Rapport final

" Etablissement et développement d'outils de Data Mining et d'aide à la décision sur les données issues des accidents domestiques et de loisirs, dans le cadre du Programme prévention des blessures "

Rapport préparé par :

Marc Nectoux - BIOSTA (Fr)

Dr Henning Bay-Nielsen - NIPH (Dk)
Dr Birthe Frimodt-Moller - NIPH (Dk)
Dr Robert Bauer - Sicher Liben (Au)
Dr Jean-Pierre Darlot - BIOSTA (Fr)
Claude Mugnier - BIOSTA (Fr)

SOMMAIRE

Les conclusions opérationnelles

- Nous sommes partis du fait que l'ancien système EHLASS de recueil d'information a permis de recueillir des données sur plusieurs millions d'accidents domestiques et de loisirs (ADL) en Europe (plus de cinq millions de cas fin 1999). Ces données ont été exploitées statistiquement au niveau national. Mais il apparaît que le potentiel informatif contenu dans les bases nationales, et a fortiori dans les bases européennes récemment développées ou en cours de développement, est largement sous-exploité et qu'il importe de mieux le valoriser en utilisant des outils de Data Mining (exploration de la « mine » des données) et des procédures spécifiques.

- A partir de ce constat, nous avons mené une enquête préalable auprès de la DG SANCO et des équipes nationales en charge de ce système d'information (SI). Elle a permis de connaître les outils logiciels utilisés (pour les logiciels statistiques : SAS - 5 citations, SPSS - 3 citations) et de constater que les outils spécifiques déjà développés sont rares, tandis que les besoins en outils de recueil, de contrôle et d'exploitation des données sont nombreux et divers.

- De plus, nous exposions dans cette enquête quatre propositions de procédures pour une meilleure exploitation des données EHLASS :

> - la **Procédure SSRD** (Score Synthétique de Dangerosité Relative) pour hiérarchiser la dangerosité potentielle des produits;
> - la **Procédure SAA** (Système d'Alerte Automatisée) pour mettre en place une alerte automatique à partir des données recueillies par le système;
> - la **Procédure NGA** (Note de Gravité de l'Accident) pour définir la sévérité d'un accident;
> - la Méthode **SCENAR** (Méthode des scénarios) pour définir des scénarios types d'accidents.

- Ces quatre procédures ont fait l'objet d'une notation. Elles possèdent en moyenne, au regard des notes obtenues concernant l'utilité décisionnelle, la validité logique et l'utilisabilité, un niveau d'acceptabilité relativement bon et comparable. Cependant, la variance des notes obtenues est assez grande, ce qui indique que certaines équipes n'approuvent pas ces méthodes, tandis qu'un plus grand nombre ont manifesté un fort intérêt pour celles-ci.

- Nous avons ensuite défini le Data Mining et exposé l'apport de ces méthodes dans le cadre du SI sur les ADL. Le Data Mining peut être vu comme un processus d'aide à la décision où les utilisateurs cherchent eux-mêmes des modèles d'interprétation dans les données. On s'accorde encore à définir le Data Mining comme un ensemble de procédures de découverte de connaissances dans les bases de données de gros volume (Knowledge Discovery in Database - KDD). Ces procédures englobent des outils statistiques mais, les méthodes statistiques classiques sont plus descriptives et confirmatives, tandis que les méthodes du Data Mining sont plus exploratoires et décisionnelles.

- Nous avons ensuite montré que le SI relatif aux ADL peut être vu comme un Data Warehouse (entrepôt de données) où les méthodes du Data Mining sont applicables. Nous avons aussi souligné l'importance du cercle vertueux du Data Mining :

> 1- Identifier les données d'intervention
> 2- Utiliser les techniques du Data Mining pour transformer les données en informations utiles
> 3- Transformer les informations en actions concrètes
> 4- Evaluer les résultats

- Nous avons voulu ensuite rendre compte de l'expérience que nous avons acquise dans l'utilisation de certaines méthodes relativement sophistiquées de Data Mining, pour mettre en évidence les qualités et les défauts de ce type d'outils. Nous rapportons ainsi les expériences faites avec les méthodes des prédicteurs neuronaux et de la segmentation, appliquées à des données EHLASS.

- Nous avons constaté que la mise en œuvre de ces outils est relativement lourde (recodage des données, détermination des axes factoriels, paramétrage fin de l'outil, interprétation délicate) et nécessite une certaine expérience statistique. D'autre part, ces méthodes générales ne peuvent répondre qu'à des questions générales. L'importance et le nombre des questions spécifiques issues du SI, nous ont amenés à vouloir mettre au point des procédures décisionnelles mieux adaptées.

- A chaque niveau possible d'exploitation des données sur les ADL correspond une plage d'outils efficaces de caractéristiques différentes. Ainsi pour l'approche macro-accidentologique, on utilisera de préférence des outils de type indicateurs à visée épidémiologique, où la qualité statistique de l'outil est primordiale. Pour l'approche micro-accidentologique, outre les outils classiques de description statistique et de Data Mining, nous avons proposé de développer des outils spécifiques à caractère pragmatique où l'aspect validité statistique est moins essentiel. Ces outils concernent l'exploitation de données non agrégées, plus particulièrement celles issues du recueil dans les services d'urgence.

- Nous avons volontairement construit des outils qui n'utilisent que des variables endogènes au système sans procédures mathématiques complexes. Nos procédures n'utilisent donc que des variables du système EHLASS dans des modèles additifs, à partir de construction de scores et de croisements de variables. Nous avons voulu développer des outils simples d'emploi, facilement compréhensibles et fonctionnant sur les données telles qu'elles sont.

- Nous avons exposé dans le détail les quatre procédures et les méthodes développées en tenant compte des remarques et des suggestions faites dans l'enquête et par le groupe d'experts Data Mining mis en place dans le cadre de ce projet avec nos partenaires danois du NIPH et autrichien de l'Institut Sicher Liben. Ainsi, plutôt que de développer la procédure NGA, comme exposée initialement, nous avons décidé de développer la procédure SSC (Severity SCale) portant sur le même sujet, telle qu'elle a été proposée par nos partenaires danois.

Nous avons ensuite développé les procédures choisies en utilisant le logiciel de statistique le plus répandu dans les équipes : le logiciel SAS (SAS Institute).

C'est ainsi que nous avons mis à disposition des équipes, via le réseau Internet, les fichiers :

- **dmtools.c** : écran d'interface permettant de choisir et de lancer les différentes procédures
- **saa.sas** : la procédure SAA (System of Automated Alert) programmée en SAS
- **ssrdd1.sas** : la procédure SSRD utilisant la distribution des variables pour la variable « Produit impliqué dans l'accident », programmée en SAS
- **ssrdd2.sas** : la procédure SSRD utilisant la distribution des variables pour la variable « Produit ayant causé l'accident », programmée en SAS
- **ssrdp1.sas** : la procédure SSRD utilisant les percentiles pour la variable « Produit impliqué dans l'accident », programmée en SAS
- **ssrdp2.sas** : la procédure SSRD utilisant les percentiles pour la variable « Produit ayant causé l'accident », programmée en SAS
- **ssc.sas** : la procédure SSC (Severity SCale) programmée en SAS

- **Intropro.doc** : présentation en anglais du projet et des procédures
- **saa.doc** : documentation en anglais relative à la procédure SAA
- **ssrd.doc** : documentation en anglais relative aux procédures SSRDxx
- **ssc.doc** : documentation en anglais relative à la procédure SSC
- **fr99v96.zip** : fichier d'essai compacté des données françaises anonymisées de 1999.

- La période de test en grandeur réelle devra être sans doute plus longue que le mois initialement prévu. De l'avis de certaines équipes, elle devrait s'étendre sur un an. Ce n'est qu'à l'issue de cette période que les équipes pourront vraiment juger de la pertinence et de l'utilisabilité des méthodes proposées.

- Quoiqu'il en soit, et quel que soit le degré d'intérêt des procédures développées, nous montrons plus largement dans ce rapport qu'il importe :

- de mieux valoriser les données existantes en renforçant la coopération transnationale;
- de partager les expériences dans l'utilisation d'outils et de logiciels;
- d'utiliser les méthodes éprouvées de Data Mining dans une perspective décisionnelle;
- de développer, par ailleurs, d'autres approches et d'autres outils plus spécifiques;
- de tester les outils simples proposés en tenant compte du contexte de leur développement.

Il est clair, toutefois, qu'aucune technique d'analyse de données ou de Data Mining ne remplacera l'expertise humaine. Mais, l'expertise humaine peut être enrichie par la confrontation aux données réelles, par l'utilisation de logiciels et de méthodes standards de Data Mining et par les résultats issus d'outils nouveaux qui viendront à leur tour guider les bons choix méthodologiques et techniques. Il y a donc fertilisation commune entre l'expertise humaine du domaine et la maîtrise d'un ensemble d'outils d'analyse que nous contribuons dans ce rapport à explorer.

1- Buts de l'étude et méthodologie utilisée

1.1- Buts et contexte

Quels sont le contexte et les buts de l'étude ?

- La Direction Générale SANCO/F/3 nous a chargés d'un travail portant sur l'établissement et le développement d'outils de Data Mining et d'aide à la décision sur les données issues des accidents domestiques et de loisirs (ADL), dans le cadre du programme Prévention des blessures (IPP - Injury Prevention Programme).

- Un tel développement s'avère nécessaire de bien des points de vue :

- L'ancien système EHLASS de recueil d'information a permis de recueillir des données sur plusieurs millions d'accidents domestiques en Europe (près de 5 500 000 de cas fin 1999). Ces données ont été exploitées statistiquement au niveau national. Mais il apparaît que le potentiel d'informations contenues dans les bases nationales et a fortiori dans les bases européennes est largement sous-exploité.

- L'apport des Nouvelles Technologies de l'Information et le développement du réseau Internet s'avèrent déterminants et renouvellent radicalement les stratégies d'utilisation de ces bases de données. Les données agrégées sur les ADL sont maintenant disponibles au niveau européen dans le cadre du réseau EUPHIN-HIEMS (Health Information Exchange and Monitoring System). Les données non agrégées vont l'être dans peu de temps. La disponibilité de ces bases de données incite donc à développer de nouveaux outils d'exploitation.

Ceci d'autant plus que, pour le moment, la plupart des outils d'exploitation mis en œuvre au niveau national sont très simples :

- sélection d'observations sur plusieurs critères
- tris simples et tris croisés de variables
- calcul de moyennes et de fréquences
- mise en forme graphique des résultats

- Or, de grands progrès ont été accomplis dans les techniques d'exploitation des données depuis l'apparition des techniques d'analyses multivariées dans les années 70 (Analyses Factorielles de Correspondances - AFC, Analyses en Composantes Principales - ACP, etc.), jusqu'à l'apparition des concepts et des outils du Data Mining dans les années 90.

- D'autre part, de nouveaux impératifs se sont faits jour dans la gestion des systèmes d'informations de la Commission. Il importe d'améliorer fortement le rapport Coût/Efficacité de ces systèmes et de mieux mettre en évidence la plus-value communautaire générée par leur utilisation.

La capacité à valoriser au mieux les masses d'informations disponibles sur les ADL est devenue une nécessité stratégique pour la Commission comme pour les autorités nationales en charge de ce système d'information.

==> Nos objectifs dans cette étude sont donc :

- de faire un bilan des outils utilisés et des besoins des équipes européennes participant au système en matière d'exploitation de données;

- de faire le point sur l'apport du Data Mining dans le contexte spécifique du Programme IPP et des données disponibles;

- de proposer et développer des outils spécifiques correspondant aux besoins identifiés;

- de mettre à disposition ces outils;

- d'inciter les équipes nationales à utiliser plus systématiquement ces outils qui permettent de mieux mettre en valeur les informations contenues dans leur base.

Il s'agit donc de contribuer à développer une " HLA prevention intelligence " comme on parle du développement d'une " business intelligence ".

Quelle est la situation présente du système d'information sur les ADL ?

Le système européen d'information sur les ADL a évolué considérablement durant ces deux dernières années. Rappelons ici ce qui nous semble être les points principaux de cette évolution :

- Intégration du système de recueil au sein de Injury Prevention Programme - IPP : l'ancien système EHLASS fait maintenant partie d'un programme plus vaste d'action communautaire sur la prévention des blessures volontaires et involontaires. Il s'est doté d'un réseau épidémiologique de correspondants dans les différents Etats membres (Injury epidemiologic network).

- Recentrage du système d'information vers des objectifs de santé publique : l'objectif du programme IPP est clairement de contribuer aux activités de santé publique qui visent à réduire l'incidence des blessures. Cette volonté est illustrée, sur le plan administratif, par la reprise de la gestion du projet par la DG V, puis par la DG SANCO, et le renouvellement de certaines équipes nationales en charge du système qui sont maintenant plus souvent liées à des structures de Santé publique.

- Le souci accru d'une valorisation de l'information : la Commission comme les autorités nationales doivent gérer au mieux le système d'information et sont soucieuses d'accroître son utilisation, son efficacité, ainsi que sa valeur ajoutée européenne.

- L'évolution technique relayant la volonté politique : l'apport des nouvelles technologies de l'information et le développement du réseau Internet ont permis la création de la base européenne de données agrégées (dans le cadre du réseau EUPHIN-HIEMS) et la mise en place prochaine d'une base européenne de données non agrégées sur les ADL.

Cependant des difficultés demeurent dans le fonctionnement de ce système d'information, difficultés liées :

- à la coexistence de 2 méthodologies de recueil (l'une par voie d'enquêtes, l'autre par recueil dans les services d'urgence);

- à la présence d'au moins deux systèmes de codage (versions V86 et V96) nécessitant un transcodage de l'information;

- au caractère représentatif à divers degrés ou non représentatif des informations recueillies selon les Etats;

- à la qualité très hétérogène des données entre Etats, rendant difficile la comparabilité directe;

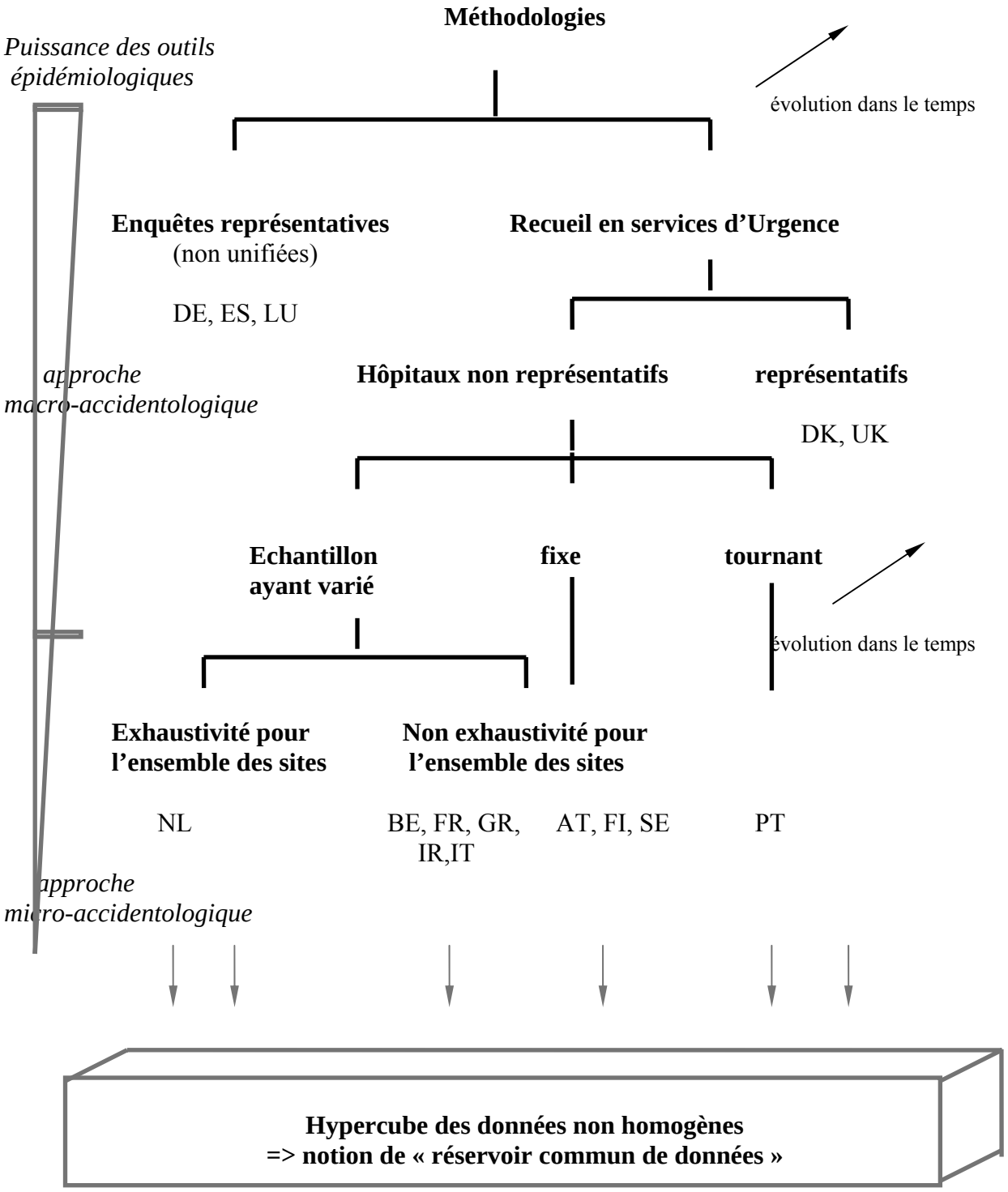
- au fait que certains pays disposent de leur propre système de recueil dans le domaine, alors que d'autres n'en possèdent pas;

- à la présence d'équipes nouvelles, encore peu expérimentées, et d'équipes très avancées dans le domaine, etc.

Il faut avoir présent à l'esprit les éléments de cette situation pour mieux évaluer la nécessité de développer des outils d'exploitation simples d'utilisation et d'interprétation et mieux comprendre les difficultés rencontrées.

Qu'est-ce que le système EHLASS d'un point de vue épidémiologique ?

- C'est un système hybride d'information sur les ADL. Ses buts et les méthodologies utilisées dans les différents Etats ont évolué avec le temps. Si l'on examine pour l'année 1998, par exemple, les méthodologies utilisées, on peut établir le schéma suivant portant sur leurs caractéristiques principales :



Pour les Etats pratiquant le recueil dans les services d'urgence, on pourrait penser que le système EHLASS est un registre d'accidents Or, ce n'est pas un « vrai » registre au sens épidémiologique. En effet, la notion de registre implique un recueil exhaustif sur l'ensemble d'un territoire. Mais, il existe des biais évidents de déclaration dans le présent système :

- tous les ADL ne passent pas par les services d'urgence (problème de l'exhaustivité globale de la méthodologie),

- les hôpitaux faisant partie du système ne sont pas nécessairement représentatifs (problème de la représentativité des hôpitaux),

- tous les ADL passant par ces hôpitaux ne sont pas forcément recueillies (problème de l'exhaustivité sur le site de recueil).

On en conclut qu'avec ce type de recueil, on ne peut connaître avec certitude ni l'incidence, ni la prévalence des ADL au niveau de l'ensemble des Etats.

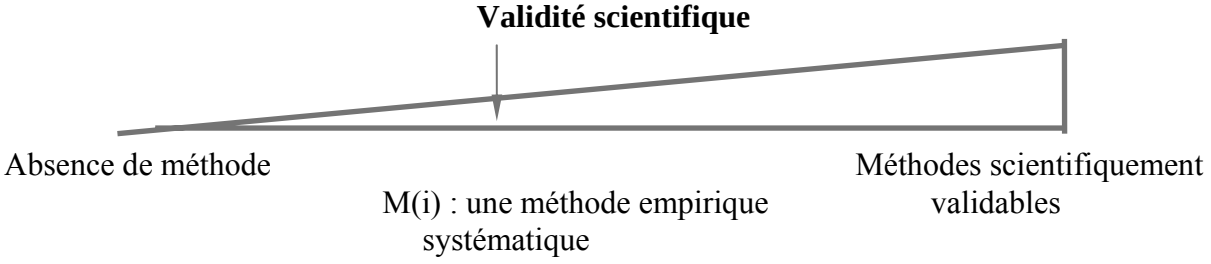
Nous sommes donc en présence d'un système de surveillance (comme il est indiqué dans l'acronyme EHLASS) à visée purement descriptive, sans que cette description soit globalement exhaustive et représentative. Quand on veut traiter les données européennes dans leur ensemble, on ne peut parler que d'un « réservoir commun de données », sachant que ces données sont hétérogènes dans leur qualification et non représentatives.

Quelles sont les implications méthodologiques pour l'exploitation des données ?

Les implications méthodologiques de cette situation sont très importantes. L'épidémiologie descriptive poursuit deux objectifs : mieux connaître le phénomène étudié pour donner des moyens d'actions, faire naître des hypothèses de recherche. Avec ce système d'information, nous sommes clairement dans un contexte où nous ne pouvons effectuer que des démarches descriptives limitées et non de l'épidémiologie analytique.

Avec les données suffisamment qualifiées du système (en termes de représentativité, d'exhaustivité, etc.) nous devons privilégier une approche macro-accidentologique utilisant des outils de type calcul de taux, avec les autres données nous devons privilégier l'approche micro-accidentologique utilisant des outils de type empiriques et systématiques.

Il faudra donc comparer les méthodes empiriques que nous proposons plus loin, non par rapport à des méthodes scientifiquement validables, mais par rapport à une absence de méthode. Ainsi, pour hiérarchiser la dangerosité potentielle des produits, par exemple, on passera de l'absence de méthode systématique (repérage par signalement individuel au cas par cas, articles de journaux, expérience personnelle, intuition, etc.) à une méthode empirique (la procédure SSRD) mais présentant un caractère systématique.



Si l'on se situe du point de vue de l'absence de méthode, la méthode M(i) constituera une amélioration intéressante. Mais, on ne peut pas demander à cette méthode d'être un outil scientifiquement validé, puisque par le contexte même de sa construction et des données auxquelles il s'applique, il ne peut l'être.

Cependant, il existe en épidémiologie, la notion de « bon sens scientifique ». Elle caractérise des outils :

- constants (plusieurs études dont les résultats vont dans le même sens sont nécessaires),
- généralisables (on peut utiliser l'outil dans différents contextes)
- en cohérence avec les connaissances

- Dans la construction de nos outils, nous nous attacherons à respecter ces critères en essayant de faire converger différentes approches (par exemple, en faisant varier les coefficients attribués arbitrairement, en utilisant une approche par percentile, puis une approche par distribution de fréquence, etc.).

Mais, il ne faut pas condamner a priori, au nom de la « rigueur scientifique », ce type d'outil qui peut se révéler très utile dans le choix de stratégies de prévention et la détermination d'actions correctives. Il faut bien sûr bien connaître les limites de leur champ d'application. En aucun cas ce ne seront des outils prédictifs ou explicatifs, mais descriptifs.

1.2- Méthodologie utilisée

Quels ont été les principes de conduite de l'étude ?

Nous avons eu deux préoccupations majeures au cours de ce travail :

Aspect coopératif du projet :

Nous avons voulu associer le plus possible les membres de réseau épidémiologique HLA à ce projet. Pour ce faire :

- nous avons d'abord constitué, en amont, une équipe internationale de projet comportant outre les membres de l'équipe française de BIOSTA, les membres de l'équipe danoise du NIPH et un membre autrichien de l'Institut Sicher Leben.

- nous avons mené une enquête auprès de tous les membres du réseau épidémiologique HLA. Le texte du questionnaire comportait une invitation explicite à nous contacter et à s'associer à notre étude.

- nous avons systématiquement cherché à avoir des contacts par mail et/ou téléphone avec les membres du réseau qui n'avaient pas répondu au questionnaire.

- nous avons constitué un petit groupe d'experts dans le domaine du Data Mining chargé de suivre plus particulièrement le déroulement des travaux.

- enfin, il nous semble avoir tenu le plus grand compte des remarques faites à différentes étapes du projet, soit directement par nos partenaires ou par l'intermédiaire de la coordination des projets IPP assurée par le CSI (Consumer Safety Institute - Amsterdam), quand ces remarques se situaient clairement dans le contexte de développement énoncé au point suivant.

Aspect opérationnel des outils à développer :

Nous avons clairement annoncé, dès la rédaction de notre proposition d'étude, que notre but n'était pas de travailler au développement d'outils statistiques de haut niveau ou de mettre en œuvre des outils accessibles aux seuls statisticiens, mais bien d'aider concrètement la Commission et les autorités nationales dans le déroulement de leur processus décisionnel concernant la prévention des ADL.

Nous avons donc voulu développer des outils opérationnels, simples d'emploi et d'interprétation, à visée pragmatique. Nous avons voulu privilégier l'utilité pratique à la rigueur scientifique, quand celle-ci empêche, en définitive, l'utilisation de données chèrement acquises, déjà disponibles, directement exploitables et recelant un fort potentiel informatif.

Quelles ont été la méthodologie utilisée et la chronologie des travaux ?

Travaux effectués

Début officiel des travaux du projetn°1999/ IPP/1006 le 01/01/2000

Conformément à la méthodologie exposée dans notre réponse à l'appel d'offres de la Commission, nous avons ordonné nos travaux en 5 phases. Voici les travaux que nous avons réalisés dans le cadre de l'accomplissement de ces phases :

Etape 1 : Analyse des outils existants

- Mise en place de l'équipe de travail et répartition des travaux - janvier 2000 :

BIOSTA - France
SICHER LEBEN - Autriche
NATIONAL INSTITUTE OF PUBLIC HEALTH - Danemark

- Prise de connaissance du contexte de l'étude et de l'évolution des outils de Data Mining - janvier 2000.

- Rédaction commune d'un questionnaire (version française et anglaise v2.2) - janvier 2000.

- Envoi par mail des questionnaires en plusieurs formats à l'ensemble des membres du réseau épidémiologique HLA - 01/02/2000.

- Mise en place du groupe d'experts sur le Data Mining - mars 2000. Le groupe est composé de :

Henning Bay-Nielsen - NIPH - DK
Birthe Frimodt-Moller - NPIH- DK
Robert Bauer -SICHER LEBEN - AT
Fons Blankendaal - CONSUMER SAFETY INSTITUTE - NL
Hugh Magee - DEPARTMENT OF HEALTH AND CHILDREN- IR
Marc Nectoux - BIOSTA - FR

- Adaptation du questionnaire pour une meilleure compréhension : corrections, introduction d'exemples => Questionnaire v2.3 - mars 2000.

- Envoi par mail du nouveau questionnaire aux équipes n'ayant pas encore répondu - mars 2000.

- Traçage du questionnaire auprès des équipes n'ayant pas répondu (téléphone et mail) mars-avril 2000.

- Contacts directs réalisés par l'équipe française :

10-11/02/2000 : Participation du Dr Christine DUVAL à la première réunion de **l'Injury Epidemiology Network** - Amsterdam
15/03/2000 : Visite à Hubert ISNARD - InVS - Paris
27/03/2000 : Visite à Bernard Le Goff -DG SANCO - Luxembourg
27/04/2000 : Visite à Fons Blankendaal - Amsterdam
27/04/2000 : Intervention devant les chercheurs du CSI - Amsterdam

Remarque : il est certain que l'utilisation systématique du courrier électronique change profondément la manière de travailler et remplace une grande partie des déplacements et des contacts directs précédemment nécessaires.

Etape 2 : Détermination et analyse des outils souhaités

- Echange d'informations, d'observations et de données avec l'équipe danoise sur le score de gravité
février-mars 2000.

- Evolution des méthodologies proposées : SSDR (Score Synthétique de Dangerosité Relative), NGA (Note de Gravité de l'Accident) et SCENAR (la méthode des Scénarios) - février-mars-avril 2000.

- Dépouillement des questionnaires reçus - avril 2000

- Détermination des outils souhaités - mai 2000 :

Réunion de travail du groupe des experts Data Mining à Luxembourg le 2 mai 2000

- Analyse des réponses aux questionnaires
- Discussion sur les outils proposés
- Détermination des outils souhaités

Etape n°3 : Développement des procédures sélectionnées

- Développement de la procédure SSDR adaptée à la V96 en SAS - mai 2000
- Développement de la procédure NGA adaptée à la V96 en SAS - mai 2000

Rédaction du rapport intermédiaire (mai 2000)

- Développement de la procédure SAA adaptée à la V96 (juin - juillet 2000)
- Mise au point de la méthode SCENAR (juillet - août 2000)
- Participation à la réunion des chefs de projet IPP (27 et 28 juin à Luxembourg)

Etape n°4 : Validation auprès d'un groupe d'utilisateurs et diffusion des outils

- Participation à la seconde réunion de l'Injury Epidemiology Network du 6/09/2000 à Paris (M. Nectoux, P. Uziel).

- Actions correctives et explicatives entreprises :

  - Diffusion lors de cette réunion de documents explicatifs, dont :
    « Epidemiologic aspect in the HLA information system »,
    « Methodologic implications », etc.

  - Consultation d'experts en épidémiologie :

    - Dr COSTE (specialist of the construction and validation of Composite Ratio Scale - Département de Biostatistique et d'information médicale - Hôpital Cochin - Paris)
    - Dr ISNARD (Institut de la Veille Sanitaire - St Maurice)

  - Attente de propositions concrètes de la part des membres réseau pour la modification des procédures proposées.

  - Proposition d'un allongement de la période de test.

  - Développement d'une nouvelle méthode de calcul du SSDR utilisant la distribution des fréquences des variables.

  - Adoption et programmation de la procédure « A three level severity scale for use in the EHLASS » dont les auteurs sont nos partenaires danois du NIPH.

- Enfin, nous avons diffusé le 11 octobre 2000, par le réseau Internet à l'ensemble des chefs de projet EHLID membres de l'Injury Epidemiology Network, les procédures SAS mises au point ainsi que les documentations afférentes en anglais.

Etape n°5 : Production du projet de rapport final

- Compléments et modifications des chapitres du rapport intermédiaire (octobre 2000).
- Rédaction des chapitres manquant au projet de rapport final (octobre 2000).

Remise du projet de rapport final
et des procédures SAS avec leur documentation en anglais
(31 octobre 2000)

Quel est le contenu du rapport final ?

Notre rapport final rend compte du travail accompli. Ainsi :

- **le Chapitre 2** fournit les résultats détaillés par Etat de notre enquête et les conclusions que nous en avons tirées;

- **le Chapitre 3** donne plusieurs définitions du Data Mining et précise la place et l'apport de ces méthodes dans le cadre du programme IPP;

- **le Chapitre 4** rend compte d'expériences liées à l'utilisation de certains outils standards des logiciels de Data Mining (réseau de neurones, segmentation);

- **le Chapitre 5** donne les principes de développement et de mise à disposition des procédures que nous avons choisies de développer;

- **le Chapitre 6** expose les principes de la procédure intitulée Score Synthétique de Dangerosité Relative - SSDR;

- **le Chapitre 7** expose les principes de la procédure intitulée Système d'Alerte Automatisée - SAA;

- **le Chapitre 8** expose les principes de la méthode des Scénarios - SCENAR;

- **le Chapitre 9** expose les principes de la procédure intitulée Note de Gravité de l'Accident - NGA et les raisons de son abandon au profit du Severity Scale (NIPH);

- **le Chapitre 10** détaille les procédures SAS développées et les réponses aux critiques et suggestions exposées au Chapitre 2;

- **le Chapitre 11** expose les conclusions de nos travaux.

Enfin, le **Chapitre " Les conclusions opérationnelles "** qui précède ces pages a pour but de souligner ce qui paraît être, à nos yeux, les principaux points de l'étude (en 3 pages).

2- Les résultats de l'enquête

2.1- Présentation de l'enquête

Quels étaient les buts de l'enquête ?

Nous voulions mener une enquête préalable auprès de l'ensemble des membres du réseau épidémiologique "Accidents domestiques et de loisirs " de l'ensemble des Etats membres.

Cette enquête comportait trois volets, correspondant à trois buts distincts :

Volet n°1 - Etat actuel : Il s'agissait de connaître les outils et les plates-formes informatiques utilisés pour l'exploitation des données de l'ancien système EHLASS.

Volet n°2 - Souhaits : Il s'agissait de connaître les outils souhaités par les différentes équipes pour une meilleure exploitation des données, dans le cadre du nouveau système d'information sur les accidents domestiques et de loisirs (EHLID).

