

Evaluation de la Performance Thérapeutique pour la Chirurgie du Rachis

Hugues GENVRIN

December 2025

Résumé

On a pratiqué l'évaluation de la performance chirurgicale au CAD entre 2001 et 2007. Cela a conduit à une évaluation longitudinale de tous les patients opérés. La méthode retenue s'appuie sur une image complète des faits. Elle permettait de réaliser des études prospectives de manière rétrospective. Ce projet a été mené par le professeur Sénégas.

Table des matières

1	Introduction	2
2	Méthodologie	3
2.1	Le cadre actuel	3
2.1.1	Historique	3
2.1.2	Les apports de la biométrie	3
2.2	Quantification de la performance	4
2.2.1	Les objectifs	4
2.2.2	Les domaines à évaluer	4
2.2.3	Les outils d'évaluation	5
2.2.4	Les scores retenus	6
2.2.5	Intégration pragmatique	10
3	Exploitation des résultats	13
3.1	Analyses des groupes de patients	14
3.1.1	Obtention des données	14
3.1.2	Valeurs de références	15
3.2	Analyses Individuelles	15

3.3	Analyses Statistiques	16
3.3.1	Le cadre des RCT (Randomized Controlled Trials) . .	17
3.3.2	Tests	20
4	Conclusions	28
4.1	Pragmatisme	28
4.2	Analyse de groupe performante	31
4.3	Coût Economique d'un telle évaluation	31

1 Introduction

Face à l'évolution des techniques chirurgicales, aux coûts engendrés par la prise en charge des patients opérés, l'évaluation doit permettre de déterminer les meilleures solutions thérapeutiques. A titre d'exemple, en 1995 on dénombrait 415 000 opérations de fusion vertébrale aux USA, soit 20% de plus qu'en 1990, le coût atteignant 49 800\$ par opération [1]. Cependant, on ne peut pas évaluer de la même façon un concept de dispositif chirurgical, une technique qui fait ses premiers tests in-vivo, ou une technique éprouvée dont on souhaiterait en faire un « gold-standard ». On peut distinguer deux grands courants méthodologiques dans l'expérimentation biologique : la méthode hypothético-déductive qu'on retrouve chez Claude Bernard [2], et dont le rationalisme critique de Karl Popper [3] apparaît la continuité. Et les méthodes numériques qui ont émergées à partir du début du XX° [4] et qui se sont accélérées à partir de 1950 avec d'une part les méthodes d'induction probabilistes développées par Ronald Fisher d'un côté, Jerzy Neymann et Karl Pearson de l'autre ; il ne faut oublier aussi le développement de l'informatique aux niveaux des ordinateurs (hardware), des logiciels (software) de gestion de bases de données, de statistiques et des tableurs Nous allons ici nous limiter à l'évaluation des dispositifs utilisés pour traiter des pathologies dans le cadre d'un exercice clinique courant. Dans un premier temps on présentera le cadre méthodologique de la quantification de la performance chirurgicale telle que nous l'avons pratiquée au Centre Aquitain du Dos. La structure se composait de quatre chirurgiens, quatre secrétaires, quatre kinésithérapeutes formés à l'évaluation, et un attaché de recherche clinique. Nous montrerons ensuite comment on peut exploiter les résultats obtenus autour de trois axes : les analyses de groupes, individuelles et les analyses statistiques. Nous dévoilerons également d'autres outils d'analyse qui complètent idéalement ceux déjà en place. En conclusion, on abordera le côté pragmatique de la quantification de la performance chirurgicale qui nous procure un indice de confiance élevé pour les résultats de groupe.

2 Méthodologie

2.1 Le cadre actuel

2.1.1 Historique

C'est en 1992 qu'a émergé l'*Evidence Based Medecine* (EBM) ; le but étant de faire reposer les décisions médicales sur des preuves ou des faits scientifiquement avérés. L'EBM fournit un cadre méthodologique largement reconnu et utilisé de nos jours. Voici les niveaux de preuves qu'elle préconise pour l'évaluation des traitements : La chirurgie présente cependant des

Niveau 1	Étude randomisée contrôlée
Niveau 2	Étude en cohorte prospective avec groupe de contrôle
Niveau 3	Étude en cohorte prospective sans groupe de contrôle
Niveau 4	Étude en cohorte retrospective
Niveau 5	Étude de cas témoins
Niveau 6	Opinion d'experts

singularités qui font que les études en double aveugle sont impossibles et les études en simple aveugle presque impossibles. En effet, dans le cas de la chirurgie du dos le chirurgien voit le dispositif qu'il pose et lors d'un contrôle radio le patient verrait également le système en place. Par ailleurs la notion de randomisation présente des difficultés qu'on expliquera plus loin.

2.1.2 Les apports de la biométrie

Pour définir des critères d'évaluation, on profite des apports de la biométrie qui s'est formidablement développée depuis la création de l'indice de masse grasseuse (Body Mass Index ou BMI) inventé par Quetelet en 1833 [5]]. Par exemple, on peut évaluer l'état de dégénérescence d'un disque grâce à des IRM avec la classification des grades de Pfirmann [6]. Parallèlement l'Organisation mondiale de la santé a été plus exigeante pour définir l'état de bonne santé : « Etat complet de bien être, physique, psychologique et pas seulement comme l'absence de douleur » (1946 à la création de l'OMS). Ainsi se sont développés des outils pour évaluer la qualité de vie, on peut citer le principal : le SF36 (®), mais aussi le SF12(®), le Nottingham Health Profile (NHP), le Duke Health Profile (DUKE), le Sickness Impact Profile (SIP). Pour l'évaluation de la douleur qui reste un critère essentiel, on va utiliser des échelles de cotation : numériques, verbales, ou bien encore visuelles. On a également à la prise en compte de la déficience, du handicap et de l'incapacité. Là encore des scores ont été développés pour la chirurgie du dos, on peut

citer : le Japanese Orthopedic Association lombaire (JOA Lombaire), le Neck Disability Index (NDI), ou encore le score d'Oswestry (Oswestry Disability Index ou ODI) ou bien le score de Beaujon-Lassale (BEAUJON). Il existe aussi le Dallas pain questionnaire (DALLAS) ou le Roland-Morris (EIFFEL dans sa version française) qui sont des scores regroupant des questions sur la qualité de vie et l'incapacité spécifiques au problème au dos. On peut déjà voir l'aspect multidimensionnel de l'évaluation qui ne se limite pas à l'évaluation de la douleur, même si elle reste le principal indicateur du résultat thérapeutique.

2.2 Quantification de la performance

2.2.1 Les objectifs

Le but est de quantifier la performance chirurgicale, la chirurgie étant un acte performatif qui se démarque de la prise d'un médicament qui elle ne requiert pas d'acte performatif. « Par exemple si un traitement médicamenteux échoue, on attribuera plus facilement l'échec au médicament, alors que si une intervention échoue on attribuera plus facilement la responsabilité au chirurgien et non au matériel » [?]. Performatif pour le dos dans le sens où elle induit une performance quelque fois concomitante à l'effet de l'implant. La performance ne s'inscrit pas uniquement dans le geste per-opératoire, mais également dans l'attribution de la pathologie, et plus généralement autour de la gestion de l'ensemble des variables qui peuvent conditionner le résultat ; le chirurgien restant l'opérateur central du système. On peut dire qu'on est en présence d'un véritable système complexe à opérateur central [7]. Nous visons ici le suivi du résultat thérapeutique pour chaque patient, qui permette également une évaluation statistique par technique, par pathologie ou par groupe de patients. Pour cela on va choisir des indicateurs qui vont quantifier des variables qualitatives représentant un domaine particulier.

2.2.2 Les domaines à évaluer

Il est assez difficile de partitionner l'espace de projection des fonctions de valeur et d'utilité pour la chirurgie du dos. Certains domaines peuvent se recouvrir partiellement, et les scores utilisés pourront évaluer les mêmes concepts en partie (par exemple la partie qualité de vie du Japanese Orthopedic Association Lombaire (JOA) et le SF36). Voici les principaux concepts qu'on a retenus : la douleur, la qualité de vie, l'incapacité, la satisfaction du patient, l'état clinique, la déficience, la durée moyenne de séjour, le coût économique, la durée moyenne de séjour (DMS). En fait on rentre dans un cadre

minimaliste suivant la nature de la pathologie et du traitement, c'est-à-dire que la liste de domaines à évaluer dépend de manière fine de la pathologie et du traitement.

2.2.3 Les outils d'évaluation

Husserl : « La mesure appartient à toute culture, du plus simple au plus complexe ».

Les outils sont des fonctions d'un domaine vers un ensemble continu ou discret. Quelquefois on obtient des ensembles images qualitatifs qu'on transformera en variables quantitatives sur la base d'une association biunivoque entre une réponse et un entier. Par exemple pour l' Index de satisfaction, le test d'Odom ou l'on retrouve 4 types de réponses qu'on notera de 1 à 4. Les outils peuvent être des observations pures et simples : mesures angulaires (locales, régionales), le délai de retour au travail, des cotations (mots-clefs), des stades, des grades (Pfirmmann), des index, des indices : Indice de Torg, des tests sensitifs ou moteurs, des échelles : Echelle Visuelle Analogique (EVA) de la douleur, ou des scores proprement dit : SF36, Oswestry Disability Index (ODI) ou encore le Neck Disability Index (NDI). Les mesures qui sont archi-évidentes sont intégrées dans des chaînes de décision, elles deviennent des paramètres comme par exemple l'angle de Cobb pour la scoliose. Les scores ont ici un statut particulier parmi les outils d'évaluation, ils sont généralement validés par une ou plusieurs études cliniques. Devant la profusion des scores, c'est l'usage par la communauté scientifique qui va en faire un outil de référence. Par exemple en 20 ans, Guillaumat [8] a recensé plus de 40 échelles de la douleur pour la lombalgie. On définit un score comme une méthode pour quantifier des variables qualitatives dans un souci d'évaluation de l'état de santé d'un patient. Les objectifs étant de comparer, classer et mesurer l'état de santé des patients pour définir les meilleurs traitements. L'intérêt est multiple, tout d'abord on ne réinvente pas la roue, cela facilite la conception de protocoles, permet la comparabilité des résultats entre différentes études cliniques (méta-analyses) [9]. Le score apparaît donc comme un véritable « instrument » d'évaluation. Dans le cas du dos, il existe des outils qui regroupent une batterie de scores comme le NASS Low Back Pain Instrument ; il recouvre l'évaluation fonctionnelle, l'histoire médicale, . . . au travers de 170 questions. On touche alors une dimension importante d'un score il doit être pratiquement utilisable, en un temps acceptable. D'un point de vue typologique, on distinguera les scores suivant la dimension de l'ensemble d'arrivée. Il est unidimensionnel pour l'EVA de la douleur, l'ODI, ou multidimensionnel pour le SF36. On peut également classer les scores suivant la nature des résultats : continue pour l'EVA de la douleur, ou discrète pour le

test d'Odom. Les qualités d'un score sont tout d'abord sa validité : il doit mesurer avec précision ce qu'il est sensé mesurer, il doit être également validé dans la langue ou l'on l'utilise si elle est distincte de la langue initiale du score [10] ; cela peut prendre beaucoup de temps et donc restreindre l'utilisation des meilleurs scores le plus rapidement possible. Par exemple pour le SF36, il existe une version en hispanique pour les américains de langue espagnol. On attend également d'un score qu'il soit fiable, c'est-à-dire qu'il n'y ait pas de dispersion aléatoire. Parallèlement il doit être sensible aux changements d'état de santé. Cependant, n'oublions pas deux critères pragmatiques : son utilisation concrète et son coût (en effet, certains scores bénéficient d'un droit intellectuel). La méthode qu'on a retenue pour sélectionner les scores est le nombre de références dans la littérature à date à partir de la base *pubmed*. Concrètement, il vaut mieux utiliser un score de référence in-

	Avantage	Inconvénient	Nbre d'occurrences pubmed (Rubrique Abstract)
EVA Cervical	Simplicité	Métriologie	47
EVA Lombaire	Simplicité	Métriologie	98
JOA Myélopathie	International	Pas validé en français	55
JOA Lombaire	International	Pas validé en français	18
Neck Disability Index	International	Pas validé	46
Beaujon-Lassale	Périmètre de marche	Français - Parisien	1
Oswestry	Troubles sexuels	Pas validé en français	85
Dallas	16 items / Traduit en français		18
Rolland-Moris	Handicap faible	Handicap fort	58
Eifel (Echelle d'Incapacité Fonctionnelle pour l'Evaluation des Lombalgies)	Traduction du Roland-Morris	Handicap fort	2
SF36	Validé en français	Valeur individuelle faible	100
Critère d'Odom		Valeur	23
Indice de Satisfaction	Simplicité	Valeur Subjective	26

FIGURE 1 – Nombre d'occurrences *Pubmed*

ternational, même si la traduction n'a pas été validée scientifiquement qu'un score validé mais qui ne resterait valable qu'en France.

