CHAPITRE 5
Gestion et contrôle de la base de données

Table des matières

CHAPITRE 5
Gestion et contrôle de la base de données

Base de population

La connaissance des dénominateurs, en particulier ceux des programmes de diagnostic et d'assistance destinés à toute la population présente (dépistage), est partie intégrante des contrôles épidémiologiques qui sont demandés aux registres. En outre, la divulgation d'enquêtes spécifiques sur des groupes de population à risque (personnes âgées, immigrés, catégories en difficulté) amène toujours plus le registre à se munir de compétences lui permettant de participer à ces études. Etant donnée la complexité (de laquelle il a déjà été question précédemment) des estimations en mesure de quantifier la population effectivement couverte par le système sanitaire et présente sur le territoire, par rapport à la simple donnée sur le lieu de résidence, il est utile que chaque registre contacte les institutions qui sur le territoire s'occupent d'évaluation démographique (communes, région).

Les registres des tumeurs font officiellement référence à la population résidente à l'intérieur du propre territoire pour les années couvertes par l'enregistrement. En accord avec les critères d'homogénéité établis par la banque de données nationale d'incidence, les données disponibles doivent être :

- sexe : 1 = homme; 2 = femme ;
- année de résidence : quatre chiffres, siècle et millénaire inclus ;
- classes d'âge : de préférence annuelle (voir plus bas) ;
- commune de résidence : code ISTAT à 6 chiffres en Italie ;
- nombre de sujets en fonction du sexe, de l'année, de l'âge, de la commune de résidence.

En ce qui concerne la classe d'âge, on recommande la séparation en année mais il est possible de fournir les classes d'intervalle non supérieur à cinq ans, en conservant de préférence la donnée pour la classe 0-1 an (ex. : 0; 1-4; 5-9; 10-14, etc.) ; les classes quinquennales doivent alors être au moins étendues à la classe "85 ans et plus", et n'importe quelle variation par rapport à ce schéma doit être signalée.

Sources des données

Les populations doivent être récupérées à partir des sources officielles, en suivant l'ordre de priorité suivant :

- ISTAT en Italie, INSEE (Institut National de la Statistique et des Etudes Economiques) en France ;
- source régionale ;
- source communale ;
- autres sources (ex. : estimations locales).

La source de données utilisée doit être précisée dans la section réservée aux méthodes.

Population résidente et population présente

Le choix de faire référence à la population officiellement résidente à l'intérieur de l'aire couverte par le registre est une règle à suivre pour garantir la qualité de l'identification démographique des cas. L'organisation de l'état civil des administrations communales garantit en effet que la donnée soit une donnée réelle. Le service sanitaire national assure la protection de la santé et l'assistance sanitaire à tous les citoyens, non seulement aux résidents mais aussi aux populations migrantes et aux personnes domiciliées à l'intérieur du territoire et inscrites sur le registre des agences de santé territoriales. Ces dernières années, cette tranche de population s'est progressivement accrue du fait de l'immigration de sujets (et familles) qui, bien qu'ils n'aient pas les pré-requis pour la reconnaissance de la nationalité, sont intégrés de façon stable au territoire duquel ils reçoivent l'assistance sanitaire. Cette population est intéressante du point de vue de l'épidémiologie en oncologie étant donné qu'elle présente souvent des niveaux de risque différents par rapport à la population d'origine. De plus, de par le principe d'équité dans le droit à la santé, ces groupes sociaux sont aussi destinataires des interventions de prévention.

A ce propos, la diffusion progressive sur le territoire national des dépistages oncologiques amène à considérer la pathologie incidente également dans les strates sociales récemment installées sur un territoire et souvent caractérisées par des risques individuels majeurs.

Pour le bon enregistrement de ces cas, il est donc important d'identifier la population de référence (**présente** et prise en charge, outre la population **résidente**) en évitant au maximum les erreurs de classification. Les communes, les agences de santé, les provinces et les régions fournissent généralement des données démographiques dont l'acquisition de la part des registres est utile déjà en phase de conception du projet de création du registre. Cette disponibilité des données doit inciter chaque registre à vérifier l'évolution des indicateurs démographiques au niveau local.

Indicateurs

De façon détaillée, les indices présentés sont :

Nombre de cas

Indique le nombre total de cas enregistrés :

$$N = \sum_i n_i$$

(n_i = nombre de cas par classe d'âge ; i = indice de la classe d'âge quinquennale)

Distribution proportionnelle

Indique la part en pourcentage des cas incidents pour chaque localisation tumorale par rapport au total des cas enregistrés.

Taux brut

Taux pour 100.000 habitants par an :

$$T = \frac{\sum_i n_i}{\sum_i p_i} \times 100,000$$

(p_i = population par classe d'âge)

Taux standardisé

Taux pour 100.000 habitants par an, standardisé sur l'âge par la méthode directe (à une population de référence ou standard). Il consent la comparaison entre différentes zones (se rapportant à la même population standard), en éliminant l'effet de la diverse composition par âge des populations :

$$T_{st} = \frac{\sum_i (T_i \times \text{standard pop.} i)}{\sum_i \text{standard pop.} i}$$

La population standard peut être choisie librement en fonction des données qui doivent être comparées ; pour une comparabilité maximale des données, on conseille toutefois la référence au modèle de population mondiale, européenne, ou italienne dans le cas de données relatives à l'Italie.

