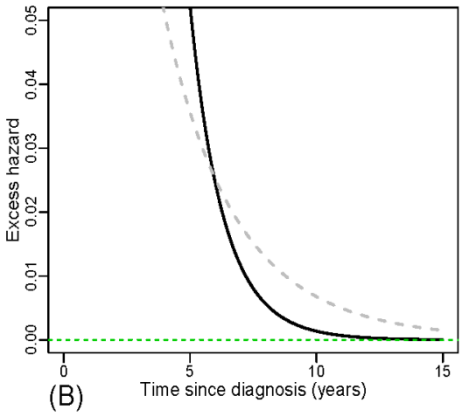
The excess hazard is high just after the diagnosis and decreases to become very close to zero within the follow-up interval. The censoring rate is about 7%.

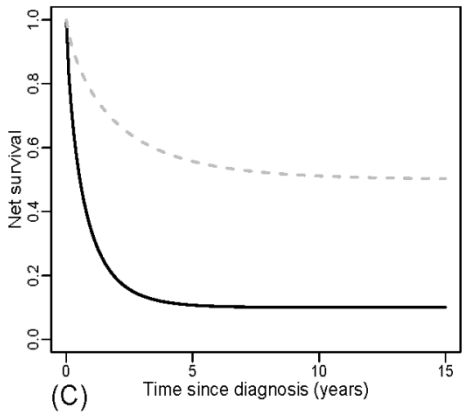
Case 2: $\mathbf{\Phi} = (0.50, 1.90, 0.80)$

The excess hazard is low excess hazard just after the diagnosis and decreases but remains less close to zero than the case 1, within the follow-up interval. The censoring rate is about 26%.

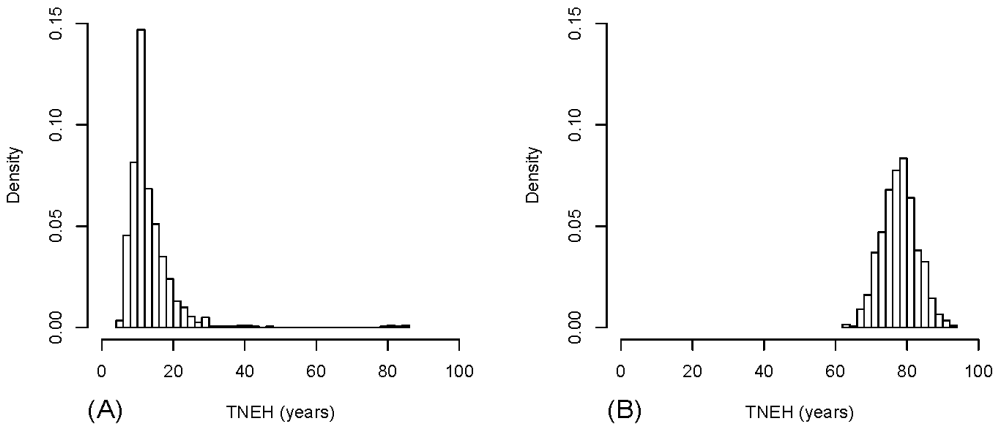
2. We fit the beta–TNEH model (Equation 6 in the paper) without covariate to the simulated data and assess its performances by computing the bias, the root mean square error and the coverage probability for the cure fraction, and the net survival at times $t = 5$, 10 and 15 years.







Excess hazards (A), excess hazards with focus on very low values (B) and net survivals (C) functions from the Weibull mixture cure model. The vector of true parameters are $\mathbf{\Phi} = (0.10, 0.70, 0.80)$ and $\mathbf{\Phi} = (0.50, 1.90, 0.80)$ corresponding respectively to the simulation case 1 (black lines) and the simulation case 2 (dashed grey lines).



Distributions of the 1000 estimates of the TNEH when the beta–TNEH model is fitted for each of the 1000 samples of size 2000 generated from the Weibull mixture cure model: (A) corresponds to Case 1 and (B) corresponds to Case 2.