2.2.4 Les scores retenus

La douleur. L'évaluation de la douleur se pratique avec une échelle qui peut être verbale, visuelle ou numérique. La plus intéressante est l'EVA de la douleur qui à un statut quasi métriologique, pour être le moins contestable

possible. L'introduction de la « *Visual Analogue Pain Rating Scale* » ayant été popularisée par Huskisson dans les années 1970 [11]. Le patient doit déplacer

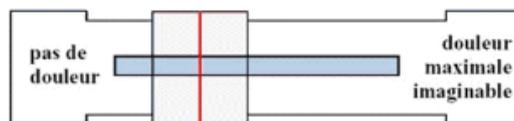


FIGURE 2 – Echelle EVA Recto-verso. Le patient déplace le curseur sur le recto, l'observateur renseigne la valeur à partir du verso.

un curseur sur une ligne horizontale ou est indiqué à l'extrémité gauche « Pas de douleur » et à l'extrémité droite « La douleur maximale imaginable ». Au dos de la règle est indiqué un niveau de douleur correspondant, la règle étant généralement graduée de 0 à 100 (0 correspond au niveau pas de douleur, et 100 la pire douleur). Le temps de saisie est d'une minute. Bien entendu c'est un score subjectif, il n'y a pas de bijection entre le niveau de douleur et un état physiologique. Il est sensible aux effets plafonds et planchers. De plus ce score peut être influencé par l'observateur qui pose la question. On peut voir ci-dessus la différence de niveau de douleur dans le premier décile entre une population où le patient est évalué par le chirurgien : 68% de la population ayant un niveau de douleur inférieur à 10 et une population avec un observateur indépendant ou 50% de la population à un niveau de douleur inférieur à 10. La différence se portant surtout sur le nombre de douleur nulle. Par ailleurs il peut être conseillé de mettre en place des notions de douleur globale et spécifique ; ou la douleur globale prendrait en compte tous les types de douleurs et les douleurs spécifiques ne prenant en compte que les douleurs locales. Cela serait intéressant pour évaluer la douleur du cou pour une opération cervicale en post-opératoire immédiat qui sera différente d'une éventuelle douleur irradiante dans l'épaule (qui peut être due à une mauvaise position per-opératoire).

Le JOA lomboradiculalgie est un score qui a été développé par la Japanese Orthopedic Association, il se décompose en plusieurs parties : une partie fonctionnelle, clinique et une partie sur la qualité de vie. On peut utiliser une version tronquée qui ne prend en compte que la partie fonctionnelle et clinique. Il doit être complété par un observateur formé à l'évaluation et non par le patient, on y évalue le test de Lasegue, les troubles sensitifs et moteurs, les problèmes urinaires, ainsi que le périmètre de marche, la douleur et les picotements. Le temps de remplissage est de 5 minutes, on calcule le score en sommant les entiers associés à chacune des réponses, l'amplitude du

EVA douleur du cou à 12 mois (Sans Observateur Indépendant)

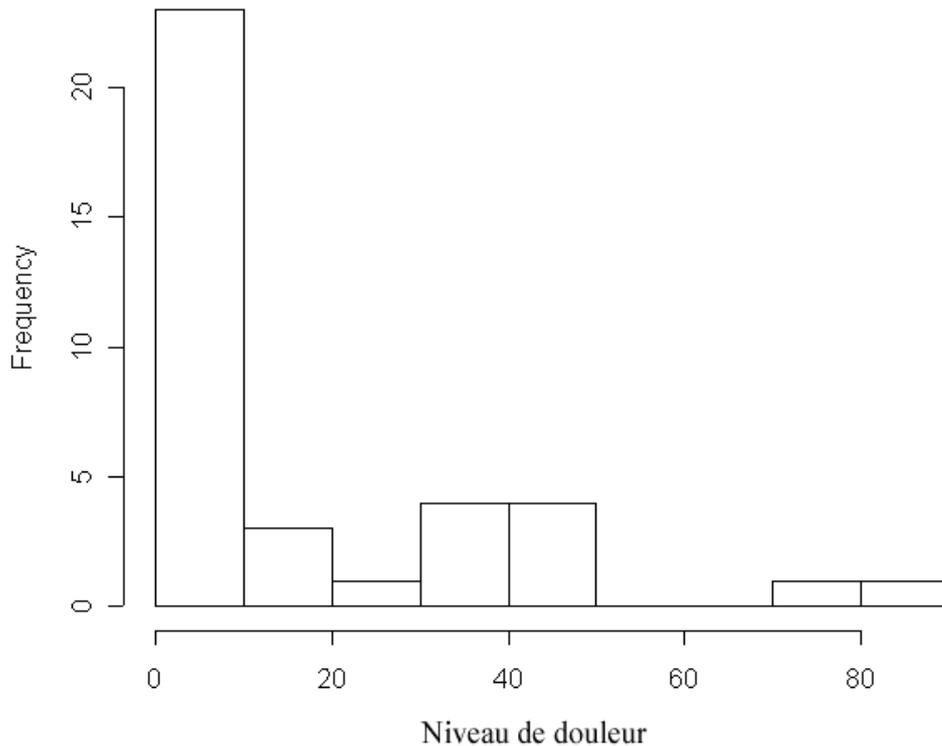


FIGURE 3 – Évaluation de la douleur pour des cages intersomatiques cervicales (Source Société Scient’x – Cages Samarys)

score est de 21, les scores pouvant aller de -6 à $+15$. Plus le score est élevé, meilleur est l'état de santé du patient.

Le JOA myélopathie. Suivant le même principe que le JOA lomboradiculalgie, le JOA myélopathie est un score clinique pour évaluer les problèmes myélopathiques. On évalue la motricité et la sensibilité des membres, et les troubles du sphincter. Il faut 5 minutes pour renseigner le score, qui nécessite un observateur formé à l'évaluation, on calcule le score en sommant les entiers associés à chacune des réponses. L'amplitude du score étant de 17, de 0 à $+17$. Plus le score est élevé, meilleur est l'état de santé du patient.

L'ODI (Oswestry disability Index) a été développé dans les années 1976-1980 par le docteur JO Brian à l'hôpital d'Oswestry en Grande Bretagne, la première version opérationnelle date de 1980. L'objectif étant d'indiquer le degré auquel une personne est restreinte par la douleur due à son problème au dos ; le score va renvoyer un taux d'incapacité. Il est composé de 10 questions reliées à des listes de choix comprenant 6 réponses possibles, qui sont associées à des entiers allant de 0 à 5 par ordre croissant. On calcule le score en faisant la somme des items qu'on multiplie par deux pour obtenir un pourcentage d'incapacité. Il y a 3 intérêts particuliers pour l'usage de ce score. Tout d'abord si une ou deux questions ne sont pas remplies, on peut tout de même calculer le score par une règle de trois. Si le nombre de réponses manquantes est supérieur ou égale à 3 le score n'est pas faisable. Par ailleurs si le patient coche plusieurs réponses pour une question, on retient celle qui est associée avec le chiffre le plus grand. Enfin, ce score prend en considération l'activité sexuelle du patient, qui reste un paramètre non pris en compte par les autres scores retenus. C'est un auto-questionnaire, il faut compter 5 minutes pour que le patient complète ce score.

Le NDI (Neck disability index). Sur le même modèle que l'ODI, on utilisera les NDI pour évaluer l'incapacité liée aux problèmes au cou. Les règles de cotation et le temps sont identiques à celles de l'ODI. Un critère de l'IDE pour la FDA pour juger de l'efficacité d'un traitement est une diminution minimum du taux d'incapacité de 15% entre le pré-opératoire et le post-opératoire. On peut également tabler sur un temps de remplissage de 5 minutes.

Le SF36. C'est le principal score d'évaluation de qualité de vie. En 1999, on recensait plus de 1400 études qui l'utilisaient. Il a été conçu le Dr John Ware en partenariat avec Glaxo Wellcome, après une enquête d'observation *Medical Outcome Study* (MOS) réalisée entre 1986 et 1990 au cours de laquelle 20000 patient ont été suivis [12]. Le score à la forme d'un QCM, avec 36 questions ou les réponses possibles sont structurées sous la forme d'une échelle sémantique du type *likert*. Le patient cochant une case (ou plusieurs par question, mais ce n'est pas conseillé). Le résultat multidimensionnel, se déploie sur 8 échelles d'analyses qui sont cotées de 0 à 100 à l'aide d'un algorithme croisé. Pour chaque échelle d'analyse 0 est le score le moins bon, et 100 le meilleur. Il y a deux types d'échelles : échelles physiques et psychologiques pour lesquelles on obtient des scores synthétiques. On peut noter que l'algorithme prend en compte la consistance des réponses.

Il a été validé en Français et le temps de remplissage est d'environ 10 minutes. Il est préférable que le patient le remplisse sous le contrôle d'un observateur formé à l'évaluation. En effet, dans ce test, 10 à 25% de la variance est due aux erreurs de mesure. Il est à noter également des variabilités inter-population dues à des facteurs ethniques (Modes de vies). De même que pour la douleur il y a un effet plancher important (beaucoup de sujets peuvent avoir 0 en préopératoire). Pour une analyse plus fine on pourrait utiliser le Sickness Impact Profile (à fin de différencier les très faibles résultats), mais il reste composé de 133 questions, donc difficilement applicable.

Un évolution minimum de 7 points pour chaque échelle du SF36 indique un changement de l'état de santé rapporté à l'échelle correspondante. On doit aussi tenir compte du fait que c'est un score de qualité de vie générique et non spécifique au mal au dos ; ainsi si le patient souffre d'une maladie quelconque sans rapport avec le problème au dos, le score sera altéré. Précisons une remarque importante, le score est la propriété intellectuelle de Quality Metric et son coût d'utilisation reste élevée (de 1 080\$ par an pour l'utilisation dans un cadre de recherche pour un chirurgien, à 2 800\$ par an pour l'utilisation dans le cadre d'une étude clinique pour un industriel).

Indicateurs de succès thérapeutique Il existe deux grands indices de succès thérapeutiques : l'indice de satisfaction du patient et le test d'Odom. Le premier est subjectif, et le patient choisit une réponse parmi 4 possibles. Pour le test d'Odom c'est l'observateur indépendant qui choisit une réponse parmi 4 possibles (différentes de celles de l'indice de satisfaction), on a un temps de remplissage d'une minute pour ces deux scores.

Le test de Hall C'est un test qui mesure la résilience du patient à la chirurgie. La résilience étant l'aptitude du patient à ne pas être déstabilisé par le traitement chirurgical. Le temps de remplissage est d'une minute.