Risque cumulé

Il exprime la probabilité de développer une tumeur dans l'intervalle de temps qui va de la naissance jusqu'à un âge donné (probabilité de tomber malade avant un âge donné si le sujet ne décède pas avant pour une autre cause). Il est souvent exprimé pour 1.000 habitants et le risque 0-74 ans est :

$$R_{cum} = 1 - \exp\left[\left(-\sum_i T_i\right) \times 5\right]$$

Rapport mortalité/incidence

Il exprime le rapport entre les cas décédés et incidents (généralement exprimé par sexe et localisation tumorale).

Nombre d'années de vie perdues/gagnées

Elles correspondent à l'estimation des années de vie perdues ou gagnées par une cohorte d'exposés à un facteur par rapport à une cohorte de non exposés. On peut calculer les années de vie perdues (Years Life Lost, YLL) ou potentiellement perdues par rapport à un nombre attendu (YPLL). A partir de ces estimations, les années de vie perdues/gagnées (moyenne de YLL/YPLL d'une population), les taux bruts de YLL, YPLL (en rapportant YLL/YPLL à la population d'âge inférieur au seuil choisi), les taux standards (standardisation directe de YLL/YPLL) ou encore les taux cumulés de YLL/YPLL peuvent être calculés.

Survie

Une vérification régulière de l'indicateur de la survie des cas incidents enregistrés devrait être effectuée, vu qu'elle sert de contrôle interne au registre pour les données du statut vital et pour l'analyse épidémiologique. Le suivi ou follow-up des patients doit être conduit à travers une source officielle (commune de résidence, agences de santé si celles-ci sont en lien avec les communes). A l'intérieur de la structure de recueil des données du cas, à la date de follow-up correspond une variable indicative du statut vital du patient (vivant, décédé, perdu de vue, etc.). La donnée synthétique finale peut être exprimée comme survie observée (selon le modèle actuariel ou de Kaplan Meier) ou relative, cette dernière exprimée comme rapport entre survie observée et survie attendue ; la survie attendue est calculée à partir des taux de mortalité générale de la population. La survie relative représente donc la survie de la cohorte de patients au net de la mortalité pour les autres causes, qui peut être différente d'une population à l'autre. Les algorithmes de ces indices et de leurs erreurs standard relatives sont généralement fournis par les divers programmes de calcul disponibles. Une méthodologie alternative à l'analyse d'une cohorte de patients incidents est constituée par la survie de période (les détails méthodologiques sont consultables à l'intérieur du programme SEERStat, voir en annexe).
Les cas enregistrés à partir de la date de décès (DCO, cas autoptiques) sont exclus de l'analyse de survie.

Prévalence

Elle représente (sous forme de proportion ou de nombre absolu de cas) le nombre de patients avec diagnostic antécédent de néoplasie (dans un intervalle de temps à déterminer) en vie au moment de l'observation. Elle est un indice important pour la pro-

grammation de la prise en charge sanitaire. Pour les patients perdus de vue (lost to follow-up) et qui sont par conséquent considérés "censurés" (censored), des estimations à partir des données de survie sont généralement utilisées. Diverses modalités de calcul des indices sont aussi disponibles pour prendre en compte les patients porteurs de tumeurs multiples.

Tendances temporelles

Les indicateurs ponctuels qui peuvent être calculés à partir de la banque de données du registre consolident nettement leur valence informative s'ils sont présentés pour un certain intervalle de temps. En effet, ces dernières années l'analyse des tendances temporelles (incidence, mortalité, survie, prévalence) a fourni de nouveaux éléments d'interprétation de l'incidence des tumeurs dans les divers contextes géographiques. Une première information intuitive sur cet argument est possible par la visualisation des taux (bruts ou standardisés) relatifs à différentes périodes calculés "petit à petit". L'oscillation aléatoire des indices, en particulier pour les petites zones pour lesquelles les intervalles de confiance sont larges, dans la plupart des cas ne permet pas d'obtenir des informations convaincantes. Une synthèse de l'évolution temporelle s'obtient par un modèle de régression : parmi les multiples solutions possibles, une sorte de standardisation de la procédure est offerte par le programme Joinpoint Regression Program [1], basé sur le repérage des segments linéaires qui s'adaptent le mieux aux taux observés (logarithme des taux), minimisant la somme des carrés des distances entre les points et les segments [2]. Le nombre maximal de segments avec lequel la tendance est décomposée est limité par le nombre k de joinpoint fixé a priori. Le joinpoint représente le point de jonction, l'année qui identifie une variation dans la tendance.
Le modèle peut être représenté par une unique équation :

$$\ln(T_{\text{stand}}) = \beta_0 + \beta_1 x_i + \delta_1 (x_i - \tau_1)^+ + \\ + \delta_2 (x_i - \tau_2)^+ + \varepsilon_i^{(k)}$$

Pour de plus amples informations sur la méthodologie on renvoie à la bibliographie (ou directement au site : srab.cancer.gov/joinpoint).

Erreurs standard et intervalles de confiance des indicateurs

Les programmes en utilisation pour le calcul des indicateurs prévoient généralement le calcul des erreurs standard (ES) et des intervalles de confiance, au seuil 95% de probabilité. Pour chaque indicateur, cette donnée est extrêmement utile et, quand cela est possible, elle ne doit pas être omise afin de reporter la fluctuation aléatoire de l'indicateur ponctuel (taux, probabilité, etc.), notamment pour des nombres restreints de cas (typiques de nombreuses zones). L'évaluation d'éventuelles réelles différences géographiques et temporelles dans les indicateurs dépend de ces indices.