Volet n°3 - Réactions et critiques à nos propositions : Nous exposions brièvement 4 procédures décisionnelles simples que nous proposions de développer et de mettre à la disposition des équipes et de la Commission. Nous voulions connaître les réactions, les critiques et les suggestions des membres du réseau à leur sujet.

Enfin, un but complémentaire mais essentiel de l'enquête était de vouloir associer concrètement l'ensemble des partenaires du réseau épidémiologique à ce travail de réflexion.

Comment s'est déroulée l'enquête ?

- Nous avons mis au point le questionnaire avec nos partenaires (version anglaise fournie en Annexe n°1).

- Une première version (V2.2) en anglais a été diffusée le 1er février 2000 par mail, avant la première réunion du réseau épidémiologique à Amsterdam les 10 et 11 février 2000.

- A la suite de remarques faites par l'équipe française, nous avons ajouté à chacune des 4 procédures décrites une page donnant un exemple concret d'utilisation. Nous avons renvoyé par mail et courrier postal cette nouvelle version (V2.3) aux équipes qui n'avaient pas répondu précédemment.

- L'Assistant du projet a été ensuite contacté systématiquement par téléphone et mails, durant le mois d'avril 2000, les correspondants n'ayant pas retourné leur questionnaire.

En définitive nous avons reçu les réponses suivantes :

Questionnaires complétés :

Etat	Contact	Organisation	Date
Autriche	Robert Bauer	Institut Sicher Leben	16/04/2000
Danemark	Marie Kruse	National Institute of Public Health	17/03/2000
France	Christine Duval	Direction Générale de la Santé	29/02/2000
Irlande	Tim McCarthy	Dpt of Health and Children	03/02/2000
Pays-Bas	Fons Blankendaal	Consumer Safety Institute	27/04/2000
Luxembourg	Yolande Wagener	Division de la Médecine Préventive et Sociale	27/04/2000
Belgique	Peter Hoft	Ministerie van de Vlaamse gemeenschap	05/05/2000
Espagne	Nelson Castro Gil	Instituto Nacional del Consumo	23/05/2000
Grèce	Eleni Petridou	Cerepri	13/06/2000
Commission	Bernard Le Goff	DG SANCO	27/03/2000

Réponses sans questionnaire :

Royaume-Uni Maria Cody 18/04/2000	Apologies. We can answer some questions such as how we have used Ehlass data but we do not actually understand all the questions and also we are in the middle of changing our computer system. Currently Ehlass data runs on an Ingress database which we send to the Commission in their chosen format. From the middle of May we will have Ehlass data running on our database which uses " Viper from Smart focus " as the search engine and outputs in windows NT format. I.e. Excel, Access or Word. At that time we should also be able to e-mail you Ehlass data up to 1998. 1999 data should be available by June. I will be attending the next IPP in Luxembourg and would be happy to talk about this then.
Allemagne Annelie Henter 27/03/2000	Die EHLASS-Daten wurden in Deutschland im Rahmen repräsentativer Haushaltsbefragungen erhoben (d. h. als Stichprobendaten), mit der Software Quanvert statistisch ein- und mehrdimensional ausgewertet und ursachenspezifisch analysiert. Die deutschen EHLASS-Daten (Stichprobe) können den Datensätzen anderer Länder nicht zugerechnet werden. Da eine derartige Auswertung hinsichtlich der Unfallprävention nur sinnvoll ist, wenn zunächst gezielt Unfallschwerpunkte ermittelt werden, diese Schwerpunkte hinsichtlich Struktur und Unfallbedingungen sowie der Unfallursachen analysiert werden, wurde auch so in der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin bei der Analyse der Unfallverletzungen in Heim und Freizeit vorgegangen. Die Ergebnisse wurden interpretiert. Vorschläge für Unfallverhütungsmaßnahmen konnten abgeleitet werden und im Rahmen gezielter Aufklärungskampagnen präsentiert werden. Die Beteiligung von Deutschland ist noch unklar.

Nous avons donc disposé de 10 questionnaires complétés sur 16 réponses possibles (15 Etats membres + DG SANCO) et de 2 réponses expliquant la non participation à l'enquête.

Nous fournissons ci-après les réponses détaillées des Etats membres pour chaque question.

2.2- Les résultats détaillés par Etats

Quelles sont les réponses portant sur le Thème n°1 ?

Description des outils et des plates-formes matérielles et logicielles utilisées pour l'exploitation de l'ancien système EHLASS.

Q.1.1 - Quels outils informatiques utilisez-vous pour exploiter les données issues de l'ancien système EHLASS ?

Logiciels statistiques :

Austria	SPSS
Belgium	DBase - SAS
Denmark	SAS
Espagne	Pulsar 2.2 sous Windows 98
France	SAS sous VMS + BMDP sous VMS
Grèce	SAS
Ireland	SAS
Luxembourg	SPSS
Netherlands	SPSS
DG SANCO	SAS

Outils graphiques :

Austria	Microsoft Excel (97)
Belgium	SAS Graph, Excel
Denmark	SAS Graph, Harvard Graphics
Espagne	Harvard Graphics
France	SAS Graph sous VMS
Grèce	Excel
Ireland	Harvard Graphics
Netherlands	Excel
DG SANCO	SAS Graph

Autres outils :

Austria	Microsoft Access & Excel (97)
Belgium	Excel
Denmark	Excel, Epi-Info
Espagne	Excel
France	Epi-Info sous DOS pour le masque de saisie et une exploitation simple
Grèce	Word, Excel, Power Point, Paradox
Ireland	Microsoft Word + Excel (97)
Luxembourg	Excel

Netherlands	Power Point

Remarque :

Luxembourg	ILReS don't exploit such data at all. We only do data collection by CATI using ODINE (from NIPO Amsterdam). For preparing control tables we use TABLES (from GESS Hamburg), but IAE Nancy used SPSS in making the annual report. The national analysis of the EHLASS data 1992-1999 will be done by the Division of Preventive Medecine and the Centre de Recherche Santé. The used system will be SPSS and EXCEL

Q.1.2 - Sur quelles plates-formes informatiques (matériel et système d'exploitation) les utilisez-vous ?

Austria	Windows 95
Denmark	Windows NT
Espagne	Windows
France	VMS sur DEC, PC DOS et Windows
Grèce	Windows 98
Ireland	Microsoft Word + Excel (97) sous Windows 95
Luxembourg	Windows 95 et Windows NT
Netherlands	Windows NT
DG SANCO	SUN Solaris et Windows 98

Q.1.3 - Avez-vous développé des outils ou des procédures spécifiques pour exploiter vos données dans le cadre de l'ancien système EHLASS ?

Q.1.4 - Pouvez-vous nous décrire brièvement les buts et le fonctionnement de ces outils ?

Austria	Yes : - SPSS macro to generate a set of standard tables and graphs (reporting tool)
Belgium	No
Denmark	No
Espagne	No
France	Oui : - Procédure d'interrogation interactive des données sous DOS - Serveur Web pour interroger dynamiquement les données structurées sous ORACLE - Programme d'édition en clair des enregistrements sous DOS - Les 4 outils d'exploitation proposés dans le questionnaire
Grèce	- Program for data entry and an error checking system
Ireland	No
Luxembourg	Non
Netherlands	Yes, not directly for EHLASS, but for our system LIS :

	- HLA Scenario method. - HLA Indicators : Absolute number, Incidence, % Fracture, % Hospital Admission, Lenght of Stay, % Mortality, Trend, Costs. - Priority model : Principal Component Analysis (PCA) to limit amount of variables. PCA will give factor values for variables (indicators). The Priority Model will structure the scenarios on the basis of indicators values.
DG SANCO	Non

Q.1.5 - Quelles sont les plates-formes utilisées par ces outils (langage, système d'exploitation, matériel) ?

Austria	- SPSS macro / Win 95
France	- Procédure d'interrogation interactive des données : programme Fortran sous DOS - Serveur Web pour interroger dynamiquement les données : ORACLE - Programme d'édition en clair des enregistrements : programme Fortran sous DOS - Les 4 outils d'exploitation proposés dans le questionnaire : - SSDR : procédure SAS - AAS : Epi-Info + programme Fortran - SNA : procédure SAS - SCENAR : programmes BMDP
Grèce	- Windows 95, Paradox, Microsoft SQ Server 7
Netherlands	- SPSS on Windows NT

Q.1.6 - A votre avis, ces outils sont-ils diffusables et exploitables par d'autres Etats membres ?

Austria	- May be to help producing a standard report
France	- Nous pensons que ces outils sont utilisables par les autres équipes. C'est le but de ce projet de proposer ces outils en tenant compte des apports des autres Etats membres.
Grèce	- Definitely, as they have been repeatedly tested to provide clean data that we have used in the analyses of data that have been published in peer-reviewed journals. - It should be also mentioned that these programs have been useful in the analysis of the descriptions (free texts) of the accidents which are only available in greek. - Finally these programs have been used in order to convert to the different coding systems that we are currently working on EHLASS, ICD-9, ICD-10, NOMESCO and Pilot-Testing ICECI Codes.
Netherlands	Yes, we think we could adapt these kind of methods to the EHLASS data.

Quelles sont les réponses portant sur le Thème n°2 ?

Description des outils souhaités pour une meilleure exploitation du nouveau système d'information sur les ADL ?

Q.2.1 - Pensez-vous que l'on puisse améliorer l'exploitation des données ? , Comment ?

Austria	- Yes, by : - a flexible query tool (e.g. MS Access forms) - a set of tools for special question (data mining catalogue); e.g. - rare or interesting case detector, - trend detector - dangerousity indicator for products or groups of products
Belgium	- Availability of source-database on " information highway "
Denmark	- Analytical : Knowledge of representativity problems, Incidence calculation - Technical : Better user interfaces, Easy search function in free text. - Other : Secure and easy comparison of countries.
Espagne	- Poseer un cuerpo minimo de datos comparables entre todos los Estados. Además este cuerpo mínimo debería estar sometido a un mínimo de cruces de variables sociodemográficas - por ejemplo sexo, edad, tamaño de habitat, mecanismo ... - Potenciar la red de trabajo de comunicación entre los miembros del grupo IPP, a través del correo electroníco; - Desarrollar sistemas que nos permitan comprobar o evaluar los peligros o riesgos que conllevan determinados productos. - Destacar la importancia de la difusión de los datos ante la opinión pública.
France	- Il faut utiliser plus souvent les données par collaboration avec les autres équipes et aussi en utilisant de nouveaux outils d'exploitation qui soient simples d'emploi. - Au niveau européen : il faut pouvoir échanger rapidement et simplement sur : - les problèmes rencontrés - les échanges d'expériences - les sujets d'étude - les alertes
Grèce	- On the planning level : 1- We conduct regular staff meetings with the hospital interviewers to discuss any « bizzare » accidents they came across and we try to monitor through special registries the frequency of these events. If there is such a suspicion we go through formal procedure to add extra questions for a specific time period aiming to substantiate the problem. After a testing period we decide whether we should modify our standard questionnaire form in order to collect information on the issue more systematically. This decision implies new training of the personnel and changes in the rubrics used but on the other hand, this practice has led to early publications on hazardous consumers products or writing of reports that have been used to raise awareness or to defend consumer protection actions.

	2- We participate in the ANEC (european secratariat for representation of consumers in the standardization) bi-annual meetings and this has provided us the opportunity to experiment with real life situations and face a series of problems that had to be solved. We regularly extract information from our databe on newly identified health risks. - On the analysis level : 1- Preparation of the annual report 1.1- During this preparation we try to identify outlying frequencies of the annual incidences of tracing conditions and accidents taking into account the seasonality patterns. 1.2- We are experimenting on deriving conclusions according to the results of principal components and factor analysis of some types of accidents. 1.3- We conduct time trend analyses for the tracer categories of accidents but due to relatively frequent changes in the coding systems used, even in our database, it is very time consuming to perform compatibility in order to have comparable data. 2- Publications-scientific presentations 2.1- Use of readily available EHLID data or combination of these data with the ones collected on ad-hoc basis has provided useful feedback on how to both improve the quality of the selected data and to exploit the system 9 see relevant list of publications in the report. 2.2- Provision of data to the increasing requests we have on behalf of the physicians working in the collaborating hospitals presents another opportunity for the exploitation of the EHLID dataset. 3- Exploitation of data on EU level 3.1- As stated above, we provide data to ANEC in order to be used in the standardization committees. 3.2- We provide data according to orders made by EU institutions. 3.3- We collaborate in projects envisioned through IPP. 3.4- According to the process followed by other EU funded actions, we suggest that a permanent committee for the analysis and exploitation of the EHLID data is established, consisting of national EHLID project leaders who have both the knowledge of the data collection and coding system as well as the responsability to make inferences from these data. Opinion experts should be invited to participate on ad-hoc basis.
Ireland	- The Irish Dataset is too small to lend itself readily to wide usefulness. It would be more readily applicable to a wider range of applications if the sample size was larger.
Luxembourg	- National analysis in a interdisciplinary approach (including other national partners for.example : national insurances, sport associations, schools, etc.).
Netherlands	- We need a web based interface on top of scenario's/priorities etc. to easily monitor developments and goals set. OLAP reporting tool (eq. Powerplay by Cognos)
DG SANCO	- Il faut absolument chercher à exploiter les données existantes des bases HLA agrégées dans le cadre de HIEMS, ainsi que les données des

	futures bases HLA non agrégées accessibles par le réseau pour : - valider ces données et les procédures mises en place - produire de la valeur ajoutée au niveau d'une exploitation européenne par l'utilisation d'outils de Data Mining intégrés au système déjà existant.

Q.2.2 - Quels types d'outils vous semble utile dans cette perspective ?

Austria	- See Q.2.1
Belgium	- Tool for an alert on EHLASS products
Denmark	- Provision of information on population/catchment area - Graphic interfaces - Search tools - A good database with sufficient information and " metadata " for comparisons
Espagne	- La utilización de un único Manual de codificación. - Establecer un sistema de comunicación rápido y seguro de intercambio de información entre los miembros del grupo IPP. - Establecer mecanismos experimentales y reservados sin difusión, que evaluen la gravedad de un producto. - Desarrollar mecanismo de representatividad de los datos recogidos a nivel de cada Estado participante en la recogida de los datos. - Elaboración de informes por parte de los Servicios de la Comisión que compendien todos los datos o informaciones de los paises - Elaboración de campañas de información o divulgacíon de los resultados con los accidentes domésticos
France	- Mettre au point des outils simples d'exploitation. - Utilisation systématique du courrier électronique et des échanges de fichiers de données entre les équipes européennes.
Grèce	1- Representativity 1.1- Typical statistical processes should be followed and validations of the hospitalisations estimated through EHLASS with population and hospital discharge data from other sources by at least : age gender, urbanisation, type of injury. 1.2- Capture-recapture techniques in order to compare data with those from other sources. 1.3- It should be taken into account. However, that no sample is strictly representative and that changes in the sampling procedure should be constantly made if there are populations movements or changes in the exposure patterns. Therefore it should first be questionned what is the purpose of data collection procedure and what level of representativity could satisfy these needs. Collecting data is always a very expensive process and wise spending is a useful guiding principle.
Netherlands	See Q.2.1
DG SANCO	- La Commission est intéressée par le développement d'outils portant sur : - la recherche des produits défectueux

	- le développement d'indicateurs sur la qualité des données : représentativité, exhaustivité, comparabilité, etc. - le développement d'indicateurs sur la sévérité des accidents - les aspects d'évaluation du coût économique des accidents - le développement d'indicateurs sur la sécurité des produits

Q.2.3 - Avez-vous déjà utilisé des logiciels de Data Mining ? , Si oui lesquels ?, Qu'en avez-vous pensé ?

Austria	- No, no dedicted Data Mining software package, just general purpose statistical tools
Belgium	No
Denmark	No
Espagne	- Logiciel de statistique Pulsar 2.2
France	- Oui. Nous avons utilisé **SPAD Data Mining**. - Nous avons fait des études multivariées utilisant la segmentation, la typologie et les réseaux de neurones. Cela semble intéressant mais pas très facilement exploitable par des équipes non spécialisées. - C'est pour cela qu'il est important de développer des outils simples pour répondre à des besoins spécifiques. Par exemple : déterminer les produits défectueux, trouver des groupes homogènes d'accidents, etc.
Grèce	- As described above -in summary these are based on the identification of any unusual time, place or time-place clustering. - The main problems we are faced with, are the ones known to be linked to this type of analysis.
Ireland	- **Business Objects** is used in this office but the Data Mining tool is not widely used nor is it required in most cases.
Netherlands	**SPSS modules**. On my opinion they are bad in presenting data on Internet.
DG SANCO	Non

Quelles sont les réponses portant sur le Thème n°3 ?

Réactions, critiques et suggestions face à l'exposé de 4 procédures décisionnelles simples.

Remarques :

Luxembourg*	- As we have not yet done a national analysis of the data, we don't know at this stage which could be the best methodology of analyse for our data. Beside this we wonder whether the comparability of the data at a European level is guaranteed. For Luxembourg at least, we think that our data, bases on household interviews can only be compared with those from Spain and Germany, using the same procedure.

** Le Luxembourg ne fournit pas d'autres commentaires ou de notes pour cette partie du questionnaire.*

P1-Procédure " Score Synthétique de Dangerosité Relative - SSDR "

L'exposé succinct de la procédure proposée dans l'enquête est fourni dans le questionnaire en Annexe.

Q.3.1.1 - Les réactions et critiques :

Austria	- We already use the Synthetic Severity Score in our annual report. However, it is very abstract and it is not very frequently cited. Some possible difficulties with the construction of the SSDR : - HR and ALS are age dependent, age should therefore be considered - HR requires a representative sample in the respect (which is not necessarily the case) - HR differs among health care systems, international comparability might be hampered - No. of deaths not available within the HLA data collection system - Rare products which might be interesting are under-represented
Belgium	- EFF may be dependant on exposure probability more than dangerosity. - RH may be dependant on other determinants rather than product : age, comorbidity, local hospitalisation politics (e.g. in case of concussion). - ADS may be dependant on other factors not directly related to the product involved. - It can be expected that only a minor percentage of the injuries result in hospitalisation. The impact of this " high number resulting in high impact of single less impact " will probably not show.
Denmark	- There may be cultural differences in the frequency of use of different products. - The measure is very product-oriented, many accidents are not product related. Also, the relevant product should perhaps be the " product causing injury " and not the product involved. - Regarding the measure itself, we find that hospitalisations and frequency are not very logic indicators of danger and we find you should ask yourself : What is danger and how is danger measured ? To us, beddays or frequency of accidents are not very good indicators of danger, since a product may be very frequently used and very frequently involved in accidents but the accidents or injuries need not be very severe. For example, we may have many accidents with roller-skates or in-liners because they are very popular, so the frequency is high but the injuries are mainly superficial. On the contrary, a toy can be considered dangerous enough to be called back from the entire world market due to one accident with the toy. Similarly with length of stay, you may have to consider whether a broken leg (long ADS) is more dangerous than a poisoning (short ADS).
Espagne	- No se debería utilizar como elemento valorable el producto que interviene en el accidente (EFF). Este elemento debería estar compuesto, de manera exclusiva, por el producto causante de la lesión

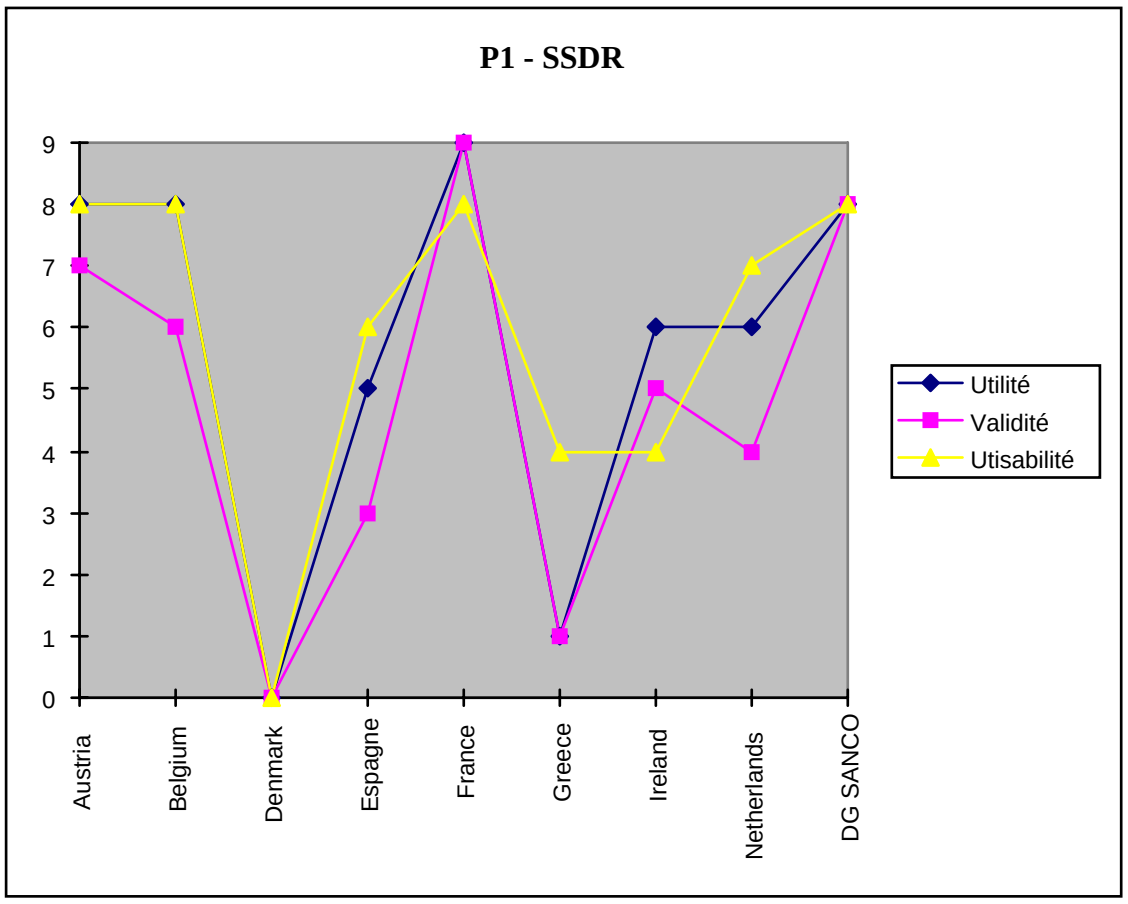
	al sujeto que lo padece. - El sistema inicialmente expuesto por BIOSTA nos parece correcto, no obstante deberá tener en cuenta sistemas de recogida de la información como el español. - Que se entiende por hospitalizados y duración media de la estancia en un hospital ? - El sistema que se elabore deberá ser lo más fiable, desde el punto de vista científico posible, no obstante deberá ser de facil aplicación.
France	- Cet outil paraît utile dans la mesure où : - Il permet de mettre en relief les informations dont on dispose sur les produits - Il permet de comparer les données entre Etats - Il a été conçu en fonction de l'expérience. Nous l'utilisons en France, sous une forme proche, depuis plusieurs années.
Grèce	- We believe that SSDR is difficult to be interpreted both by professionals and lay people, because as a rule the frequency of product use is unknow. Nevertheless, one can get information on the amount of injuries attributed to a specific product, although this amount is a function of both the frequency of use and the severity of the injury. There is no justification for the proposed weights in the respective formula. Although the frequency of use can be estimated through some sort of a control group. - Some minor points : - Are you going to consider the second or the third object ? For example : in sports injuries, let's say in basketball, an object can hit a player and cause a fall injury. In this instance, the first object will coded will be the ground but, the most important cause is the second object involved i.e. the basketball. - The frequency of the category « other object » can be very high. - How are you going to handle missing data in any of three variables ? - Do you include cases with outcome 'death' ? If yes, will you separate death that happened before any hospitalisation than these that happened after small or long duration of hospitalisation ? - Examining just objects can produce false score. Will it be worthwhile to introduce any other variable (mechanism or type of injury ?). - Some rare events will take small score even they are very dangerous.
Ireland	- This is an interesting proposition but there is a risk that results can be skewed by extraordinary events. E.g. a very prolonged hospital stay for what would normally be a minor event that would not result in hospitalisation. This is more harmful, obviously, when you have a small size data sample as is the case in Ireland.
Netherlands	- How validate the multiplication coefficients ? We use Principal Component Analysis (PCA) to derive coefficients.
DG SANCO	- Comment justifier le système de pondération des variables (2 pour Effectif et 2 pour RH) ?

Q.3.1.2 - Les suggestions :

Austria	- The preconditions for the SSDR and possible difficulties should be mentioned in the description of the tool; for frequent products age should be considered.
Belgium	- At least correction of RH and ADS for age.
Espagne	- Que su aplicación se desarrolle de forma experimental durante un periodo de tiempo que permite evaluar su fiabilidad y utilidad. - Que se respecte los principicios de la confidencialidad. - Sistema de fácil aplicación. - Elaboración de una manual de trabajo que clarifique cada uno de los conceptos a utilizar para la elaboración del SSRD.
France	- Il faut mieux expliquer le fonctionnement du score. - Il faudrait pouvoir sortir facilement, pour un produit donné, un tableau donnant le score de gravité dans les différents Etats.
Grèce	- This is a special project that needs to be developped.
Ireland	- Determine an approach where the skewage is kept to a minimum.
Netherlands	- It's necessary to have validate tools (human expertise ?)

Q.3.1.3- Les notes :

SSRD	Utilité décisionnelle	Validité logique	Utilisabilité
Austria	8	7	8
Belgium	8	6	8
Denmark	0	0	0
Espagne	5	3	6
France	9	8	9
Grèce	1	1	4
Ireland	6	5	4
Netherlands	6	4	7
DG SANCO	8	8	8
Moyenne	**5,67**	**4,78**	**5,89**
Ecart-type	3,02	2,90	2,60



P2-Procédure " Système d'Alerte Automatisée - SAA "

L'exposé succinct de la procédure proposée dans l'enquête est fourni dans le questionnaire en Annexe.

Q.3.2.1- Les réactions et critiques :

Austria	- This seems a very useful tool ; in our practice it would be used primarily for trend analysis – what is now, and how was it last year (or last season) ?.
Belgium	- The alert system may be triggered by single " mass accidents " of accidents with normally low incidence figures or by seasonal variations in exposure to some products.
Denmark	- The measure is interesting. It is crucial that you have up-dated data for this type of analysis. - I am very interested to know more about how you are going to analyse the free text. - From our perspective it is not so relevant but for a consumer safety perspective is may prove very useful.
Espagne	- El comportamiento estocácisto o probabilisto de los accidentes domésticos y de ocio habría que analizarlo a lo largo de un periodo largo, por ejemplo, un año; de otra forma podría estar sometido a un gran error estadístico.
France	- Cela permet de se poser des questions sur les variations de fréquence. - C'est intéressant surtout pour les codes produits. - On peut comparer les variations entre Etats. - Il ne faut pas oublier que, pour le moment, il y a un décalage de plusieurs semaines entre le recueil des données dans les hôpitaux et leur mise à disposition dans la base => alerte différée. - L'alerte détectée n'est pas forcément due à une augmentation de la gravité des accidents.
Grèce	- Use of proportional indicators is satisfactory under certain assumptions, for example : - Has the absolute number of injuries increased or decresased during the respective period ? - Is there an indication that the total number of accidents may vary due to specific reasons such as seasonability or increase in use of a specific product ? - How are you going to treat injuries of low frequency ? i.e. the frequency of injuries due to a specific product may be found 1% during Period 1 and 2% during Period 2. Can you say that the frequency has doubled and therefore further consideration is needed ? - Are all numerical equal increases going to be treated in the same way ? For example a 5% increase in poisonings due to medicinal products is of much higher importance and needs further investigation, whereas the same increase in basketball injuries may be somewhat unimportant. - Are you going to assume that the exposure time remains constant during the period under consideration ?

Ireland	- This effort is very much in keeping with the EHLASS principles. Text searching on EHLASS fields can be unreliable if events are poorly described.
Netherlands	- You must use trends over a longer period of time : Stepwise linear regression over a period of 10 years. Trends are very important !
DG SANCO	- Ce système a surtout une utilité s'il est utilisé sur des données HLA récentes, pour mettre en valeur les variations. Ce sont les Etats membres qui possèdent ces données. Au niveau de la Commission, la mise à jour des bases nationales ne pourra se faire qu'un rythme assez lent (annuel ou semestriel). L'aspect alerte est donc moins important du fait du décalage dans la mise à jour des données. Cet outil est à mettre à la disposition des équipes nationales.

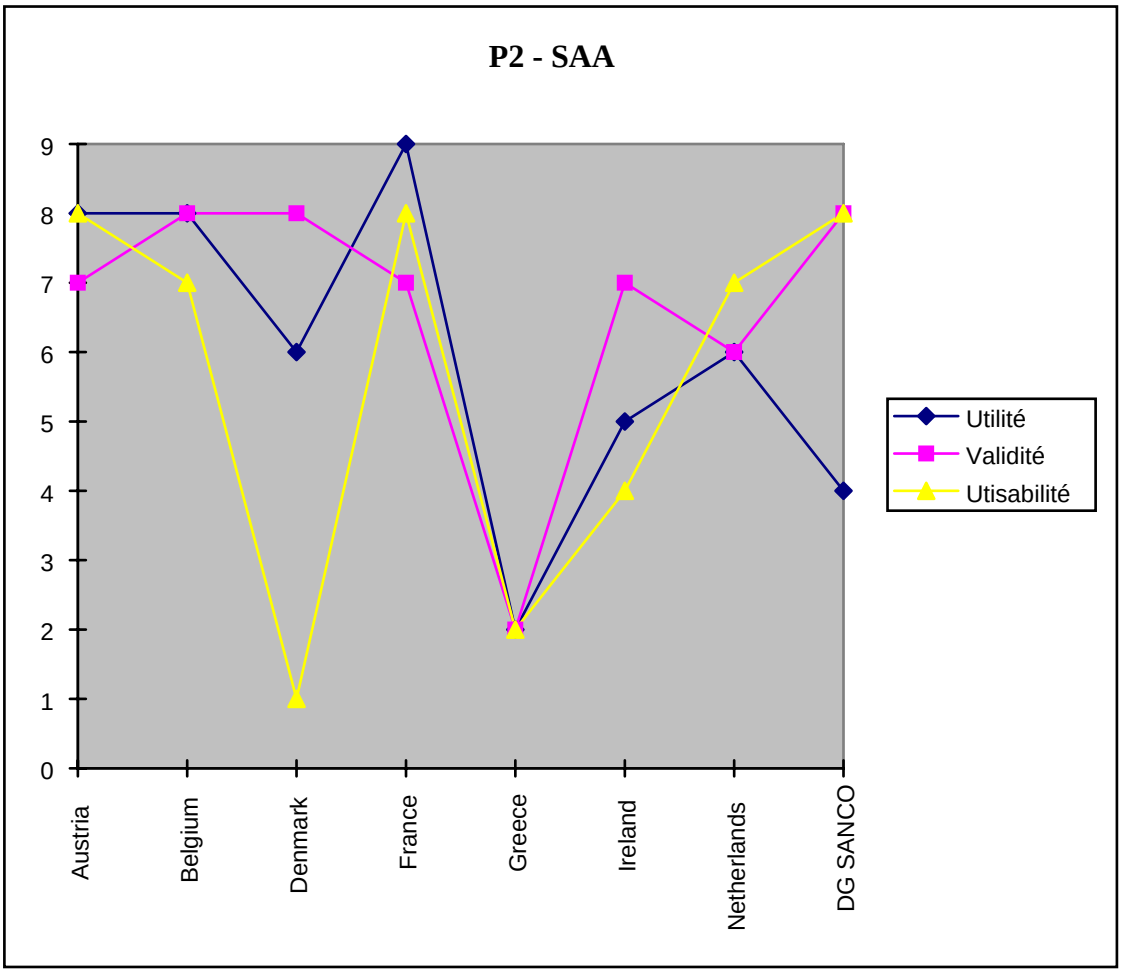
Q.3.2.2- Les suggestions :

Austria	- Are newly emerging products also discovered, at which treshold ?
Belgium	- Try to construct a " reference period " which is not a fixed point-estimation but a mix of historical variation.
Denmark	- Needs more refinement - Feasibility analysis of the free text as analysis variable
Espagne	- Descartar este sistema de evaluación
France	- L'utilisation de cette alerte devrait être systématique. - Il faudrait faire un schéma pour mieux expliquer le fonctionnement du système. - Dans le système NEISS, mis au point aux Etats-Unis, il semble qu'il y ait un outil comparable. - Il faudrait coupler ce système avec d'autres outils pour un examen approfondi des données. Cela ne donne que des pistes d'interrogation.
Grèce	- In general, the statistical problem is that time, person and time-person clustering should be examined and this is among the most demanding statistical procedures, particularly when referring to proportional indicators. Our department has a special expertise in this field as shown by a series of publications on time-place clustering of childhood leukemia and case cross over studies on injuries and testing of abrupt increases in the context of environmental studies.
Netherlands	Find a way to pinpoint to new scenario's i.e. yet unknown field's for injury prevention

Q.3.2.3- Les notes :

SAA	Utilité décisionnelle	Validité logique	Utilisabilité
Austria	8	7	8
Belgium	8	8	7
Denmark	6	8	1
France	9	7	8
Grèce	2	2	2
Ireland	5	7	4
Netherlands	6	6	7
DG SANCO	4	8	8
Moyenne Ecart-type	**6,00** 2,18	**6,63** 1,87	**5,63** 2,69

*Pas de notes attribuées par l'Espagne pour cet outil



P3-Procédure " Note de Gravité de l'Accident - NGA "

L'exposé succinct de la procédure proposée est fourni dans le questionnaire en Annexe.