2.2.5 Intégration pragmatique

« Tous les essais randomisés ayant pour but d'évaluer un traitement doivent s'appuyer sur la validité, la fiabilité de la chaîne : évaluation, diagnostic, traitement, résultat » [13] La mise en place pragmatique d'un système d'évaluation nécessite l'élaboration d'une taxinomie des traitements, des indications et des complications. Chaque patient étant évalué lors de la consultation préopératoire, juste après l'intervention et éventuellement lors des consultations post-opératoires (à 3 mois, 12 mois et 24 mois). Cela a impliqué de travailler sur des systèmes d'informations hétérogènes. Tout d'abord il y a le logiciel de gestion du cabinet médical (Surgica édité par Medialog) qui

Indications	Dégénératif Trauma Déformations Tumeur Rachis rhumatisal Neuro Chirurgie itérative Clinique Malformation SPL Autres	Traitement	Discectomie Discectomie endoscopique Laminectomie Recalibrage Fils souslamaires GPL Ligamentoplastie Réparation isthmique Vissage pédiculaire Cages Ciment Lombotomie Thoracophréno-lombotomie Thoracoscopie Thoracotomie Voie de la crête Voie de Wiltse Ablation de matériel Arthroèse intersomatique Foraminotomie Osteotomie Plaque antérieure Plaque Centaur Rachis cervical Cage Rachis cervical Discectomie Rachis cervical Laminoplastie Rachis cervical Ostéosynthèse Post Rachis cervical type A Rachis cervical type B Rachis cervical type C Rachis cervical vissage antérieur Reprise chirurgicale Pinsie vertébrale
Traitements	Liste		
Complications	Complications Péroperatoires		

FIGURE 4 – Aperçu de la taxinomie (CHU de Bordeaux – Dr P.Mangione)

permet de coter les mots-clefs, d'encoder les données socio-démographiques, et les comptes-rendus. Ensuite il y a le logiciel de la clinique (PMSI) qui nous permet de récupérer les données sur le coût économique, le type d'implant utilisé, et la durée moyenne de séjour (DMS). Enfin, il y a le logiciel d'évaluation (Scoring édité par Medscore). On a dès 2001 constitué des protocoles d'évaluation où étaient inclus les différents scores à réaliser, à différentes périodes. Il y avait des consultations dédiées à l'évaluation réalisées par des observateurs indépendants : des kinésithérapeutes (4 demi-journées par semaine). Lors de la consultation préopératoire, c'est-à-dire la veille des consultations le kinésithérapeute passait dans les chambres des patients pour le recueil des données sur support papier, les formulaires étaient ensuite ré-encodées au niveau informatique par l'attaché de recherche clinique. Ceci dans un souci de faisabilité, l'idée d'un ordinateur utilisé au pied du lit du patient ne satisfaisant pas des critères de productivité. Pour les consultations post-opératoire, l'observateur indépendant complétait les scores sur un support informatique dans une salle de consultation. Le patient ayant rempli les questionnaires subjectifs dans la salle d'attente. On pouvait ainsi s'assurer de la cohérence des réponses.

Il y avait trois protocoles au Centre Aquitain du Dos.

Le rôle de l'observateur indépendant étant de contrôler les données subjectives remplies par le patient, d'assurer l'examen clinique. On peut estimer le temps consacré à l'évaluation (dans la salle de consultation) entre 10 et

Protocole Lombaires	Préop	3 Mois	12 Mois	24 Mois	Temps de remplissage
SF36	x	x	x	x	10'
VAS	x	x	x	x	1'
ODI	x	x	x	x	5'
TEST D'ODOM		x	x	x	1'
INDICE SATISFACTION		x	x	x	1'
JOA LOMBAIRE	x	x	x	x	5'
TEST DE HALL		x	x	x	1'

Protocole Cervicales	Préop	3 Mois	12 Mois	24 Mois	Temps de remplissage
SF36	x	x	x	x	10'
VAS	x	x	x	x	1'
NDI	x	x	x	x	5'
TEST D'ODOM		x	x	x	1'
INDICE SATISFACTION		x	x	x	1'
TEST DE HALL		x	x	x	1'

Protocole Myélopathie	Préop	3 Mois	12 Mois	24 Mois	Temps de remplissage
SF36	x	x	x	x	10'
VAS	x	x	x	x	1'
NDI	x	x	x	x	5'
TEST D'ODOM		x	x	x	1'
INDICE SATISFACTION		x	x	x	1'
JOA MYELOPATHIE	x	x	x	x	5'
TEST DE HALL		x	x	x	1'

15 minutes en moyenne. En effet, les scores les plus longs étaient complétés dans la salle d'attente, il n'y avait qu'un travail de saisie à réaliser ainsi que les tests cliniques. À la fin de la consultation préopératoire, le chirurgien rencontrait le patient et l'observateur indépendant qui lui présentait les résultats d'évaluation individuel du patient sous forme d'un graphique en radar. Compte tenu du calage entre la fin de la consultation de l'observateur indépendant et la sortie du chirurgien qui sort de ses propres consultations, on peut estimer le temps d'une consultation de révision à 20 minutes. Le coût pour le patient étant celui d'une consultation de l'observateur indépendant (22 €), le chirurgien ne percevait rien.

On pouvait bien sûr créer autant de protocoles que l'on souhaitait en créant des composants d'observation dans le logiciel d'évaluation. Si bien qu'on pouvait intégrer des études cliniques prospectives parallèlement au suivi longitudinal, par exemple en prenant en compte des données radiologiques (évaluation

tion de fusion ou de mobilité vertébrale) ou bien des données per-opératoires (temps opératoire, saignements, type et taille d'implants)...plus spécifiques à une étude.

Pour les révisions post-opératoires, on utilisait le logiciel d'évaluation qui fournissait une fenêtre des révisions, et permettait d'organiser les convocations.

On avait une visualisation des programmations des évènements sous le logiciel d'évaluation, mais on l'a exporté ici sous un tableur pour préserver l'anonymat des patients. A partir de ces données on pouvait ainsi communiquer aux secrétaires les rendez-vous à programmer (avec les fenêtres de révision).

3 Exploitation des résultats

Le système en place a permis de réaliser trois types d'évaluation : par groupe, individuelle ou bien des études prospectives dans le cadre d'études cliniques pour le compte des chirurgiens ou d'industriels. Elle a également permis l'émergence de valeurs de références pour certaines pathologies. Au niveau de l'analyse graphique on a élaboré un graphique en radar obtenu en re-normalisant les données, à fin que plus l'étoile soit grande, meilleur est l'état de santé du patient correspondant. Par exemple pour une douleur qui vaut x on associe la « douleur modifiée » qui vaut $100-x$; pour un JOA Lombarde qui vaut y , on construit le JOA Lombarde modifié qui vaut $\frac{y+6}{21} \times 100$. On peut alors définir des valeurs de références à 95% ; par rapport à la moyenne si les distributions sont normales ou par rapport à la médiane si elles ne le sont pas ; par traitements, par pathologies. On les appelle valeurs de référence relatives par rapport à la pathologie, qu'on oppose aux valeurs de références absolues qui concernent les résultats par rapport à la population générale (dite normale). A ce titre il semble intéressant d'évaluer régulièrement des échantillons de la population générale de la même tranche d'âge pour permettre ce rapprochement. Il faut bien saisir que l'image de la population générale n'a pas les scores au maximum ; « la population n'est pas une population en bonne santé, mais une population dotée d'états de santé divers » [14]. La Sofres a ainsi évalué 4 000 sujets de 15 à 65 ans et 500 sujets de plus de 65 ans représentatifs de la société Française [14]. Autant pour le SF 36 nous avons des résultats obtenus grâce à la SOFRES, autant il manque des indicateurs de douleurs qui seraient primordiaux. On peut tout de même avancer que 80% de la population est atteinte de douleur lombaire à un moment de la vie, c'est la cause la plus importante d'incapacité chez les

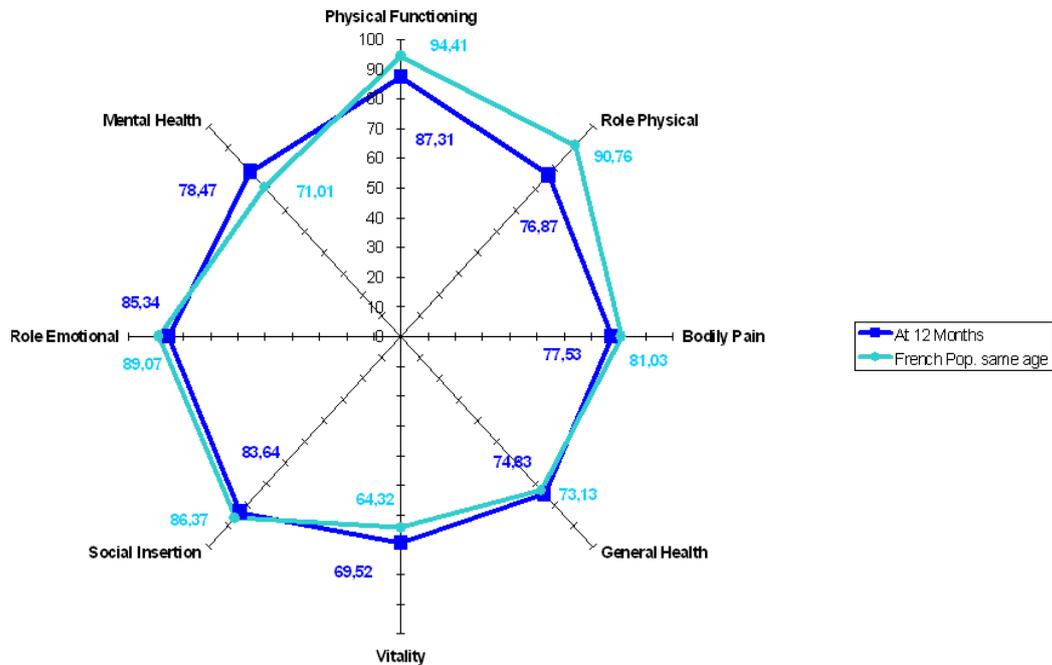


FIGURE 5 – Etude longitudinale Vissage selon Magerl (Centre Aquitain du Dos)

personnes de plus de 45 ans [Diagnostic evaluation of low back pain].

3.1 Analyses des groupes de patients

Les analyses de groupes doivent permettre des générer des valeurs de référence par pathologies, traitements ou bien encore par chirurgien, sur la même démarche qu'en biologie clinique [15].

3.1.1 Obtention des données

A partir de l'onglet analyse du logiciel de cotation des mots clefs (Surgica-Medialog) qu'on exporte sur un tableur, on peut alors enregistrer un fichier de patients sélectionnés.

On utilise alors un tableur ayant la possibilité de construire des tableaux croisés dynamiques pour extraire une matrice à coefficient binaire ou les colonnes représentent les mot-clefs et les lignes les identifiant patients. Cela donne une matrice de 6 562 lignes sur 183 colonnes ou sont renseignées 28 809 cotations.

On construit alors la sélection par l'usage de filtres sur les mots-clefs, puis

on exporte la liste obtenue de patients sous forme de fichiers textes avec des tabulations comme limites d'enregistrements. A l'aide du logiciel d'évaluation, on importe cette même liste pour récupérer les résultats d'évaluation correspondants, le rapprochement se faisant sur l'identifiant patient unique.

A partir de là, on sélectionne les scores à visualiser et on lance la requête. L'onglet des résultats se présente sous la forme ci-dessous avec tous les champs nécessaires à un traitement statistiques :

On obtient une grille des résultats qui peut s'exporter sous forme de fichier pour les exploiter sous un tableur ou sous un logiciel de statistiques.

Voici quelques résultats obtenus :

On peut bien entendu comparer deux traitements différents pour une même pathologie. En fait, la méthode est très souple et permet de réaliser des analyses prospectives.

3.1.2 Valeurs de références

La détermination des valeurs de références dépend du type de distribution de données. Pour décider de la normalité des données on peut s'appuyer sur un test de normalité du type Kolmogorov-Smirnov ou bien s'appuyer sur une représentation visuelle, dans la pratique on utilise les deux. D'un point de vue empirique, pour l'évaluation du dos, 50% des courbes de distribution des données sont Gaussiennes. Dans le cas où la distribution des résultats est normale on détermine l'intervalle de confiance à 95% autour de la moyenne à partir de la formule bien connue $[\mu - 2\sigma; \mu + 2\sigma]$ où μ désigne la moyenne et σ l'écart-type de la série. Par contre, si la courbe n'est pas Gaussienne, on détermine l'intervalle de confiance à 95% autour de la médiane de la série. Pour sa construction, on classe les valeurs du caractère par ordre croissant (en associant les probabilités correspondantes) et on inclut de part et d'autre de la médiane les valeurs prise par les caractères jusqu'à l'obtention d'un intervalle comprenant 95% de l'échantillon. Bien entendu si la variable prend plusieurs fois un caractère, on prend également le même nombre de caractères de l'autre côté de la médiane. Si on bute sur une borne de l'amplitude, on piochera alors uniquement les valeurs de l'autre côté de la médiane.