ES du taux brut

Le calcul présuppose que le nombre de cas suit une distribution de probabilité de Poisson :

$$ES_{\text{brut}} = \sqrt{N} \,/\, population \times 100{,}000$$

ES du taux standardisé

L'ES est calculé en supposant que le nombre de cas suit une distribution de probabilité de Poisson. En supposant que le taux ajusté sur l'âge soit composé de groupes d'âge qui vont de X à Y :

$$ES_{TSD} = \left[\sum_{i=x}^{y} \left(\frac{Ns_i}{Ps} \right)^2 \times \left(\frac{N}{\text{population}^2} \right) \right]^{\frac{1}{2}} \times \\ \times 100{,}000$$

Ns_i = nombre de personnes de la i-ème classe d'âge de la population standard
Ps = population standard

ES du taux cumulé

$$SE = 5 \times \sqrt{\sum_i \frac{n_i}{P_i^2}}$$

Programmes de calcul des indices

Ils sont une ressource importante du fait qu'ils incident de façon positive sur les temps et les coûts de réalisation des rapports produits par le registre. Parmi les multiples solutions offertes par les logiciels, commerciaux et non, on retient utile d'en signaler deux.

CanReg5

C'est un programme conçu pour la gestion globale du registre, disponible gratuitement auprès du CIRC ou de l'IACR. Il comprend une fonction d'analyse avec distributions de fréquence, tableaux d'incidence et une interface avec EpiInfo ® pour les élaborations plus complexes [3]. Dans la version Canreg5 il existe aussi une interface avec le logiciel R (http://www.r-project.org). Il permet également l'exportation de fichiers d'incidence à utiliser directement sur SEERStat.

SEERStat

C'est pour le moment le logiciel d'analyse le plus complet pour les registres des tumeurs, disponible gratuitement sur le site internet du gestionnaire [4]. Il prévoit une phase de chargement des données à travers un logiciel spécifique (SEERPrep) [5]. Ses principaux avantages sont :

- la grande expertise et compétence des programmeurs,
- la conception spécifique aux exigences d'un registre des tumeurs,
- la disponibilité de toutes les statistiques habituelles et la possibilité de réaliser des analyses complexes,
- la facilité d'utilisation,
- la gratuité,
- la disponibilité de mises à jour périodiques,
- la possibilité de l'installer sur n'importe quel ordinateur.

Le logiciel, à partir d'une base de données sur laquelle des transformations de variables sont possibles avant et après l'importation, comprend des modules d'analyse pour :

- les distributions de fréquence,
- les taux bruts et standardisés,
- l'analyse de la survie, observée et relative, de cohorte et de période,
- le calcul de la prévalence (complète ou partielle), avec comme option l'évaluation des patients avec tumeurs multiples.

Une procédure d'exportation de données et tableaux vers d'autres logiciels (ex. : Office) permet de personnaliser les analyses et la construction de graphiques. Le réseau SEER offre également des procédures d'analyse pour des données propres ou à partir de tableaux produits par SEERStat :

- DevCan : probabilité de développer une tumeur ou de mourir à cause de celle-ci,
- Joinpoint : analyse des tendances temporelles,
- CanSurv : analyse de la survie avec graphique et modèles (paramétrique standard, Cox, mixte),
- ComPrev : prévalence complète,
- ProjPrev : projections de la prévalence de durée limitée de SEERStat sur d'autres populations.

Les programmes sont accompagnés d'une documentation complète.

Un support encore plus complet, comprenant des exemples et une partie théorique sur les modèles, est disponible sur le site www.seer.cancer.gov.

Contrôle des erreurs

Procédure de saisie des données

Les diverses modalités des systèmes d'archivage des registres présupposent divers critères de contrôle des erreurs en phase de saisie des données. En effet, par rapport aux modalités traditionnelles d'identification des cas (à travers l'examen systématique des dossiers médicaux et l'insertion des cas sur support informatique), il est toujours plus fréquent l'acquisition de la part des registres de sections entières de bases de données sanitaires (SDO en Italie, rapports d'anatomo-pathologie, certificats de décès), dont les données sont plus facilement contrôlables a posteriori grâce à des programmes automatisés comme ceux qui sont décrits plus bas.

Il est en revanche très important, pour la prévention des erreurs de procédure dans les nombreuses étapes d'acquisition et de gestion de la base de données, que chaque registre cherche à formaliser toutes les phases de sa propre activité à l'intérieur d'un manuel des procédures qui, d'une part, peut constituer une garantie pour une reproductibilité et qualité constantes dans le temps du processus d'élaboration des données, et d'autre part, peut stimuler une implication plus active et une remise à niveau du personnel en ce qui concerne toutes les étapes de définition des données.

Programmes de contrôle des erreurs de données

Le CIRC délivre de façon périodique des mises à jour des programmes de contrôle en mesure d'identifier la plupart des possibles erreurs de cohérence interne de chaque cas. Ces programmes sont mis à disposition gratuitement pour les registres et sont directement disponibles sur internet.