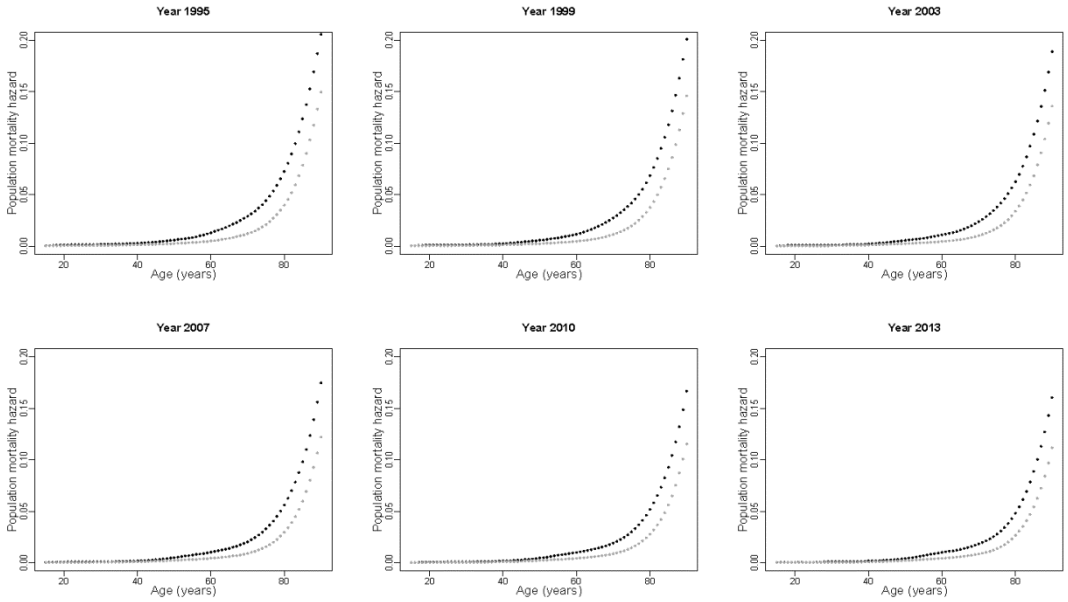
Table: Beta–TNEH model performances when fitted to data generated from a Weibull mixture cure model (1000 repetitions of samples of size 2000): "True" is the value from the true model which generate the data, "Bias" is the mean of the biases from the beta–TNEH estimates, "RMSE" is the root mean square error of the estimates and "cp" is the coverage probability.

Case 1		$S_{\text{net}}(5 years)$	$S_{\text{net}}(10 years)$	$S_{\text{net}}(15 years)$	*Cure rate*
	True	0.107256	0.100203	0.100008	0.100000
	Bias	−0.001039	0.001092	0.001268	0.001276
	RMSE	0.007982	0.008429	0.008461	0.008463
	cp	0.944000	0.941000	0.941000	0.941000
Case 2		$S_{\text{net}}(5 years)$	$S_{\text{net}}(10 years)$	$S_{\text{net}}(15 years)$	*Cure rate*
	True	0.557170	0.511460	0.502697	0.500000
	Bias	−0.025030	0.003064	0.011230	0.013911
	RMSE	0.028640	0.016508	0.019877	0.021511
	cp	0.583000	0.920000	0.855000	0.815000

Web Appendix F: Population mortality hazard in France

Expected mortality rates were derived from the observed mortality rates available by sex, annual age, year of death (1975 to 2012), Department of residence and provided by the Institut National de la Statistique et des Etudes Economiques. For a given sex and a given Department, these observed mortality rates were smoothed for ages above 15 using a Poisson regression model that included a bidimensional smoothing spline of year and age. Mortality rates were projected for the years 2013 to 2017 using this same model. Expected mortality rates were also derived for the whole France. This work has been done by the biostatistical unit of the Hospices Civils de Lyon, using *mgcv* package in R software.