Q.3.3.1- Les réactions et critiques :

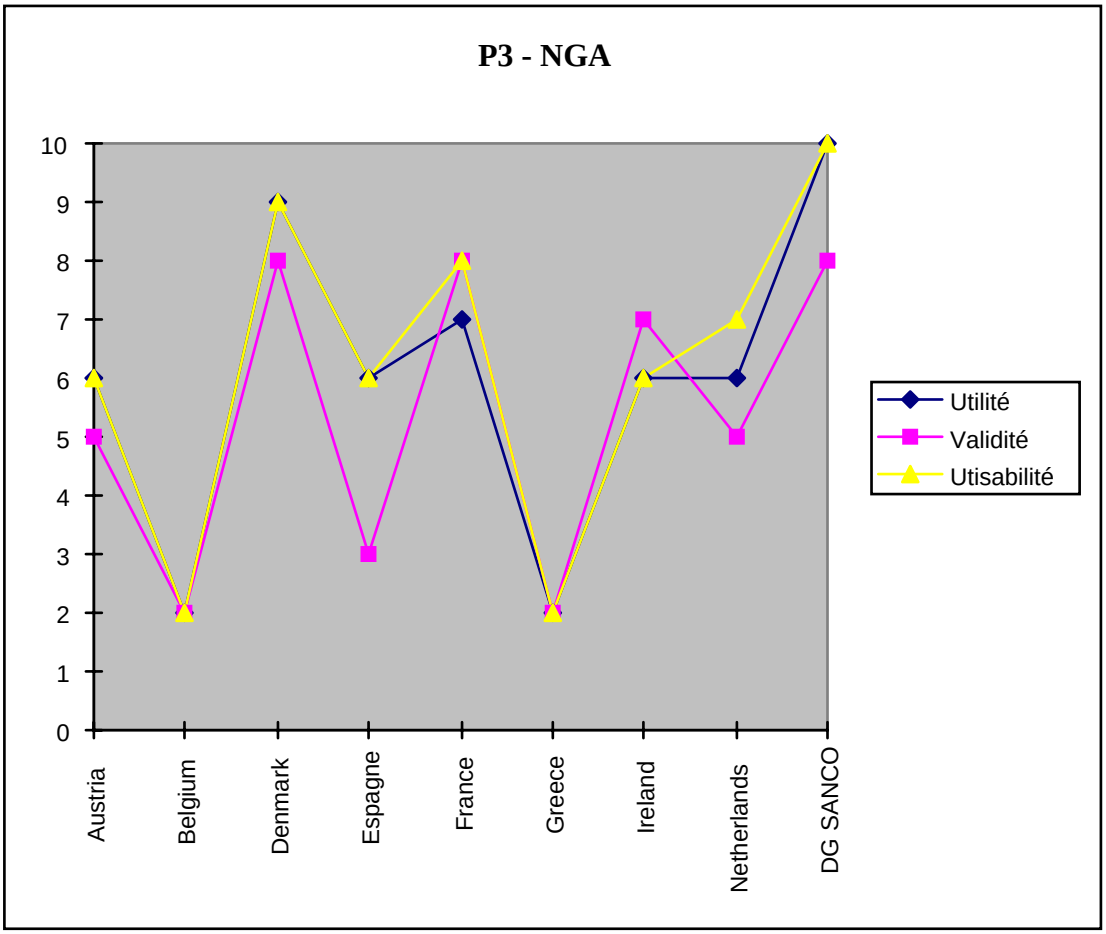
Austria	- The SNA seems to serve purpose as the SSDR, however, it is even more abstract as the scores do not come " naturally ".
Belgium	- TRT may be dependant on local health services politics. At least in Belgium the TRT scores will not reflect severity in a valid way (5 and 9 will predominate 3 as some control examination appointment is common pratice even in relatively minor injuries). - DH : see comment Q.3.1.1 - TI is not specific enough.
Denmark	- The most interesting and useful measure in our view. - " Other types of injury " should be excluded
Espagne	- Inicialmente este sistema presenta una característica que con anterioridad defendiamos, su sencillez para la aplicación. - Creemos que falta un elemento para su posterior correlación, el producto causante del daño. Habría que incluir este factor y establecer su peso en el conjunto del proceso.
France	- C'est intéressant pour les comparaisons transnationales. - Ce score n'est pas relié à d'autres scores reconnus internationalement, mais ce n'est pas gênant. Il est spécifique du système. On sait bien qu'avec ce système HLA on ne recueille pas de données socio-économiques, de données sur les coûts des accidents et sur les séquelles. Il faut donc utiliser les données existantes. - C'est un objectif interne.
Grèce	- There are several problems associated with this score : 1- Deaths are missing. 2- The first variable in the formula is culture and country specific and this can be taken into account. 3- If « duration of hospitalisation » is one of the examined variables, this assumes that all injuries where hospitalised : this introduces essential interaction (see David Clayton's terminology) and creates substantial analytical difficulties.
Ireland	- The inclusion of the field Body Part is very important.
Netherlands	- It's necessary to validate the scores. We like the idea but validation is very important.
DG SANCO	- Ce type d'indicateur nous paraît utile. Il a fait l'objet de critiques lors d'une réunion à Copenhague dont il faudra tenir compte. Il faut notamment envisager comment on pourrait faire intervenir les variables Age et Partie du corps lésée dans la détermination de cette note.

Q.3.3.2- Les suggestions :

Austria	- Same as for the SSDR
Belgium	- Without some knowledge of physicological parameters (RTS) and real severity estimate (AIS or ISS) this kind of severity note will be of marginal use.
Denmark	- Needs validation/testing with data
France	- Donner des exemples. - Voir comment on pourrait introduire le facteur âge : une fracture chez une personne âgée est plus grave qu'une fracture chez un enfant.
Grèce	- There are already too many severity scores that have been developped to try to add another, imperfect one. The suggestion would be to establish an IPP Committee that would choose among the existing ones and try to validate them in EU MS.
Ireland	- The Age of the patient may also need to be factored in.

Q.3.3.3- Les notes :

NGA	Utilité décisionnelle	Validité logique	Utilisabilité
Austria	6	5	6
Belgium	2	2	2
Denmark	9	8	9
Espagne	6	3	6
France	7	8	8
Grèce	2	2	2
Ireland	6	7	6
Netherlands	6	5	7
DG SANCO	10	8	10
Moyenne	**6,00**	**5,33**	**6,22**
Ecart-type	2,54	2,40	2,62



P4-Procédure " Méthode des scénarios - SCENAR "

L'exposé succinct de la procédure proposée est fourni dans le questionnaire en Annexe.

Q.3.4.1- Les réactions et critiques :

Austria	- In general this seems to be a very interesting tool as it takes a lot of information into account. Without examples, however, it is hard the follow and to judge this procedure.
Belgium	- see comments on SNA.
Denmark	- While the Data Mining process and the grouping of variables are good ideas and the process seem useful, we find that the overall categorisation of variables into " causes ", " circumstances " etc. may give rise to some conceptual criticism or that it may not be completely consistent. - Generally we find the process interesting.
Espagne	- Procedimiento interesante. No obstante nos parce más descriptivo que un sistema que genere un parámetro medible, el de la peligrosidad. Habría que profundizar más en este procedimiento o ampliar la información.
France	- Sur le principe, cet outil paraît intéressant pour cibler des actions. - Il devrait permettre de mettre en évidence des scénarios oubliés comme " ingestion de corps étrangers chez les personnes âgées ". - Les explications données sont trop dans un jargon statistique, donc pas suffisamment claires.
Grèce	- This approach is useful only when large numbers of diverse injuries from diverse population groups are accounted for.
Ireland	- There seems to be a lot of permutations while the sample size may not be very large.
Netherlands	- Scenario's are very important ! - Ask people who use the data to suggest topics for scenario's - Develop a method to adjust or supplement the set of scenario's (flexible set)
DG SANCO	- Cette méthode nous paraît utile pour la détermination des groupes homogènes d'accidents.

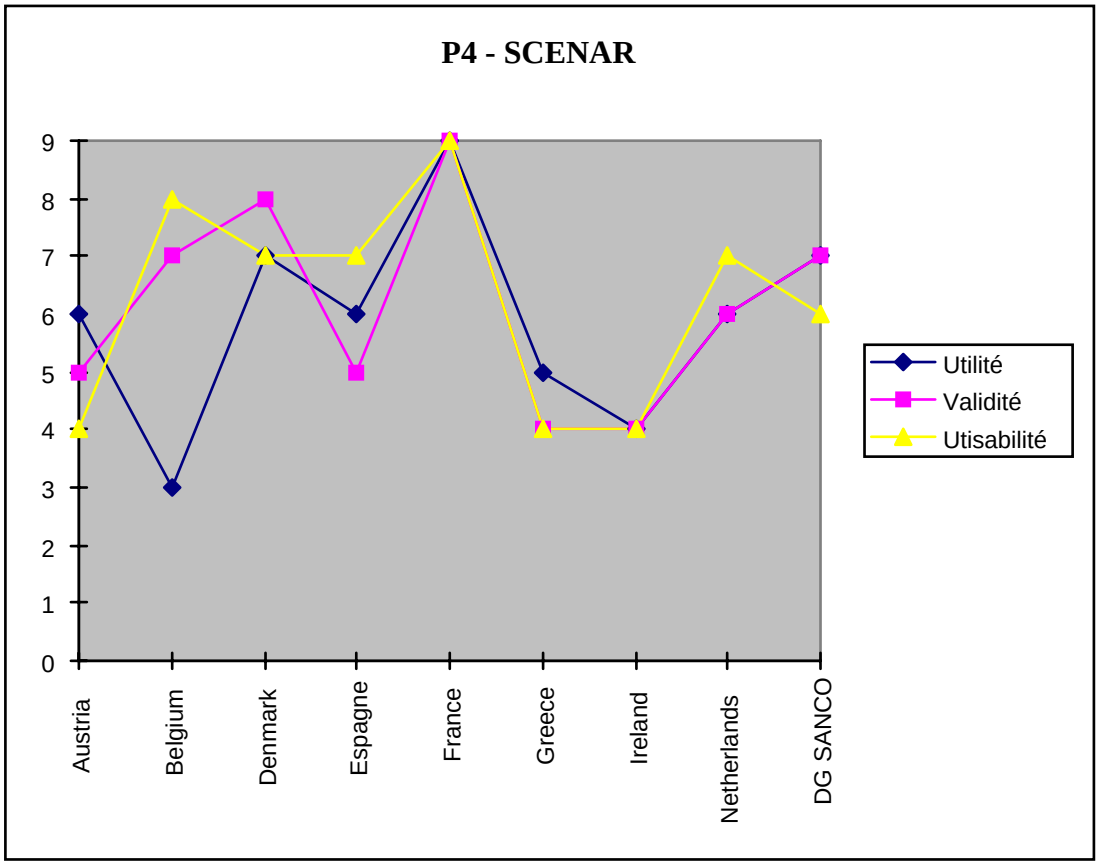
Q.3.4.2- Les suggestions :

Austria	- The added value to a standard reporting tools (e.g. simple cross tabulations) should demonstrated
Belgium	- This may be useful for hypothesis-inducing descriptive analysis only, not for decision-making. It should be complemented with more in-depth study.
France	- Il faudrait visualiser le processus avec un schéma. - Il faudrait donner des exemples.
Grèce	- It should be first tested by public health workers in other population

	settings to increase the reliability of the process. - It could be also useful to produce sets of combinations of mechanism by place and then to distinguish major categories, after you have set a limit for less frequent injuries.
Netherlands	- Start with requirements for scenario's. - Total set of scenario's : - basic set : Product x Mechanism (how does HLA happen ?) - on top of it : Place x Activity (for interventions much more meaningful)

Q.3.4.3- Les notes :

SCENAR	Utilité décisionnelle	Validité logique	Utilisabilité
Austria	6	5	4
Belgium	3	7	8
Denmark	7	8	7
Espagne	6	5	7
France	9	9	9
Grèce	5	4	4
Ireland	4	4	4
Netherlands	6	6	7
DG SANCO	7	7	6
Moyenne	**5,89**	**6,11**	**6,22**
Ecart-type	1,66	1,66	1,75



2.3- Conclusions tirées de l'enquête

Quelles conclusions peut-on tirer de cette enquête concernant ?

Le taux de réponse :

Le nombre des questionnaires complétés n'est pas très élevé (10/16), ceci malgré les nombreuses relances. Plusieurs explications peuvent être avancées :

- Le questionnaire implique une bonne connaissance du système d'information et une pratique de son exploitation. On peut comprendre que des équipes nouvelles ne puissent répondre facilement au thème 3, par exemple, portant sur les outils proposés.

- Les méthodes proposées dans le thème 3 s'adressent plutôt aux équipes utilisant le recueil dans les services d'urgence.

- Certaines équipes disposent apparemment de peu de temps pour répondre. Elles ont pu aussi trouver le questionnaire trop complexe. Il était cependant facile d'envoyer des mails pour expliquer ces raisons.

- La coopération européenne dans le domaine est un phénomène récent et la réactivité des équipes est encore à construire. Les études menées dans le cadre du programme IPP participent à la mise en place de cette nouvelle forme de coopération active.

Le thème de l'existant :

- Les logiciels statistiques utilisés sont : SAS (5 citations dont la DG SANCO), SPSS (3), BMDP (1), Pulsar 2.2 (1).
- Les logiciels graphiques utilisés sont : SAS Graph (4), Excel (4) et Harvard Graphics (3).
- Les autres outils logiciels cités sont : Excel (7), Epi-Info (2), Power Point (2), Paradox (1).
- Les systèmes d'exploitation cités sont : Windows 95, 97 ou NT (8), Solaris (1), VMS (1).

- Le nombre de procédures spécifiques déjà développées dans les Etats membres est assez faible, plus faible que ce que nous pensions au début de cette étude. Des procédures spécifiques ont été développées surtout en France, aux Pays-Bas et en Grèce.

Le thème des souhaits :

- Les souhaits exprimés concernent :

- la disposition d'un noyau minimum de données comparables (ES) et de croisements avec des données socio-économiques (ES);
- des outils de mesure de la qualité (DG), de la représentativité (DK, ES, GR) et de la comparabilité des données;
- l'interrogation des données : disponibilité sur le réseau (BE) et interrogation dynamique,

- recherche dans le texte libre (AT, DK), interface graphique (DK);
- des outils simples de reporting (AT, FR);
- des outils de détection des cas rares (AT, GR);
- des outils d'analyse de la dangerosité des produits (AT, DG, ES);
- des outils d'alerte sur les produits et de recherche de produits défectueux (BE, FR, DG);
- des outils de mesure de la sévérité des accidents (DG);
- des outils d'analyse en tendance (AT);
- des outils de calcul d'incidence (DK);
- des outils de construction de scénarios et de hiérarchisation des priorités (NL);
- des outils concernant l'évaluation du coût économique des ADL (DG);
- l'échange rapide d'informations (sujets d'études, alertes, etc.) (ES, FR);
- des outils de diffusion de l'information et des résultats (ES).

- Il faut noter que ces souhaits portent sur des thèmes globaux très divers et n'ont pas été accompagnés de propositions concrètes de développement, sauf dans deux cas :

- l'équipe danoise a développé sa propre proposition concernant l'établissement d'une échelle de sévérité des accidents;

- l'équipe hollandaise a développé ses propres outils d'évaluation du coût des ADL qui font appel à des informations externes au système EHLASS et de détermination des scénarios d'accidents et d'un « Priority model ».

- Concernant l'utilisation de logiciels de Data Mining : peu d'équipes en ont l'expérience. Sont cités : Business Object en Ireland, le module SPSS de Data Mining aux Pays-Bas et SPAD Data Mining en France.

Le thème des outils proposés :

Tableau récapitulatif des notes attribuées :

	Utilité décisionnelle	**Validité logique**	**Utilisabilité**
SSRD			
Moyenne	**5,67**	**4,78**	**5,89**
Ecart-type	3,02	2,90	2,60
NGA			
Moyenne	**6,00**	**5,33**	**6,22**
Ecart-type	2,54	2,40	2,62
SAA			
Moyenne	**6,00**	**6,63**	**5,63**
Ecart-type	2,18	1,87	2,69
SCENAR			
Moyenne	**5,89**	**6,11**	**6,22**
Ecart-type	1,66	1,66	1,75

Au vu des notes moyennes, l'acceptation des différents outils est relativement bonne, mais l'écart-type est important, indiquant par là une forte variabilité des notes. En effet, l'équipe grecque (Mme Petridou et associés) et le secrétariat de projet (M. Rogmans) ont critiqué fortement les outils SSRD et SAA. Mais, ils ne semblent pas avoir tenu compte du contexte de développement spécifique que nous avons choisi dès la rédaction de notre proposition d'étude. Nous reviendrons sur ce point dans le chapitre 5.

Procédure P1- SSRD :

L'utilité décisionnelle est reconnue par l'ensemble des équipes ayant répondu (moyenne 5,67), sauf par les équipes danoise et grecque. Les critiques portent essentiellement sur les points suivants :

- le taux d'hospitalisation et la durée moyenne de séjour sont des variables dépendantes de l'âge et de la structure du système de soins;

- le système d'information ne donne qu'une idée très partielle du nombre des décès puisque seuls les décès durant le séjour à l'hôpital sont recueillis;

- les produits rares sont sous-estimés (fréquence peu élevée d'apparition). Ils peuvent cependant être très dangereux;

- la variable « Effectif » est liée à la fréquence d'utilisation plus qu'à la dangerosité. Il faudrait avoir des informations sur les fréquences d'utilisation des classes de produits;

- il y a des différences culturelles entre Etats dans la fréquence d'utilisation des produits;

- beaucoup d'accidents n'ont pas de lien de causalité directe avec le produit impliqué;

- l'affectation des coefficients multiplicateurs paraît arbitraire;

- comment faire intervenir ensemble les trois codes produits ?;

- la fréquence du code « Autre produit » peut être forte;

- il faudrait faire intervenir d'autres variables (mécanismes, type de lésion ?);

- il faudrait créer un groupe de travail sur ce seul projet;

- il faut un mode d'emploi explicitant bien la méthode et les concepts utilisés.

Procédure P2- SAA :

L'utilité décisionnelle et la validité logique sont en moyenne reconnues (moyenne respectivement à 6,00 et 6,63). Les critiques portent essentiellement sur les points suivants :

- on pourrait analyser l'évolution non pas par comparaison entre une période donnée et une période de référence, mais entre une période donnée et plusieurs périodes de référence (analyse en tendance);

- il peut y avoir des variations saisonnières dans la survenue de certains types d'accidents;

- il faut utiliser cette méthode sur des données assez récentes pour garder le caractère d'alerte de la procédure. Cette méthode est donc plus adaptée à l'utilisation dans les Etats membres qu'au niveau européen;

- on peut commettre des erreurs statistiques si l'on n'utilise pas les mêmes périodes ou des périodes trop restreintes;

- le nombre absolu d'accidents peut varier entre les 2 périodes;

- comment traiter les fréquences faibles d'accidents ?

- une augmentation de 5% des cas d'intoxication médicamenteuse est plus importante qu'une augmentation de 5% des accidents de basket-ball.

- la durée d'exposition n'est pas forcément la même entre les deux périodes.

Procédure P3- NGA :

L'utilité décisionnelle et l'utilisabilité sont en moyenne reconnues (moyenne respectivement à 6,00 et 6,22). Les critiques portent essentiellement sur les points suivants :

- la variable « Traitement » est liée à la politique de soins appliquée localement et au comportement culturel;

- la variable « Durée d'hospitalisation » est dépendante de la structure du système de soins;

- tous les accidents ne donnent pas lieu à une hospitalisation, cela introduit des interactions;

- le but de cette Note semble proche de la méthode du SSRD;

- il faut exclure la modalité « Autre type de lésion »;

- il faudrait inclure la variable « Partie du corps lésée ».

- il faudrait faire intervenir la variable « Produit causant la lésion »;

- les données de mortalité sont exclues du calcul;

- il est important de valider ce score.

- beaucoup de scores de sévérité ont déjà été développés. Il vaudrait mieux créer un Comité au sein de IPP pour choisir lequel utiliser.

Procédure P4- SCENAR :

La validité logique et l'utilisabilité sont en moyenne reconnues (moyenne respectivement à 6,11 et 6,22). Les critiques portent essentiellement sur les points suivants :

- la méthode peut conduire à un nombre très élevé de combinaisons de circonstances;

- la valeur ajoutée de la méthode par rapport à un simple croisement de variables doit être démontrée;

- il faudrait pouvoir introduire des scénarios externes fournis par l'expertise humaine et combiner ces informations avec des données de coûts.

- cette approche est utile quand on a affaire à un grand nombre d'accidents très divers.

- on pourrait utiliser le croisement des variables « Mécanisme » et « Lieu », puis distinguer les principales combinaisons en fonction de leur fréquence.

Nous avons tenu compte des remarques exposées dans la mesure où elles s'inscrivaient dans le cadre du développement que nous avons choisi et clairement explicité dès le début du projet. Nous répondons à l'ensemble de ces objections et suggestions dans le Chapitre 10 , une fois exposées les procédures développées après leur prise en compte.

3- Définitions et apports du Data Mining

« *Comment trouver un diamant dans un tas de charbon sans se salir les mains* »

3.1- Définitions du Data Mining

Qu'est-ce que le Data Mining ?

Plusieurs définitions ont été proposées (*voir l'ouvrage : Le Data Mining - René Lefébure et Gilles Venturi - Editions Eyrolles*). Le Data Mining serait :

> **- " la découverte de nouvelles corrélations, tendances et modèles par le tamisage d'un grand nombre de données " (John Page);**
>
> **- " un processus d'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans les données " (Kamran Parsaye);**
>
> **- " l'extraction d'informations originales, auparavant inconnues, potentiellement utiles à partir des données " (Frawley et Piateski-Shapiro).**
>
> **- "un processus de mise à jour de nouvelles corrélations, tendances et de modèles significatifs par un passage au crible des bases de données volumineuses, et par l'utilisation de modèles d'identification technique aussi bien statistiques que mathématiques." (Erick Brethenoux, Gartner Group)**

- SAS Institute définit le Data Mining comme " le processus d'exploration et de modélisation des gisements de données permettant de découvrir des informations/indicateurs inconnus pour obtenir des avantages concurrentiels ".

Généralement, on s'accorde à définir le Data Mining comme un ensemble de procédures de découverte de connaissances dans les bases de données (Knowledge Discovery in Database - KDD).

- Ces procédures englobent des outils statistiques mais, les méthodes statistiques classiques sont plus descriptives et confirmatives, tandis que les méthodes du Data Mining sont plus exploratoires (recherche de modèles sous-jacents inconnus) et décisionnelles. Les outils actuels de Data Mining reprennent souvent des outils statistiques parfaitement connus depuis longtemps (comme l'Analyse Factorielle des Correspondances - AFC, la segmentation, etc.) mais les incluent dans des démarches à valeur ajoutée décisionnelle. Ces outils sont théoriquement accessibles aux " utilisateurs métiers ", non-spécialistes de la statistique, par l'emploi de logiciels spécifiques relativement conviviaux.

- Le but est de découvrir des tendances cachées dans l'amas des données (la " mine " de données) et les modèles qui les traversent. Ces outils servent à déterminer des profils de comportement, à découvrir des règles, à évaluer des risques.

Quels sont les objectifs des méthodes de Data Mining ?

On peut regrouper les objectifs des méthodes de Data Mining en 4 grandes fonctions :

- **classifier** : on examine les caractéristiques d'un nouvel objet pour l'affecter à une classe prédéfinie. Les classes sont bien caractérisées et on possède un fichier d'apprentissage avec des exemples préclassés. On construit alors une fonction qui permettra d'affecter à telle ou telle classe un nouvel individu.

- **estimer** : la classification se rapporte à des événements discrets (le patient à été ou non hospitalisé). L'estimation, elle, porte sur des variables continues (par exemple : la durée d'hospitalisation).

- **segmenter** : il s'agit de déterminer quelles observations vont naturellement ensemble sans privilégier aucune variable. On segmente une population hétérogène en un certain nombre de sous-groupes plus homogènes (les clusters). Dans ce cas, les classes ne sont pas prédéfinies.

- **prédire** : cette fonction est proche de la classification ou de l'estimation, mais les observations sont classées selon un comportement ou une valeur estimée futurs. Les techniques précédentes peuvent être adaptées à la prédiction au moyen d'exemples d'apprentissage où la valeur à prédire est déjà connue. Le modèle, construit sur les données d'exemples et appliqué à de nouvelles données, permet de prédire un comportement futur.

Le choix de la technique dépendra de la nature du problème posé et du type de données dont on dispose.

Quelles sont les techniques spécifiques du Data Mining ?

Notre but n'est pas ici de fournir une présentation des techniques du Data Mining. De nombreux ouvrages spécialisés y sont consacrés. Rappelons juste que les techniques de Data mining comprennent des outils comme :

- les raisonnements à base de cas ou raisonnements basés sur la mémoire (RBM),
- le traitement analytique en ligne (TAEL),
- les arbres de décision,
- les segmentations,
- les méthodes des associations,
- les algorithmes génétiques,
- les réseaux baysiens,
- les réseaux de neurones, etc.

Ces techniques sont regroupées dans des logiciels ayant des niveaux de fonctionnalité, de coût, de performance et de souplesse d'utilisation très variés (Datamind, SPAD-D, Predict, Classpad, Knowledge Seeker, etc.). Nous verrons dans le chapitre suivant des applications concrètes de certaines de ces techniques aux données sur les ADL.

Qu'est-ce que le cercle vertueux du Data Mining ?

On peut résumer la méthodologie du cercle vertueux du Data Mining à la mise en œuvre dans les systèmes d'informations de la succession des tâches suivantes :

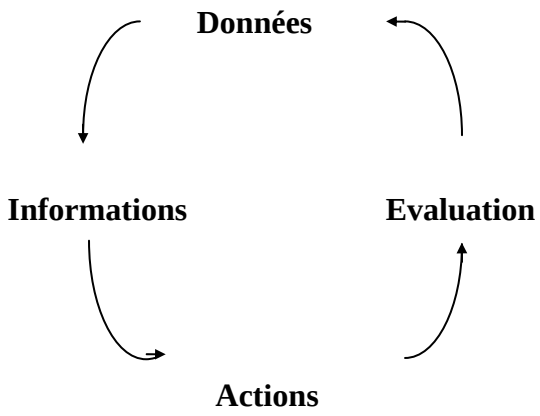
1- Identifier les données d'intervention

2- Utiliser les techniques du Data Mining pour transformer les données en informations utiles

3- Transformer les informations en actions concrètes

4- Evaluer les résultats

On peut illustrer ce cercle vertueux par le schéma ci-dessous :



Le Data Mining est donc plus un processus qu'un ensemble d'outils épars.

C'est dans cet esprit qu'il faut comprendre le titre notre étude « Etablissement et développement d'outils de Data Mining... ». Les outils spécifiques que nous voulons développer, s'ils n'ont pas toutes les caractéristiques techniques des outils de Data Mining, participent de ce processus et de cette volonté de construire des outils pragmatiques d'aide à la décision.

Le Système d'Information sur les ADL est-il un " Data Warehouse " ?

Les outils du Data Mining sont réputés être applicables sur des ensembles de données de volume important : les " Data Warehouse " (entrepôts de données) qui sont des bases de données décisionnelles, détaillées, orientées sujet, non volatiles et historisées.

Or, les bases de données sur les ADL déjà constituées (bases nationales, base agrégée européenne et base non agrégée) peuvent être vues comme un véritable " Data Warehouse " européen, car elles répondent à ces critères de définition.

Ainsi, nous pouvons inventorier les différents types de fichiers :

- Les fichiers de **données brutes** ou non agrégées : ce sont les fichiers par année et par Etat de l'ensemble des Etats participant au système. Ces fichiers ont été contrôlés et transcodés au format V96 par l'Institut allemand LOëGD;

- les fichiers de **données historisées** : ce sont ces mêmes fichiers complets avant transformation. On a donc des fichiers au format V86, au format V96 issus du recueil dans les services d'urgence, et les fichiers issus des enquêtes.

- les fichiers de **données agrégées** : ce sont les fichiers en sortie des procédures d'agrégation développées par LOëGD. Ces fichiers ont été créés en vue de leur introduction dans l'application HIEMS, permettant l'interrogation en ligne des données agrégées européennes.

- Les **données de pilotage** : ce sont les informations décrivant les méthodes de recueil, leurs caractéristiques méthodologiques et les performances du système (coût de fonctionnement, nombre de cas, qualité des données, etc.).

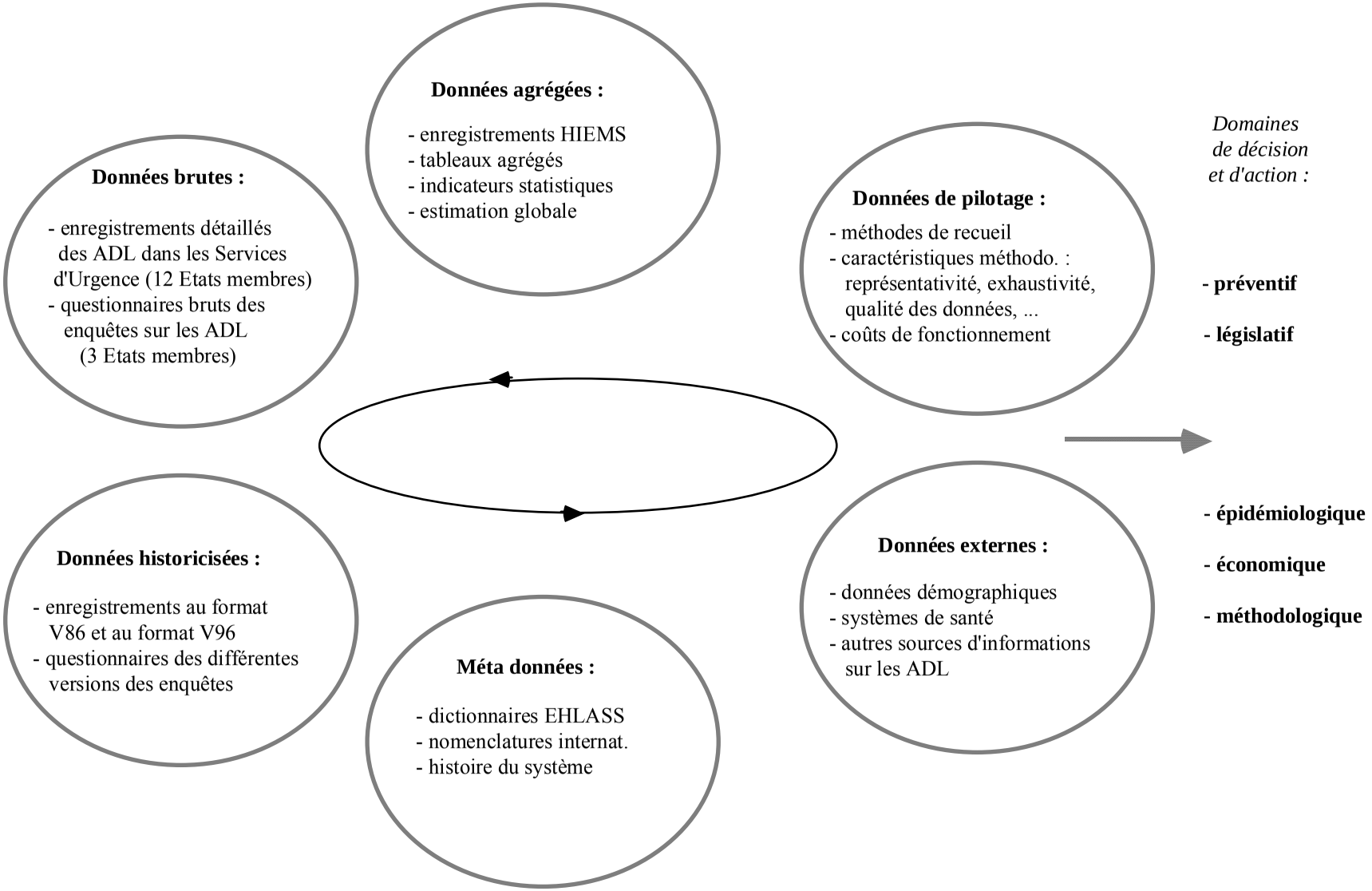
- Les **Méta données** : ce sont des fichiers décrivant les données : dictionnaires des variables, nomenclatures utilisées, les nomenclatures internationales liées au système, etc.