Indépendamment de l'utilité de chacun de ces logiciels à l'activité d'enregistrement des cas des registres, lors de la participation à des études ou à des publications externes, il est demandé aux registres de contrôler leurs propres données d'incidence avant de les envoyer à la banque de données nationale (ex. : Cancer incidence, EUROCARE). Dans l'encadré au bas de cette page sont présentés les principaux logiciels à disposition des registres qui garantissent des procédures de contrôle complémentaires sur les données (l'utilisation des deux est donc nécessaire). De plus, ces logiciels offrent la possibilité de conversion entre les divers systèmes de classification (CIM et versions successives de la CIM-O) et le contrôle des tumeurs multiples.

Logiciels pour le controle des données

IARCcrgTools*

Logiciel utilisable avec les versions de Windows 95/98/Me (de préférence NT/2000/XP), qui comprend :

- un programme de conversion entre les classifications :
  - de CIM-9 (1975) à CIM-O-2 (1990)
  - de CIM-9 (1975) et CIM-O-1 morphologie (1976) à CIM-O-2 (1990)
  - de CIM-10 (1992) à CIM-O-2 (1990)
  - de CIM-10 (1992) et CIM-O-2 morphologie (1990) à CIM-O-2 (1990)
  - de CIM-O-1 (1976) - CIM-O-2 (1990)
  - de CIM-O version d'essai (1988) à CIM-O-2 (1990)
  - de CIM-O-2 (1990) à CIM-9 (1975)
  - de CIM-O-2 (1990) à CIM-10 (1992)
  - de CIM-O-2 (1990) à CIM-O-3 (2000)
  - de CIM-O-3 (2000) à CIM-10 (1992)
- un programme de contrôle de validité et cohérence entre les variables :
  - âge, incidence et date de naissance
  - âge, siège anatomique et morphologie (CIM-O-3)
  - sexe et siège anatomique
  - sexe et morphologie (CIM-O-3)
  - comportement et siège anatomique (CIM-O-3)
  - comportement et morphologie (CIM-O-3)
  - grade et morphologie (CIM-O-3)
  - base de diagnostic et morphologie (CIM-O-3)
  - un programme de contrôle des tumeurs multiples, en accord avec les règles IARC/IACR 2004 [7].

Le programme est simple d'utilisation et de lecture et est accompagné d'une bonne documentation, également consultable à travers le site www.iacr.com.fr/iarccrgtools.htm [6]*version 2.05, juillet 2008

DEPedits*

Comme le logiciel IARCcrgTools, ce logiciel comprend des programmes de conversion entre les différentes versions des CIM et CIM-O et le traduction de celles-ci en CIM-O-3 (avec des limitations pour les nouvelles entités prévues par cette dernière qui intéressent principalement le système lymphatique et hématopoïétique) [8].

Un programme de contrôle basé sur la CIM-O-3 est fourni, et son utilisation est fortement recommandée avant l'envoi des données au groupe d'épidémiologie descriptive (CIN) du CIRC. Ce dernier met en évidence les erreurs d'attribution des codes relatifs à la topographie, à la morphologie et au sexe (problèmes de validité du codage), tout comme (dans un autre document) les cas relatifs aux combinaisons "inhabituelles" à soumettre à une ultérieure vérification (problème de cohérence entre les codes).

Toutes les conversions et les contrôles disponibles dans le logiciel IARCcrgTools sont également présents dans ce logiciel, qui inclut en outre :

- contrôles spécifiques sur les variables de survie,
- contrôles spécifiques sur les tumeurs de l'enfant,
- validation (optionnelle) selon les critères de combinaisons topographiques et morphologiques de la CIM-O-3 en accord avec la liste adoptée par le programme SEER (Surveillance Epidemiology and End Results, National Cancer Institute, USA) [9].

Les critères qui s'appliquent à cette dernière opération sont énumérés dans la documentation fournie en annexe au programme, ainsi que sur le site internet du réseau SEER. Dans le détail, les contrôles réalisés par DEPedits vérifient :

- validité :
  - sexe
  - date d'incidence
  - date de naissance
  - date de suivi ou de follow-up
  - topographie selon la CIM-O-3
  - morphologie selon la CIM-O-3 (4 premiers chiffres de la morphologie)
  - comportement selon la CIM-O-3 (5ème chiffre de la morphologie)
- cohérence - consistance :
  - sexe ↔ CIM-O-3 T
  - sexe ↔ CIM-O-3 M
  - CIM-O-3 T ↔ CIM-O-3 M (CIRC)
  - CIM-O-3 T ↔ CIM-O-3 M (SEER)
  - CIM-O-3 M ↔ base de diagnostic
  - morphologies spécifiques généralement admises avec le niveau de confirmation histo-cytologique, à l'exception du cerveau et du foie, également acceptées sur la base d'examens paracliniques ou techniques
  - date d'incidence ≥ date de naissance
  - date d'incidence ≤ date de follow-up
  - âge ↔ CIM-O-3 T
  - âge ↔ CIM-O-3 M
  - statut vital ↔ base de diagnostic

Ce programme est assez simple d'utilisation et est associé à une description détaillée des procédures et à des instructions complètes. Le logiciel et le manuel peuvent être récupérés sur le site de l'ENCR (http://www.encr.com.fr/encr_resources.htm).

Programme CHILD-CHECK

Il s'agit d'un programme directement disponible auprès du CIRC, adapté aux tumeurs de l'enfant et conçu pour contrôler la cohérence des enregistrements individuels et pour la conversion des données de la CIM-O à l'International Classification of Childhood Cancer (ICCC) [10].