3.2 Analyses Individuelles

Les analyses individuelles sont des rapports de consultation remises au chirurgien et éventuellement au patient ou au médecin généraliste qui suit

EVA Lombaire à 12 mois- Discectomies sans Vissage Pédicul

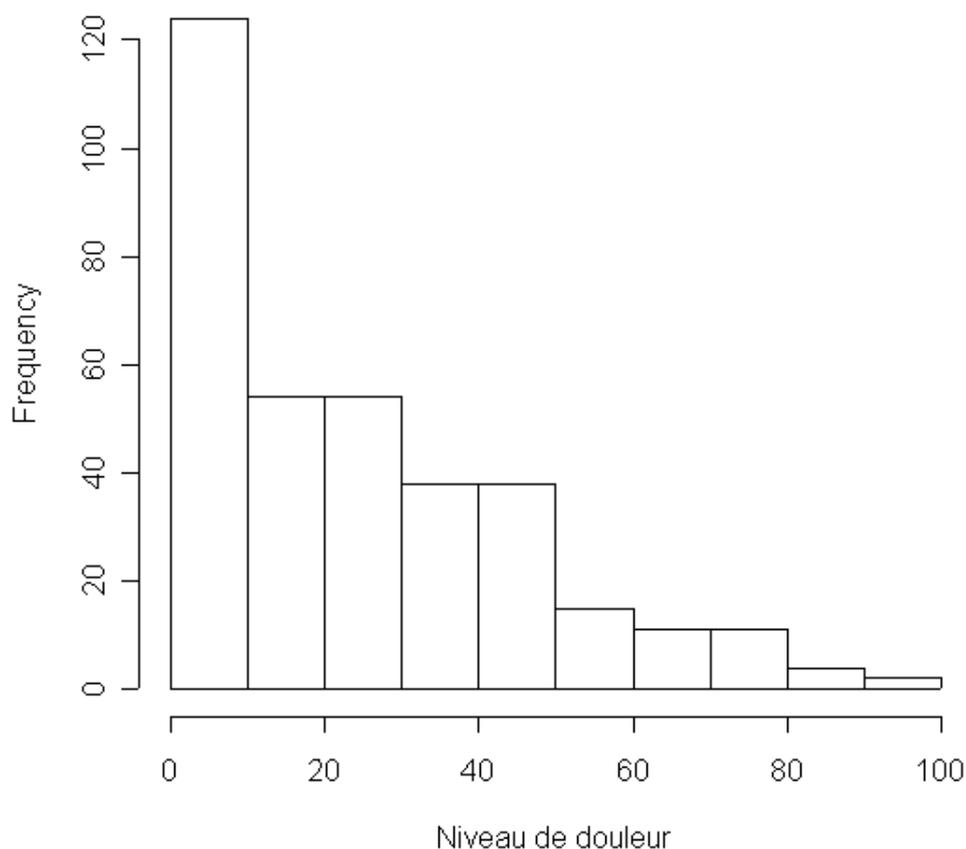


FIGURE 6 – CAD Discectomie sans Vissage Pédiculaire N=352

le patient. On représente sous la forme d'un graphique en radar renormalisé le polygone des résultats du patient, on peut alors préciser si l'équilibre du patient est relatif ou absolu.

3.3 Analyses Statistiques

« La fonction la plus connue de l'expérience est celle qui consiste à mettre les hypothèses à l'épreuve. L'expérimentateur lit par dessus l'épaule du théoricien et il élabore des expériences dans le but d'éprouver si l'on n'est pas allé trop loin dans la réduction, au point de rendre la théorie empiriquement inadéquate. Cette caractérisation de l'expérience est simple et attrayante ;

EVA Lombaire à 12 mois - Vissage pédiculaire 1 niveau

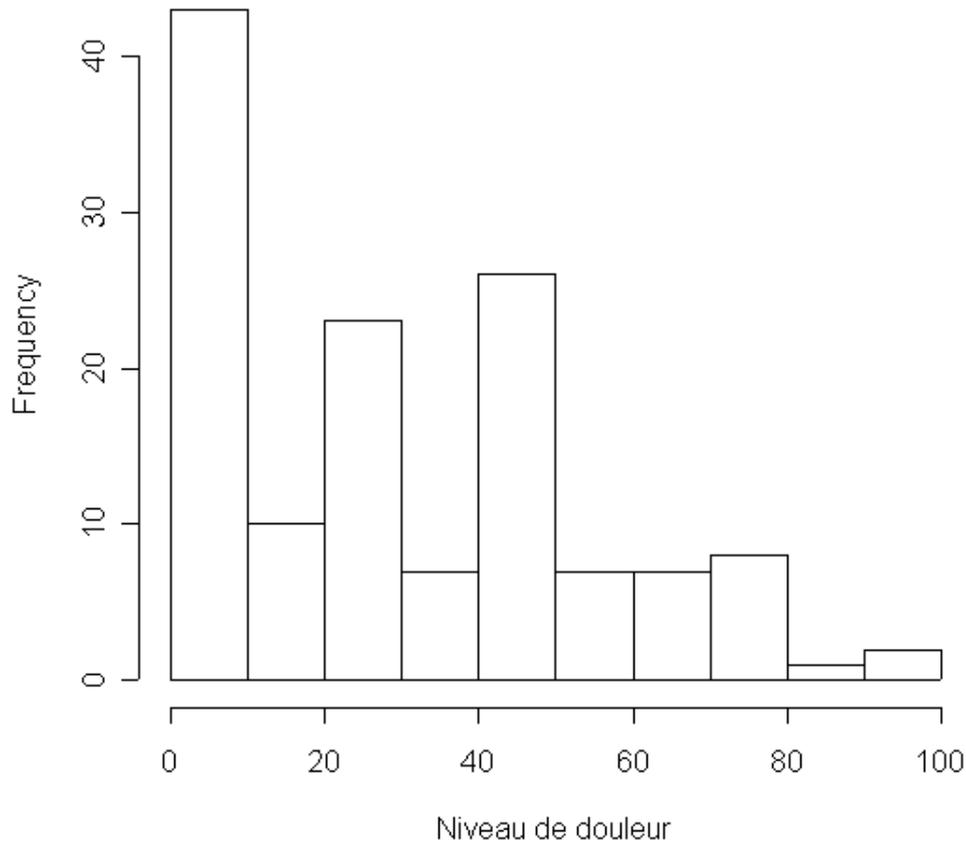


FIGURE 7 – CAD – Vissage Pédiculaire 1 niveau N=683

malheureusement elle est aussi trop simple. Elle conduit en premier lieu à négliger le fait que la procédure de mise à l'épreuve est en général une procédure comparative qui ne s'achève pas avec le renforcement ou la réfutation d'une hypothèse unique, mais avec le renforcement d'une hypothèse au dépend d'une autre. Elle conduit aussi à négliger un second aspect : le fait que la procédure d'épreuve a tout autant à voir avec le contenu informatif » [16].

3.3.1 Le cadre des RCT (Randomized Controlled Trials)

Le cadre des RCT permet d'atteindre le plus haut niveau de preuve dans le cadre de l'EBM ; cependant Sackett reprenant Mc Leod distingue 8 biais

EVA Lombaire à 12 mois - Vissage pédiculaire 2 niveaux

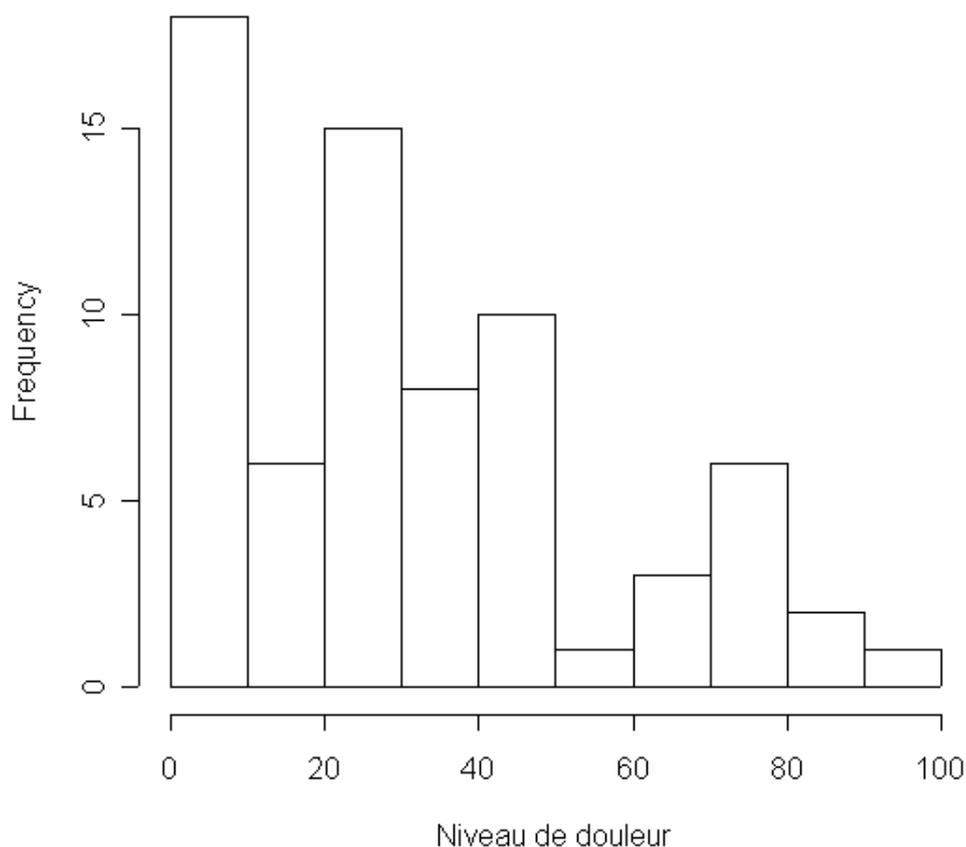


FIGURE 8 – CAD – Vissage Pédiculaire 2 niveaux N=71

pour l'appliquer à la chirurgie [17] :

1. Le chirurgien apprend à améliorer l'effet d'une technique
2. Le médicament produit un résultat identique quelque soit la compétence du médecin
3. Il est difficile de standardiser une intervention, il y a cadre minimaliste (niveau anatomique par exemple).
4. Impossibilité de pratiquer le double aveugle, et difficulté à pratiquer le simple aveugle.
5. Un essai thérapeutique en chirurgie ne peut statuer de l'inéligibilité après la randomisation.