- Les **données externes** : ce sont des données utiles pour l'interprétation des résultats, mais non directement liées au SI lui-même : données démographiques, informations sur les systèmes de santé dans les différents Etats, informations sur les autres sources de données sur les ADL.

Ces différents fichiers sont ou devront être mis en réseau et être accessibles par les utilisateurs agréés du réseau épidémiologique HLA et les experts internationaux.

Le schéma de la page suivante montre comment le SI peut être structuré comme un Data Warehouse.

Injury Prevention Programme - HLA
Le Système d'Information vu comme un DataWarehouse - Les fichiers



Quels outils techniques sont en relation avec le " Data Warehouse " ?

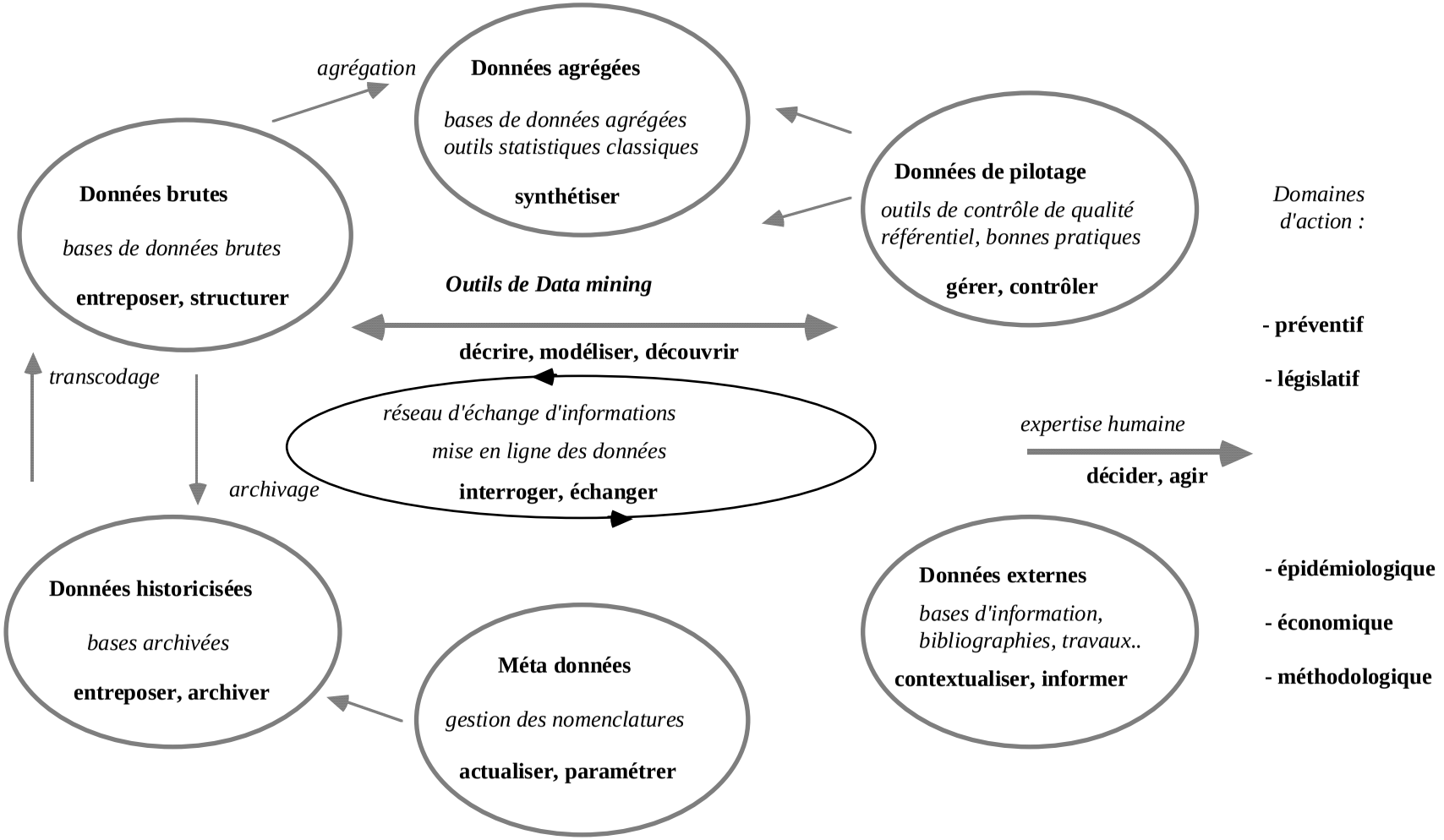
Nous pouvons maintenant situer dans l'optique du Data Warehouse l'ensemble des actions à accomplir et des outils techniques à mettre en œuvre :

Type de fichiers	Actions	Types d'outils
Données brutes	entreposer, structurer, exploiter	SGBD + outils statistiques classiques **+ outils de Data Mining**
Données historisées	entreposer, archiver	SGBD
Données agrégées	synthétiser, exploiter	SGBD + outils de reporting et de statistiques classiques
Données de pilotage	gérer, contrôler	Outils de contrôle de qualité,
Méta données	paramétrer le SI	Dictionnaires de données
Données externes	contextualiser, informer	Bases d'informations, Bases bibliographiques
Mise en réseau	interroger, échanger	Outils réseau

Le schéma de la page suivant complète le schéma précédent en situant ces nouveaux éléments.

- Il ne faut pas oublier qu'à l'issue de ces différentes actions, l'expertise humaine est incontournable pour sélectionner l'information pertinente, décider et mettre en œuvre les actions.

Injury Prevention Programme - HLA
Le Système d'Information vu comme un DataWarehouse - Les outils



3.2- Apports des méthodes du Data Mining

Pourquoi le Data Mining est-il un apport pour l'exploitation des données sur les ADL ?

Pour répondre à cette question, il nous faut d'abord identifier les points faibles de l'ancien système d'information (SI) EHLASS :

Identification des points faibles du système d'information sur les ADL

- Nous avons déjà fréquemment fait remarquer qu'il existe un déséquilibre important dans le fonctionnement de l'ancien SI EHLASS entre les efforts (budgétaires, humains, intellectuels, etc.) consacrés à la phase de recueil des données et ceux consacrés à la phase d'exploitation de ces données. Mais, il s'agit plus encore d'un déficit de connaissances. En effet, le SI gère beaucoup de données, mais a généré relativement peu de connaissances. Ceci confirme que les efforts des acteurs ont beaucoup porté, en amont, sur l'acquisition des données et peu, en aval, sur leur exploitation ou plutôt sur leur transformation en connaissances opérationnelles, puis en décisions. C'est l'image de la montagne de données accouchant de la souris de la connaissance...

- Le gain en informations opérationnelles du SI est donc globalement faible, car on n'a pas suffisamment fait fructifier le capital de connaissances potentielles contenues dans les données. Il existe deux voies d'analyse d'un SI :

> - l'une « orientée problème » : il s'agit de répertorier tous les problèmes posés par le système de recueil et son exploitation et de proposer des solutions pour le futur. Cette voie a été largement explorée et de nombreuses études d'évaluation ont été menées pour mettre en évidence les défauts du SI sur les ADL;
>
> - l'autre « orientée solution » : il s'agit de mettre en évidence les domaines de compétence des données, d'exploiter les qualités propres du SI.

Aucun système d'information n'atteindra jamais une rigueur scientifique absolue. Il y a toujours des approximations, des zones d'incertitude dont il faut être conscient. Mais, ces manquements à l'orthodoxie statistique, sous prétexte de « rigueur scientifique » ne doivent pas empêcher de faire émerger des solutions pragmatiques, des points de vue utiles. Il faut alors voir les bases de données existantes comme des réservoirs de données utiles et se poser la question « que pouvons-nous apprendre de ces données ? » et non pas toujours les questions « qu'est-ce qu'on ne peut pas faire avec ces données ? » et « comment devraient être les données ? ».

- On sait bien que la donnée n'est pas la connaissance, encore moins la décision. C'est de cet oubli que naît sans doute le déficit constaté. La connaissance procède d'un processus d'abstraction, de modélisation, de réflexion à partir des données accumulées et exploitées et se nourrit aussi de l'expérience et de l'expertise humaine antérieure.

- Nous avons voulu illustrer le mécanisme de ces processus dans le schéma de la page suivante. Il montre aussi que la connaissance, par une sorte de « rétropropagation », favorise à son tour le meilleur paramétrage du SI lui-même et la meilleure exploitation des données recueillies.

Dans ce contexte, les faiblesses du SI se situent à notre avis en 4 points :

Point **A** : les données recueillies sont peu exploitées et/ou avec des méthodes rudimentaires. La plupart des outils d'exploitation mis en œuvre actuellement par les équipes nationales sont très simples (sélection multicritère d'observations, tris simples et croisés de variables, calcul de moyennes et de fréquences, mise en forme graphique des résultats). Il faut donc diversifier et **enrichir les méthodes statistiques** utilisées.

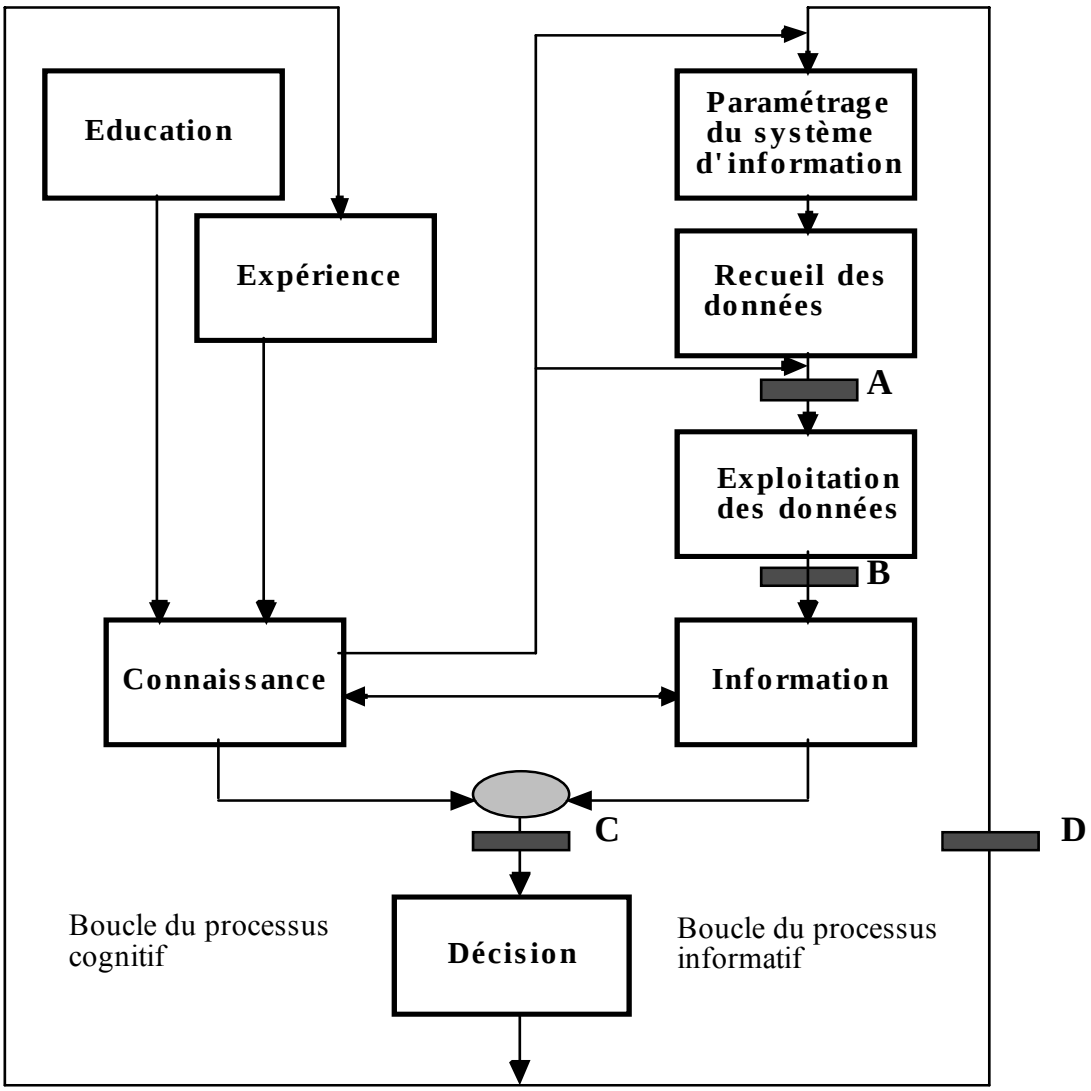
Point **B** : l'exploitation des données ne se suffit pas en elle-même si elle n'est pas guidée par le souci de produire une information à valeur ajoutée opérationnelle. C'est par **la mise en œuvre des procédures de Data Mining** que nous pourrons améliorer cette situation.

Point **C** : les connaissances et les informations issues de l'exploitation du SI ne sont pas suffisamment traduites en termes de décision. Une des solutions consiste à créer **un petit noyau d'experts** (épidémiologistes, juristes, médecins, spécialistes de la prévention, experts en système d'information) chargé de préparer des propositions de décision en rapport avec les connaissances et les informations acquises. C'est ce qui s'est déroulé récemment avec la création du réseau épidémiologique HLA

Point **D** : les décisions précédemment prises ont été peu nombreuses et lentes à mettre en application. Elles ont peu servi à mieux paramétrer le SI lui-même. **Il faut donc améliorer le niveau organisationnel**, notamment par la mise en place d'une structure voulant et pouvant exercer un rôle moteur et coordinateur fort. C'est ce qui se passe actuellement avec les nombreuses évolutions administratives et techniques qui se mettent en place sous l'impulsion de la DG SANCO.

Schéma des processus informatif et cognitif

dérivé de celui de René Lefébure et Gilles Venturi : Le Data Mining - Editions Eyrolles



points de connexion faible du SI

En définitive, pour améliorer les performances du SI, il faudrait de notre point de vue :

- enrichir les méthodes statistiques utilisées,
- les inclure dans des procédures de Data Mining adaptées, pour " mettre de l'intelligence dans les données ",
- exploiter les connaissances et les informations opérationnelles acquises dans la préparation des décisions, par le recours à une expertise externe spécifique,
- et enfin, revoir le niveau organisationnel pour améliorer la réactivité et le paramétrage du système d'information lui-même.

Comment positionner le Data Mining parmi l'ensemble des outils techniques ?

Dans le cadre du système d'information sur les ADL, comme dans celui de tout système d'information, il est possible de positionner les différents outils techniques du processus informatif qu'il faut mettre en œuvre, tour à tour, dans une optique de stratégie décisionnelle. Ces outils techniques ont pour fonction :

- d'entreposer et de structurer les données : ce sont les gestionnaires de bases de données relationnelles - SGBDR (ex : Oracle, Sybase, etc.);

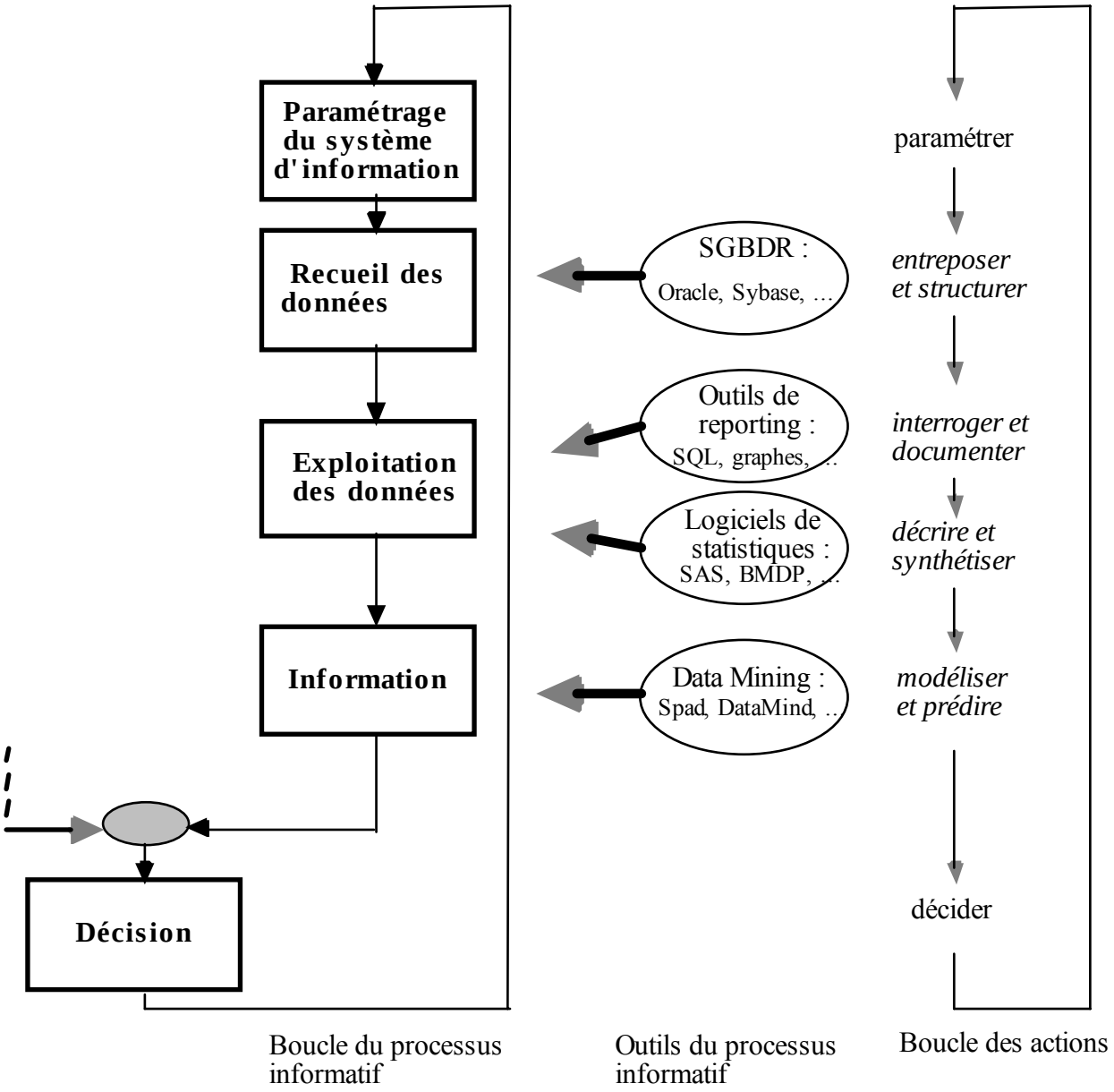
- d'interroger facilement ces bases : ce sont les outils d'interrogation simple et de " reporting " (ex : langages SQL, générateurs d'états simples, de graphiques, etc.);

- de décrire et de synthétiser les données : ce sont les logiciels de statistiques classiques (ex : modules statistiques de SAS, BMDP, SPSS, etc.);

- de modéliser (classifier, estimer, segmenter) et de prédire pour décider : ce sont les outils du Data Mining.

Nous illustrons par le schéma suivant cette correspondance entre les outils et les étapes du processus informatif :

Schéma des outils du processus informatif



Il est clair, toutefois, qu'aucune technique d'analyse de données ou de Data Mining ne remplacera l'expertise humaine. Mais, comme on l'a vu dans le schéma précédent, l'expertise humaine peut être enrichie par des résultats issus d'outils nouveaux qui viendront à leur tour guider les bons choix techniques. Il y a donc fertilisation commune entre l'expertise humaine du domaine et la maîtrise des outils d'analyse.

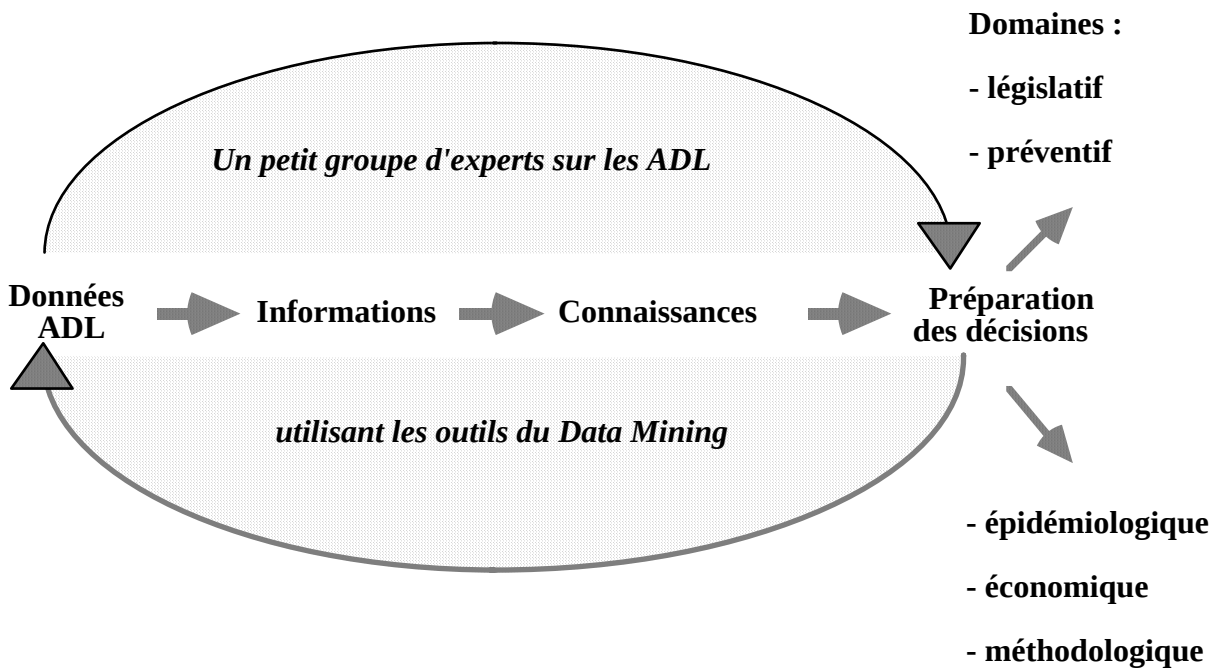
Peut-on donner un exemple concret d'application de ce cercle vertueux ?

Nous détaillons ici un exemple concret d'application du cercle vertueux du Data Mining dans le cadre du processus décisionnel que nous proposons par ailleurs et qui s'intitule **" Recherche de produits à risques dans la base IPP "**

Etapes du cercle vertueux	Application au processus " Recherche de produits potentiellement à risques "
1- Identifier les données d'intervention	- Rechercher les produits potentiellement à risques dans les bases de données ADL non agrégées en utilisant les codes produits et le texte libre. - Identifier les différences de pratiques de codage entre Etats et les problèmes de nomenclature. - Tenir compte du fait que le produit est rarement la cause directe de l'accident.
2- Utiliser des outils de Data Mining pour transformer les données en informations utiles	- Utiliser deux procédures spécifiques de Data Mining : le Score Synthétique de Dangerosité Relative (SSRD) que nous détaillerons au Chapitre 6 et une procédure d'analyse du texte libre (ANATEXT). - Identification des produits potentiellement à risques. - Construction des tableaux fournissant les circonstances, les causes, les conséquences et la population concernée pour chacun des produits identifiés. - Examen par expertise humaine de ces résultats et des observations complètes pour déterminer la liste de produits potentiellement à risques.
3- Transformer les informations en propositions d'actions concrètes	- Préciser les mesures préventives ou législatives possibles par expertise humaine, du type : - rappel de produits - modifications du marquage des produits - modification des normes de fabrication - mise en garde - campagnes de prévention spécifiques, etc.
4- Evaluer les résultats	- Mesurer l'évolution de la fréquence et de la gravité des accidents liés aux produits ayant fait l'objet de mesures préventives ou législatives : - évolution des indicateurs globaux - évolution du score de SSRD - évolution de l'échelle de sévérité

Le chaînage décisionnel synthétisé appliqué au système d'information peut se résumer ainsi :

Schéma synthétique du chaînage décisionnel



C'est dans cette optique que nous avons inscrit notre proposition de développer des outils spécifiques d'aide à la décision.

4- Exemples d'utilisation d'outils standard

Dans ce Chapitre, nous avons voulu rendre compte de l'expérience que nous avons acquise dans l'utilisation de certaines méthodes relativement sophistiquées de Data Mining pour mettre en évidence les qualités et les défauts de ce type de méthodes.

4.1- La méthode des prédicteurs neuronaux

Nous avons voulu mettre à l'épreuve des faits les considérations théoriques évoquées dans le chapitre précédent et mettre en œuvre concrètement un des outils phares du Data Mining : les prédicteurs neuronaux. Nous exposerons d'abord les fondements de la méthode, sa mise en œuvre sur les données EHLASS France et les résultats obtenus.

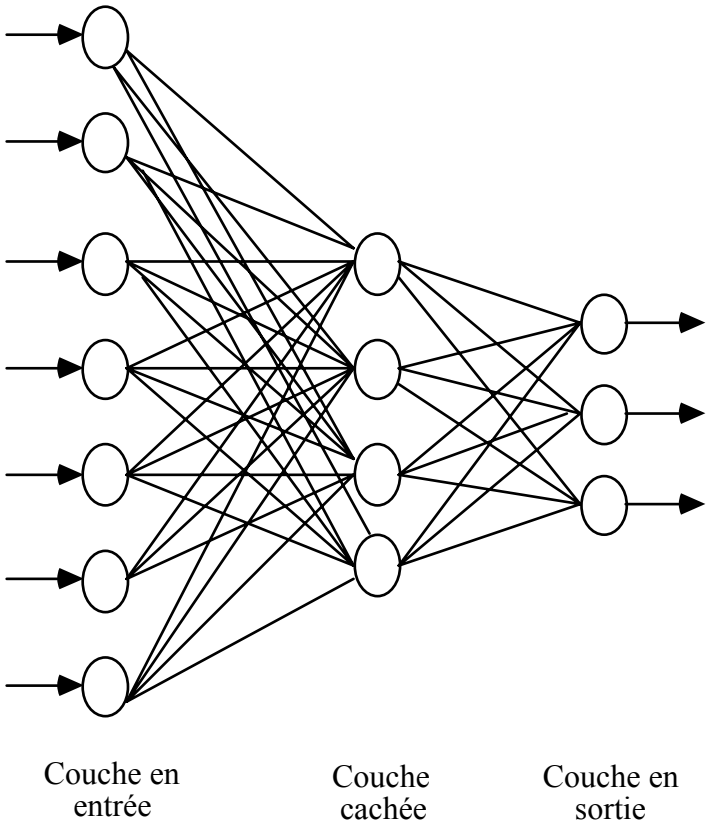
Quels sont les principes des méthodes à base de réseaux de neurones ?

Les méthodes à base de réseaux de neurones permettent la recherche d'un modèle dans les données, par l'analyse d'un ensemble d'exemples où les entrées et les sorties sont connues. Comme l'indiquent René Lefébure et Gilles Venturi dans leur livre *Le Data Mining - Editions Eyrolles* : " le réseau est un système non linéaire qui associe aux états de la couche des entrées, des états de la couche des sorties. Chaque configuration de poids d'un réseau d'une architecture déterminée définit une modélisation spécifique " et plus loin, " une des particularités du réseau de neurones est cette capacité à s'organiser sans qu'aucun agent extérieur n'intervienne dans ce processus d'optimisation ".

On utilise le plus souvent le réseau de neurones comme une méthode d'analyse discriminante non paramétrique. La procédure neuronale permet de discriminer K groupes. Elle est particulièrement adaptée à l'analyse de relations non-linéaires. Dans le logiciel SPAD-D que nous avons utilisé, le réseau mis en œuvre est un " réseau de perceptrons multicouches - un neurone ne peut être connecté qu'à des neurones d'une couche limitrophe - utilisant l'algorithme de rétropropagation du gradient ".

On peut comprendre le principe de rétropropagation comme celui jugeant de la validité du résultat pour améliorer la pertinence du modèle. Le meilleur réseau étant choisi après avoir fait varier différents paramètres (nombre de couches cachées, coefficient d'apprentissage, pourcentage de répartition entre fichier d'apprentissage et fichier test, etc.), on peut soumettre des observations anonymes (sans classement initial) pour leur attribuer une classe de sortie prévue. C'est donc de ce point de vue un outil de prévision.

Schéma d'un réseau de neurones perceptrons multicouches



Quelles sont les étapes de la construction d'un prédicteur neuronal ?

Nous décrivons rapidement les étapes de cette construction dans le cadre des données EHLASS France 1997, en nous appuyant sur la procédure NEURO du logiciel SPAD-D.

La question initiale : au départ, nous voulions discriminer la variable " Traitement " suivant 3 classes : " Pas de traitement ultérieur - Trt1 " (codes EHLASS 1 et 2), " Traitement ultérieur - Trt2 " (codes 3 et 4), " Hospitalisation - Trt3 " (code 5). Ces 3 classes nous semblaient fortement corrélées à la gravité de l'accident. Le but est de construire un modèle prédisant au mieux le classement d'une observation dans l'une de ces 3 catégories, puis d'expliquer ce classement.

Principes du calcul : dans la couche d'entrée, il y a autant de neurones que de variables explicatives, dans la couche de sortie, autant de neurones que de classes à discriminer (ici 3). Les nombres de couches cachées et de neurones dans ces couches sont choisis par l'utilisateur au cours du déroulement de la procédure.

Les neurones d'une couche inférieure sont reliés aux neurones d'une couche supérieure par des connexions appelées POIDS et chaque neurone de chaque couche est affecté d'un BIAIS. C'est le calcul de ces poids et de ces biais qui permet l'affectation des individus dans les classes. Les poids et les biais sont pris au hasard lors du premier passage des individus.

Les poids et les biais sont ajustés par un apprentissage utilisant la méthode de " rétropropagation du gradient ", dont voici le principe :

- on calcule les valeurs de chaque neurone avec le réseau de connexion calculé à l'itération N. Les neurones de sortie seront des combinaisons linéaires des neurones d'entrée, poids et biais en étant les coefficients;

- la rétropropagation consiste à modifier les poids et les biais en fonction de l'erreur observée, afin que les données d'entrée fournissent de meilleurs résultats, par exemple un pourcentage d'individus bien classés le plus fort possible.

L'estimation des poids est donc un processus itératif avec correction en fonction de l'erreur rencontrée. L'utilisateur doit choisir un certain nombre de paramètres :

- nombre maximum de cycles de l'apprentissage;
- choix de la stratégie de minimisation de l'erreur;
- seuil d'arrêt quand l'amélioration de l'erreur devient trop faible.

La procédure s'arrête quand ce seuil est atteint et le classement est sauvegardé.

Etapes de la construction de notre prédicteur neuronal :

- On a d'abord constitué une base d'exemples en prenant 3 000 enregistrements au hasard dans le fichier EHLASS France 1997. Cette base comporte volontairement 1/3 d'observations sans traitement ultérieur, 1/3 avec traitement ultérieur et 1/3 avec hospitalisations, pour mieux équilibrer l'échantillon. Puis, nous avons réparti cette base en un fichier d'apprentissage (70% des observations) et un fichier test (30%). Dans la procédure NEURO les variables explicatives doivent être continues, or les variables explicatives EHLASS sont surtout des variables nominales (Mécanisme, Lieu, Type de lésion, Partie lésée, Produit impliqué, etc.). C'est pour cela que :

- Nous avons ensuite effectué une analyse des correspondances multiples (ACM) de façon à créer des axes factoriels qui sont des combinaisons linéaires des variables explicatives. Ces axes factoriels, qui décrivent au mieux la forme initiale du nuage de points, sont ensuite traités comme des variables continues.