Les procédures de contrôle vérifient :

- les codes définis pour les dates, l'âge et le sexe
- la compatibilité des codes morphologiques et topographiques
- les combinaisons improbables entre tumeur, sexe et âge
- les codes topographiques inhabituels pour certaines morphologies spécifiques
- l'utilisation inappropriée de codes morphologiques non spécifiques

Les procédures de conversion transforment les codes :

- de la CIM-O-1 (1976) à la CIM-O-2 (1990)
- de la CIM-O-2 (1990) à l'ICCC (1996).

Programmes de contrôle de cohérence des données

CheckRT : logiciel AIRTUM-CCM pour le contrôle de qualité

(Dr. Ivan Rashid)

CheckRT est un programme qui consent aux opérateurs des registres d'effectuer des contrôles de qualité sur les données de façon rapide et approfondie. La fonction de CheckRT est de formaliser, divulguer et élargir les critères à la base du processus de qualification des registres des tumeurs italiens auprès de l'AIRTUM en fournissant un instrument également utilisable pour l'auto-évaluation des données des registres déjà qualifiés.
CheckRT consent l'importation de fichiers au format Microsoft Access des cas incidents, de la mortalité et de la population et fournit un compte-rendu détaillé au format texte sur le résultat des contrôles qui ont été réalisés.

Contrôles effectués

Le logiciel CheckRT contient les contrôles suivants sur les données des registres :

- analyse de la sous-catégorie topographique
- analyse du pourcentage de sièges mal définis et génériques
- analyse du pourcentage de DCO
- analyse du pourcentage de vérifications microscopiques
- analyse du pourcentage de morphologies spécifiques par siège
- analyse du pourcentage de morphologies génériques
- analyse de la valeur du rapport mortalité/incidence
- analyse de l'évolution temporelle du rapport mortalité/incidence
- analyse de la stabilité du rapport mortalité/incidence
- analyse du taux standardisé d'incidence par siège
- analyse du rapport homme/femme
- analyse de la distribution du taux âge-spécifique
- analyse de la stabilité des vérifications microscopiques
- analyse des tumeurs en âge infantile
- analyse de la survie à un an du diagnostic
- analyse de la stabilité du taux standardisé d'incidence
- analyse des incohérences siège-morphologie-contrôles EUROCARE

pour un total de plus de 1.300 contrôles. A chaque contrôle est associé un score indicatif du poids à l'intérieur du groupe de contrôles. La somme des points, définie comme référence ou benchmark, représente un indicateur synthétique de la qualité du registre du cancer.
On renvoie à l'annexe pour la consultation du manuel d'utilisation.

Vérification de la complétude de l'enregistrement

La force d'un registre général du cancer repose sur la couverture complète de l'enregistrement à l'intérieur du territoire étudiée. A ce pré-requis méthodologique doit suivre une vérification des procédures d'enregistrement et de contrôle avant la publication des données. Outre les contrôles d'erreurs et de cohérence interne des données déjà présentés, la communauté scientifique des registres des tumeurs a depuis un certain temps identifié des indicateurs de la qualité du flux informatif en terme d'efficacité et d'efficience. Ceux-ci doivent être associés à la production des données des registres, dont elles signalent la fiabilité en terme de procédures d'acquisition et de contrôle des cas incidents. La vérification qualitative des données du registre (niveau élevé de spécificité, dans le respect des règles d'enregistrement) implique l'évaluation de la complétude de l'enregistrement (niveau élevé de sensibilité dans l'identification des cas incidents). Une série de variables, certaines déjà vues dans les chapitres précédents, notamment dans celui sur les indicateurs de fréquence, peut être analysée d'un point de vue critique, en se basant sur les règles d'enregistrement communes (reproductibilité, objectivité), de manière à garantir un important détail d'informations associé à un bon niveau de complétude (dans l'identification de tous les cas effectivement incidents) et de couverture de la population d'étude [12]. Ce type de contrôle doit faire partie des procédures classiques d'un registre et ne pas se limiter à représenter une opération occasionnelle réalisée lors des rendez-vous scientifiques.

Complétude de la couverture

Ce contrôle cherche surtout à identifier l'éventuelle perte de cas (mais aussi la duplication accidentelle ou l'inclusion de cas qui ne remplissent pas les critères d'éligibilité) que peuvent engendrer les problèmes liés au flux informatif. Les procédures de contrôle de qualité et cohérence reportées plus haut, outre les diverses approches spécialement conçues pour cette évaluation, peuvent également être adaptées à contrôler la complétude.

Proportion de cas avec vérification microscopique

Elle est un indicateur de la qualité de l'incidence enregistrée, qui correspond au pourcentage de cas enregistrés avec une base de diagnostic histopathologique ou cytopathologique. Les diagnostics hématologiques obtenus à travers la ponction médullaire (cytologie) ou la biopsie médullaire (histologie) doivent être codés à l'aide de ce niveau de diagnostic. Ces dernières années des techniques paracliniques (scanner, IRM, marqueur) se sont également affirmées, qui atteignent presque le niveau de qualité microscopique, consentant l'identification de la lésion et la successive planification thérapeutique. En particulier, les néoplasies hépatiques et cérébrales sont des tumeurs pour

lesquelles ces méthodes garantissent une qualité diagnostique comparable à l'histopathologie : en présence de ce niveau de diagnostic, il est possible d'indiquer un code morphologique de tumeur spécifique, normalement consenti uniquement dans le cas de confirmation microscopique. Le pourcentage de cas munis de diagnostic microscopique (total et siège-spécifique) devrait respecter une bonne homogénéité entre les zones géographiques limitrophes, en l'absence de facteurs de risque évidents ; celui-ci peut s'abaisser en présence d'une consultation de la documentation cyto-histopathologique incomplète de la part du registre, tandis qu'un pourcentage trop élevé peut cacher un sous-enregistrement dû à la perte de cas confirmés à l'aide d'un niveau diagnostique inférieur.
L'examen attentif de cet indicateur peut fournir des renseignements sur les vérifications et corrections à apporter au flux informatif dès les premières phases de production des données d'incidence.