Here for illustration purposes, we plot the population mortality hazard at years 1995, 1999, 2003, 2007, 2010 and 2013.



French population mortality hazard in men (dark dots) and in women (grey dots).

ANNEXE E : Résultats des 1000 échantillons de taille N=2000 simulés à partir du modèle TNEH

Tableau E-1 - Scénario de bon pronostic : Résultats des 1000 échantillons de taille N=2000 simulés à partir du modèle TNEH

	Biais			Erreur-type empirique			Erreur-type moyenne			Taux de couverture		
	MGM	MGF	TNEH	MGM	MGF	TNEH	MGM	MGF	TNEH	MGM	MGF	TNEH
<45 ans												
$S_n(3) = 0,897$	<0,001	0,014	<0,001	0,012	0,011	0,012	-0,013	0,011	0,012	96,0%	77,7%	95,6%
$S_n(5) = 0,897$	<0,001	0,003	<0,001	0,012	0,012	0,012	-0,013	0,012	0,012	96,0%	95,6%	95,6%
$S_n(10) = 0,897$	<0,001	-0,003	<0,001	0,012	0,013	0,012	-0,013	0,013	0,012	96,0%	93,8%	95,6%
$\pi = 0,897$	<0,001	-0,003	<0,001	0,012	0,013	0,012	0,013	0,013	0,012	96,0%	93,1%	95,6%
[45-59] ans												
$S_n(3) = 0,842$	<0,001	0,015	-0,001	0,013	0,012	0,013	-0,013	0,012	0,013	95,4%	78,7%	95,5%
$S_n(5) = 0,842$	-0,001	-0,001	-0,001	0,013	0,013	0,013	-0,013	0,013	0,013	95,6%	94,6%	95,6%
$S_n(10) = 0,842$	-0,001	-0,011	-0,001	0,013	0,014	0,013	-0,013	0,014	0,013	95,6%	85,8%	95,6%
$\pi = 0,842$	-0,001	-0,011	-0,001	0,013	0,014	0,013	0,013	0,014	0,013	95,7%	84,5%	95,6%
≥60 ans												
$S_n(3) = 0,808$	-0,001	0,011	-0,001	0,016	0,015	0,016	-0,017	0,016	0,017	96,0%	92,8%	96,4%
$S_n(5) = 0,796$	-0,002	0,002	-0,001	0,018	0,017	0,018	-0,019	0,018	0,019	95,2%	95,8%	95,2%
$S_n(10) = 0,796$	-0,002	-0,009	-0,002	0,018	0,018	0,018	-0,019	0,019	0,019	95,3%	93,5%	95,0%
$\pi = 0,796$	-0,002	-0,010	-0,002	0,018	0,018	0,018	0,019	0,019	0,019	95,4%	92,4%	95,0%

Tableau E-2 - Scénario de moyen pronostic : Résultats des 1000 échantillons de taille N=2000 simulés à partir du modèle TNEH