- Puis, nous avons lancé la procédure NEURO en prenant les 5 premiers axes factoriels comme variables explicatives continues et la variable Traitement en 3 classes comme variable à expliquer. Nous avons effectué de nombreux essais en faisant varier les paramètres de la procédure pour obtenir le meilleur réseau, c'est-à-dire celui donnant le pourcentage de " bien classés " le plus fort - pourcentage de " bien classés " = (Nb des Trt1 en Classe de sortie prévue 1 + Nb de Trt2 en Classe de sortie prévue 2 + Nb de Trt3 en Classe de sortie prévue 3) / Nb total des observations.

Quels sont les résultats obtenus ?

- Avec les 3 classes initiales de la variable Traitement, le pouvoir prédictif du modèle n'est pas bon. Il n'arrive qu'à classifier les observations en 2 classes de sortie prévues, avec une troisième classe vide. Ceci donne un pourcentage de " bien classés " faible, aux environs de 33%.

Le modèle n'arrivant pas à distinguer clairement entre la classe " Pas de traitement ultérieur - Trt1 " et la classe " Traitement ultérieur - Trt2 ", nous avons donc regroupé les 2 classes Trt1 et Trt2 pour analyser en définitive les " Hospitalisés " (Trt3) par rapport aux " Non hospitalisés " (Trt1+Trt2).

- Avec ce modèle, nous arrivons à un pouvoir prédictif bien meilleur (voir tableau suivant), dans la mesure où le pourcentage de " bien classés " est de l'ordre de 77%.

Pourcentage de " bien classés " dans l'échantillon d'apprentissage :

	Classe 1	Classe 2	Total	% Bien classés
Non hospital.	1284	131	1415	90,74
Hospitalisés	354	331	685	46,03
Total	1638	462	2100	**76,90**

- On constate que le modèle est meilleur pour classer les Non hospitalisés (plus de 90% de succès) que les Hospitalisés. Globalement, le réseau de neurones arrive donc à classifier correctement plus de 3 observations sur 4. Ce pourcentage est bon sans être exceptionnel.

- Il faut bien voir que cette méthode ne permet pas d'analyser directement le processus de classification et d'identifier clairement le pouvoir classifiant de chaque variable. Le réseau de neurones agit de ce point de vue comme une " boîte noire ", ne permettant de connaître que le résultat sans connaître l'explication du classement.

- Il y a cependant une manière de répondre en partie à cette interrogation en croisant la variable " Classe de sortie prévue " avec l'ensemble des variables explicatives et en isolant les modalités caractérisant au mieux les individus de chaque classe. Nous fournissons dans le tableau suivant, par ordre décroissant, les modalités des variables caractérisant le mieux les observations de la Classe 1, avec le pourcentage d'observations appartenant à la Classe 1 dans l'effectif total de la modalité correspondante :

Caractérisation de la Classe de sortie prévue 1 : Les " Non hospitalisés "

Variable EHLASS	Modalité de la variable	% Classe 1 / modalité
Age	5-14 ans	94,9
Lieu	zones de sport	99,6
Produit	sport	99,8
Sexe	masculin	85,6

Mécanisme	coup, collision	95,9
Activité	sport	100,0
Lésion	entorse	99,2
Lésion	plaie ouverte	92,1
Age	15-24 ans	94,5
Mécanisme	effort physique	96,6
Activité	scolaire	99,4
Activité	ménagère	93,1
Lieu	zones scolaires	98,3
Produit	véhicules : vélo,	95,9
Produit	éléments construction	91,9
Partie lésée	extrémités supérieures	83,9
Lésion	contusion	82,2
/...		

On constate donc que, dans la Classe 1 regroupant les accidents ne nécessitant pas d'hospitalisation, il y a prédominance : des accidents de sport chez les garçons dans la tranche d'âge 5-14 ans, accidents se produisant lors d'activités scolaires ou de loisirs et entraînant des entorses ou des plaies ouvertes, ainsi que des chutes de vélos.

Caractérisation de la Classe de sortie prévue 2 : Les " Hospitalisés "

Variable EHLASS	Modalité de la variable	% Classe 2 / modalité
Age	65 et +	78,5
Lieu	domicile	50,0
Partie lésée	organisme total	96,7
Lésion	intoxication	93,9
Produit	escalier, plancher	78,3
Mécanisme	expo pdt chimiques	72,7
Activité	besoins personnels	54,6
Sexe	féminin	33,0
Lésion	fracture	40,0
Mécanisme	chute	28,9
Partie lésée	partie inf. du dos	55,1
Activité	jeux, activités de loisirs	28,9
Age	45-64 ans	36,5
Activité	bricolage	38,7
Produit	outils	46,5
/...		

On constate donc que, dans la Classe 2 regroupant les accidents nécessitant très fréquemment une hospitalisation, il y a prédominance : chez les femmes âgées de 65 ans et plus des chutes au domicile entraînant des fractures, chez les 45-64 ans des accidents de bricolage, ainsi que des intoxications.

Quelles conclusions peut-on tirer de cette expérience ?

En ce qui concerne la méthode :

- La procédure des prédicteurs neuronaux utilisée permet d'opérer une classification en K groupes à partir de variables explicatives quantitatives. Cela implique, quand les variables explicatives sont nominales, d'utiliser une analyse des correspondances multiples comme procédure intermédiaire, ce qui conduit à une perte d'information.

- Cette méthode ne permet pas une explication directe de la façon dont s'opère le classement. La procédure est une " boîte noire ", qui peut cependant conduire à des résultats intéressants en matière de prédiction.

En ce qui concerne les résultats :

- Sur notre échantillon de 3 000 observations issues du fichier EHLASS France 1997, les 2 classes d'accidents ayant les modalités " Pas de traitement ultérieur - Trt1 " et " Traitement ultérieur - Trt2 " ne sont pas statistiquement différentes au regard de cette méthode.

- Nous obtenons un pourcentage de " bien classés " de 77% à l'issue de la procédure neuronale en discriminant entre la classe des " Hospitalisés " (Trt3) et celle des " Non hospitalisés " (Trt1+Trt2). Le réseau de neurones arrive donc à classifier correctement plus de 3 observations sur 4.

- La classe des " Hospitalisés " se caractérise surtout par la prédominance des chutes au domicile, des accidents de bricolage et des intoxications. La classe des " Non hospitalisés " se caractérise par la prédominance des accidents de sport lors des activités scolaires ou de loisirs et des chutes de vélos.

On constate que la mise en œuvre de ce type de méthodes est relativement lourde (recodage des données, détermination des axes factoriels, choix fin du paramétrage, interprétation délicate des résultats) et nécessite une certaine expérience statistique. C'est en partie pour cela que nous proposons de développer des outils plus simples d'emploi et d'interprétation et qui répondent plus directement aux questions spécifiques engendrées par le système d'information.

4.2- La segmentation

Nous allons maintenant exposer la méthode de segmentation appliquée aux accidents EHLASS France 1996. Dans notre exemple, nous avions pour but d'identifier les variables qui interviennent le plus pour expliquer :

- la durée d'hospitalisation (variable continue);
- la répartition des patients entre ceux qui nécessitent un suivi et les autres (variable nominale).

Quels sont les principes de la méthode de segmentation* ?

Cette méthode permet la construction d'un arbre de décision binaire complet en effectuant une régression non paramétrique d'une variable à expliquer sur un ensemble de variables explicatives de nature quelconque.

Les méthodes paramétriques (analyse linéaire discriminante, régression logistique, etc.) fournissent des règles de décision algébriques difficilement interprétables. La segmentation permet de mettre en évidence des règles simples de décision (des règles binaires, i.e. en oui/non) permettant de classer ou d'estimer au mieux les observations au regard de la variable à expliquer.

Ainsi, dans notre exemple, on pourra expliquer ou mieux encore prédire (avec une certaine marge d'erreur bien entendu), la durée d'hospitalisation en fonction de certaines caractéristiques de l'accidenté (par exemple : l'âge, le type de lésion, etc.) décrites dans le recueil EHLASS. De même, on pourra expliquer ou prédire l'appartenance de l'accidenté au groupe où un suivi médical sera nécessaire ou non.

La méthode est basée sur la construction d'un arbre binaire obtenu à l'aide de divisions successives de sous-ensembles de l'échantillon en 2 descendants. L'idée fondamentale est de sélectionner chaque division d'un segment de telle sorte que les segments descendants soient plus « purs » que le segment parent au regard du classement ou de l'estimation de la variable explicative.

Quelle est la méthodologie utilisée ?

Nous avons réalisé nos deux segmentations à l'aide du logiciel SPAD version 3 (Système Pour l'Analyse des Données - CISIA - France) sur le fichier EHLASS France 1996.

La durée d'hospitalisation

La première segmentation porte uniquement sur les hospitalisés (code traitement = 5) dont la durée d'hospitalisation est connue, soit 3 975 observations. La variable à expliquer est la durée d'hospitalisation.

** Les propos suivants sont inspirés de ceux du manuel du logiciel SPAD que nous utilisons pour ces travaux (SPAD : logiciel de la Société CISIA - France).*

Les variables explicatives que nous avons sélectionnées sont :

- l'âge;
- le sexe;
- le mécanisme de l'accident;
- le lieu de l'accident;
- l'activité;
- la partie lésée;
- le type de lésion.

Afin d'obtenir des résultats plus facilement interprétables, nous avons agrégé les modalités des variables en un nombre de modalités inférieur à celui d'origine, le plus souvent en ne considérant que les modalités regroupées au premier niveau du système de codage (i.e. suivant le premier caractère des variables de codes). Nous avons réalisé ensuite la segmentation.

Répartition des patients entre ceux qui nécessitent un suivi et les autres

La seconde segmentation porte sur un échantillon au 1/5ème du fichier EHLASS France 1996 total, soit 9 138 observations. Nous avons sélectionné les patients ayant eu un « Traitement sans suivi » (code Traitement = 1 ou 2) et ceux ayant un « Traitement avec un suivi » par un généraliste, en ambulatoire ou une hospitalisation (code Traitement = 3 ou 4 ou 5). La variable à expliquer est l'appartenance à l'une ou l'autre de ces deux catégories.

Les variables explicatives choisies sont également :

- l'âge;
- le sexe;
- le mécanisme de l'accident;
- le lieu de l'accident;
- l'activité;
- la partie lésée;
- le type de lésion.

Comme précédemment, pour obtenir des résultats plus facilement interprétables, nous avons agrégé les modalités des variables en un nombre de modalités inférieur à celui d'origine, le plus souvent en ne considérant que les modalités regroupées au premier niveau du système de codage. Puis, nous avons réalisé la segmentation.

Quels sont les résultats obtenus par cette méthode ?

La durée d'hospitalisation

La population totale des hospitalisés dont on connaît la durée d'hospitalisation (n=3 975) est découpée en un échantillon de base (n=2 660) et un échantillon test (n=1 315). Après la procédure d'élagage, l'arbre optimal est constitué de 5 segments terminaux. La description des coupures sur l'échantillon de base qui figure à la page suivante montre que pour le segment 1 :

- la taille de l'échantillon de base est de 2 660 accidents
- la durée moyenne d'hospitalisation est de 6,2 jours et l'écart-type de 9,0 jours

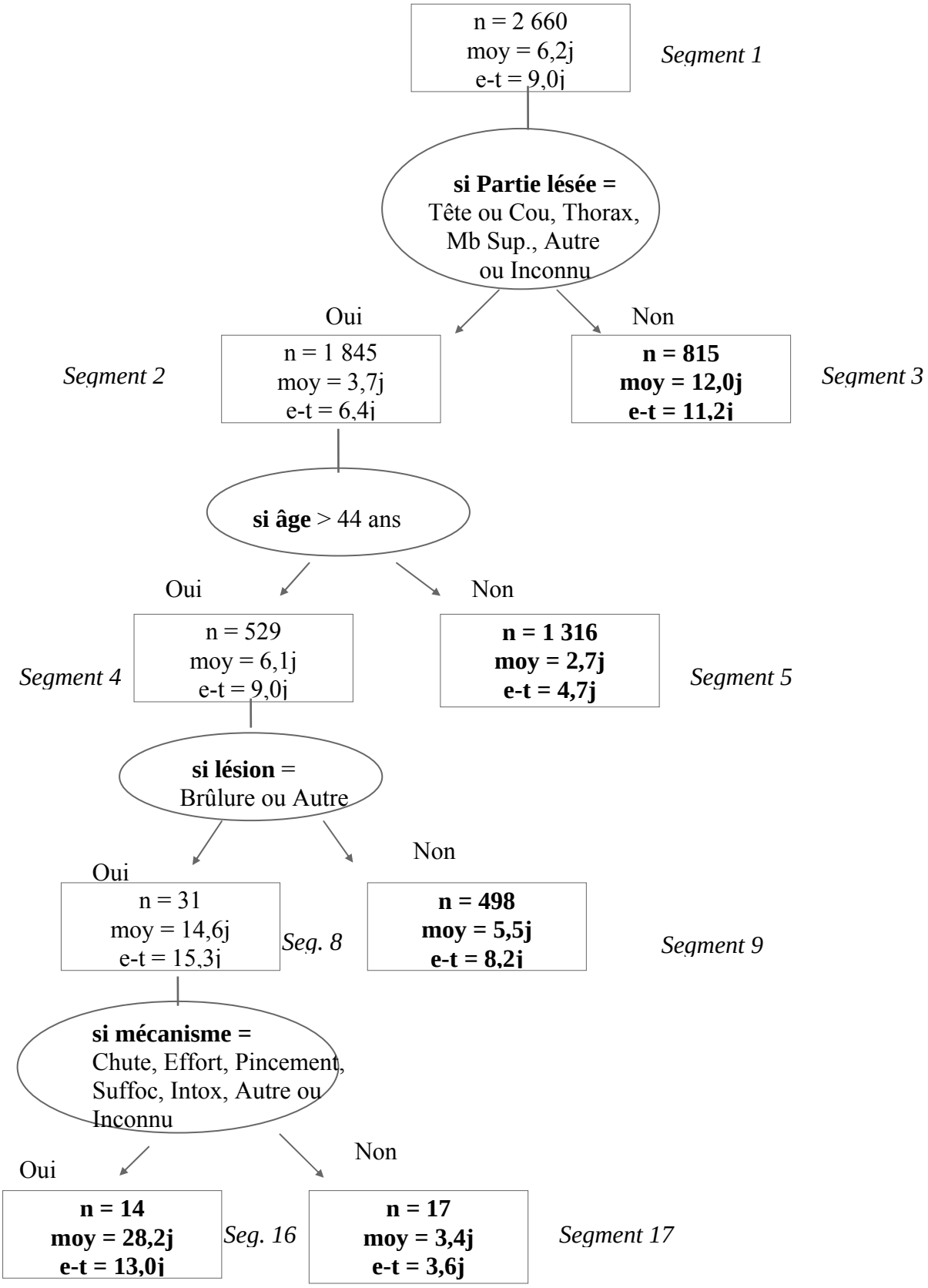
R1- La première règle de décision concernant le segment 1 est la suivante :

> # Si la variable « Partie lésée » prend les modalités = Tête ou Cou, Thorax, Membre supérieur, Autre ou Inconnu, alors :
>
> la durée moyenne d'hospitalisation est de 3,7 jours et l'écart-type de 6,4 jours
>
> sinon (si « Partie lésée » = Partie inférieure du dos/abdomen ou membre inférieur) :
>
> la durée moyenne d'hospitalisation est de 12,0 jours et l'écart-type de 11,2 jours
>
> La population est répartie en 2 segments : le segment 2 contenant 1 845 accidents, le **segment 3** en contenant 815. Ce dernier segment constitue un *segment terminal* dans la mesure où il n'est pas divisé dans la suite de la procédure.

R2- La seconde règle de décision concernant le segment 2 est la suivante :

> # Si la variable « Age » est supérieure à 44 ans, alors :
>
> la durée moyenne d'hospitalisation est de 6,1 jours et l'écart-type de 9,0 jours
>
> sinon (si « Age » est inférieur ou égal à 44 ans) :
>
> la durée moyenne d'hospitalisation est de 2,7 jours et l'écart-type de 4,7 jours
>
> La population est à nouveau répartie en 2 segments : le segment 4 contenant 529 accidents, le **segment 5** en contenant 1 316. Ce dernier segment constitue aussi un *segment terminal* dans la mesure où il n'est pas divisé dans la suite de la procédure.

Arbre de décision



R3- La troisième règle de décision concernant le segment 4 est la suivante :

> # Si la variable « Type de lésion » = Brûlure ou Autre, alors :
>
> la durée moyenne d'hospitalisation est de 14,6 jours et l'écart-type de 15,3 jours
>
> sinon (si la variable « Type de lésion » est différente de Brûlure ou Autre) :
>
> la durée moyenne d'hospitalisation est de 5,5 jours et l'écart-type de 8,2 jours
>
> La population est à nouveau répartie en 2 segments : le segment 8 contenant 31 accidents, le **segment 9** en contenant 498. Ce dernier segment constitue aussi un *segment terminal* dans la mesure où il n'est pas divisé dans la suite de la procédure.

R4- La quatrième règle de décision concernant le segment 8 est la suivante :

> # Si la variable « Mécanisme » = Chute ou Effort, Pincement, Suffocation, Intoxication, Autre ou Inconnu, alors :
>
> la durée moyenne d'hospitalisation est de 28,2 jours et l'écart-type de 13,0 jours
>
> sinon (si la variable « Mécanisme » prend une modalité différente de celles ci-dessus) :
>
> la durée moyenne d'hospitalisation est de 3,4 jours et l'écart-type de 3,6 jours
>
> La population est à nouveau répartie en 2 segments : le **segment 16** contenant 14 accidents, le **segment 17** en contenant 17. Ces deux derniers segments constituent la fin de la procédure.

La population de départ (segment 1) est donc répartie en 5 segments terminaux (3, 5, 9, 17, 16).

Echantillon de base	Moyenne en J	Ecart-type en J
Segment 1 (n=2 660)	6,2	3,0
Segment 3 (30,6%)	12,0	11,2
Segment 5 (49,5%)	2,7	4,7
Segment 9 (18,7%)	5,5	8,2
Segment 17 (0,6%)	3,4	3,6
Segment 16 (0,5%)	28,2	13,0

L'échantillon test sert à évaluer la pertinence de cette segmentation construite à partir de l'échantillon de base. Le passage des individus de l'échantillon test dans l'arbre binaire construit avec les règles exposées ci-dessus permet d'obtenir une estimation de la durée d'hospitalisation. On obtient ainsi :

Echantillon test	Moyenne en J	Ecart-type en J
Segment 1 (n=1 315)	6,2	3,0
Segment 3 (29,5%)	12,5	11,6
Segment 5 (52,4%)	2,5	3,7
Segment 9 (16,9%)	5,5	7,2
Segment 17 (0,9%)	13,8	14,9
Segment 16 (0,3%)	39,0	0,0

Il est certain que pour des segments à effectif faible et/ou écart-type fort par rapport à la moyenne (segments 16 et 17, par exemple), l'estimation de la durée d'hospitalisation n'est pas bonne. Elle est bien meilleure pour les autres segments.

En termes de caractérisation de la durée d'hospitalisation par les modalités des variables explicatives, on obtient les résultats suivants :

Modalités	Moyenne en J	Ecart-type en J
Partie lésée = Peau	23,48	13,34
Lésion = Brûlure	18,39	14,34
Mécanisme = Expo therm.	14,23	14,12
Partie lésée = Inf. dos	11,47	11,07
Age = 65 ans et +	11.45	11.10
Partie lésée = Mb inf.	10,66	10,00
Lésion = Autre	9,58	12,81
Activité = Bricolage	8,44	10,45
...		

En termes de hiérarchisation du pouvoir discriminant des variables, les variables qui influent le plus sur la durée d'hospitalisation sont, par ordre décroissant :

1- la partie lésée;
2- le type de lésion;
3- l'âge;
4- le mécanisme de l'accident;
5- le lieu de l'accident;
6- l'activité;
7- le sexe.

Alors qu'il était généralement admis que le mécanisme était la variable la plus importante pour expliquer la durée d'hospitalisation, **la segmentation prouve que c'est la variable « Partie lésée ».**

Répartition des patients entre ceux qui nécessitent un suivi et les autres

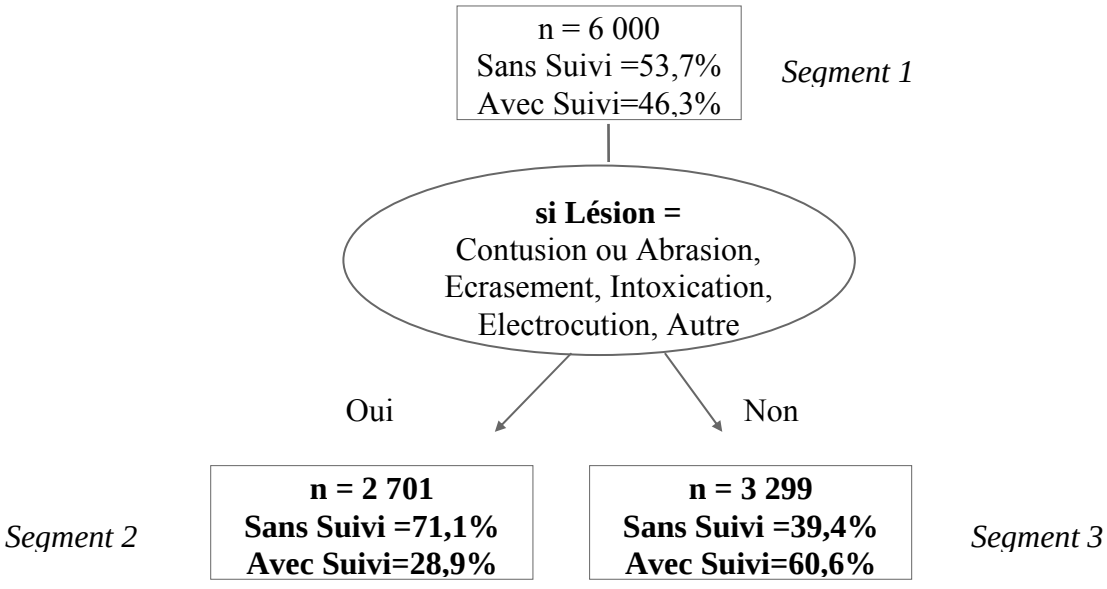
La population sélectionnée (patients ayant un « Traitement sans suivi » et ceux ayant un « Traitement avec un suivi », soit n=8 955) est découpée en un échantillon de base (n=6 000) et un échantillon test (n=2 955). La variable à expliquer est l'appartenance à l'une ou l'autre de ces deux catégories. Après la procédure d'élagage, l'arbre optimal est constitué de 2 segments terminaux. La description de la coupure sur l'échantillon de base, qui figure à la page suivante, montre que pour le segment 1 :

- la taille de l'échantillon de base est de 6 000 accidents
- le pourcentage de la modalité « Traitement avec suivi » - TRT3 = 46,3%

R1- La règle de décision concernant le segment 1 est la suivante :

> # Si la variable « Type de lésion » prend la modalité = Contusion ou Abrasion, Ecrasement, Intoxication, Electrocution ou Autre, alors :
>
> le pourcentage de la modalité « Traitement sans suivi » - TRT1 = 71,1%
>
> sinon (si « Type de lésion » = Plaie ouverte ou Fracture, Luxation, Entorse, Lésion nerveuse, Brûlure, Inconnu) :
>
> le pourcentage de la modalité « Traitement avec suivi » - TRT3 = 60,6%
>
> La population est donc répartie en 2 segments : le **segment 2** contenant 2 701 accidents, le **segment 3** en contenant 3 299. Ces segments constituent des *segments terminaux* dans la mesure où ils ne sont pas divisés.

Arbre de décision



La population de départ (segment 1) est donc répartie en 2 segments terminaux (2, 3).

Echantillon de base	Sans Suivi	Avec suivi
Segment 1 (n=6 000)	53,7%	46,3%
Segment 2 (n=45,0%)	71,1%	28,9%
Segment 3 (n=55,0%)	39,4%	60,6%

L'échantillon test sert à évaluer la pertinence de cette segmentation construite à partir de l'échantillon de base. Le passage des individus de l'échantillon test dans l'arbre binaire construit avec la règle exposée ci-dessus permet d'obtenir une estimation de la répartition des accidents. On obtient ainsi :

Echantillon test	Sans Suivi	Avec suivi
Segment 1 (n=2 955)	53,7%	46,3%
Segment 2 (n=44,8%)	70,9%	29,1%
Segment 3 (n=55,2%)	39,6%	60,4%

Dans ces 2 échantillons, le pourcentage de « bien classés » (i.e. des patients sans suivi classés dans le groupe des « sans suivi » et des patients avec suivi classés dans le groupe des « avec suivi ») est de l'ordre de 65%.

La segmentation montre que la variable la plus importante pour expliquer et prévoir le type de traitement avec ou sans suivi d'un patient accidenté était **la variable « Type de lésion ».**

Quelles sont nos conclusions ?

Nous avons choisi d'appliquer deux procédures de segmentation, l'une sur une variable continue (la durée d'hospitalisation), l'autre sur une variable nominale (appartenance ou non au groupe des traités avec suivi). Nous aurions pu, bien évidemment, déterminer d'autres problématiques au sein du système EHLASS. Nous voulions juste tester le pouvoir explicatif de la méthode.

Avec les segmentations choisies, nous avons montré que :

- En termes de hiérarchisation du pouvoir discriminant des variables, les variables qui influent le plus sur la durée d'hospitalisation sont, par ordre décroissant :

> 1- la partie lésée;
> 2- le type de lésion;
> 3- l'âge;
> 4- le mécanisme de l'accident;
> 5- le lieu de l'accident;
> 6- l'activité;
> 7- le sexe.

Alors qu'il était admis généralement que la variable « Mécanisme » était la plus importante pour expliquer la durée d'hospitalisation, **la segmentation prouve que c'est en fait la variable « Partie lésée ».**

Les règles de décision mises en évidence permettent d'estimer la durée moyenne d'hospitalisation d'un patient accidenté.

- De même, nous avons montré que la variable la plus importante pour expliquer et prévoir l'appartenance ou non d'un patient au groupe des traités avec suivi était **la variable « Type de lésion ».**

Dans ce Chapitre, nous espérons avoir montré que :

> **La segmentation peut être un instrument statistique intéressant pour mieux comprendre l'appartenance à certains groupes à risques ou encore expliquer ou estimer des variables essentielles. Il conviendrait donc de généraliser son emploi pour mieux fertiliser les données EHLASS déjà recueillies.**

Mais, on remarquera qu'avec ces méthodes l'on ne peut pas répondre à toutes les sortes de questions engendrées par le SI, comme par exemple celle relative au repérage de produits potentiellement dangereux ou celle relative à la sévérité de l'accident.

4.3- Les autres méthodes du Data Mining

En dehors de la segmentation et des réseaux de neurones, nous avons vu qu'il existe bien d'autres méthodes de Data Mining présents dans les logiciels spécialisés disponibles sur le marché.

Quelles sont les autres méthodes du Data Mining ?

Les méthodes de réduction de dimension (méthodes factorielles) :

Elles permettent de visualiser simultanément des variables et des individus selon autant de dimensions que l'utilisateur le souhaite.

Parmi ces méthodes, on peut citer :

- L'analyse en composantes principales (*p* variables quantitatives)
- L'analyse des correspondances multiples (*p* variables qualitatives)
- L'analyse factorielle discriminante (*p* variables quantitatives, 1 variable qualitative)

Les méthodes de modélisation :

Elles consistent à déterminer une structure déduite de l'analyse d'une partie des données et à vérifier si cette structure est applicable à l'ensemble des données. Parmi ces méthodes, on peut citer :

- La régression linéaire multiple
- La régression logistique
- La régression par voisinage

Les méthodes de classification :

Elles ont pour but de regrouper des individus ayant des caractéristiques proches au regard d'un ensemble de variables. On peut citer :

- Les classifications hiérarchiques
- Les classifications non hiérarchiques
- La segmentation

Quels sont les logiciels de Data Mining les plus connus ?

En dehors du logiciel français SPAD que nous avons utilisé, nous pouvons citer parmi les logiciels les plus connus :

- SPSS Chaid et Neural Connection de SPSS

statistiques, classification, réseaux de neurones

- DataMind de Datamind SA

classification, modèles fonctionnels, statistiques

- SAS Enterprise Miner de SAS Institute

arbres de décision, réseaux de neurones

- Intelligent Miner d'IBM

analyse relationnelle, classification, modèles fonctionnels

Transformer les données en informations est le but de toutes ces méthodes de Data Mining. Mais, la découverte d'informations masquées est très dépendante des besoins de chaque utilisateur. Les méthodes standards ne pourront répondre qu'à des questions générales. Les questions spécifiques issues du système de recueil de données sur les ADL nous amènent à vouloir mettre au point des procédures décisonnelles mieux adaptées.

5- Les procédures décisionnelles spécifiques proposées

5.1- Principes de développement

Quel est le périmètre d'efficacité du système d'information sur les ADL ?

Pour trouver les meilleures procédures adaptées à la spécificité du système d'information sur les ADL, il faut d'abord déterminer le périmètre d'efficacité du SI :

- la méthodologie par enquête fournit une vision globale de l'ensemble des ADL, en permettant des calculs d'incidence et une approche socio-économique des populations à risques : c'est ce que nous avons appelé, dans des travaux précédents, l'approche macro-accidentologique.

- la méthodologie par recueil dans les services d'urgence permet de disposer d'un grand nombre d'observations et de données médicalement fiables sur les ADL. Ce recueil d'informations va permettre de repérer les populations, les situations, les comportements ou les produits à risques en rapport avec des lésions précises : c'est ce que nous avons appelé l'approche micro-accidentologique.

Comme nous l'indiquions dans une précédente étude sur l'évaluation du système EHLASS :

« Pour des produits précis, il importe peu de connaître l'incidence exacte. Le nombre absolu de survenue d'un accident où est impliqué un produit particulier (et non sa fréquence relative) suffit à décider de la dangerosité potentielle d'un produit (exemple : les aérosols). Il importe peu de savoir que tel produit représente 0,00X % d'un ensemble représentatif d'ADL, il suffit de savoir qu'il est à l'origine d'au moins N accidents graves répertoriés pour justifier une réflexion, une étude approfondie, puis, par exemple une éventuelle évolution de la réglementation ».

- Avec les deux méthodologies on ne peut pas connaître avec certitude les accidents très rares, survenus, par exemple, avec des produits précis (une marque commerciale, un numéro de série dans une production donnée, etc.) : ce serait le niveau pico-accidentologique.

Le périmètre d'efficacité du SI peut donc être ainsi déterminé suivant l'approche et la méthodologie utilisée :

Périmètre d'efficacité du SI selon l'approche	Recueil par enquête	Recueil dans les services d'urgence
Macro-accidentologique	**Elevé**	Faible
Micro-accidentologique	Faible	**Elevé**
Pico-accidentologique	Très faible	Très faible

A chaque approche correspond une plage d'outils efficaces ayant des caractéristiques différentes. Ainsi pour l'approche macro-accidentologique, on utilisera des outils de type indicateurs à visée épidémiologique, où la qualité statistique de l'outil est primordiale. Pour l'approche pico-accidentologique, on privilégiera les recherches d'informations directes sur les cas signalés.
Pour l'approche micro-accidentologique, outre les outils classiques de description statistique et de Data Mining, nous proposons de développer des outils spécifiques à caractère pragmatique où l'aspect validité statistique est moins essentiel. Ces outils concernent l'exploitation de données non agrégées, plus particulièrement celles issues du recueil dans les services d'urgence.

Quels ont été nos principes de développement ?

L'expérience de l'utilisation des logiciels de Data Mining que nous avons, la nécessité de disposer d'outils spécifiques au domaine de notre action (les ADL) et à l'approche choisie (la micro-accidentologie), ainsi que les résultats de notre enquête font que nous avons confirmé notre hypothèse de départ, à savoir **la nécessité de développer des outils décisionnels spécifiques** et ceci sur des bases précises.

- Nous avons voulu répondre à quelques questions simples exprimées dans l'enquête :

1- Peut-on et comment hiérarchiser la dangerosité potentielle des produits ?