DCI/DCO

A la fin du processus de couplage des diverses sources d'informations du registre, une part des cas incidents créés n'est définie que sur la base du certificat de décès du patient. Il s'agit, à moins d'erreurs dans la phase de création du certification de décès, de cas qui échappent à l'enregistrement durant la période où ils sont en vie (à travers les examens cliniques, hospitalisations, etc.), et qui signalent en général des problèmes inhérents au flux d'informations. Comme déjà dit précédemment, un premier contrôle génère une classe de cas appelés DCI (ou DCN : Death Certification Notification), sur lesquels doit être conduit un contrôle rétroactif (trace back) pour la récupération de la documentation qui a éventuellement échappé à l'enregistrement durant le parcours diagnostique et d'assistance du patient. Au terme de cette phase, une partie des DCI est réintégrée dans l'incidence avec un meilleur niveau diagnostique, tandis que le reste des DCI devient des cas DCO (Death Certification Only), du fait que les contrôles n'ont relevés aucune trace du diagnostic posé lorsque le patient était en vie. DCI et DCO représentent d'importants indicateurs de qualité du flux informatif du registre et leur nombre doit être enregistré séparément. Les cas DCI indiquent une difficulté du registre à atteindre une bonne complétude du recueil des cas ; les cas DCO mettent également en évidence les possibles carences internes des systèmes informatifs sanitaires. L'inclusion finale dans l'incidence des cas DCI/DCO corrige cependant seulement de façon partielle le défaut de complétude, qui comprend vraisemblablement un certain nombre de cas perdus car non décédés. A titre d'exemple, en admettant une proportion de 10 perdus sur 100 et une mortalité de 60%, six de ces cas seraient certainement récupérés comme DCI/DCO, tandis que quatre resteraient inconnus, attestant le niveau de complétude du recueil à 96%. La précision des certificats de décès, bien qu'elle soit généralement inférieure aux données d'incidence, peut être tenue sous contrôle par le registre à travers la comparaison croisée des informations provenant des sources cliniques.

Rapport mortalité/incidence (M/I)

Sa stabilité dans le temps et les diverses zones géographiques fait de cet indicateur un des principaux et plus fiable standard d'évaluation des données des registres. La référence aux données de la littérature consent aux registres (en particulier à ceux en phase d'activation) d'identifier rapidement la perte de cas incidents (avec l'augmentation du rapport M/I), l'inclusion accidentelle de doublons et de cas prévalents ou encore une base de données des cas décédés incomplète (rapport M/I bas). Il convient de souligner qu'il est indispensable, pour chaque vérification des données d'incidence à partir des données de mortalité, que celles-ci soient certifiées et proviennent d'une source externe au registre (en Italie l'ISTAT, région, agence de santé, etc.).

Date d'enregistrement du cas

Elle est une importante variable de contrôle de la complétude des données et par conséquent des temps de publication. En effet, ces dernières années, la traditionnelle publication des données, souvent caractérisées par un important retard entre l'année d'incidence des cas et l'année de leur enregistrement, a été progressivement abandonnée en faveur d'une production de données plus rapide, souvent sollicitée par le besoin d'évaluer les programmes de dépistage ou par l'exigence de contribuer à la programmation sanitaire dans le cadre de certaines urgences en matière de santé publique et d'épidémiologie.
Il est néanmoins évident que plus les temps de production des données s'écourtent, plus le risque d'une incomplète reconstruction de l'ensemble des cas augmente, ces problèmes de complétude se traduisant par une augmentation des cas "récupérés" après la fermeture de l'incidence.
La nécessité de contrôler de façon dynamique la complétude des données d'incidence en fonction du temps, ainsi que les temps de production eux-mêmes, est remplie par la variable (obligatoire) qui signale la date d'enregistrement du cas (insertion dans la base de données). De cette manière, il est possible de développer une fonction (avec le pourcentage de cas enregistrés en ordonnée et le temps en abscisse) qui indique, à partir du temps T0 (début de l'enregistrement de l'incidence) la rapidité avec laquelle est atteint un bon niveau de complétude (représenté par le plateau supérieur de la courbe dans la phase de stabilisation des insertions des nouveaux cas). Cette opération rend possible un monitorage continu dans le temps de la vitesse requise pour atteindre la complétude (et par conséquent, de l'intervalle minimum qui peut s'écouler entre l'incidence d'un cas et son enregistrement),

qui représente la fiabilité des données d'incidence du registre en terme de précision et de ponctualité.

Incidence par sous-aires d'enregistrement

Elle consent de relever une différence d'homogénéité de recueil des cas à l'intérieur de la zone d'étude (différence entre diverses agences de santé, villes/provinces/départements, etc.).