EVA Lombaire à 12 mois - Vissage pédiculaire 3 niveaux

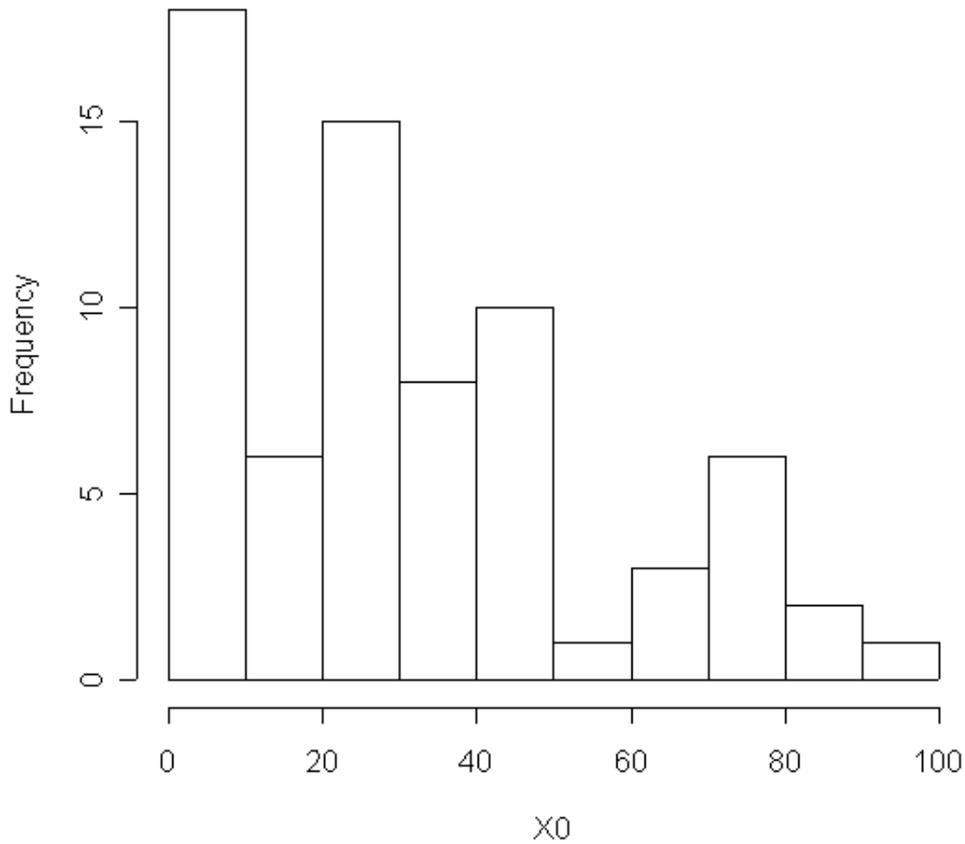


FIGURE 9 – CAD – Vissage Pédiculaire 3 niveaux N=22

6. Une expérience médicamenteuse peut débuter dans les minutes qui suivent une randomisation. Pour la chirurgie c'est impossible et l'état de santé du patient peut évoluer.
7. Co-intervention, le chirurgien opère avec une équipe.
8. Problème de recrutement des patients, le patient ne peut obtenir un nouveau médicament que dans le cadre d'un RCT, alors que pour un dispositif médical il peut l'obtenir en changeant de chirurgien.

EVA Lombaire à 12 mois - Vissage pédiculaire 3 niveaux

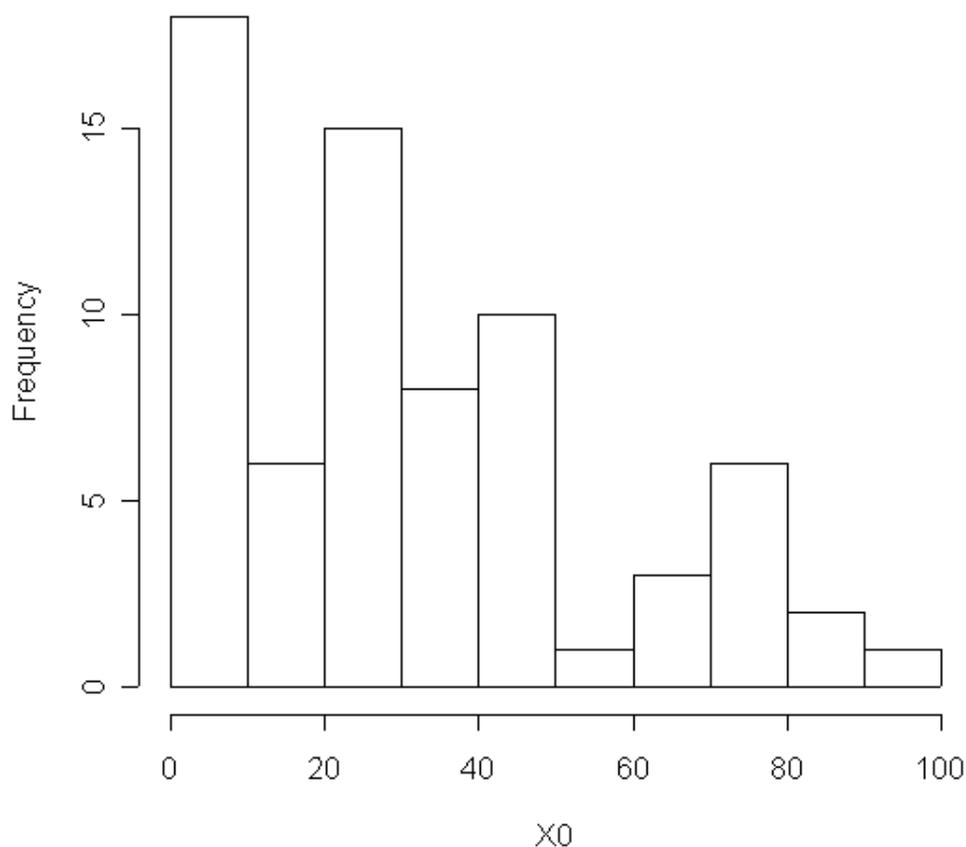


FIGURE 10 – CAD – Vissage Pédiculaire 3 niveaux N=22

3.3.2 Tests

Test entre deux dispositifs à une date t post-opératoire La typologie des courbes laisse apparaître plusieurs types de courbes de distribution : courbes de Gauss, courbes de Poisson notamment (pour la douleur). On a représenté ci-dessus les deux principales formes des courbes de la douleur, qu'on retrouve tant au niveau lombaire qu'au niveau cervical, on a rapporté à $[0; 100]$ l'amplitude des niveaux de douleurs. Ce sont des courbes de Poisson de paramètres respectifs 1.7 et 1. On peut remarquer qu'on est très éloigné du cadre de deux courbes en cloches.

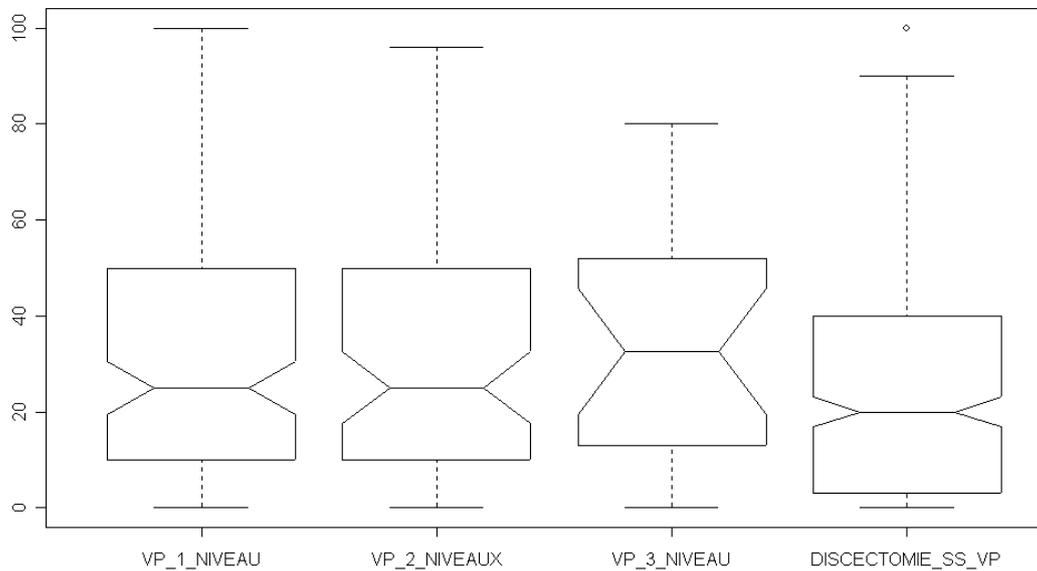


FIGURE 11 – Centre Aquitain du Dos - Valeurs de Références de la Douleur lombaire et Comparaison entre les différents Traitements à 12 mois

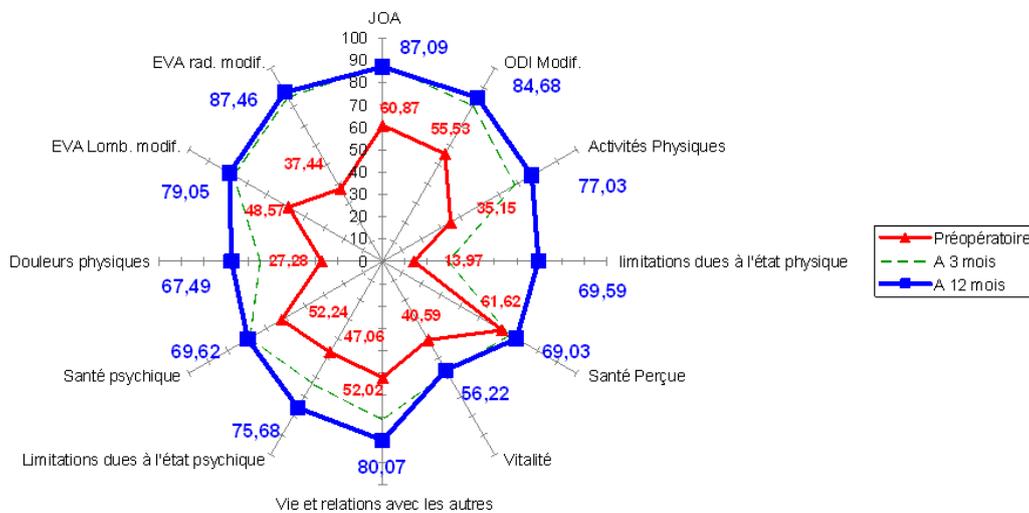


FIGURE 12 – Ligamentoplastie (Centre Aquitain du Dos) N=38 à 12 mois

Voici ce que donne le lissage de ces deux courbes avec les paramètres retenus pour des courbes de Poisson. Voici ce que donneraient les fonctions densité et de répartition :