=> nous proposons la **Procédure SSRD (Score Synthétique de Dangerosité Relative) ou SSRD (Synthetic Score of Relative Dangerosity)**

2- Quelle alerte automatique peut-on mettre en place à partir des données recueillies ?

=> nous proposons la **Procédure SAA (Système d'Alerte Automatisée) ou AAS (Automated Alert System)**

3- Comment définir la gravité d'un ADL dans le système ?

=> nous avons proposé la **Procédure NGA (Note de Gravité de l'Accident) puis adopté la Procédure SSC (Severity SCale)**

4- Peut-on définir des scénarios types d'accidents ?

=> nous proposons la **Méthode SCENAR (Méthode des scénarios) ou SCENAR (Method of Scenario)**

- Nous voulions utiliser les données issues de l'ancien système EHLASS telles qu'elles sont et non telles qu'elles auraient dû être ou telles qu'elles seront dans l'avenir. **Notre approche est donc pragmatique**. Nous connaissons les défauts de l'ancien recueil de données sur les ADL, mais nous voulions pouvoir utiliser au mieux les millions de données existantes et disponibles.

- Nous avons volontairement construit **des outils qui n'utilisent que des variables endogènes** au système sans procédures mathématiques complexes. Nous n'utilisons donc que les variables du système EHLASS dans des modèles additifs à partir de construction de scores et de croisements de variables.

- Nous avons voulu développer **des outils simples et facilement compréhensibles** dans leurs principes et dans leur utilisation. Nous avons vu que les méthodes sophistiquées de Data Mining comportent souvent un aspect " boîte noire " qui empêche la compréhension des mécanismes explicatifs des résultats trouvés. Nos outils voulaient être exploitables par des équipes qui ne possèdent pas de statisticiens. Ils s'appuient sur un certain bon sens scientifique et sont simples d'emploi.

- Nos outils n'apportent pas la vérité définitive sur les différents sujets. Mais, ils permettent, nous l'espérons, de sélectionner des **pistes d'interrogation**, de délimiter des territoires pour des investigations plus précises et de donner aussi des débuts de réponses. Leur but n'est pas de figurer dans des publications scientifiques confidentielles, mais **d'aider les autorités en charge des politiques de prévention à se poser les bonnes questions**.

- L'apport des experts des autres équipes nous a permis de définir pour trois méthodes (SSDR, SCENAR et SAA) deux niveaux d'utilisation : **un niveau simple et un niveau évolué**.

- Nous avons voulu utiliser l'outil de développement le plus répandu dans les Etats membres et à la Commission. C'est ainsi que nous avons développé les procédures en utilisant **le langage SAS** (SAS Institute) et une interface utilisateur utilisant un programme en langage C.

- Enfin, nous avons voulu développer des procédures qui soient les plus indépendantes possibles du système de codage qui est susceptible d'évoluer rapidement. Ainsi, les tables de codes des variables sont dans des fichiers externes séparés du noyau de la procédure.

Ces outils répondent-ils aux souhaits exprimés dans l'enquête ?

Si l'on reprend les souhaits exprimés dans la Partie 2 de notre enquête concernant l'exploitation des données, nous pouvons dresser le tableau suivant :

Outils souhaités concernant	Outils proposés
la validation des données	concerne d'autres projets IPP
la mesure de la qualité et de la comparabilité des données	concerne d'autres projets IPP
l'interrogation des données : disponibilité sur le réseau et interrogation dynamique	concerne d'autres projets IPP
le reporting	module standard de logiciels statistiques
la détection des cas rares	la procédure SAA répond en partie
l'analyse de la dangerosité des produits	la procédure SSDR
les outils d'alerte sur les produits et de recherche de produits défectueux	les procédures SSDR et SAA répondent en partie
la mesure de la sévérité des accidents	la procédure SSC
l'analyse en tendance	module standard de logiciels

	statistiques
le calcul d'incidence	module standard de logiciels statistiques
la construction de scénarios et de hiérarchisation des priorités	la méthode SCENAR
l'évaluation du coût économique des ADL	pas de procédure proposée car cela nécessite des données complémentaires.

Comment s'est effectué le choix des procédures ?

Nous avons proposé au départ 4 procédures qui relèvent de l'approche micro-accidentologique. Après consultation directe de certaines équipes et le dépouillement du questionnaire, il nous est apparu que :

- Les 4 procédures possèdent en moyenne, au regard des notes obtenues, un niveau d'acceptabilité relativement bon et comparable. Cependant, la variance est grande, ce qui indique que certaines équipes n'approuvent pas ces méthodes, tandis qu'un plus grand nombre manifeste leur intérêt.

- L'avis de la Commission nous a paru prépondérant dans la mesure où ces outils pourront être utilisés au niveau de la base européenne des données non agrégées. Quoiqu'il en soit ces outils doivent être testés en vraie grandeur et leur utilisation ne présente aucun caractère d'obligation.

- Aucune autre procédure n'a été proposée pour remplacer celles que nous exposions, même si d'autres besoins ou idées de procédures ont été exprimés.

Nous avons donc choisi de développer les procédures proposées en les inscrivant dans une notion **de catalogue de procédures** d'exploitation. Nous avons exposé en détail les 4 procédures en question et commencé à créer ainsi un début de catalogue qui devra être complété par la suite avec d'autres outils, développés par d'autres équipes.

Comment s'est déroulée la mise à disposition des procédures ?

Au cours du déroulement du projet et lors des contacts pris, il est apparu que :

- La période de test en grandeur réelle devait être plus longue que le mois initialement prévu. De l'avis de certaines équipes, elle doit s'étendre sur un an au moins. Ce n'est qu'à l'issue de cette période que les équipes pourront vraiment juger de la pertinence et de l'utilisabilité des méthodes. Nous avons maintenu cependant notre période de test préliminaire sur un mois pour vérifier que les procédures étaient compréhensibles et pouvaient fonctionner dans différents environnements humains et techniques.

- Il y a deux niveaux d'utilisation : le niveau national et le niveau européen. Certains outils semblent mieux convenir au niveau national (le SAA par exemple), d'autres aux deux niveaux.

- Nous devions mettre à disposition des équipes des procédures informatisées sous forme de fichiers et décrire pour chacune des procédures ses principes de fonctionnement et un exemple d'utilisation.

	SSRD	SAA	SCENAR	SSC
Principes de la méthode	X	X	X	X
Programmes SAS	X	X		X
Exemples d'utilisation	X	X	X	X

- L'intégration de certaines de ces procédures au système global en cours de développement dans le cadre du projet EUPHIN-HIEMS (accès à la base européenne des données ADL non agrégées via le réseau EUPHIN) pourra se faire ultérieurement, si nécessaire. Elle n'est pas du ressort de ce projet.

5.2- Difficultés rencontrées

Quelles ont été les difficultés rencontrées ?

- Nous aurions souhaité recevoir des réponses plus nombreuses à notre questionnaire et avoir des demandes de collaboration ou de précision de la part de l'ensemble des partenaires du projet. Cependant, nous avons eu des échanges fournis avec les équipes danoise, autrichienne, hollandaise, française et grecque. Mais, certaines équipes nationales ne semblent pas encore convaincues du fait qu'une meilleure exploitation des données accumulées est une condition à la fois de la survie du système et de l'amélioration de la qualité du service rendu.

- Nous avons été soumis à deux tendances contraires : construire des outils simples mais relativement « naïfs » ou promouvoir des outils plus sophistiqués, mais plus difficiles d'emploi et d'interprétation. Nous avons rapidement et clairement opté pour la première voie, sachant que nous exposerons une étape plus évoluée pour trois procédures : SSDR, SAA et SCENAR.

- Les équipes hollandaise et grecque avaient un point de vue assez différent de celui que nous avons adopté dès la rédaction de notre proposition de travail : nous souhaitions développer des outils pragmatiques mais sans doute empiriques; ils souhaitaient des outils fortement validés sur un plan épidémiologique. Nous avons à plusieurs reprises exprimé les raisons de notre choix et exposé clairement le cadre de nos développements, sans avoir l'impression d'être compris ou complètement entendus. C'est sans doute la longue fréquentation des données réelles issues du système EHLASS et la confrontation avec leur utilisation qui nous ont fait adopter ce point de vue et nous y tenir, avec le soutien notamment des équipes française et autrichienne.

- Enfin, les données non agrégées sur les ADL ne sont pas sans poser des problèmes à qui veut les utiliser concrètement. Rappelons-en quelques-uns :

> - coexistence de 2 méthodologies de recueil : l'une par voie de questionnaire auprès des ménages, l'autre par recueil dans les services d'urgence des hôpitaux;
>
> - multiplicité des systèmes de codage : nous sommes en présence d'au moins deux systèmes de codage de données : format V86 et format V96, sachant que l'institut allemand LOëGD a transcodé les données de V86 vers V96. Mais ce transcodage donne forcément lieu à une perte d'information du fait qu'il n'y a pas de bijection complète de code à code. Par ailleurs, il faut savoir qu'un nouveau système de codage est en cours de définition.
>
> - hétérogénéité de la qualité des données : certains systèmes de recueil sont représentatifs au niveau national, d'autres non, certains ont des caractéristiques très spécifiques (recueil dans les hôpitaux d'enfants), tandis que d'autres couvrent l'ensemble du champ des ADL.

C'est en ayant à l'esprit ces principes de développement et les difficultés rencontrées qu'il faut comprendre la logique des procédures décisionnelles proposées.

6- Le Score Synthétique de Dangerosité Relative - SSDR

(Synthetic Score of Relative Dangerosity - SSRD)

Quels sont les objectifs d'un tel score ?

Un des buts du système de recueil sur les ADL est de mettre en évidence la "dangerosité" de certains produits. Cette dangerosité peut se lire suivant les deux axes de la gravité et la fréquence de l'accident. On voudrait pouvoir repérer à la fois des produits causant des accidents peu fréquents mais graves et des produits causant des accidents fréquents mais moins graves.

Il est donc apparu utile de tenter de construire un Score Synthétique, à partir de critères simples et endogènes au système EHLASS, permettant de hiérarchiser la dangerosité potentielle des produits et qui, de plus, faciliterait les comparaisons entre les Etats membres et permettrait de suivre l'évolution de ce score au cours du temps.

Quelles sont les méthodes de calcul utilisées ?

- Nous avons caractérisé la dangerosité potentielle d'un produit en utilisant 3 variables endogènes au système : **l'effectif** (EFF = nombre des observations indiquant le code produit pour une variable "Produit", **le taux d'hospitalisation** (TH = nombre des hospitalisés pour ce code/EFF), **la durée moyenne de séjour** (DMS = somme totale des durées de séjour/nombre des hospitalisés pour ce code).

- On a choisi d'exclure les produits avec un effectif très faible (< 3), du fait du caractère non significatif du TH ou de la DMS (pour un effectif de « 1 » accident entraînant une hospitalisation, le TH serait de 100%).

Méthode n°1 utilisant les percentiles (SSRDP)

- On affecte aux critères EFF, TH, DMS un coefficient (coeff) variant de 1 à 20 en se fondant sur l'appartenance du produit à l'un des **percentiles.** Ainsi, pour la variable Effectif : après avoir classé les produits par ordre croissant d'effectif, on attribue :

> - le coefficient **1** pour les produits classés entre l'effectif minimum et le 5ème percentile (les effectifs les plus faibles),
> - le coefficient **2** pour les produits classés entre le 6ème percentile et le 10ème percentile,
> ...
> - le coefficient **20** pour les produits dont l'effectif est supérieur au 95ème percentile (les effectifs les plus forts).

- On agit de même pour les percentiles des variables DMS et TH, *mais uniquement sur les produits ayant au moins une hospitalisation (de façon à " lisser " les coefficients*). Ainsi, les coeff EFF, DMS, TH varient de 1 à 20.

- Ce calcul par percentile présente l'avantage de n'être pas sensible à la variabilité des valeurs extrêmes et aux points aberrants. Il permet aussi de ne pas être soumis à la variation du nombre absolu d'observations d'une année sur l'autre ou encore à la différence entre Etats du taux moyen d'hospitalisation et de sa durée moyenne, puisqu'il s'agit de coefficients fondés sur les classements des valeurs des variables et non sur leurs valeurs absolues.

Méthode n°2 utilisant les écarts-types (SSRDD)

- Nous avons choisi de développer aussi une autre méthode qui utilise la distribution des effectifs de chaque variable pour affecter les coefficients. Ainsi, pour chacune des 3 variables, on calcule la moyenne (m) et l'écart-type (s), puis on attribue un coefficient à un produit donné en fonction de sa place dans la distribution :

- coeff = 0 pour une valeur $X < m - 2\,s$
- coeff = 5 pour une valeur comprise entre $m - 2\,s \leq X < m - s$
- coeff = 10 pour une valeur comprise entre $m - s \leq X < m + s$
- coeff = 15 pour une valeur comprise entre $m + s \leq X < m + 2s$
- coeff = 20 pour une valeur $X \geq m + 2\,s$

Dans la dernière version du calcul du SSDR et après consultation des experts, nous avons décidé de choisir **un modèle additif équipondéré** : nous ajoutons les différents coefficients et nous leur attribuons le même poids.

Formule générale de calcul du SSDR :

SSDR = coeff EFF + coeff DMS + coeff TH

SSDR minimum = 0 (produit avec un effectif faible, sans hospitalisation).

SSDR maximum = 60 (produit avec un effectif très élevé, un TH très élevé, une DMS forte).

Nous fournissons en Annexe les programmes SAS de calcul du SSDR suivant ces 2 méthodes et appliqués tour à tour à la variable « Produit impliqué dans l'accident » puis à la variable « Produit ayant causé la lésion ».

Exemple d'utilisation

- Par exemple, pour les données EHLASS France 1997, les scores les plus élevés concernaient les "produits" suivants, en utilisant la méthode avec percentile :

- échelles (54)
- eau chaude (53)
- chien (51)

Cette liste fournit des pistes de réflexion pour mener des enquêtes complémentaires sur la sécurité des produits ou les conduites à risques dans des domaines précis. Elle est applicable distinctement pour les deux codes produits utilisés dans le système de codage : " Produit impliqué dans l'accident " et " Produit ayant causé la lésion ".

Quelles sont les utilisations possibles du SSDR ?

Remarques préliminaires :

- La liste des codes produits classés par ordre décroissant de SSDR n'est pas exploitable directement. Il faut sélectionner parmi ces produits ceux qui sont susceptibles de faire l'objet d'études spécifiques et d'actions de types réglementaires ou informatifs. On privilégiera un produit bien défini (par exemple : "Chaise haute pour bébé") plutôt que des produits trop génériques ("Toit", par exemple).

- Ce travail d'expertise est indispensable et ne peut être conduit que par un spécialiste des questions de sécurité des consommateurs et/ou de santé publique. Un programme informatique ne peut pas s'y substituer.

- L'énumération hétéroclite des produits renvoie à la richesse (ou à la pauvreté relative) de la nomenclature actuelle des produits dans le système de codage. Il faut donc aussi, parallèlement, progresser dans la direction d'une remise à jour et d'un enrichissement continu de la nomenclature des produits.

L'analyse temporelle :

Nous avions calculé à partir des données EHLASS France les SSDR (selon une ancienne méthode) de l'ensemble des produits pour chaque année entre 1990 et 1996. Nous avons obtenu un tableau qui permet de suivre l'évolution du SSDR de chaque produit. C'est donc un tableau à double entrée croisant le code produit et l'année. Nous avions ensuite établi à partir de ce tableau la courbe d'évolution du SSDR pour chaque produit.

Pour " expertiser " le tableau dans sa dimension temporelle, il faut repérer les produits qui présentent une évolution remarquable :

- soit la baisse du SSDR est continue au sens large sur plusieurs années (il faut pour cela que le SSDR de l'année N+1 soit inférieur ou égal au SSDR de l'année N),

- soit la hausse du SSDR est continue au sens large sur plusieurs années (il faut pour cela que le SSDR de l'année N+1 soit supérieur ou égal au SSDR de l'année N),

- soit une évolution particulièrement forte d'une année sur l'autre du SSDR (perte ou gain de plus de 30 points, par exemple),

- on peut encore utiliser la technique de la régression linéaire pour mesurer la tendance de l'évolution du SSDR : un coefficient directeur positif de la droite d'ajustement indiquera un SSDR en tendance haussière sur les années prises en compte, un coefficient négatif une baisse, en tendance, sur ces mêmes années.

La mesure de l'efficacité :

Nous pouvons aussi regarder ce tableau en fonction des mesures réglementaires prises et essayer de visualiser leurs effets sur les courbes du SSDR en fonction du temps. Une baisse dans la courbe devrait correspondre à la prise d'une mesure efficace réduisant de façon sensible la fréquence et/ou la gravité des accidents liés à cette classe de produits.

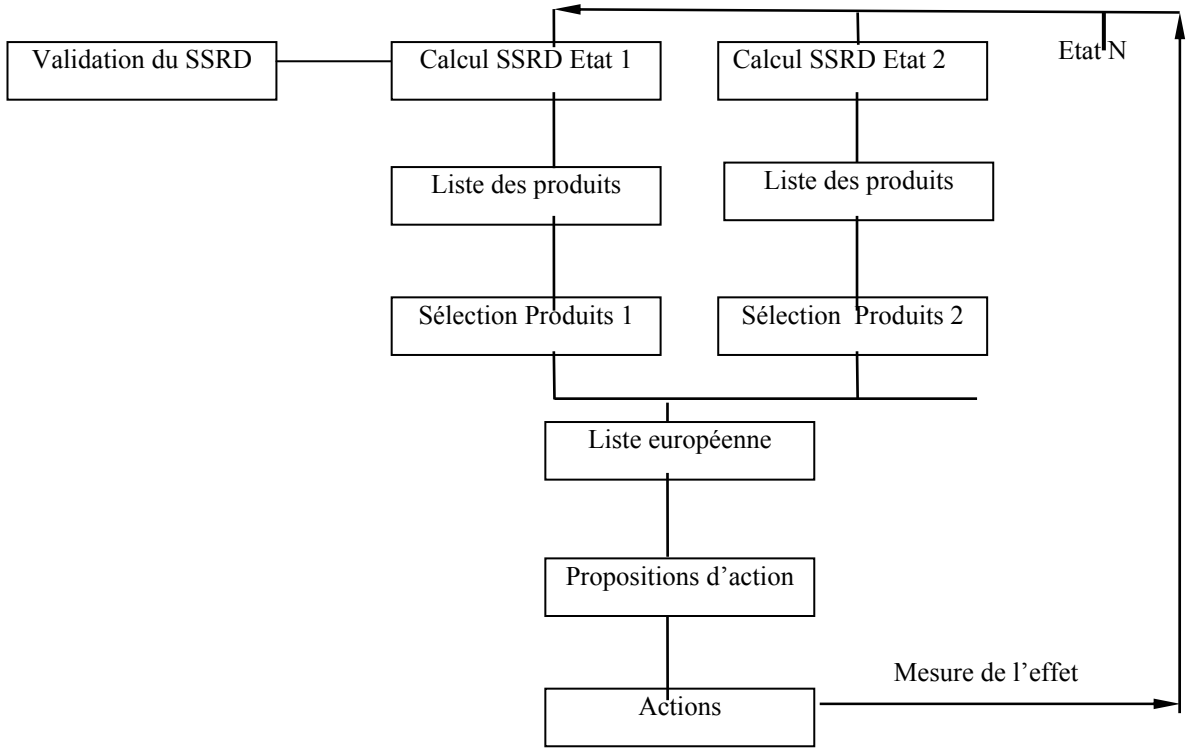
Les indications sur les futures mesures réglementaires à prendre :

A l'inverse, une hausse continue sur plusieurs années du SSDR ou un score continûment élevé peut conduire à une réflexion sur des mesures réglementaires à prendre pour ces types de produits.

La comparaison entre Etats :

On peut aussi utiliser le SSDR pour comparer la dangerosité potentielle des produits entre Etats. Une fois la procédure de calcul validée au niveau de l'ensemble des Etats, on peut envisager le calcul et l'exploitation du SSDR suivant le schéma qui suit :

Schéma d'utilisation du SSDR :



Autres utilisations :

- On peut aussi publier cette liste de produits en utilisant la nomenclature agrégée, sur 3 positions par exemple, pour travailler sur des groupes de produits plus larges.

- On peut aussi stratifier le SSDR en fonction de classes d'âge spécifiques (cas du rapport EHLASS Portugal) et repérer ainsi des produits potentiellement " dangereux " uniquement pour certaines classes d'âge (les bébés, les enfants, les adolescents, les personnes âgées, etc.) ou encore repérer les produits potentiellement " dangereux " pour certaines caractérisations d'accidents (les fractures, les chutes, les brûlures, etc.).

A-t-on déjà utilisé le SSDR ?

L'équipe française l'utilise depuis plusieurs années. Nous rapportons ici l'essai de l'utilisation du SSDR avec une ancienne méthode de calcul (les principes de calcul étaient identiques mais le système de pondération était différent - variation de 1 à 100) utilisée sur les données françaises codées en V86 :

- Evolution de la moyenne du SSDR :

	Nb de produits	Moyenne
1990	542	38.55
1991	540	36.97
1992	504	36.91
1993	507	36.44
1994	544	35.08
1995	573	37.15
1996	508	34.54

D'après ce tableau, chaque année environ 540 produits avaient un SSDR sur les 1100 présents dans la nomenclature des produits de la version de codage V86. Le SSDR moyen est relativement stable et se situe autour de 36. Ceci illustre que c'est bien une dangerosité potentielle " moyenne " qui est calculée et que le score de chaque produit est un score relatif par rapport aux autres produits et non une mesure absolue.

- Des SSDR décroissants :

Nous avons pu repérer un certain nombre de produits dont le SSDR est décroissant sur les 4 ou 5 dernières années. **Pour la plupart, nous avons pu mettre en relation cette baisse avec des mesures réglementaires prises pour ces produits.** Ainsi pour :

- 01305 Berceau - hamac
- 02090 Poussette pliante

- 04010 Liquide vaisselle
- 04130 Liquide lave linge (avec exemple de régression linéaire : coeff. corrélation : -0.91, coeff. directeur : -7.64)

- 07999 Sol, plancher autre non spécifié (sol des aires de jeux ?)
- 22099 Revêtement de plancher non spécifié

- 12780 Solvants
- 37788 Pesticides, insecticides autres spécifiés

- 51388 Ski autre spécifié

- Des SSDR croissants :

Nous avons aussi pu repérer un certain nombre de produits dont le SSDR est croissant sur les 3 ou 4 dernières années et qui nécessiteraient donc un examen particulier. Par exemple **:**

- 26060 Casserole

- 28620 Feu d'artifice
- 52150 Saut de cheval (résultats de la régression linéaire) :
coefficient de corrélation : +0.58, coefficient directeur : +3.86)

- Des SSDR constamment élevés :

Nous avons pu aussi repérer un certain nombre de produits dont le SSDR est constamment élevé et qui ont nécessité un examen particulier. Comme **:**

- 16450 Eau chaude
- 37260 Chien

Quelles sont les limites de la méthode ?

La causalité du produit dans l'accident :

Nous avons mené ce travail sur la variable " Produit impliqué dans l'accident ". Or le produit impliqué n'est pas forcément (et même pas souvent !) la cause directe de l'accident. Il peut s'agir d'une mauvaise utilisation, d'un concours de circonstances (inattention, intervention d'un tiers, etc.) ou encore d'autres facteurs (alcoolémie, malaises, etc.). Le facteur déclenchant de l'accident est rarement le produit lui-même. Le SSDR ne mesure qu'une **dangerosité potentielle** qui n'est pas forcément liée au produit lui-même, mais qui peut être liée aux circonstances de son utilisation ou aux caractéristiques de ses utilisateurs. Il est cependant utile de hiérarchiser cette dangerosité potentielle et de fournir ainsi des pistes pour examiner en priorité certaines situations à risques.

La nomenclature des produits :

On a souvent relevé que la nomenclature des produits devait évoluer rapidement. Le passage du système de codage V86 à V96 a déjà permis de passer d'une nomenclature d'environ 1100 postes à une nomenclature d'environ 1700 postes. Cette nomenclature a besoin elle-même d'être renouvelée pour accroître l'efficacité du SSDR. Les produits très rarement cités ne sont pas pris en compte (si la fréquence < 3). La dangerosité potentielle de nouveaux produits, de ceux qui n'ont pas de code produit correspondant ou qui font partie d'une large classe de produits déjà existants ne sera pas mise en évidence.

La fréquence d'utilisation des produits :

Cette méthode ne prend pas en compte, bien entendu, la fréquence d'utilisation des différentes classes de produits dans la vie quotidienne. Il ne s'agit pas de comparer des dangerosités potentielles à même fréquence d'utilisation. Ainsi un produit comme " couteau " d'usage fréquent aura plus de chance de figurer en tête dans notre liste que le produit " bétonnière ". De ce point de vue, les différences culturelles entre Etats vont aussi agir sur la hiérarchie des dangerosités potentielles. Nous trouverons plus d'accidents avec des barbecues dans les pays de l'Europe du Sud que dans ceux de l'Europe du Nord. Il faut tenir compte de ces faits dans les analyses comparatives entre Etats.

L'arbitraire de la définition du SSDR :

Nous avons construit deux formules de calcul du SSDR. On pourrait en imaginer d'autres en travaillant avec un découpage encore plus fin des percentiles, en attribuant d'autres coefficients de pondération (nous avons ici choisi un modèle additif équipondéré) ou en utilisant autrement la distribution des effectifs, etc.

Cependant, nous avons testé la modification du poids des coefficients. Il apparaît que, quel que soit le système de pondération utilisé, on obtient globalement les mêmes produits potentiellement les plus " dangereux ", les mêmes produits potentiellement les moins " dangereux ". Les variations de classements ne seraient donc que " locales " et la méthode relativement stable vis-à-vis des coefficients de pondération utilisés.

Les pratiques hétérogènes de codage des produits :

Le SSDR pourrait être utilisé pour des comparaisons entre Etats. Cependant les pratiques de codage ne sont pas homogènes :

Actuellement, le système de codage comprend trois codes produits : le produit qui est impliqué dans l'accident, celui qui est à l'origine de la lésion et un code produit complémentaire. En examinant les données recueillies, on constate que les pratiques de codage sont très différentes d'un Etat à l'autre :

- Certains Etats utilisent un code " Aucun produit " pour signifier qu'aucun produit n'est en cause et un code " Autre produit " pour signifier qu'un produit est impliqué dans l'accident mais qu'il ne fait pas partie de la nomenclature. D'autres Etats n'utilisent pas ces codes qui sont donc agrégés avec la modalité " Non Spécifié ".

- Donnons quelques exemples de ces différences de stratégie de codage (données 1995) :

Le Danemark distingue pour la variable " Produit impliqué dans l'accident " entre les modalités :

- Aucun produit impliqué : 82,6%
- Produit impliqué non spécifié : 0,2%
- Autre produit impliqué : 1,2%

L'Italie distingue pour la variable " Produit impliqué dans l'accident " entre les modalités :

- Aucun produit impliqué : 20,3%
- Non Spécifié : 2,8%
- Autre produit impliqué : 9,3%

La Belgique distingue pour " Produit impliqué dans l'accident " entre les modalités :

- Aucun produit impliqué : 67,3%
- Autre produit impliqué : 6,7%

mais n'a pas de modalité " Non Spécifié " explicite

Ces différences rendent difficiles les comparaisons transnationales.

En définitive que mesure-t-on avec le SSDR ?

Le SSDR est une méthode pour hiérarchiser la dangerosité potentielle relative des produits dans le système de codage. Cette dangerosité est rarement liée au produit lui-même mais plus au contexte de son utilisation (habitudes, circonstances, populations, etc.). Cette méthode permet cependant de mettre en évidence pour certains produits des baisses significatives de " dangerosité " ou au contraire des hausses qui doivent attirer l'attention des autorités.

L'information sur les produits réellement défectueux ne peut figurer que dans le texte en clair. Cette recherche est redevable d'une méthodologie plus fine utilisant l'expertise humaine.

Malgré les limites évoquées plus haut, il nous semble, pour l'avoir utilisée sur les fichiers EHLASS France, que cette méthode permet une approche intéressante de la notion de dangerosité potentielle relative des produits dans chaque Etat.

Rappelons encore que ce score est endogène, c'est-à-dire interne au système et qu'il est relatif, par rapport aux autres produits. Il présente l'immense avantage d'être simple dans son principe, immédiat dans son calcul, déjà automatisé et mis en œuvre. Rappelons aussi que le recours à l'expertise humaine est indispensable pour sélectionner les produits susceptibles de faire l'objet des mesures préventives ou correctives.

7- Le Système d'Alerte Automatisée - SAA

(Automated Alert System - AAS)

Quelle est la méthode utilisée ?

Il s'agit d'une procédure statistique simple qui consiste à calculer la fréquence relative d'apparition, d'un produit par exemple, sur une période choisie (un mois, un trimestre, etc.) d'une année (Période 1) et de la comparer à la fréquence relative d'apparition sur la même période d'une autre année (Période 2). Si la fréquence relative de la Période 1 est supérieure (ou inférieure) à un seuil donné à la fréquence relative de la Période 2, « l'alerte automatisée » se déclenche : on édite les modalités des variables sélectionnées pour examiner la cause de cette variation de fréquence relative.

Cette procédure peut être mise en œuvre systématiquement tous les mois ou tous les trimestres sur l'ensemble des codes produits (niveau à 5 caractères du code produit), ainsi que sur toutes les autres variables (mécanisme, type de lésion, etc.), le but étant de détecter des variations brusques dans les fréquences relatives de distribution des modalités de ces variables. Les différents paramètres (périodicité, seuil) sont paramétrables par l'utilisateur.

Cette procédure complète l'alerte humaine, puisqu'elle examine systématiquement et automatiquement l'ensemble des fréquences relatives d'apparition des modalités. Elle pourrait aussi s'appliquer à l'ensemble des mots signifiants du texte en clair : par exemple le mot « LEGO » serait apparu 21 fois durant le premier trimestre 1998 contre 9 fois durant le premier trimestre 1999.

Quand ces variations (augmentation ou diminution) sont détectées et isolées automatiquement, il faut ensuite examiner, par une expertise humaine, si ces variations proviennent de problèmes liés au système de codage, au recueil de données, à des changements de comportement des consommateurs, à l'apparition d'une nouvelle référence dans une gamme de produits existants, d'un nouveau type de produit ou encore d'une importation ponctuelle d'un certain produit, etc.

Comment fonctionne la procédure développée ?

Les paramètres de sélection :

> 1- Sélection de la Période 1 comprise entre deux dates sous la forme jjmmaa :
> ex : **010198 310398** (on sélectionne les données du 1er trimestre 1998)
>
> 2- Sélection de la Période 2 comprise entre deux dates sous la forme jjmmaa :
> ex : **010199 310399** (on sélectionne les données du 1er trimestre 1999)
>
> 3- Choix du seuil de sélection de la variation (en fréquence) : **0.2** (on sélectionne les modalités dont l'effectif relatif varie d'au moins 20% - en plus ou en moins)

La procédure va donc :

1- sélectionner dans la base des enregistrements EHLASS en format V96 les 2 sous-populations (dans notre exemple, celles du 1er trimestre 1998 et du 1er trimestre 1999);

2- construire les tableaux des effectifs pour l'ensemble des variables (sexe, âge, mécanisme, traitement, lésion 1 et 2, activité, partie lésée 1 et 2, produit impliqué et produit ayant causé la lésion) croisés par les 2 périodes;

3- comparer les fréquences relatives des modalités des 2 périodes;

4- si cette différence est supérieure ou égale au seuil choisi, éditer la modalité correspondante avec le pourcentage d'augmentation ou de diminution.

Il nous a semblé nécessaire dans l'algorithme de sélection :

- de travailler avec des effectifs relatifs (pour s'affranchir de la variation d'effectif total d'une période à l'autre);

- de distinguer les modalités dont l'effectif est faible (< 30) des autres : quand l'effectif est faible - 2 ou 3, par exemple - une augmentation d'effectif de 1 ou 2 est très forte exprimée en pourcentage et pas nécessairement significative;

- de faire jouer un rôle symétrique aux 2 périodes : nous calculons le pourcentage de variation de P1 à P2 et celui de P2 à P1;

- de tenir compte du cas où l'un des effectifs est nul.