Contrôles sur échantillon

Ils s'effectuent en sélectionnant un échantillon de patients oncologiques à l'intérieur de séries de cas sur le territoire (hôpitaux, consultations ambulatoires, flux sanitaires) et en contrôlant si le recueil opéré par le registre est advenu correctement.

Survie

L'exécution périodique de cette analyse de la part des registres consent d'identifier, à travers les écarts par rapport aux valeurs attendues, les possibles problèmes de complétude ou de recueil des cas incidents.

Tendances temporelles d'incidence

L'étude de la tendance temporelle de l'incidence représente un instrument de contrôle utile pour vérifier au cours du temps d'éventuelles variations qualitatives et quantitatives dans les flux informatifs qui sont fournis au registre. Dans le passé, les variations temporelles d'incidence trop brusques ont généralement représenté des problèmes de sous- ou sur-enregistrement ; ces dernières années, l'introduction et la diffusion de techniques de diagnostic plus sensibles (endoscopie, échographie, marqueurs tumoraux, radiologie) ou l'activation des programmes de dépistage dans la population (spontané ou organisé) ont déterminé des augmentations significatives et rapides de l'incidence de certaines formes tumorales, parfaitement cohérentes avec ce qui était attendu sur la base d'expériences précédentes. Chaque variation temporelle de l'incidence (sexe et siège-spécifique) doit dans tous les cas toujours être considérée et interprétée avec rigueur et attention. La récente utilisation de programmes automatiques chargés de ce type de contrôle a nettement contribué à améliorer et à standardiser les procédures d'évaluation de la qualité des registres émergents tout comme celle des registres de tradition plus longue.

Techniques spéciales

Mise en place de modèles de capture-recapture ou d'autres techniques statistiques.

Complétude de détail

Les informations manquantes influencent négativement, autant que les erreurs, la précision de l'enregistrement. Il peut en effet ne pas être toujours possible de compléter toutes les données prévues pour chaque patient ; certaines variables peuvent donc être définies manquantes ou missing, représentant un déficit d'information plus ou moins grave (informations essentielles comme le sexe, le lieu de résidence, les dates de naissance ou d'incidence ou informations facultatives). La fréquence des informations manquantes doit faire l'objet d'une attention particulière, du fait qu'elles représentent une partie importante du contrôle de qualité de l'enregistrement. Deux variables essentielles recouvrent généralement un rôle important dans les protocoles internationaux sur les contrôles de qualité : le pourcentage de cas dont la localisation topographique est un siège inconnu ou mal défini et le pourcentage de patients d'âge inconnu.

Le premier groupe de cas, indice de la qualité de l'information diagnostique, se réfère aux codes CIM-10 et CIM-O C26, C39, C76, C77 et C80. Le pourcentage de patients avec un âge au diagnostic inconnu (à cause de l'absence de la date de naissance) signale en général une grave carence dans le processus d'identification du patient ; dans les registres des pays développés, cet indice doit être maintenu largement sous le seuil de 1%. Il existe enfin d'autres données manquantes dont le pourcentage doit être surveillé avec attention, et en particulier :

- le sexe,
- la commune/province de naissance (contrôle des migrants),
- l'histologie tumorale (pourcentages de néoplasies avec code générique "SAI"),
- le lieu de résidence,
- la date d'incidence.

La toujours plus étroite collaboration des registres avec l'organisation sanitaire locale permet également le recueil d'informations utiles à l'évaluation de l'impact diagnostique (ex. : dépistage) ou de celui d'autres procédures d'assistance, comme :

- des détails sur les aspects topographiques et morphologiques de la tumeur,
- la désagrégation des variables qui déterminent le stade (dimensions, nombre de ganglions, sièges de métastases à distance, etc.),
- l'histoire du patient par rapport aux programmes de dépistage,
- les thérapies,
- les variables socio-économiques,

pour lesquelles des circuits de recueil des informations doivent être organisés ad-hoc.

Les registres doivent vérifier périodiquement la complétude de la propre activité d'enregistrement en utilisant les indicateurs de qualité classiques et les systèmes de contrôle prévus par ce manuel. La banque de don-

nées centrale peut évidemment réaliser ces vérifications à travers d'autres systèmes de contrôle, parmi lesquels celui de la North American Association of Central Cancer Registries (NAACCR) [13].

Programmes de couplage des données

Logiciel pour le couplage de données automatisé en Italie SALI

(Dr. Luigino Dal Maso, CRO Aviano)