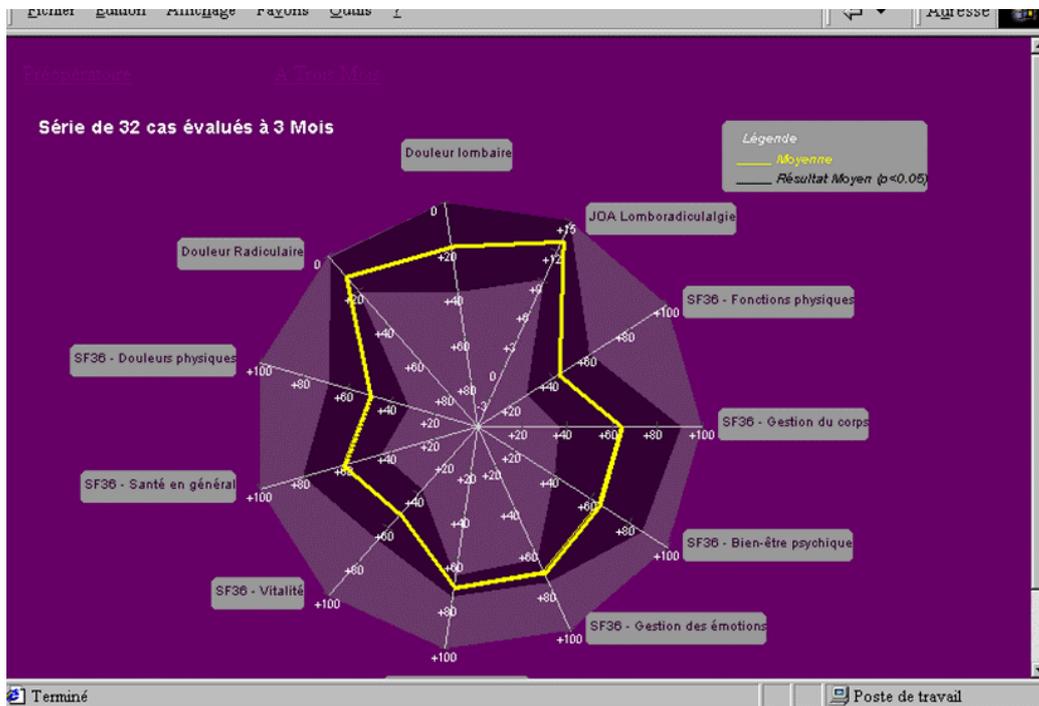


FIGURE 13 – Application java

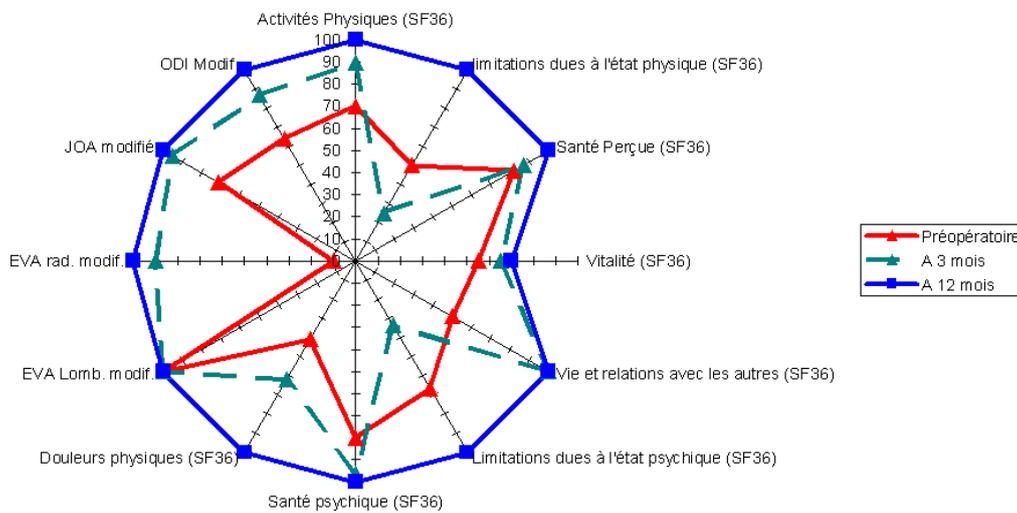


FIGURE 14 – Centre Aquitain du Dos - Monsieur Durand (âge 35 ans, opéré d'une hernie discale)

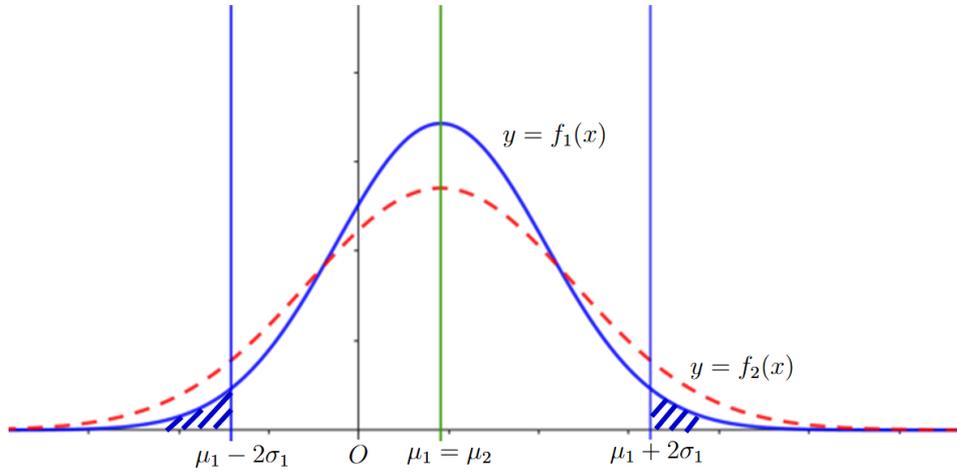


FIGURE 15 – Distribution - Série 1 sous l'hypothèse de normalité des données

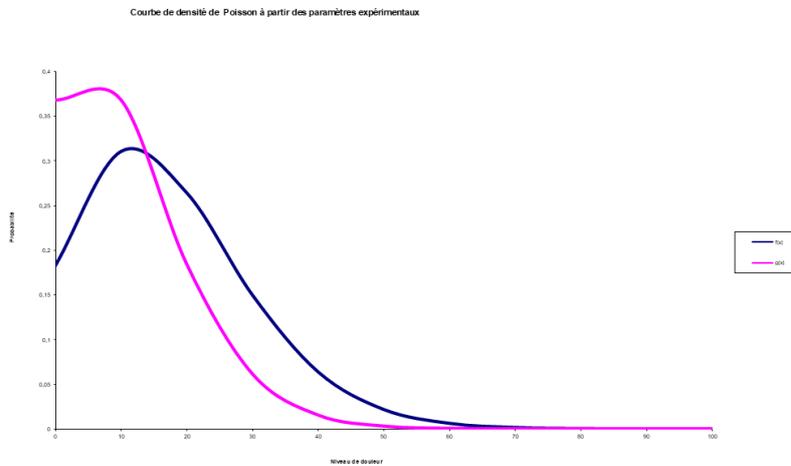


FIGURE 16 – Distribution - Série 1 sous l'hypothèse d'une distribution de Poisson.

Cela pose un problème pour comparer deux paramètres dans le cadre d'un test d'hypothèse. Même l'utilisation d'un test non paramétrique comme celui de Mann-Whitney peut être biaisé par la mé-

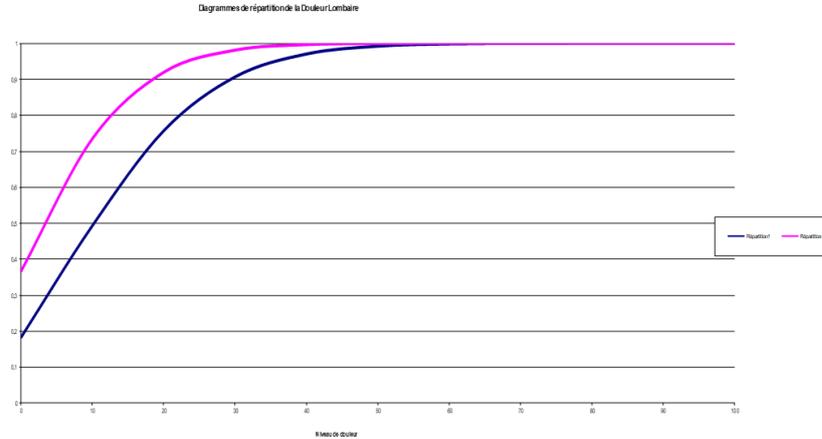


FIGURE 17 – Répartitions - Série 2

thodologie du recueil de données. Rappelons que ce test est basé sur les rangs des valeurs des deux populations réunies en une seule série. Il faut également une méthodologie adaptée pour traiter le cas où les séries ont une modalité nulle (ce qui est souvent le cas pour la douleur). Par ailleurs l'écart-type de la série ne tend pas systématiquement vers zéro, ainsi on peut remettre en cause le nombre d'erreurs de type II qu'on retrouverait en chirurgie [18]. Ceci est renforcé par une puissance statistique nécessaire qui serait aberrante.

Soit une série de données $(x_i)_{1 \leq i \leq n}$. On note \bar{x} la moyenne arithmétique, $\bar{x}_q = \sum_{i=1}^n \frac{x_i^2}{n}$ la moyenne quadratique, et on appelle $V_n = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n}$ la variance. Montrons que la variance ne converge pas systématiquement vers zéro.

Démonstration. Montrons tout d'abord que (V_n) converge.

$$\text{Soit } V_{n+1} = \sum_{i=1}^{n+1} \frac{(x_i - \bar{x})^2}{n} = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n} \times \frac{n}{n+1} + \frac{(x_{n+1} - \bar{x})^2}{n+1} = V_n \times \frac{1}{1 + \frac{1}{n}} + \frac{(x_{n+1} - \bar{x})^2}{n+1}.$$

Comme $(x_i)_{1 \leq i \leq n}$ est bornée, lorsque n tend vers l'infini, l'expression de la variance $V_{n+1} \rightarrow V_n \Leftrightarrow V_{n+1} - V_n \rightarrow 0$. Donc, $(V_i)_{1 \leq i \leq n}$ est convergente vers une valeur réelle positive. En conséquence l'écart-type $\sigma_n = \sqrt{V_n}$ est convergent.

Montrons que l'expression de $(V_i)_{1 \leq i \leq n}$ ne converge pas systématiquement vers zéro. Soit $V_n = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n} = \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \sum_{i=1}^n \frac{x_i^2}{n} - 2\bar{x} \sum_{i=1}^n \frac{x_i}{n} + \bar{x}^2 = \bar{x}_q - 2\bar{x} \bar{x} + \bar{x}^2 = \bar{x}_q - \bar{x}^2$. Donc, une condition nécessaire pour que la variance converge vers zéro est que la moyenne quadratique vaut la moyenne arithmétique au carré. Ce qui est presque systématiquement impossible. A fortiori, l'écart-type ne tend pas systématiquement vers zéro.

Prenons un exemple, considérons un nombre pair de données. On suppose que $x_i = Sup$ pour la moitié des données et $x_i = Inf$ pour l'autre moitié (pour maximiser la variance). Alors $\bar{x} = \frac{1}{2}(Sup + Inf)$, tandis

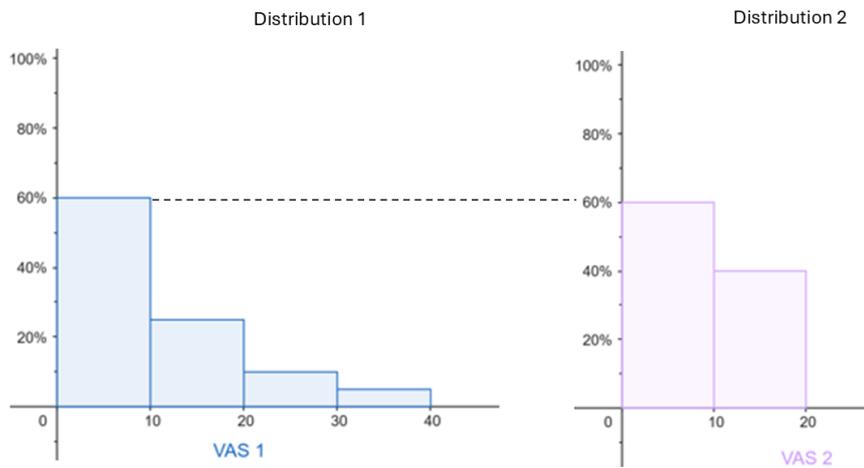
$$\text{que } \sqrt{\bar{x}_q} = \sqrt{\frac{n(Sup^2 + Inf^2)}{2n}}.$$

En élevant au carré de part et d'autre, et en simplifiant les expressions. Il vient $\frac{1}{4}Sup^2 + \frac{1}{4}Inf^2 + \frac{1}{2} \times Sup \times Inf = \frac{1}{2}Sup^2 + \frac{1}{2}Inf^2 \Leftrightarrow Sup \times Inf = \frac{1}{4}Sup^2 + \frac{1}{4}Inf^2$, ce qui est manifestement faux pour une multitude de cas. On conclut que la variance et donc l'écart-type ne convergent pas systématiquement vers zéro. Ils sont même convergent la plupart du temps vers des valeurs non nul. On est proche du contre exemple qui sert de cas de base ici. \square

Interprétons ce contre exemple qui nous servira de cas de base. Montrons que la variance est systématiquement fini dans le cas de la chirurgie. Interprétons le cas d'une distribution où des effectifs avec $\frac{N}{2}Inf$ ou $\frac{N}{2}Sup$ (Où N définit le nombre de patients inclus) n'est pas exceptionnel. On pense au cas d'une évaluation post-opératoire avec un recul de 3 ou 12 mois. La distribution des résultats peut témoigner d'un mode à 0, renchérit par un effet plancher. Prenons le cas où il y aurait 60 % des effectifs qui n'auraient plus de douleur post-opératoire. Alors c'est légèrement supérieur à $\frac{N}{2}Inf$. Par ailleurs, l'amplitude de la série ne sera pas 100, car on suppose qu'aucun patient n'est 100 en post opératoire. Si bien que Sup est plus proche de Inf en pratique.