C'est pourquoi nous avons choisi la méthodologie suivante :

Pour une modalité d'une variable :

Si l'effectif de la Période 1 (Eff(P1)) et l'effectif de la Période 2 (Eff(P2)) >= 30

et si la variation de la fréquence relative de la modalité de P1 à P2 ou de P2 à P1 est supérieure ou égale au seuil choisi (ex : 20% ou 0.20), i.e. si :

(Eff(P2) x Eff Total (P1)) / (Eff(P1) x Eff Total (P2)) – 1 > **0.20** (augmentation de P1 à P2)
ou si
(Eff(P1) x Eff Total (P2)) / (Eff(P2) x Eff Total (P1)) - 1 >= **0.20** (diminution de P1 à P2)

alors la modalité est sélectionnée;

Si 0 < Eff(P1) < 30 ou 0 < Eff(P2) < 30

et si la variation de l'effectif « normalisé » entre les 2 périodes est supérieure ou égale au nombre (30 x % du seuil de sélection), soit pour notre exemple : 30 x 0.20 = **6**, i.e. si : la variation de fréquence relative de la modalité de P1 à P2 ou de P2 à P1 est supérieure ou égale à ce seuil (ici 6),

alors la modalité est sélectionnée;

Si Eff(P1) = 0 ou Eff(P2) = 0

et si la variation de l'effectif brut entre les 2 périodes est supérieure ou égale au nombre (30 x % du seuil de sélection), soit pour notre exemple : 30 x 0.20 = **6**.

alors la modalité est sélectionnée.

On édite ensuite, variable par variable, la liste des modalités sélectionnées avec leurs fréquences relatives et les pourcentages de variation.

A-t-on déjà utilisé cette procédure ?

Nous avons appliqué cette procédure aux fichiers EHLASS France avec les paramètres de sélection suivants :

- Première période : du 010196 au 310396 (1T96)
- Deuxième période : du 010197 au 3/0397 (1T97)
- Seuil de sélection de la variation : 0.2 (20%)

Les augmentations

Parmi les modalités dont la fréquence relative a augmenté entre le premier trimestre 1996 et 1997 on relève :

- Pour la variable « Produit impliqué dans l'accident » :

- les outils autres spécifiés (en 1T96 : 33/12810, en 1T97 : 39/10020) : + 51%
- les médicaments (en 1T96 : 100/12810, en 1T97 : 114/10020) : + 46%
- les bicyclettes enfants (en 1T96 : 138/12810, en 1T97 : 155/10020) : + 44%
- les patins à roulettes (en 1T96 : 66/12810, en 1T97 : 74/10020) : + 43%
- le ski (en 1T96 : 45/12810, en 1T97 : 51/10020) : + 45%
- les balles et ballons (en 1T96 : 90/12810, en 1T97 : 109/10020) : + 55%

- Pour la variable « Produit ayant causé la lésion », on trouve :

- les outils autres spécifiés (en 1T96 : 33/12810, en 1T97 : 39/10020) : + 51%
- eau chaude (en 1T96 : 38/12810, en 1T97 : 46/10020) : + 55%
- glace, neige, gel (en 1T96 : 250/12810, en 1T97 : 379/10020) : + 94%
- trottoirs, neige, glace (en 1T96 : 42/12810, en 1T97 : 51/10020) : + 55%
- pistes de ski (en 1T96 : 97/12810, en 1T97 : 270/10020) : + 256%
- skier autre spécifié (en 1T96 : 58/12810, en 1T97 : 88/10020) : + 94%

Ces quelques exemples de variation sur les produits peuvent conduire aux réflexions suivantes :

- Les accidents avec les « outils autres spécifiés » sont en augmentation. Il faut aller voir dans le texte en clair si l'on retrouve la description des outils en cause. Sinon, on peut envisager une étude spécifique avec un retour aux sites de recueil pour interrogation des accidentés.

- Le nombre d'accidents par intoxication avec des médicaments est toujours préoccupant. Mais, il faut aussi vérifier si la structure par âge de nos 2 échantillons est strictement comparable : un accroissement du pourcentage de jeunes enfants (1-9 ans) pouvant expliquer l'accroissement du nombre de ce type d'intoxication. Ce n'est pas le cas ici.

- Il en est de même pour l'évolution du nombre des accidents impliquant des bicyclettes d'enfants et des patins à roulettes. On peut alors se demander s'il y a eu apparition de nouveaux modèles plus dangereux ou si les conditions climatiques ont été plus nettement favorables à la pratique de ces activités de plein air au cours du 1er trimestre 97.

- Le fort accroissement du nombre des accidents sur le sol gelé est sans doute lié aux chutes de neige plus abondantes au cours du premier trimestre 97. Mais ce phénomène climatique explique-t-il à lui seul l'augmentation considérable du nombre des accidents sur les pistes de ski ? Un décalage dans les périodes de vacances scolaires ou encore une fréquentation accrue des stations de sport d'hiver peuvent-il expliquer une partie de cette variation ?

- Pour la variable « Traitement », on trouve :

> - une forte augmentation de la modalité « traitement par le généraliste »
> (en 1T96 : 1789/12810, en 1T97 : 2197/10020) : + 57%,
> - une augmentation de la modalité « aucun traitement »
> (en 1T96 : 452/12810, en 1T97 : 432/10020) : +22% et de la modalité « décès ».

- Pour la variable « Age », on trouve :

> - une augmentation de la classe « < 1 an » (en 1T96 : 213/12810, en 1T97 : 219/10020)
> soit + 31%.

- Pour la variable « Sexe », on ne trouve pas de variation forte.

- Pour la variable « Mécanisme », on trouve :

> - un fort accroissement des chutes (+ 154%), qu'il faut mettre en relation avec les conditions climatiques (trottoirs glissants du fait des chutes de neige);
> - une augmentation des coupures (+ 26%), qu'il faut peut-être mettre en relation avec l'augmentation des accidents impliquant des outils;
> - les accidents par flammes et liquides chauds augmentent aussi (+ 54% et + 29%). Ce phénomène est-il lié encore aux conditions climatiques ou à la distribution d'une eau domestique légèrement plus chaude ?

- Pour la variable « Lieu », on note :

> - une augmentation de la modalité « alentours de la maison » (+ 63%) sans doute en relation avec l'augmentation des chutes;
> - une augmentation des « lieux inconnus » (+ 28%) : est-ce un problème d'imprécision de la table de codes ou une équipe de codeurs qui remplit moins correctement cette variable ?

- Pour la variable « Activité », on note :

> - une augmentation forte des accidents liés à une pratique sportive : éducation physique (+ 29%), sport (+ 54%), sport non organisé (+ 225%).

- Pour la variable « Lésion », on note :

> - une augmentation de la modalité « abrasion » (+ 40%) qui pourrait être liée à l'augmentation des accidents impliquant des outils ou encore des sports;
> - une augmentation de la modalité « pas de lésion diagnostiquée» (+ 90%), du fait du peu de gravité des chutes (?).

- Pour la variable « Partie lésée », on note :

> - une augmentation des accidents au niveau de la clavicule (+ 33%), de la partie inférieure du dos (+ 24%) et de l'épaule (+ 41%);
> - une augmentation de « partie inconnue » (+ 87%), ce qui confirmerait qu'une équipe de codeurs remplit moins correctement cette variable.

Les diminutions

Parmi les modalités dont la fréquence relative a diminué entre le premier trimestre 1996 et 1997, on relève :

- Pour la variable « Produit impliqué dans l'accident » :

> - sols autres non spécifiés (en 1T96 : 935/12810, en 1T97 : 43/10020) : - 1601% (!) Il semble que cette très forte baisse procède d'un effort de certaines équipes de codage pour mieux utiliser les codes spécifiés;
> - goudron (en 1T96 : 99/12810, en 1T97 : 55/10020) : - 41% (?)

- Pour la variable « Produit ayant causé la lésion » :

> - inconnu (en 1T96 : 1340/12810, en 1T97 : 428/10020) : - 145%
> - sports autres non spécifiés (en 1T96 : 79/12810, en 1T97 : 42/10020) : - 47%
> - autres (en 1T96 : 1039/12810, en 1T97 : 616/10020) : - 32%
> Ces diminutions confirment l'effort de codage que nous évoquions précédemment.

- Pour la variable « Traitement », on trouve :

> - une diminution de la modalité inconnue ou tansfert (en 1T96 : 251/12810, en 1T97 : 18/10020).

- Pour les variables « Age » et « Sexe » on ne trouve pas de forte variation.

- Pour la variable « Mécanisme », on trouve :

> - une forte diminution de la modalité inconnue (en 1T96 : 495/12810, en 1T97 : 60/10020) soit - 545% (!) et de autre mécanisme (en 1T96 : 119/12810, en 1T97 : 38/10020) : - 145%
> - une diminution de corps étranger (en 1T96 : 553/12810, en 1T97 : 135/10020) : -220%

- Pour la variable « Lieu », on note :

> - une forte diminution de la modalité zone sportive (en 1T96 : 200/12810, en 1T97 : 94/10020) : - 66% et autre zone de sport (en 1T96 : 1043/12810, en 1T97 : 575/10020) :
> - 42%. Ce phénomène peut être mis en relation avec le fait que l'on a relevé précédemment une forte augmentation des accidents de sport non organisé (+ 225%);
> - la modalité « domicile non précisé » diminue (en 1T96 : 1319/12810, en 1T97 : 851/10020) : - 21%, ce qui participe aussi de l'effort de précision.

- Pour la variable « Activité », on note :

> - une forte diminution de la modalité « sport organisé » (en 1T96 : 312/12810, en 1T97 : 61/10020) : - 300%, qui est symétrique de l'augmentation de la modalité « sport non organisé »;
> - les modalités autres activités et inconnu diminuent (- 144% et - 31%), ce qui participe du même effort de précision.

- Pour la variable « Lésion 1», on note :

> - une diminution des intoxications (- 30%)
> - là aussi, une diminution de la modalité inconnue (- 33%)

- Pour la variable « Lésion 2», on note :

> - une diminution des fractures (- 27%)

- Pour la variable « Partie lésée 1», on note :

> - une diminution des parties lésées « œil » (- 90%), « oreille » (- 49%) et « cavité buccale »
> (- 51%).

Que peut-on conclure ?

A la fin de ce cheminement, on constate que le SAA génère des questions touchant à :

La méthodologie même du système EHLASS :

Le recrutement des hôpitaux est-il suffisamment stable pour ne pas biaiser les variations de ce seul fait ? Comment évoluent les pratiques de codage ? Quelle est la pertinence des tables de code utilisées actuellement ?

On a vu, en ce qui concerne les données françaises, que des efforts de précision ont été faits entre 1996 et 1997 dans le codage des variables « Produits », « Traitement », « Mécanisme », « Activité », « Lésion », alors qu'au contraire les variables « Lieu » et « Partie lésée » étaient globalement moins bien codées. Cela conduit à renforcer l'attention des équipes de codage sur ces variables et/ou à envisager une révision des tables de codage.

La méthode d'analyse des variations exhibées :

Il convient de mettre au point une méthode d'analyse des variations mises en évidence par le SAA. On a vu qu'une simple variation climatique entre les 2 années peut expliquer le fort accroissement des chutes sur sol gelé et enneigé, sans que cela nécessite une alerte spécifique. Il faut donc prendre en compte tous les facteurs exogènes de variation qui peuvent « expliquer » une part de la variabilité. Cette méthode fondée sur l'expertise humaine pourra se construire au cours du temps avec l'expérience de l'outil (établissement d'une liste logique de questions à se poser avant de conclure à une variation endogène).

Malgré toute la bonne volonté et le talent de l'expert, il reste sans doute une part de ce que les statisticiens nomment « la part de variabilité inexpliquée ». Cependant, il faut avoir épuisé tout l'arsenal explicatif pour cerner cette part.

La santé publique :

L'augmentation du nombre d'accidents par intoxication avec des médicaments est préoccupante (+ 46%) à structure d'âge égale entre les 2 échantillons (sauf pour les < 1 an), alors qu'il y a, par ailleurs, une diminution du nombre total des intoxications (- 30%).

La sécurité des produits :

Les accidents avec les « outils autres spécifiés » sont en augmentation (+ 51%), la croissance du nombre des abrasions (+ 40%) pourrait lui être liée. La croissance du nombre d'accidents de patins à roulettes est aussi spectaculaire entre les 2 périodes (+ 43%). Il faut examiner si la sécurité de certains de ces produits est directement en cause dans ces accidents.

Les comportements :

Par exemple, l'augmentation du nombre d'accidents de ski (+ 45%) conduit à se poser la question du comportement des skieurs sur les pistes.

Tels sont les types de questions générées par l'utilisation du SAA.

Quelles sont nos recommandations concernant l'utilisation cette procédure ?

Nous recommandons d'utiliser le SAA :

- avec des échantillons d'observations les plus stables possibles au regard du recrutement des patients (âge, types de pathologie prise en charge, exhaustivité du recueil, etc.);

- à intervalles réguliers pour confirmer ou infirmer les évolutions sur un plus grand nombre de périodes;

- en prenant des périodes de référence distantes de 12 mois pleins, de façon à s'affranchir des variations saisonnières (augmentation des accidents de ski en hiver, etc.);

- en tenant le plus grand compte des facteurs de variation exogènes au système de recueil (conditions climatiques, événements ponctuels, dates de vacances, ...) de façon à ne pas créer de « fausses questions » relatives à la santé publique ou à la sécurité des produits;

- en construisant peu à peu une méthodologie fine d'analyse des variations.

On voit que le système d'alerte automatisée (SAA) que nous avons développé est plus un « générateur de questions » qu'un générateur de réponses. Les réponses sont apportées par l'irremplaçable expertise humaine. Mais, pour que les réponses soient pertinentes, il faut que le système d'information génère les bonnes questions : c'est le rôle du SAA. Une méthodologie solide d'analyse des variations exhibées par cette procédure se construira peu à peu pour fonder une alerte systématique efficace.

8- Une échelle de sévérité de l'accident

(Severity SCale - SSC)

Nous avions d'abord proposé la création d'une Note de Gravité de l'Accident (NGA). Parallèlement l'équipe danoise a construit une échelle de sévérité de l'accident (Severity Scale). C'est cette dernière que nous avons finalement choisie de développer dans le cadre de nos procédures SAS. Toutefois, nous fournissons ci-dessous les éléments de construction de notre Note de gravité qui sont proches de la méthode danoise. Nous fournissons en Annexe le document du NIPH relatif à l'établissement de l'échelle de sévérité.

La Note de Gravité de l'Accident

Quel a été le contexte de la création de cet outil ?

- On sait que le système de recueil n'est pas représentatif dans certains Etats, en France par exemple, de l'ensemble des ADL passant par les services d'urgence. Dès lors, il n'est pas possible de suivre l'évolution du nombre de ces accidents à partir de cette seule source d'information. Si on ne peut pas retracer l'évolution quantitative des ADL à partir des seules données EHLASS, on peut tenter de construire un indicateur de gravité de l'accident qui permettrait d'avoir une idée de l'évolution de la moyenne de ce facteur et de concentrer les actions de prévention sur les accidents les plus graves.

- Nous avons proposé, dans un chapitre précédent, la construction d'un Score Synthétique de Dangerosité Relative (SSDR) permettant de hiérarchiser la "dangerosité potentielle" des produits à partir de critères simples et endogènes au système. Mais, le mode de calcul du score fait qu'il mesure une dangerosité relative d'un produit par rapport à un autre et non une dangerosité absolue. Il ne permet donc pas de mesurer l'évolution de la " dangerosité potentielle moyenne " des produits dans le temps. De plus, ce score ne concerne, par définition, que les produits et ne caractérise donc pas chaque accident en propre. C'est pourquoi, nous proposons la construction d'une Note de Gravité absolue calculable pour chaque accident.

- Par ailleurs, dans le cadre d'un autre projet IPP (n°1999/IPP/1022) " COCOL ", l'équipe danoise a proposé une échelle de sévérité en 3 niveaux sur ces mêmes données. En accord avec le groupe d'experts Data Mining du présent projet, nous avons estimé que, dans un premier temps :

> - nous pouvions développer les deux approches pour choisir ensuite la plus adéquate;
> - il fallait rendre les plus cohérentes possibles les deux méthodologies pour les comparer.

- C'est pourquoi, nous avons transformé assez profondément le calcul initialement proposé de la NGA dans notre projet pour le rendre le plus compatible possible avec la méthodologie proposée dans le document en date du 14 Avril 2000 " A three level severity scale for use in EHLASS by Henning Bay-Nielsen and Birthe Frimodt-Moller - National Institute of Public Health - Denmark " (voir Annexe).

Quelle était la méthode utilisée ?

Nous avons construit cette note de gravité (NGA) suivant les principes exposés ci-dessous :

- Chaque observation se voit attribuer une note. Pour construire cette note nous ne voulions prendre en compte que des informations issues des variables du système EHLASS.

- La « gravité au sens EHLASS » se mesurera donc à partir des 3 variables suivantes, auxquelles nous associerons un score :

> - variable " Traitement "
> - variable " Durée d'hospitalisation "
> - variable " Type de lésion "

Exclusions

- Comme dans la démarche danoise et sur la base des mêmes justifications, nous avons exclu du calcul les observations caractérisées par :

> - un décès, car ces cas ne sont pas représentatifs dans le système (code Traitement = 7),
> - un code Traitement " Inconnu " (code Traitement = 9),
> - le code lésion " Pas de lésion diagnostiquée " ou " Inconnue " ou " Autre " (code Type de lésion = 97, 98 et 99), car il est alors difficile d'apprécier la gravité de l'accident;
> - une partie lésée " Dents " (code Partie du corps lésée = 15), si elle n'est pas associée à une fracture ou une luxation-dislocation (code Type de lésion = 05 ou 06). En effet, ces cas ne relèvent pas à proprement parler du domaine des ADL.

Mode de calcul de la NGA :

1- la variable " Traitement " :

A la variable " Traitement ", nous associons le score Trt :

Traitement - V96	Score Trt
Examiné et renvoyé sans traitement	1
Renvoyé après traitement initial	2
Traité, puis traitement par le généraliste	3
Traité, puis traitement en ambulatoire	4
Hospitalisation	8
Transfert vers un autre hôpital	8
Décès	Exclus
Inconnu	Exclus

- Rappelons que nous ne calculons pas de NGA pour une observation ayant une modalité " Décédé " ou " Inconnu" pour la variable Traitement.

2- la variable " Durée d'hospitalisation " :

A la variable " Durée d'hospitalisation ", nous associons le score DH :

Durée d'hospitalisation - V96	Score DH
pas d'hospitalisation	0
de 1 à 3 jours	2
de 4 à 10 jours	4
plus de 10 jours	7

3- la variable " Type de lésion " :

A la variable " Type de lésion ", nous associons le score Lés que nous avons hiérarchisé en 3 niveaux :

Type de Lésion – V96	Score Lés
01- Commotion	1
02- Contusion, ecchymose	1
03 - Ecorchure	1
04 - Plaie ouverte	3
05 - Fracture	5
06 - Luxation, déboîtement	5
07 – Foulure, entorse	3
08 - Lésion des nerfs	5
09 - Lésion des vaisseaux sanguins	5
10 - Lésions des tendons et des muscles	5
11 – Ecrasement	5
12 - Amputation	5
13 - Intoxication	1
14 - Brûlure	1
15 - Corrosion	1
16 - Electrocution	1
17 - Irradiation	1
18 - Engelures	1
19 - Asphyxie	1
Pas de lésion diagnostiquée	Exclus
Autre type de lésion	Exclus
Inconnu	Exclus

- Nous ne calculons pas de NGA pour une observation ayant une modalité de la variable « Type de lésion » égale à " Inconnu", " Autre " ou " Pas de lésion diagnostiquée ".

- Quand 2 types de lésion sont signalés conjointement (variables Lésion 1 et Lésion 2 renseignées), nous retenons le score maximum.

- Il s'agit bien entendu d'une estimation " en moyenne " de la gravité relative d'un type de lésion. Généralement, la contusion est moins grave que la fracture, la lésion des vaisseaux plus grave que

l'entorse, etc. Des contre-exemples existent et cette hiérarchie peut être contestée. Cependant, nous pouvons avancer deux types d'arguments :

> - par l'amplitude de variation des scores et l'emploi d'un modèle additif, nous avons privilégié les 2 critères objectifs (Traitement : score variant de 1 à 8; Durée d'hospitalisation : score variant de 0 à 7) et minoré ce critère de gravité plus difficilement appréciable (Type de lésion : score variant de 1 à 5);
>
> - dans la mesure où l'on utilise cette note surtout pour des comparaisons historiques ou géographiques, l'erreur commise est constante d'une année sur l'autre ou d'un Etat à l'autre. Il est clair qu'il ne s'agit pas de mesurer une " gravité absolue ", mais une forme de gravité au sens du système d'information.

- En définitive, nous calculons la Note de Gravité suivant un modèle additif, comme somme des 3 scores précédents :

Note de Gravité (NGA) = Score Trt + Score DH + Score Lés

Exemples de NGA :

- Voici quelques exemples d'accidents avec leur NGA respectif :

Caractérisation de l'accident	NGA
Chute dans un escalier entraînant une légère commotion - examiné et renvoyé sans traitement	2
Au rugby - contusion - renvoyé après traitement initial	3
Chute avec fracture - 4 jours d'hospitalisation	15
Chute de cheval - plaie ouverte - traitement initial	5
Intoxication par liquide de frein - 1 jour en observation	11
Brûlure par un barbecue - 20 jours d'hospitalisation	16

- Le tableau suivant donne la variation de la NGA suivant les principaux types de traitement :

Scores		Trt	DH	Lés	NGA
Non traité	minimum	1	0	1	**2**
	maximum	1	0	5	**6**
Traité	minimum	2	0	1	**3**
	maximum	4	0	5	**9**
Hospitalisé	minimum	8	2	1	**11**
	maximum	8	7	5	**20**

Remarques :

- Les autres variables ne nous ont pas paru exploitables du point de vue du calcul de la NGA.

- La NGA varie en fait de 2 à 20.

- Il y a un certain arbitraire dans le mode de calcul de cette note, mais il nous semble que la hiérarchie de la gravité des accidents est globalement respectée. Il s'agit bien de construire une variable ordinale pour rendre compte de la gravité de l'accident.

A-t-on un exemple d'utilisation ?

Nous avons appliqué ce calcul à l'ensemble des observations du fichier EHLASS France 1986-1998 (soit 477 445 observations). Les résultats sont les suivants :

Distribution de la NGA :

NGA	Effectif	%
01	0	0,0
02	18 151	3,8
03	101 781	21,3
04	23 862	5,0
05	92 202	19,3
06	49 601	10,4
07	60 236	12,6
08	9 842	2,1
09	35 779	7,5
10	0	0,0
11	15 754	3,3
12	0	0,0
13	8 678	1,8
14	0	0,0
15	13 588	2,8
16	2 246	0,5
17	7 764	1,6
18	504	0,1
19	0	0,0
20	7 013	1,5
Exclus	30 444	6,4
Total	**477 445**	**100**

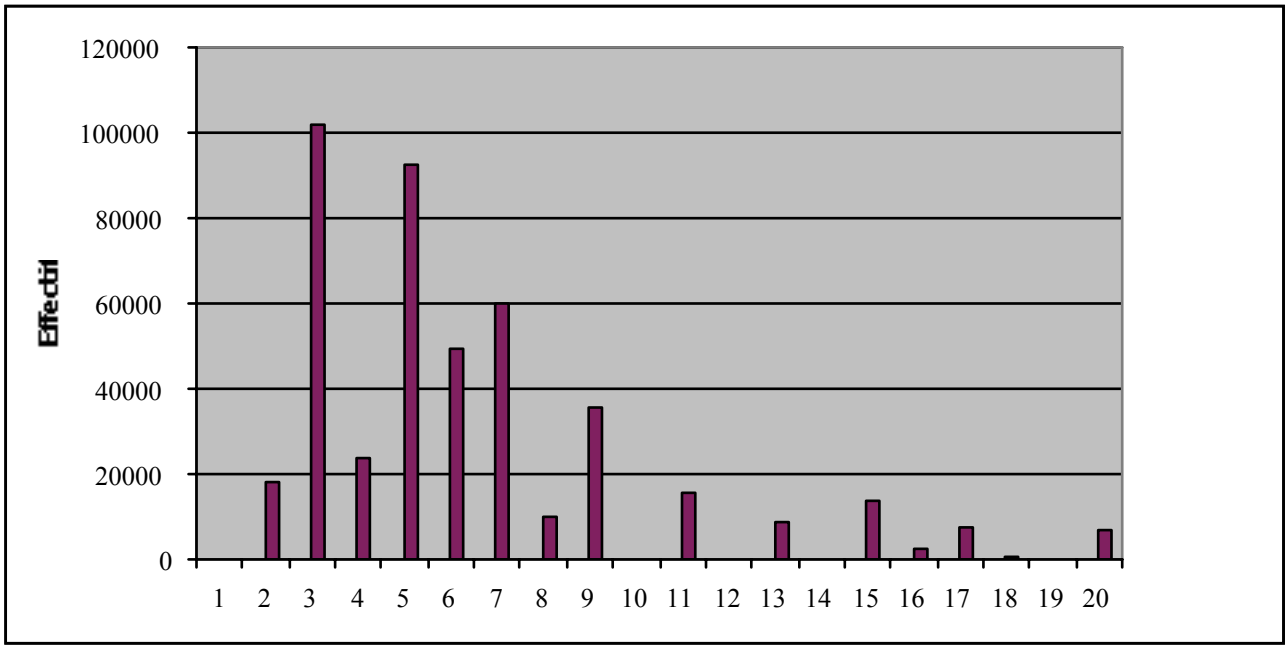
Remarques :

- On constate que certaines notes ne sont pas accessibles du fait du mode de calcul (1, 10, 12, 14 et 19).

- 447 001 observations ont une NGA renseignée sur un total de 477 445, soit 93,6 %.

- La NGA moyenne est de 6,32 (± 3,75).

Distribution des effectifs de la NGA



Remarques :

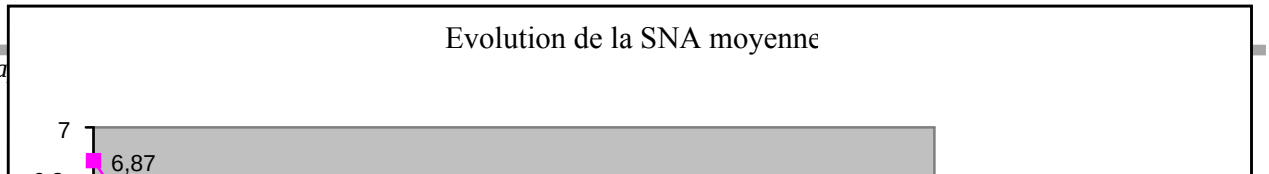
- L'effectif présente un maximum pour une NGA égale à 3, puis il décroît ensuite en tendance jusqu'à la NGA égale à 20.

- On peut découper la gravité en 3 catégories, comme pour le score danois :

NGA ≤ 3 - Les accidents sans gravité
4 ≤ NG ≤ 6 - Les accidents de gravité moyenne
NG ≥ 7 - Les accidents les plus graves

Evolution de la NGA moyenne par an :

	NGA moyenne
1987	6,87
1988	6,43
1989	6,56
1990	6,63
1991	6,34
1992	6,46
1993	6,32
1994	6,17
1995	6,38
1996	5,98
1997	6,06
1998	6,19
Total	**6,32**



Remarques :

- On constate une tendance à la décroissance de la NGA moyenne au cours du temps. Si l'on excepte 1988 qui présente un creux, la décroissance est relativement constante et forte entre 1990 et 1998.

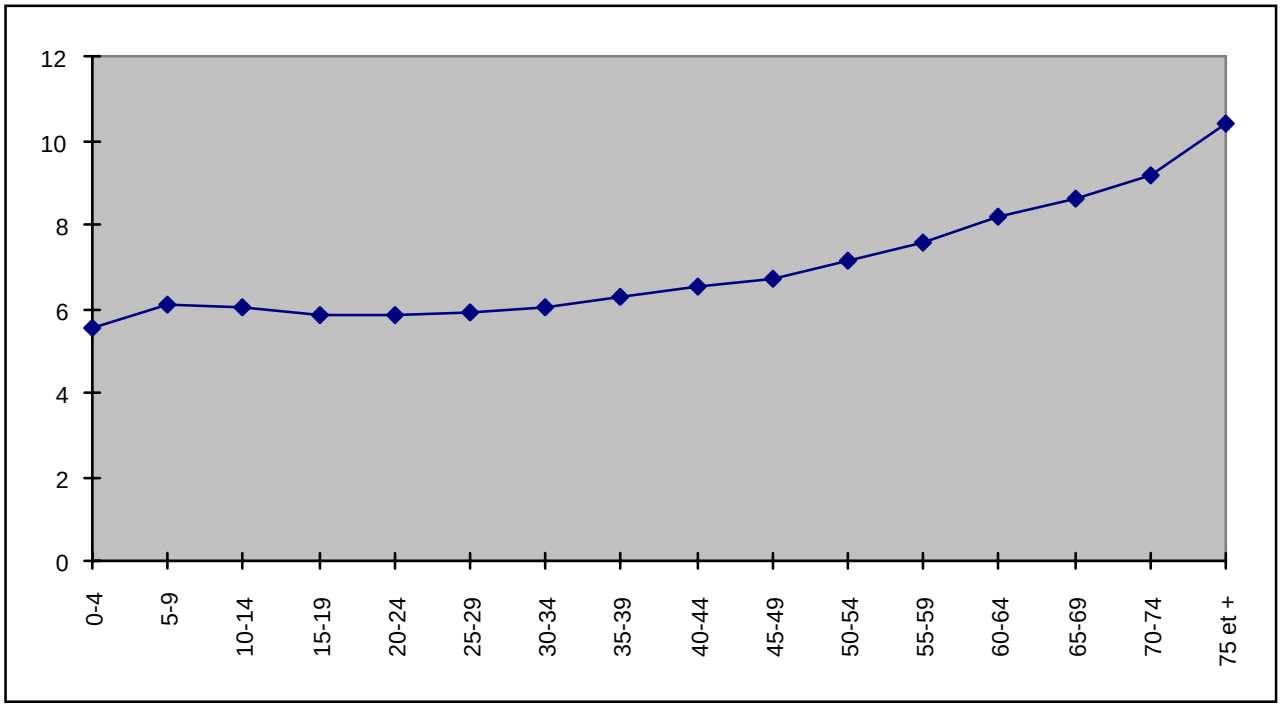
- La constatation est la même quand on sélectionne les observations EHLASS sur un échantillon stable d'hôpitaux entre 1988 et 1998 (Hôpitaux de : Besançon, Bordeaux, Vannes, Béthune, Reims).

Evolution de la NGA moyenne en fonction de l'âge :

Classes d'âge	NGA moyenne
0-4 ans	5,55
5-9 ans	6,08
10-14	6,04
15-19	5,87
20-24	5,82
25-29	5,93
30-34	6,05
35-39	6,25
40-44	6,51
45-49	6,73
50-54	7,11
55-59	7,59
60-64	8,18
65-69	8,64
70-74	9,18
75 et +	10,38

Remarques :

- La NGA moyenne présente une croissance quasi continue avec l'âge.



- Classiquement, on retrouve les accidents les plus graves chez les personnes âgées. Ce qui est moins classique, c'est de constater que les accidents chez les jeunes enfants ont une gravité moyenne assez faible. Ceci s'explique en partie par le fait que c'est pour cette classe d'âge que le coefficient de variation (écart-type / moyenne) est le plus fort : il existe à la fois des accidents sans gravité et des accidents assez graves dans cette population. Les parents amènent facilement leurs enfants aux urgences pour des accidents bénins.

- Cette étude souligne l'importance de la gravité des ADL chez l'adulte. Les enfants et les personnes âgées ont souvent constitué les 2 pôles de populations les plus étudiées. Nous montrons ici qu'il convient de ne pas négliger les accidents chez l'adulte (accidents de bricolage, de sport, etc.).

Quelles sont les conclusions ?

Méthode :

- Nous avons construit une note de gravité (NGA) à partir de 3 variables du système EHLASS : Traitement, Durée d'hospitalisation, Type de lésion. Il y a un certain arbitraire dans le mode de calcul de la NGA, mais il nous semble que la hiérarchie de la gravité des accidents est globalement respectée. Il s'agissait de construire une variable ordinale pour rendre compte de la gravité relative des ADL au sens EHLASS.