Les procédures de couplage entre les cas présents dans diverses bases de données constituent une opportunité particulièrement utile pour le partage d'informations provenant de sources différentes, comme il est souvent requis dans les études épidémiologiques ou pour la surveillance des patients enregistrés dans la banque de données des comptes-rendus de pathologie. En présence de codes univoques de couplage (critère déterministe), ces procédures sont réalisables par la plupart des programmes de gestion de bases de données. Cependant, l'indisponibilité assez fréquente de clés primaires oblige l'utilisation de logiciels en mesure de traiter des archives de taille moyenne-élevée à partir des données démographiques classiques. Même dans les cas de non parfaite correspondance entre ces données, une perte minimale des couplages possibles peut être garantie.
Développé au sein de l'unité d'Epidémiologie et Biostatistique du Centre de référence oncologique d'Aviano (département de Pordenone), le programme SALI est né avec l'objectif de coupler des enregistrements individuels d'archives de grandeur moyenne (de l'ordre de 100.000 enregistrements), consentant la possibilité d'une révision manuelle du résultat et la protection, à chaque phase de l'opération, de la confidentialité des données personnelles des individus traités [11].
SALI utilise les archives sous format de base de données à partir de clés primaires (nom, prénom et date de naissance) en tenant compte de possibles erreurs d'imputation de ces clés primaires (ex. : CASTAING peut être couplé à CASTAIGN). La mise à jour du logiciel a permis d'élargir le nombre de champs délimités pour consentir le traitement des noms composés et doubles prénoms utilisés en Suisse (ou en France). La procédure a atteint une sensibilité de 99% et une spécificité de 100% par rapport à la vérification manuelle retrouvée lors d'une étude pilote menée à Genève.
La procédure de couplage est basée sur sept niveaux, deux automatiques et cinq interactifs, dans lesquels l'opérateur peut décider à travers des fenêtres spécifiques d'accepter ou de refuser le couplage proposé.
SALI peut être utilisé dans chaque système opératif IBM-compatible (DOS ou Windows). On renvoie à l'annexe pour la consultation du manuel d'utilisation.

Révisions et mises à jour

Suivi ou follow-up

La mise à jour du statut vital des patients oncologiques a depuis longtemps trouvé sa place dans les procédures habituelles du registre, aussi bien dans le but d'effectuer le contrôle qualitatif des données que pour la production des données de survie. A intervalles réguliers (souvent tous les deux ans), le registre peut procéder à cette mise à jour en indiquant la date du dernier signalement du patient et le statut vital correspondant.
Il est ainsi possible, à travers l'utilisation de deux variables seulement, de représenter les trois scénarios suivants :

- patient en vie au terme du suivi ou follow-up (date de fermeture du follow-up et patient en vie),
- décès du patient (date de décès et statut vital indiquant le décès),
- censoring ou censure du patient avant la fermeture du follow-up (date de dernier signalement du patient et patient signalé en vie) ; cette situation se vérifie à deux occasions :
  - mort du patient pour une cause non attribuable à la néoplasie,
  - patient perdu de vue pour une raison quelconque ; dans ce cas, toute la période de survie effective atteinte par le patient jusqu'à la dernière date disponible est considérée.

La variable indiquant le statut vital peut être déclinée différemment pour indiquer, par exemple, la proportion de patients décédés, émigrés, perdus de vue, etc., mais en conservant toujours une unique modalité pour la cause en question (décès pour cause tumorale). Cette dernière fonction est remplie par la variable "cause du décès", qui peut garantir (à condition de la fiabilité de la donnée) la construction de survies causes-spécifiques en ne considérant pas l'effet de la mortalité compétitive. Il est évident que la source de la donnée sur le statut vital doit être fiable et généralement identifiée de la même façon que les autres variables démographiques du patient (état civil communal ou archives équivalentes), tandis que dans le cas de follow-up qui vise à d'autres endpoint ou critères de jugement finaux (reprise de maladie, apparition de complications), la source des données doit être explicitement déclarée (divisions cliniques, systèmes informatifs sanitaires) et soumis à des procédures de contrôle de complétude analogues à celles expérimentées pour les autres données du registre.

Bibliographie

1. Statistical research and application branch, Division of Cancer control and population sciences, National Cancer Institute, USA. (srab.cancer.gov/joinpoint/).
2. Lerman PM. Fitting segmented regression models by grid search. *Applied Statistics* 1980; 39: 77-84.
3. Cooke AP, Parkin DM, Ferlay J. *CanReg 4, Descriptive epidemiology unit*, IARC/IACR. Lyon 2005. (www.iacr.com.fr/canreg4.htm).
4. Surveillance research program. *National Cancer Institute SEER*Stat software* (www.seer.cancer.gov/seerstat) version 6.1.4, 2005.
5. Information Management Services Inc. *The SEER program*. National Cancer Institute. v.2.3.2, 2005.
6. Ferlay J, Burkhard C, Whelan S, Parkin DM. Check and conversion programs for cancer registries (IARC/IACR Tools for cancer registries). *IARC Technical report* N. 42, Lyon 2005.
7. IARC/IACR/ENCR Working group. International rules for multiple primary cancers (CIM-O Third edition). *IARC Internal report* 2004/02, Lyon 2004.
8. Ferlay J. *DEPedits* 1.00, IARC, Lyon 2006.
9. *Surveillance Epidemiology and End Results*, National Cancer Institute, USA. (seer.cancer.gov).
10. Kramarova E, Stiller CA, Ferlay J et al. International Classification of Childhood Cancer. *IARC Technical report* n. 29, IARC Lyon 1996 (dischetto incluso).
11. Dal Maso L, Braga C, Franceschi S. Methodology used for "Software for automated linkage in Italy" (SALI). *Journal of Biomedical Informatics* 2001; 34: 387-95. Description of upgraded version of SALI software available athttp://www.registri-tumori.it/PDF/AIRTUM2009HANDBOOK/Chapter_Appendix2.pdf.
12. Jensen OM, Parkin DM, MacLennan R et al. Cancer registration, principles and methods. *IARC Scientific publication* n. 95, Lyon 1991.
13. Havener LA. *Standards for cancer registries Vol. III*. Standards for completeness, quality, analysis and management of data. North American Association of Cancer Registries Inc, 2004. 131-139_AIRTUM_Cap05 30-10-2007 15:28 Pagina 139