On rappelle qu'une erreur de type II est de conclure qu'il n'y a pas de différence entre les deux traitements, alors qu'en réalité il en existe une qui ne serait décelable qu'en augmentant le nombre de patients de l'étude. Fisher s'étant toujours démarqué de Neyman à propos de ce type d'erreur pour le domaine biologique.

Voici dessous deux courbes hypothétiques de la douleur à 12 mois : $f(x) = \frac{4x}{x!}e^{-4}$ et $g(x) = \frac{0.8x}{x!}e^{-0.8}$ ou on montre une supériorité d'une moyenne sur l'autre dans le cadre de deux distributions de Poisson



La variance de la série 1 est supérieure à la variance de la série 2, car la dispersion par rapport à la moyenne est plus élevée dans la série 1 que dans la série 2, où l'effectif total est constant.

On observera que la fréquence de la première classe des deux distributions est constante .

FIGURE 18 – Exemple d'une distribution du cas de base.

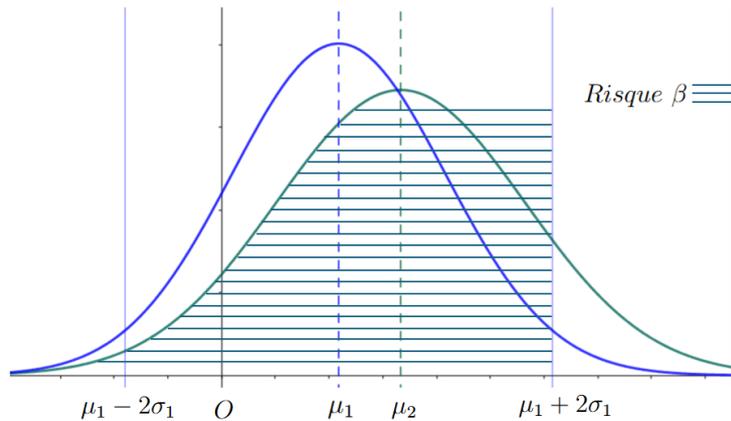


FIGURE 19 – Risque de deuxième espèce.

avec une « p-value » inférieure à 0.05.

Voici la courbe de répartition correspondante qui nous renseigne sur le niveau de douleur correspondant à 95% de l'aire de la courbe. Pour $f(x)$, l'espérance mathématique valant 8, et pour $g(x)$ elle vaut

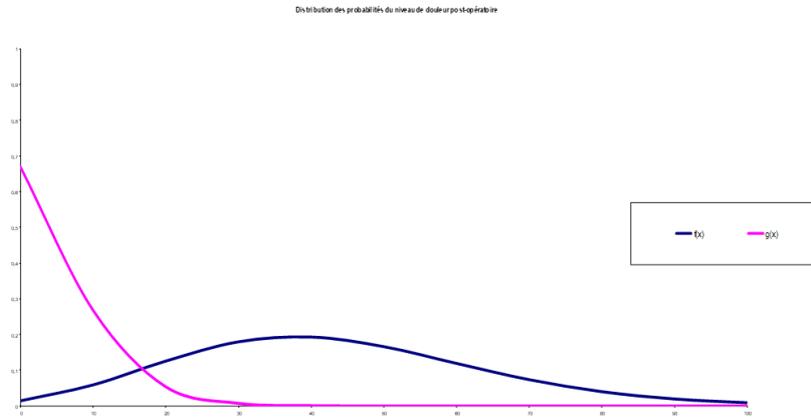


FIGURE 20 – Distribution - Série 2

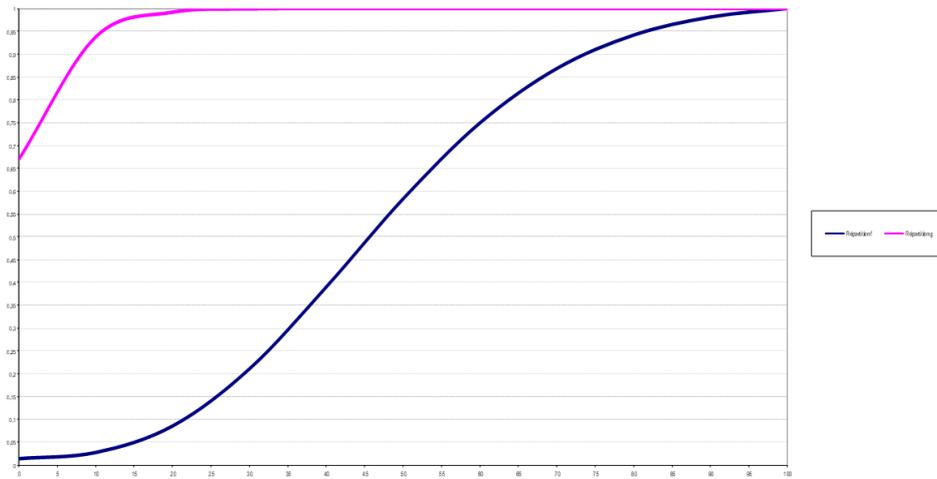


FIGURE 21 – Répartition - Série 2

40. C'est un cadre idéal assez éloigné de la réalité dans l'état actuel des traitements. En effet, en reprenant la littérature on retrouve des niveaux de douleur moyens à 12 mois entre 20 et 30. On est pour l'instant véritablement confronté à une indécidabilité d'un point de

vue statistique.

Test pour un dispositif entre le préopératoire et le postopératoire. Là encore il faut utiliser les tests adaptés à la distribution des données, cependant on doit ici aboutir à une différence significative avec une variation de l'estimateur (moyenne ou médiane selon les cas) du score d'au moins 30 % (qui est l'effet placebo) si la consultation postopératoire se situe à 3 mois de l'intervention ; ainsi qu'une « p-value » inférieure à 1%.

Classification des chirurgiens. La méthode des tests d'hypothèses est soumise à la l'égalité des chances d'un bon résultat indépendamment des chirurgiens. Or en tant que système complexe à opérateur central, en pratique n'est pas forcément le cas. Voici ce que peut donner une analyse par chirurgien suivant une échelle d'analyse du SF36 : Nous proposons de classer les chirurgiens par résultat médian suivant l'échelle d'analyse en 4 groupes qui correspondent au quantile auquel appartient la médiane de ses résultats. A partir de ce classement, une étude expérimentale doit pouvoir indiquer le groupe auquel appartient le chirurgien. Ainsi, on peut maîtriser l'ensemble des variables connexes qui peuvent avoir un effet positif ou négatif sur une opération, qui conduisent soit à du bruit soit à une amplification du signal.

4 Conclusions

4.1 Pragmatisme

Il faut se méfier d'un passage systématique de la « significativité » statistique à la « significativité » clinique. En effet, la statistique raisonne sur des groupes de patients alors que le médecin est confronté à raisonner sur un patient. A la consultation qui se réalise en un temps fini (en moyenne $\frac{3}{4}$ d'heure) le médecin ajuste les probabilités par des probabilités inverses, c'est une approche bayésienne. Un exemple typique d'inférence bayésienne sont les concepts de valeurs prédictives positives et de valeurs prédictives négatives.

Où « sens » désigne la sensibilité, « spec » la spécificité et « prev » la prévalence de la maladie [19]. On peut aussi mesurer la complexité du jugement clinique attribuant la valeur d'un niveau de douleur lombaire (subjectif) à une raison purement ou partiellement clinique ou pas du tout, ainsi que son interprétation bayésienne. D'autre part il peut y avoir un décalage entre la

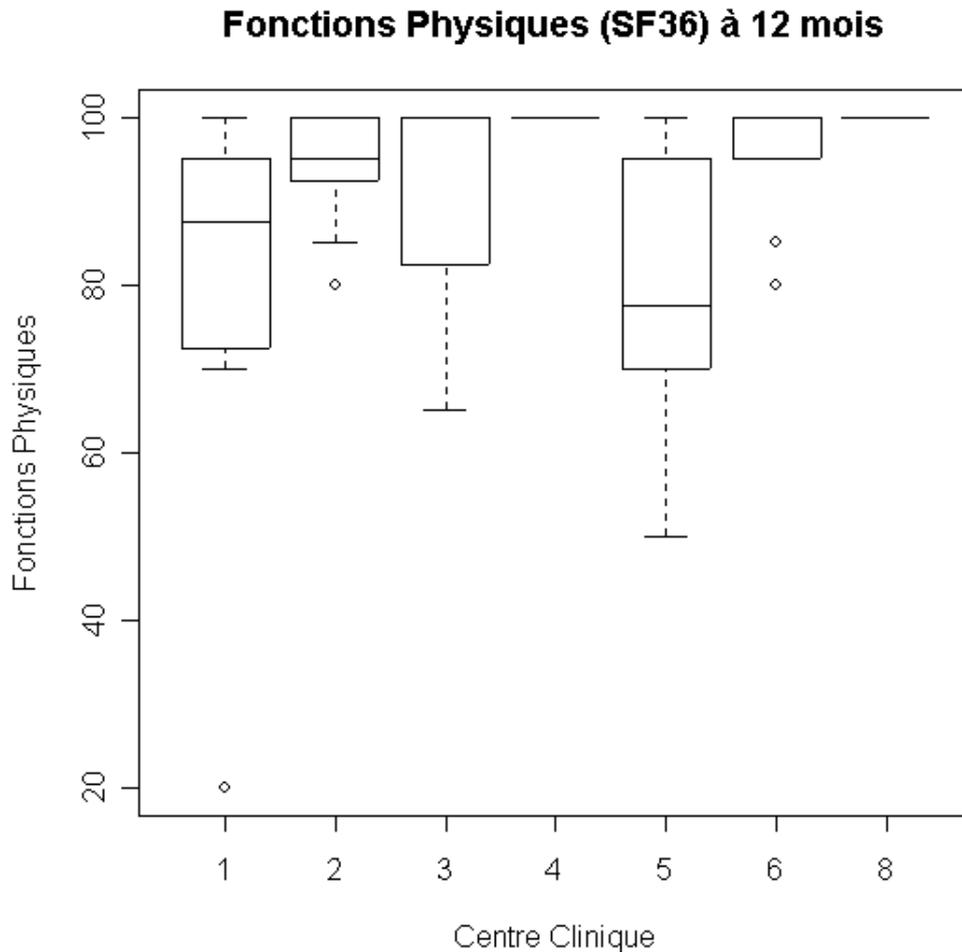


FIGURE 23 – Etude multicentrique Wallis Source Abbot Spine

perspective du patient et celle du chirurgien [20], la pondération des axes d'évaluation et même quelques fois des parties d'un score peuvent être un bon indicateur pour évaluer la réussite thérapeutique. Il existe d'autres indicateurs de performances comme la durée de survie d'un implant : Ou bien, la matrice de transition de l'état d'un disque intervertébral :en fonction des grades de Pfirrmann sur 12 mois. A titre exploratoire on peut également faire des analyses multidimensionnelles mais qui restent délicates compte-tenu que nous avons à faire à un système biologique complexe. On peut voir par exemple ci-dessus que les indications de spondylolisthésis sont plus près de bons résultats (résultats supérieurs à la médiane) que les cages associées

« Wallis 1° Génération »

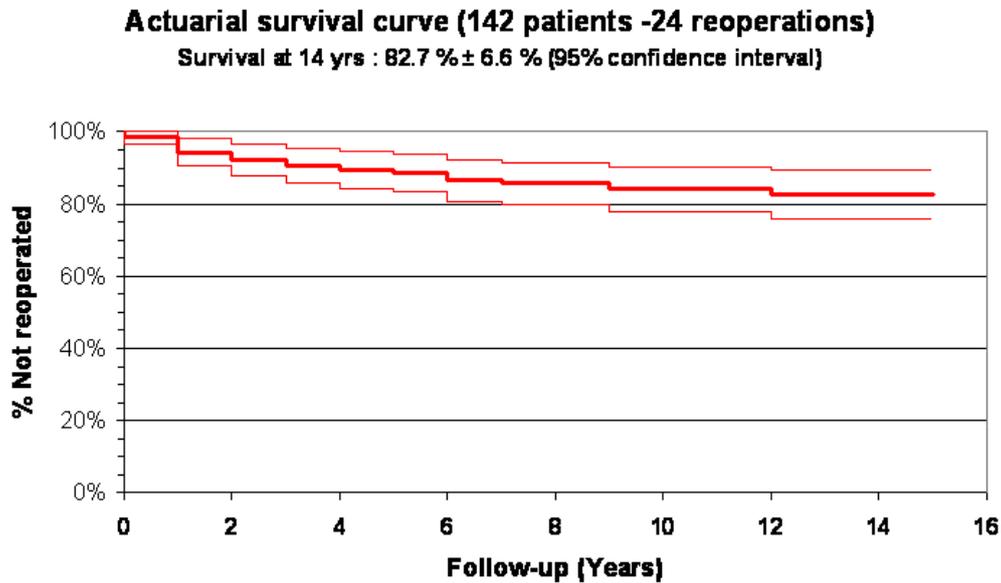


FIGURE 24 – Étude Wallis - Abott Spine (Mr Meunier).