Résultats :

- On constate une tendance à la décroissance de la NGA moyenne au cours du temps en rapport sans doute avec la décroissance de la durée moyenne de séjour. La constatation est la même quand on sélectionne les observations EHLASS sur un échantillon stable d'hôpitaux français entre 1988 et 1998 (Hôpitaux de : Besançon, Bordeaux, Vannes, Béthune, Reims).

- La NGA moyenne présente une croissance quasi continue avec l'âge. Cette étude souligne l'importance de la gravité des ADL chez l'adulte. Les enfants et les personnes âgées ont souvent constitué les 2 pôles de populations les plus étudiées et les cibles privilégiées des politiques de prévention. Nous montrons ici qu'il convient aussi de ne pas sous-estimer la gravité des ADL chez l'adulte (accidents de bricolage, de sport, etc.).

- L'intérêt de la NGA réside dans le travail de comparaison historique que l'on peut effectuer (baisse espérée - et constatée dans le cas français - de la gravité moyenne au cours du temps pour mettre en évidence une efficacité des politiques de prévention fondées sur le système EHLASS et/ou des pratiques médicales) et le travail de comparaison géographique. Il serait en effet intéressant de pouvoir comparer les évolutions de la NGA dans les différents Etats.

Le choix de l'échelle de sévérité proposée par le NIPH :

L'équipe danoise a proposé une approche légèrement différente en utilisant un arbre de décision aboutissant à une hiérarchisation en 3 classes de sévérité (injuries of minor severity, injuries of less severity, severe injuries). Cet arbre est construit à partir des informations issues des variables : Type de lésion, Traitement, Durée d'hospitalisation et Partie du corps lésée.

Cette méthode présente l'avantage de ne pas utiliser des pondérations arbitraires pour déterminer le niveau de sévérité.

C'est cette méthode, exposée en Annexe, que nous avons utilisée dans notre procédure SAS intitulé SSC (Severity SCale).

9- La méthode des scénarios (SCENAR)

(Method of scenario - SCENAR)

Quel a été le contexte de création de cet outil ?

- Après avoir utilisé des outils statistiques relativement complexes avec les logiciels de Data Mining, nous avons voulu construire une méthode d'analyse techniquement simple, mais pouvant apporter des résultats intéressants quant à la connaissance des accidents domestiques et de loisirs : **la méthode des scénarios**.

- Chaque variable EHLASS peut constituer une porte d'entrée pour une analyse de données : l'âge (ex : les accidents chez les jeunes enfants), le traitement (ex : les hospitalisés), le mécanisme (ex : les chutes), le lieu (ex : les accidents dans les aires de jeux), l'activité (ex : les accidents de bricolage), les lésions (ex : les fractures), la partie lésée (ex : les accidents de la main) et bien entendu les produits (ex : les accidents impliquant des caddies). Mais la méthode des scénarios vise à couvrir l'ensemble des ADL et doit servir à repérer des ensembles cohérents d'accidents types où l'ensemble des variables décrivant l'accident serait pris en compte. Elle doit fournir une **typologie fine des ADL sans aucune perte d'information** par rapport aux données initiales, puisque que l'on va travailler directement sur une partition des données brutes sans transformation mathématique.

- Les politiques de prévention à mettre en place varient fortement selon les types d'accidents repérés : le contenu, la forme et la cible des messages de prévention seront très différents selon que l'on vise à diminuer la fréquence des accidents par brûlure des jeunes enfants dans les cuisines ou celle des accidents de snow-board sur les pistes de ski impliquant des adolescents. Il est donc important de distinguer un ensemble des scénarios types d'accidents, pour en connaître, si possible, la cause directe, d'évaluer leur gravité et de décrire les caractéristiques des populations impliquées.

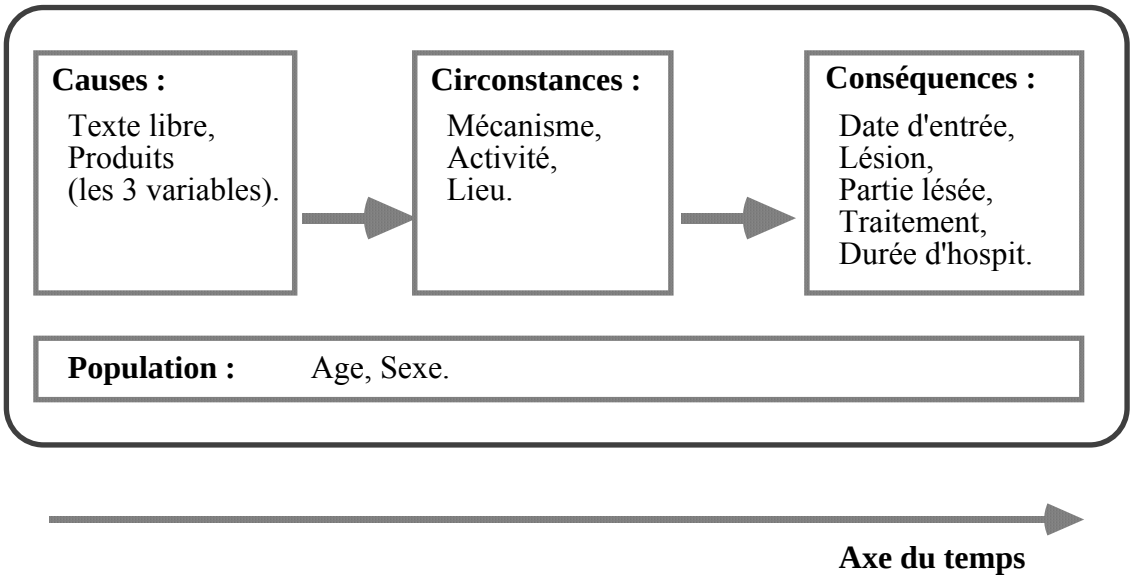
Quelle est la méthode proposée ?

Nous avons tenté de regrouper l'ensemble des variables EHLASS en fonction de leur rôle et de leur place dans la logique et la chronologie du déroulement de l'accident.

Pour nous, **un scénario d'accident au sens EHLASS résulte de la conjonction d'une *cause* et de *circonstances* entraînant des *conséquences* pour une *population donnée***.

En répartissant les variables disponibles dans le système EHLASS entre ces différentes catégories logiques et chronologiques, nous obtenons le schéma suivant :

Un scénario d'accident au sens EHLASS



A partir de cette analyse, pour repérer de façon simple les différents scénarios types, nous proposons d'opérer de la façon suivante :

1- Déterminer les différentes circonstances de l'accident en croisant les variables Mécanisme (24 modalités dans le système de codage V86), Activité (13 modalités) et Lieu (39 modalités). Il y a ainsi 12 168 combinaisons possibles de circonstances (24 x 13 x 39).

2- Trier par ordre décroissant les fréquences de ces classes;

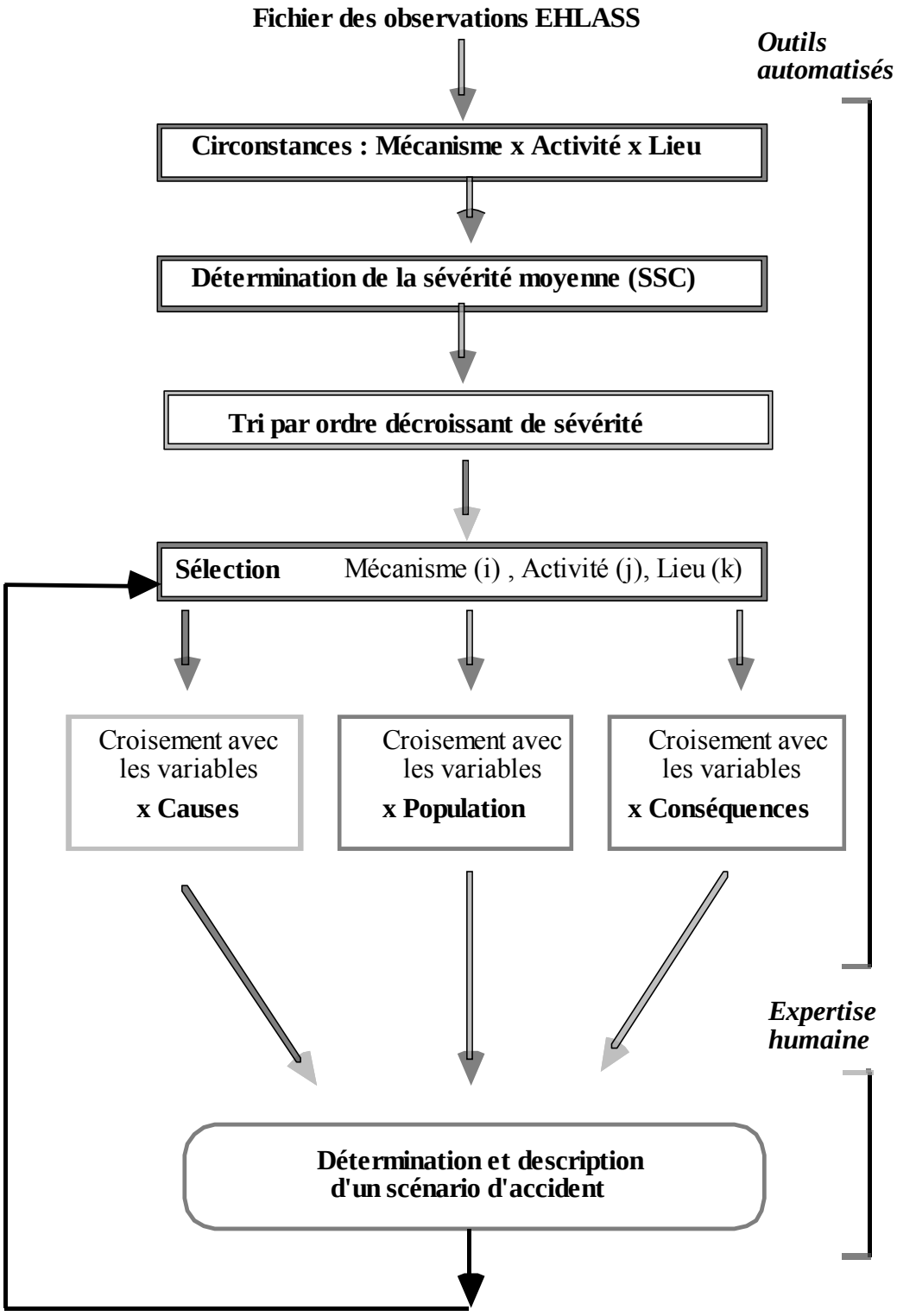
3- Sélectionner une sous-population donnée (selon l'intérêt de la combinaison, un seuil de fréquence, la gravité des accidents, etc.) c'est-à-dire les observations correspondant à une certaine combinaison d'un Mécanisme (i), d'une Activité (j) et d'un Lieu (k). Nous proposons d'utiliser l'échelle de sévérité estimant la sévérité des observations d'une sous-population comme un des critères de sélection.

4- Croiser cette sous-population avec les variables appartenant aux catégories Causes (les 3 variables Produits et le contenu du texte libre où la description de la cause directe de l'accident peut figurer), Population (avec les variables Age et Sexe) et Conséquences (avec les variables Date d'entrée, Lésion, Partie lésée, Traitement et durée d'hospitalisation).

5- Analyser les résultats de ces différents croisements, ainsi que le contenu des textes libres, pour déterminer et décrire un scénario type d'accident.

Cette procédure de détermination de scénarios peut se résumer selon le schéma suivant :

Schéma de détermination des scénarios



Cette méthode a-t-elle été mise en œuvre ?

Nous avons mis en œuvre cette procédure sur le fichier EHLASS France 1997 (50 527 observations). Sur les 12 168 combinaisons possibles (Mécanisme x Activité x Lieu), nous avons dénombré 1 955 combinaisons avec un effectif d'au moins une observation. On constate donc que ce n'est que 16% des combinaisons possibles qui sont observées.

Ceci pourrait d'ailleurs constituer une approche pour la construction d'un programme commun de contrôle des données EHLASS incluant des contrôles de cohérence : on signalerait en erreur possible les combinaisons de variables très rarement ou pas du tout observées auparavant.

Sur ces 1 955 combinaisons classées par ordre décroissant de fréquence, on constate que les 50 premières, constituées par celles dont l'effectif est supérieur ou égal à 200, regroupent plus de 47% du total des observations.

Les critères pour ne pas sélectionner une combinaison pourraient être les suivants :

- effectif trop faible : il est certain qu'il ne paraît pas très pertinent de vouloir construire un scénario d'accident pour une combinaison regroupant 5 observations par exemple, ou pour un pourcentage très faible par rapport au fichier total (< 1/1 000 par exemple).

- faible sévérité moyenne des observations : les combinaisons conduisant à des pourcentages très faibles d'hospitalisation, à des traitements peu lourds (code Traitement à 1 ou 2) et/ou des lésions peu graves ne sont pas à analyser en priorité. Nous pouvons aussi utiliser le pourcentage d'observations jugées « sévères » dans l'échelle de sévérité pour juger de la gravité des sous-ensembles d'observations.

- présence d'une modalité « Inconnu » dans la combinaison : une modalité « Inconnu » dans une variable de la combinaison ne permet pas de déterminer précisément un scénario d'accident. Nous proposons dans un premier temps d'éliminer ces combinaisons.

Dans le cadre du présent exposé, nous avons procédé de la manière suivante :

- sélection des combinaisons avec un effectif > 200;

- élimination des combinaisons avec une modalité « Inconnu »;

- classement des combinaisons par ordre décroissant du % d'accidents « sévères »

A l'issue de l'application de ces critères, nous obtenons un classement des combinaisons de circonstances à analyser en priorité.

Pour une combinaison sélectionnée, nous déterminons le scénario correspondant par analyse, pour la population concernée, des ventilations de toutes les variables EHLASS en fonction de l'âge, des textes en clair, de la durée moyenne d'hospitalisation et de la liste des produits impliqués.

Quels sont les résultats sur les données France 1997 ?

Examinons d'abord des combinaisons avec un fort pourcentage d'observations « sévères » :

1- Scénario « 146010 » : intoxication (code 14), jeux-loisirs (code 60), domicile non précisé (code 10) (N=237)

68% de cette sous-population appartient à la classe d'âge 1-4 ans. Il s'agit clairement d'un scénario d'intoxications par prise de médicaments, de produits chimiques (désinfectants, pesticides, térébenthine, etc.) ou des intoxications au gaz (chauffe-eau). Il en résulte une hospitalisation dans 31% des cas, avec une durée moyenne de séjour de 1,7 jours.

Les textes libres fournissent assez souvent le nom du produit impliqué, plus rarement les circonstances exactes de la prise (« a fait bouillir de l'eau avec de l'eau de Javel qui par erreur a servi pour le biberon », « aurait ingéré de l'eau de Javel - un berlingot dans une bouteille d'eau minérale », « solvant dans une bouteille de coca », « eau dans un biberon mélangée avec de l'eau de Javel », « ingestion d'un demi-flacon de NOPRON posé sur la table à langer », etc.).

2- Scénario « 026021 » : chute d'une hauteur (code 02), jeux-loisirs (code 60), jardin (code 21) (N=465)

75,7% de cette sous-population appartient à la classe d'âge 1-14 ans. Ce sont des garçons dans 61% des cas. Il s'agit de scénario de chutes dans des jardins impliquant des balançoires (97 cas), des toboggans (34 cas) ou des échelles (30 cas). Il en résulte des contusions (49% des cas) ou des fractures (29% des cas) surtout des membres supérieurs (37%) ou de la tête (31/%). Ces accidents conduisent à une hospitalisation dans 25,4% des cas, avec une durée moyenne de séjour de 4,6 jours.

Les textes libres fournissent le plus souvent une information sur le produit depuis lequel s'est effectuée la chute (« chute d'une balançoire », « chute d'une échelle », « chute d'un toboggan », sans donner plus de précision sur les causes de la chute.

Donnons un autre exemple :

3- Scénario « 052021 » : coupure, bricolage, jardin (N=304)

67,8% de cette sous-population appartient à la classe d'âge 25-64 ans. Ce sont des hommes dans 88% des cas. Les produits impliqués sont les scies circulaires (24 cas) ou à chaîne (21 cas), les taille haies électriques (18 cas) et autres équipements de jardin (77 cas). La coupure provoque une plaie ouverte (77% des cas) aux membres supérieurs (72%), surtout aux mains (15%) et aux doigts (49%). Ces accidents conduisent à une hospitalisation dans 21,1% des cas, avec une durée moyenne de séjour de 3,4 jours.

Les textes libres fournissent une information plus précise sur les produits impliqués : tronçonneuse (51 cas), tondeuse (33 cas), hache (12 cas), sécateurs (12 cas) ou sur les circonstances : « en voulant débourrer sa tondeuse, plaie à la main droite », « a voulu nettoyer sa tondeuse en marche », « motoculteur sans sécurité d'arrêt ».

Quelles sont nos conclusions ?

- La méthode des scénarios nous semble assez fructueuse. Elle fournit une **typologie fine des ADL sans aucune perte d'information** par rapport aux données initiales, puisque que l'on travaille directement sur une partition des données non agrégées sans transformation mathématique. Par contre, certains regroupements qui peuvent s'effectuer par des méthodes factorielles ne sont pas faits. Ces regroupements restent à déterminer au fur et à mesure de l'analyse des différentes combinaisons.

- Avec une **centaine de scénarios types**, on couvre une grande partie de l'ensemble des ADL.

- Le recours à l'expertise humaine est nécessaire dans la phase de détermination des scénarios pour, dégager la cohérence des observations, exploiter le texte libre et faire une synthèse des caractéristiques du groupe d'observations et ainsi déterminer le scénario type. On pourrait éditer **une fiche d'analyse par scénario type** que l'on pourrait enrichir d'année en année et qui pourrait servir de base à la détermination des actions de prévention.

- On pourrait, par la suite, essayer **d'apprécier l'efficacité des mesures prises** en examinant l'évolution de l'échelle de sévérité moyenne au cours du temps, pour un scénario donné, sur un échantillon stable d'hôpitaux.

C'est en effectuant des travaux de ce type que nous pourrons acquérir l'expérience nécessaire pour extraire de la montagne des données EHLASS la connaissance utile et améliorer ainsi le rapport Coût/Efficacité du système d'information.

Nous n'avons pas développé de procédure SAS spécifique du fait que la méthode est fondée sur l'analyse de sous-populations où l'expertise humaine intervient fortement. Mais, nous avons exposé une marche à suivre qui peut inspirer des équipes travaillant sur ces données.

10- Les procédures SAS mises à disposition

10.1- Les éléments mis à disposition

Nous avons diffusé le 11 octobre 2000, via le réseau Internet, à l'ensemble des chefs de projet EHLID membres de l'Injury Epidemiology Network, les procédures SAS mises au point, ainsi que les documentations afférentes en anglais.

Quels sont les éléments mis à disposition des membres du Network ?

Nous avons mis à disposition le fichier :

- **dmtools.c** : un écran d'interface développé en langage C et permettant de choisir et de lancer les différentes procédures

Nous avons mis à disposition les procédures SAS suivantes :

- **saa.sas** : procédure SAA (System of Automated Alert - voir chapitre 7)
- **ssrdp1.sas** : procédure SSRD utilisant les percentiles, appliquée à la variable « Produit impliqué dans l'accident » (voir Chapitre 6)
- **ssrdp2.sas** : procédure SSRD utilisant les percentiles, appliquée à la variable « Produit ayant causé l'accident » (voir Chapitre 6)
- **ssrdd1.sas** : procédure SSRD utilisant la distribution des effectifs des variables, appliquée à la variable « Produit impliqué dans l'accident » (voir Chapitre 6)
- **ssrdd2.sas** : procédure SSRD utilisant la distribution des effectifs des variables, appliquée à la variable « Produit ayant causé l'accident » (voir Chapitre 6)
- **ssc.sas** : procédure SSC (Severity SCale - voir Chapitre 8)

Nous avons mis à disposition les documentations suivantes en anglais (format Word) :

- **Intropro.doc** : présentation du projet et des procédures
- **saa.doc** : documentation relative à la procédure SAA
- **ssrd.doc** : documentation relative aux procédures SSRDxx
- **ssc.doc** : documentation relative à la procédure SSC

Enfin , nous avons mis à disposition un fichier d'essai de données EHLASS au format V96 :

- **fr99v96.zip** : données françaises anonymisées de l'année 1999 compactées (41307 observations sans texte en clair)

Nous avons développé et testé ces procédures sous notre environnement de développement : une plate-forme SUN avec le logiciel SAS version 6.12.

Comment utiliser ces procédures ?

Une fois installés le programme en langage C et les procédures SAS sous leur environnement correct, pour lancer l'écran d'interface il faut taper, sous DOS :

>dmtools

L'écran suivant apparaît alors :

Project n° 1999/IPP/1006 - Data Mining tools
V00.1/BIOSTA
SAS processes applied to EHLASS data format V96
Version test - September 2000

1) Automated Alert System - AAS

Synthetic score of Relative Dangerosity - SSRDP (Using percentile)
2) for product involved in the accident
3) for product causing injury

Synthetic score of Relative Dangerosity - SSRDD (Using standard deviation)
4) for product involved in the accident
5) for product causing injury

6) Severity Scale - SSC

7) New data file

Return to stop

Your choice :

En tapant l'un de ces choix (1-6), le système demande

Instruction to execute sas system (default=sas) :

(Test file France 1999=fr99.v96)
Enter the name of EHLASS file format V96 :

Vous pouvez donc entrer la localisation (le « string ») du système SAS et le nom du fichier EHLASS au format V96 sur lequel va s'exécuter la procédure.

Le résultat de la procédure est, classiquement pour l'exécution d'un programme SAS, dans le fichier :

xxx.lst (par exemple : ssc.lst,)

le compte-rendu d'exécution dans :

xxx.log (par exemple : ssc.log,)

Si l'on lance immédiatement une autre procédure, elle se déroule sur le même fichier que la procédure précédente. Pour changer de fichier de données, il faut passer par le choix 7 qui permet de changer le nom du fichier en entrée.

10.2- Les réponses aux critiques et suggestions

Nous avons rassemblé ici nos réponses aux critiques et suggestions qui se sont faits jour lors du déroulement du projet, via l'enquête et les réunions de projet, concernant les différentes procédures.

Quelles sont nos réponses aux critiques et suggestions concernant le SSRD ?

Critiques et suggestions	Nos réponses
Le taux d'hospitalisation et la durée moyenne de séjour sont des variables dépendantes de l'âge et de la structure du système de soins.	En utilisant les percentiles, on ne tient compte que du classement des valeurs. On s'affranchit donc en partie des différences absolues entre les structure de soins. Pour l'âge, on peut calculer des SSRD par classes d'âge.
Le système d'information ne donne qu'une idée très partielle du nombre des décès puisque seuls les décès durant le séjour à l'hôpital sont recueillis.	C'est pourquoi nous avons supprimé la variable décès du calcul du SSRD.
Les produits rares sont sous-estimés (fréquence peu élevée d'apparition). Ils peuvent cependant être très dangereux.	Notre score prend en compte à la fois la fréquence (EFF) et la sévérité (DMS et TH). Si la sévérité est élevée et l'effectif faible le score sera quand même assez élevé. Mais il est certain que les produits très rares (< 3) ou nouveaux échappent.
La variable « Effectif » est liée à la fréquence d'utilisation plus qu'à la dangerosité. Il faudrait avoir des informations sur les fréquences d'utilisation des classes de produits.	Voir la ligne précédente. Etant donné que l'on ne dispose pas d'informations sur la fréquence d'utilisation de l'ensemble des produits, doit-on pour autant renoncer à toute méthode et abandonner le projet ?
Il y a des différences culturelles entre Etats dans la fréquence d'utilisation des produits.	Oui, on comparera les SSRD dans chaque Etat en tenant compte de ces différences qui ne sont pas « mesurables ».
Beaucoup d'accidents n'ont pas de lien de causalité directe avec le produit impliqué.	C'est certain. C'est pourquoi nous ne parlons que d'une dangerosité « potentielle », qu'il faut confirmer par expertise humaine et examen au cas par cas.
L'affectation des coefficients multiplicateurs paraît arbitraire.	Nous avons supprimé ces coefficients dans la nouvelle version du SSRDD.
Comment faire intervenir ensemble les trois codes produits ?	On offre la possibilité de calculer séparément le SSRD pour les 2 codes produits principaux. La variable « Autre produit » est peu usité.
La fréquence du code « Autre produit » peut être forte.	Cela est lié à la précision du système de codage et à sa pratique, non pas au calcul du Score.
Il faudrait faire intervenir d'autres variables (mécanismes, type de lésion ?).	C'est sans doute une bonne idée mais comment la mettre en œuvre ?

Il faudrait créer un groupe de travail sur ce seul projet.	Nous avons créé le groupe de travail Data Mining et sollicité les avis de tous les membres du réseau, sans beaucoup de succès quant aux idées concrètes.
Il faut un mode d'emploi explicitant bien la méthode et les concepts utilisés.	C'est ce que nous croyons avoir fait dans ce rapport et le document **ssrd.doc.**

Quelles sont nos réponses aux critiques et suggestions concernant le SAA ?

Critiques et suggestions	Nos réponses
On pourrait analyser l'évolution non pas par comparaison entre une période donnée et une période de référence, mais entre une période donnée et plusieurs périodes de référence (analyse en tendance).	Oui, nous proposons de le faire dans le niveau avancé. Cette analyse en tendance risque aussi de gommer beaucoup de variations. Pour le niveau simple nous maintenons une comparaison simple entre 2 périodes
Il peut y avoir des variations saisonnières dans la survenue de certains types d'accidents.	C'est pourquoi nous recommandons plutôt de comparer les données du même mois sur 2 années différentes.
Il faut utiliser cette méthode sur des données assez récentes pour garder le caractère d'alerte de la procédure. Cette méthode est donc plus adaptée à l'utilisation dans les Etats membres.	Il est vrai que le caractère d'alerte n'est pas très prononcé, étant donné que l'on ne peut comparer au mieux que les données du mois écoulé à celles du même mois de l'année précédente par exemple.
On peut commettre des erreurs statistiques si l'on n'utilise pas les mêmes périodes ou des périodes trop restreintes.	C'est pourquoi nous recommandons de comparer des périodes de même durée sur au moins un mois.
Le nombre absolu d'accidents peut varier entre les 2 périodes.	C'est pourquoi nous utilisons des fréquences relatives dans notre comparaison.
Comment traiter les fréquences basses d'accidents ?	Nous avons prévu une procédure particulière pour les effectifs < 30.
Une augmentation de 5% des cas d'intoxications médicamenteuses est plus importante qu'une augmentation de 5% des accidents de basket-ball.	C'est l'expertise humaine qui détermine ces importances relatives. La procédure ne prétend fournir que des pistes d'interrogation.
La durée d'exposition n'est pas forcément la même entre les deux périodes.	Il faut que les périodes soient de même amplitude. Pour le reste, on fait l'hypothèse classique en statistique du « toute chose égale par ailleurs ». Encore une fois, nous ne prétendons pas à la rigueur scientifique absolue mais à l'utilité pratique.

Quelles sont nos réponses aux critiques et suggestions concernant le SSC ?

Critiques et suggestions	Nos réponses
Les variables « Traitement » et « Durée d'hospitalisation » sont liées à la politique de soins appliquée localement et au comportement culturel.	Il est vrai, mais quelles autres variables endogènes plus informatives et objectives peut-on utiliser ?
Tous les accidents ne donnent pas lieu à une hospitalisation, cela introduit des interactions.	A développer.
Le but de cette Note semble proche de la méthode du SSRD.	Le SSRD vise à hiérarchiser la dangerosité potentielle des produits et non à attribuer une position sur une échelle de sévérité absolue à chaque accident.
Il faut exclure la modalité « Autre type de lésion »;	Cela est fait en partie dans l'arbre de décision du Severity SCale.
Il faudrait inclure la variable « Partie du corps lésée ».	Cela est fait dans l'arbre de décision du Severity SCale.
Il faudrait faire intervenir la variable « Produit causant la lésion ».	Pourquoi ?
Les données de mortalité sont exclues du calcul.	C'est volontaire, car les données de mortalité sont beaucoup trop parcellaires dans le recueil EHLASS.
Il est important de valider ce score.	C'est ce que nous voulons faire en vous le proposant de l'utiliser.
Beaucoup des scores de sévérité ont déjà été développés. Il vaudrait mieux créer un Comité au sein de IPP pour choisir lequel utiliser.	Sans doute, mais les données EHLASS sont très spécifiques. Elles méritent sans doute leur propre score. Par ailleurs, nous avons déjà créé un groupe d'experts Data Mining.

Quelles sont nos réponses aux critiques et suggestions concernant SCENAR ?

Critiques et suggestions	Nos réponses
La méthode peut conduire à un nombre très élevé de combinaisons de circonstances.	Mais en fait, il n'y a qu'un nombre assez faible de combinaisons correspondant à un nombre d'accidents significatif (environ 200).
La valeur ajoutée de la méthode par rapport à un simple croisement de variables doit être démontrée.	C'est l'ensemble du processus qui constitue la méthode pas seulement l'étape du croisement des variables. L'exemple donné dans le rapport démontre à notre avis cette utilité.
Il faudrait pouvoir introduire des scénarios externes fournis par l'expertise humaine et combiner ces informations avec des données de coûts.	C'est ce qui est proposé dans le niveau avancé et ce qui est réalisé au CSI d'Amsterdam.
Cette approche est utile quand on a affaire à un grand nombre d'accidents très divers.	C'est bien le cas avec les dizaines de milliers d'accidents dans les bases EHLASS.
On pourrait utiliser le croisement des variables « Mécanisme » et « Lieu », puis distinguer les principales combinaisons en fonction de leur fréquence.	Cela est assez proche de ce que nous proposons. Il faudrait tester cette simplification pour voir si l'on arrive aussi à un ensemble de scénarios cohérents.

11- Conclusions

A la fin de ce cheminement, que pouvons-nous conclure ?

- Les données EHLASS telles qu'elles existent recèlent encore un fort potentiel informatif qu'il faut exploiter au mieux et valoriser.

- Nous avons tenté de montrer que les méthodes du Data Mining et encore plus l'esprit même du processus mis en jeu dans le Data Mining pouvaient constituer une approche intéressante pour l'exploitation des données sur les ADL.

- Nous avons essayé de regarder, pour une fois, du côté de la compétence des données déjà accumulées et du périmètre d'efficacité du système d'information et non uniquement du côté de ses faiblesses et de ses possibilités d'amélioration dans le futur.

- Nous avons appliqué deux méthodes de Data Mining, celles des prédicteurs neuronaux et de la segmentation, à des données EHLASS réelles et examiné les résultats obtenus.

- Transformer les données en informations est le but des méthodes du Data Mining. Mais, la découverte d'informations masquées est très dépendante des besoins de chaque utilisateur. Les méthodes standards ne pourront répondre qu'à des questions générales. Les résultats de notre enquête et les questions spécifiques posées par le système de recueil de données sur les ADL nous ont amenés à vouloir mettre au point des procédures décisonnelles mieux adaptées.

- Nous avons donc proposé, exposé et développé quatre procédures pouvant être utilisées pour répondre à des questions spécifiques. Nous avons choisi délibérément une approche pragmatique (adaptée à l'action sur le réel) et empirique (fondée sur l'expérience) pour la constitution de ces procédures. Nous avons développé des outils simples et facilement compréhensibles dans leurs principes et dans leur utilisation.

- Nous avons diffusé par le réseau Internet, à l'ensemble des chefs de projet EHLID membres de l'Injury Epidemiology Network, les procédures SAS mises au point, ainsi que les documentations afférentes en anglais.

- Ces procédures doivent être testées et sans doute améliorées en tenant compte des choix initiaux de développement (approche pragmatique et empirique, utilisation des variables endogènes au système, etc.) ou complétées par d'autres procédures utilisant d'autres approches.

- Dans ce domaine, l'échange d'expériences concrètes sur l'utilisation des données entre les équipes européennes est essentiel. Nous sommes au début de ce processus et la possibilité, via les bases européennes récemment mises en place ou en cours de développement, d'exploiter l'ensemble des données recueillies depuis de nombreuses années va renforcer cette coopération et la nécessité d'utiliser de nouveaux outils et de nouvelles procédures.

C'est en promouvant les outils de Data Mining et/ou des procédures spécifiques et en les utilisant que nous pourrons acquérir l'expérience nécessaire pour extraire de la montagne des données EHLASS la connaissance utile et améliorer ainsi le rapport Coût/Efficacité du système d'information.