	Biais			Erreur-type empirique			Erreur-type moyenne			Taux de couverture		
	MGM	MGF	TNEH	MGM	MGF	TNEH	MGM	MGF	TNEH	MGM	MGF	TNEH
<60 ans												
$S_n(3) = 0,665$	0,001	0,024	0,015	0,019	0,018	0,018	-0,020	0,018	0,019	95,6%	76,5%	88.0%
$S_n(5) = 0,651$	0,001	0,006	0,009	0,020	0,019	0,019	-0,020	0,020	0,020	95,4%	94,5%	93.7%
$S_n(10) = 0,651$	<0,001	-0,010	0,006	0,020	0,020	0,019	-0,020	0,020	0,021	95,2%	90,7%	94.6%
$\pi = 0,651$	<0,001	-0,012	0,006	0,020	0,020	0,019	0,020	0,020	0,021	95,3%	90,0%	94.6%
[60-69] ans												
$S_n(3) = 0,619$	<0,001	0,016	0,012	0,017	0,017	0,017	-0,018	0,017	0,017	95,6%	85,3%	89.2%
$S_n(5) = 0,601$	<0,001	-0,001	0,003	0,019	0,018	0,019	-0,019	0,018	0,019	94,8%	95,2%	94.6%
$S_n(10) = 0,600$	-0,001	-0,018	-0,002	0,019	0,018	0,019	-0,019	0,019	0,020	94,7%	82,1%	95.1%
$\pi = 0,600$	-0,001	-0,020	-0,002	0,019	0,018	0,020	0,019	0,019	0,020	94,6%	79,7%	95.1%
≥70 ans												
$S_n(3) = 0,549$	-0,003	0,016	0,012	0,021	0,020	0,020	-0,022	0,021	0,020	95,5%	91,1%	90.9%
$S_n(5) = 0,506$	-0,001	0,020	0,003	0,024	0,021	0,024	-0,024	0,022	0,023	96,1%	88,5%	94.2%
$S_n(10) = 0,502$	-0,003	0,005	-0,009	0,025	0,022	0,027	-0,025	0,023	0,027	95,4%	95,7%	93.0%
$\pi = 0,502$	-0,003	0,003	-0,009	0,025	0,022	0,027	0,025	0,023	0,027	95,5%	96,2%	93.1%

Tableau E-3 - Scénario de mauvais pronostic : Résultats des 1000 échantillons de taille N=2000 simulés à partir du modèle TNEH

	Biais			Erreur-type empirique			Erreur-type moyenne			Taux de couverture		
θ_{TNEH}	MGM	MGF	TNEH	MGM	MGF	TNEH	MGM	MGF	TNEH	MGM	MGF	TNEH
<60 ans												
$S_n(3) = 0,226$	-0,001	0,013	0,011	0,016	0,015	0,015	-0,016	0,015	0,015	94,6%	86,6%	90,7%
$S_n(5) = 0,204$	0,001	0,006	0,008	0,016	0,015	0,015	-0,016	0,015	0,016	93,7%	92,4%	92,8%
$S_n(10) = 0,202$	<0,001	-0,007	0,005	0,016	0,014	0,016	-0,016	0,014	0,017	93,5%	92,5%	94,3%
$\pi = 0,202$	<0,001	-0,008	0,005	0,016	0,014	0,016	0,016	0,014	0,017	93,7%	91,9%	94,4%
[60-69] ans												
$S_n(3) = 0,183$	-0,003	0,013	0,005	0,013	0,013	0,013	-0,014	0,013	0,013	94,3%	84,2%	94,3%
$S_n(5) = 0,157$	<0,001	0,012	0,002	0,014	0,012	0,013	-0,014	0,012	0,014	95,1%	86,4%	95,6%
$S_n(10) = 0,154$	<0,001	0,002	-0,002	0,014	0,012	0,014	-0,014	0,012	0,015	95,1%	94,9%	95,5%
$\pi = 0,154$	-0,001	0,001	-0,003	0,014	0,012	0,014	0,014	0,012	0,015	95,4%	95,2%	95,5%
≥70 ans												
$S_n(3) = 0,138$	-0,004	0,022	0,002	0,016	0,014	0,014	-0,015	0,015	0,014	94,4%	68,9%	95,4%
$S_n(5) = 0,106$	0,001	0,029	-0,001	0,016	0,013	0,014	-0,016	0,014	0,015	94,9%	40,4%	96,3%
$S_n(10) = 0,100$	0,002	0,024	-0,006	0,016	0,013	0,015	-0,016	0,013	0,017	95,3%	53,3%	94,7%
$\pi = 0,100$	0,002	0,023	-0,006	0,016	0,013	0,015	0,016	0,013	0,017	94,9%	55,7%	94,7%