	Pf.I	Pf.II	Pf.III	Pf.IV	Pf.V
Pf.I	0	0	0	0	0
Pf.II	0	0	0	0	0
Pf.III	0	0,222	0,666	0,111	0
Pf.IV	0	0,143	0,5	0,375	0
Pf.V	0	0	0	0	1

FIGURE 25 – Matrice de transition (Genvrin).

dans certains cas à cette même indication. Dans une autre optique, on peut aussi prendre en compte la compatibilité des différents implants utilisés avec l’horizon technique, en ce sens les systèmes de renormalisation mécaniques qui laissent le disque intact ou participent à sa re-génération seraient compatibles avec d’éventuels traitements par les cellules souches, tandis que les traitements par arthrodèses ou prothèses de disques ne permettent pas cette compatibilité. Ainsi dans une perspective Popérienne ou l’horizon technique (représenté par le traitement par des cellules souches) servirait de falsificateur, on peut décider de la supériorité d’un traitement sur un autre ; étant entendu qu’il n’y a pas de différences significatives d’un point de vue numé-

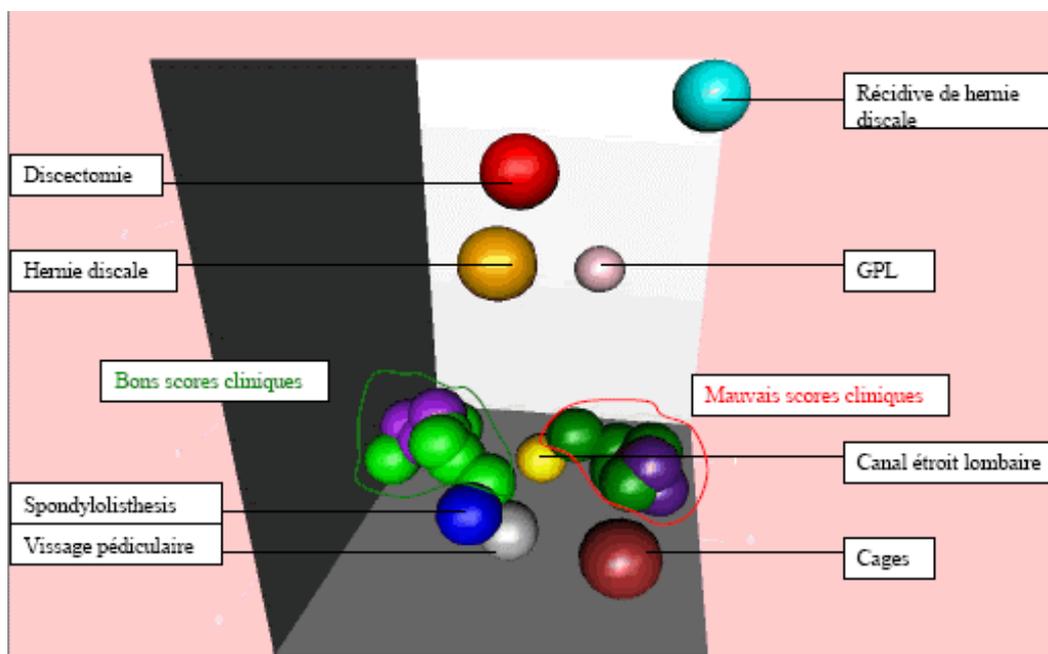


FIGURE 26 – Matrice de transition (Lesbats).

rique d'un traitement sur un autre.

4.2 Analyse de groupe performante

Si l'analyse individuelle peut conduire à des erreurs ou des variabilités importantes, l'analyse de groupe offre une représentation fidèle de l'activité du chirurgien. En l'étendant à tous les opérateurs, on disposerait d'une base de données solide qui permettraient l'établissement de valeurs de références pour les pathologies du rachis à l'échelle d'un pays et non plus d'un centre clinique. Elles permettraient également de classer les chirurgiens suivant les quantiles du critère de jugement retenu. Alors à partir de cette classification en groupes cliniquement homogène, on pourrait repartir sur les bases de la méthode EBM.

4.3 Coût Economique d'un telle évaluation

On peut apprécier le coût économique pour la mise en place d'un tel système, on l'évalue à 110.50 € par patient. Voici le détail : on considère 4 consultations à 22 € réalisées par l'observateur indépendant ; un attaché de recherche clinique qui revient à 2 700 € par mois charges comprises pour

travailler sur une base de 60 patients en post-opératoire et 60 patients en préopératoire par semaine, soit 480 évaluations par mois, ce qui implique un coût de revient de 5.62 € par consultation. Soit sur la base de 4 consultations : préopératoire, 3 mois, 12 mois, à 24 mois, on aboutit à un coût de revient approximatif par patient de 110.50 € par patient, ce qui est dérisoire si on le rapporte au coût d'une intervention comme une arthrodeèse. Bien sûr le coût de l'évaluation rapporté à une opération plus bénigne comme une hernie discale resterait plus important. On ne tient pas compte des frais d'immobilisation engendrés par la mise en place de la structure : cabinet de consultation mis à la disposition de l'observateur indépendant pour les révisions, bureau de l'attaché de recherche clinique et le matériel informatique nécessaire. Le logiciel d'évaluation ayant été développé en interne. Cependant on peut préciser que dans le cadre de son activité globale le travail consacré à la saisie, et aux analyses de l'ARC ne représentaient pas 100% de son activité, on peut l'évaluer à 50%. Sa tâche se résumant à la saisie des dossiers post-opératoires, la programmation des consultations de révision, la récupération de données issues des différents systèmes d'information et la mise au point de présentations pour les congrès aux quels les chirurgiens se rendaient. La tâche la plus délicate étant la récupération des données iconographiques à partir des appareils photographiques des chirurgiens ou de planches numérisées. Au niveau du logiciel d'évaluation, on a opté pour la création d'un dossier physique (à partir de l'identifiant patient) sur le serveur pour chaque patient ou étaient enregistrées les photos (un sous dossier par consultation), les photos n'étaient donc pas enregistrées dans la base de données SQL. On peut également préciser que les mesures radiographiques étaient faites à partir des clichés et non à partir des sources informatiques.

Table des figures

1	Nombre d'occurrences <i>Pubmed</i>	6
2	Echelle EVA Recto-verso. Le patient déplace le curseur sur le recto, l'observateur renseigne la valeur à partir du verso.	7
3	Évaluation de la douleur pour des cages intersomatiques cervicales (Source Société Scient'x – Cages Samarys)	8
4	Aperçu de la taxinomie (CHU de Bordeaux – Dr P.Mangione)	11
5	Etude longitudinale Vissage selon Magerl (Centre Aquitain du Dos)	14
6	CAD Discectomie sans Vissage Pédiculaire N=352	16
7	CAD – Vissage Pédiculaire 1 niveau N=683	17

8	CAD – Vissage Pédiculaire 2 niveaux N=71	18
9	CAD – Vissage Pédiculaire 3 niveaux N=22	19
10	CAD – Vissage Pédiculaire 3 niveaux N=22	20
11	Centre Aquitain du Dos - Valeurs de Références de la Douleur lombaire et Comparaison entre les différents Traitements à 12 mois	21
12	Ligamentoplastie (Centre Aquitain du Dos) N=38 à 12 mois .	21
13	Application java	22
14	Centre Aquitain du Dos - Monsieur Durand (âge 35 ans, opéré d'une hernie discale)	22
15	Distribution - Séries 1 sous l'hypothèse de normalité des données	23
16	Distribution - Séries 1 sous l'hypothèse d'une distribution de Poisson.	23
17	Répartitions - Séries 2	24
18	Exemple d'une distribution du cas de base.	26
19	Risque de deuxième espèce.	26
20	Distribution - Séries 2	27
21	Répartition - Séries 2	27
23	Etude multicentrique Wallis Source Abbot Spine	29
24	Étude Wallis - Abott Spine (Mr Meunier).	30
25	Matrice de transition (Genvrin).	30
26	Matrice de transition (Lesbats).	31

Références

- [1] R. M. B. B. KJ Blount, WJ Kropinger, “Moving toward a standard for spinal fusion outcomes assessment,” *Journal of spinal disorders and techniques.*, vol. 15, no. 1 (February 2002).
- [2] C. Bernard, *Introduction à la médecine expérimentale*, 1st ed. (1833).
- [3] R. Bouveresse, *Karl Popper ou le rationalisme critique*, 1st ed.
- [4] J. P. Vandenbroucke., “Clinical investigation in the 20th century : the ascendancy of numerical reasoning,” *Lancet*.
- [5] I. Hacking, *La qualité*.
- [6] M. J. N. C.Pfirmann, A.Medtzdorf, “Magnetic Resonance Classification of Lumbar Intervertebral Disc Degeneration,” *Spine*, vol. 26, no. 17, pp. 1873–1878 (september 1, 2001).

- [7] H.Genvrin, “Sociobiologie macroscopique.” .
- [8] M.Guillaumat, “Etude critique de l’évaluation des résultats.” .
- [9] F. C.Bombardier, MD, “Outcome assessment in the evaluation of treatment of spinal disorders,” *Spine*, vol. 25, no. 24, pp. 3100–3103 (2000).
- [10] “Validation en turque du SF36.” .
- [11] S. M.K Schaufele, “Outcome research in patients with chronic low back pain,” *Orthopedic Clinics of North-America*, , no. 34, pp. 231–237 (2003).
- [12] M. K. B. J.E Ware, K.K Snow, *SF-36 Health Survey – Manual Interpretation Guide* (Quality Metric Inc, The Health Assessment Lab), 1st ed. (1993,2000).
- [13] M. H. E.J Carragee, “Diagnostic evaluation of low back pain,” *Orthopedic Clinics of North-America*, , no. 35, pp. 7–16 (2004).
- [14] A.Lepège, *Le questionnaire MOS – SF36, manuel de l'utilisateur et guide d'interprétation des scores* (ESTEM), 1st ed.
- [15] X.-A. JP Henny, C Petitclerc, “Réviser le concept de valeurs de référence : une nécessité,” *Annales Biologies Cliniques*, , no. 59, pp. 383–392 (2001).
- [16] B. C. V. Fraassen, *Lois et Symétrie* (Vrin, traduction de Catherine Chevalley).
- [17] G. G. P. R.B Haynes, D.L Sackett, *Clinical Epidemiology – How to do clinical Practice Research* (Lippincott Williams Wilkins), 3rd ed. (2006).
- [18] C. S.Bailey, al, “A high potential for type II error,” *Spine*, , no. 29, pp. 383–392 (2004).
- [19] T. B. M. G. D.G Altman, D Machin, *Statistics with confidence* (BMJ Books), 2nd ed. (2000).
- [20] J.Jacquemin, “Mismatch between patient and provider treatment expectation,” *The back Letter* (2002).