Modélisation de la Qualité de Gestes Chirurgicaux Laparosopiques

Arthur Derathé

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **MBS – Modèles, méthodes et algorithmes en biologie, santé et environnement**

Arrêté ministériel : 25 mai 2016

Présentée par

Arthur Derathé

Thèse dirigée par **Sandrine Voros**, Communauté Université Grenoble Alpes,
et codirigée par **Bernard Gibaud**, Université Rennes 1

préparée au sein du **Laboratoire Techniques de l'Ingénierie Médicales et de la Complexité – Informatique, Mathématiques et Applications**
dans **l'École Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement.**

Modélisation de la Qualité de Gestes Chirurgicaux Laparosopiques

Modelling the Quality of Surgical Gestures in Laparoscopy

Thèse soutenue publiquement le **4 Juin 2020**,
devant le jury composé de :

Madame Marie-Christine Jaulent
Directrice de Recherche, INSERM U 1142, Présidente du Jury
Monsieur Germain Forestier
Professeur, Université de Haute-Alsace, Rapporteur
Monsieur Eric Vibert
Professeur des Université – Praticien Hospitalier, Hôpital Paul Brousse, Rapporteur
Monsieur Jean-Luc Faucheron
Professeur des Université – Praticien Hospitalier, CHU Grenoble-Alpes, Examinateur
Madame Sandrine Voros
Chargé de Recherche, Communauté Université Grenoble Alpes, Directeur de thèse
Monsieur Bernard Gibaud
Chargé de Recherche, UMR LTSI U1099 INSERM, Co-directeur de thèse
Monsieur Pierre Jannin
Directeur de Recherche, UMR LTSI U1099 INSERM, Co-encadrant de thèse
Monsieur Alexandre Moreau-Gaudry
Professeur des Université – Praticien Hospitalier, Communauté Université Grenoble Alpes, Co-encadrant de thèse

Remerciements

Tout d'abord un grand merci à Sandrine Voros et Bernard Gibaud pour m'avoir permis de mener cette thèse à son terme, alors même que je ne savais pas du tout dans quoi je mettais les pieds, et que le démarrage de cette thèse fut tout sauf simple. Je les remercie tout particulièrement pour la confiance qu'ils m'ont accordée dans les choix et les orientations que j'ai pu leur proposés, malgré mes doutes. Cela m'a permis d'explorer de nombreuses pistes, pas toujours fructueuses, mais qui m'ont néanmoins permis de me former à de nombreux sujets.

Je remercie également tout particulièrement le chirurgien Fabian Reche, sans qui ma thèse n'aurait simplement pas pu se faire. Son travail et son expertise ont permis de mettre en place l'étude clinique et la récolte des données essentiels en amont de ma thèse, de modéliser et annoter ces données au début de ma thèse (que ce fut long !), et d'évaluer la qualité des résultats de mon travail à la fin de ma thèse. Fabian, ta présence en tant qu'expert en laparoscopie fut sans prix.

Pour le soutien qu'ils nous ont apporté afin de réaliser les protocoles d'étude dans le respect des réglementations (obscures et sans cesse mouvantes), je remercie vivement le CIC-IT Grenoble et les juristes du CHU Grenoble-Alpes.

Pour leur éclairage techniques et leurs conseils avisés sur des points de méthodologie critiques, je remercie la cellule statistique du TIMC, et plus particulièrement Hugo Terrisse et Sophie Lambert-Lacroix.

A mes collègues de l'équipe GMCAO de Grenoble qui m'ont permis de m'épanouir tout au long de cette thèse autour d'un café, d'une discussion à bâton rompu ou d'un shot de montagne salvateur, je dis merci : aux actuels Adrien, Matthias, Clément, Guillaume, Momo, Tamara, Hatem et Maxime, et aux anciens Matthieu, Kazmi, Jerem, Sonia Fanny, Jérôme et Jean-Loup. A mes collègues de l'équipe LTSI de Rennes je dois également un énorme merci pour la place qu'ils m'ont faite dans leur microcosme, l'accueil chaleureux vous m'avez prodigué et l'expertise que vous avez partagée avec moi ont rendu précieux mon séjour à Rennes : Marie, Ehouarn, Noémie, Fabien, Maxime, Arnaud, Thibaut, Marie-Steph, Bouba et John.

Et comme une vie professionnelle fructueuse s'accompagne d'une vie personnelle tout aussi pleine et entière, je remercie infiniment ma famille pour tout ce que je leur dois. Et pour l'EVUG, Lionel et sa bande d'affamés, ces 4 dernières années n'auraient pas eu la même saveur si je n'avais pas pû m'évader en planche à voile et récidiver dès que possible avec toutes ces belles personnes. Merci les coupaings !

Enfin à toi Pauline, je n'ai pas les mots pour te remercier, si ce n'est que nous continuyons ensemble notre aventure avec la même simplicité et la même joie de vivre.

Publications

International Journal [published]

Derathé, Arthur, Fabian Reche, Alexandre Moreau-Gaudry, Pierre Jannin, Bernard Gibaud, and Sandrine Voros. October 2019. “Predicting the Quality of Surgical Exposure Using Spatial and Procedural Features from Laparoscopic Videos.” *International Journal of Computer Assisted Radiology and Surgery*, https://doi.org/10.1007/s11548-019-02072-3.

International Journal [submitted]

Derathé, Arthur, Fabian Reche, Pierre Jannin, Alexandre Moreau-Gaudry, Bernard Gibaud, and Sandrine Voros. April 2020. “Clinical Interpratation of Relationships between Video-based Spatial Features and Quality of Surgical Practice.” *Artificial Intelligence in Medicine*.

International Journal [submitted]

Derathé, Arthur, Fabian Reche, Alexandre Moreau-Gaudry, Pierre Jannin, Bernard Gibaud, and Sandrine Voros. Mai 2019. “A Dataset for Surgical Context Recognition and Surgical Practice Assessment in Laparoscopic Surgery.” *International Journal of Medical Informatics*.

National Conference [published]

Derathé, Arthur, Fabian Reche, Bernard Gibaud, and Sandrine Voros. 2017. “Extraction of Data for Surgical Assessment: Projection and Preliminary Results.” *Surgetica*.

National Conference [published]

Derathé, Arthur, Sandrine Voros, Fabian Reche, Pierre Jannin, Alexandre Moreau-Gaudry, and Bernard Gibaud. 2019. “Analyzing the Practice of Expertsurgeons Based on Video Spatialfeatures.” *Surgetica*.

Résumé

La chirurgie laparoscopique est une pratique de plus en plus communément utilisée dans différentes spécialités chirurgicales, du fait des grands avantages pour le patient en termes de complications et de temps d'hospitalisation. En revanche, cette pratique est très différente de la chirurgie dite « ouverte », et présente ses propres difficultés, notamment dans la manipulation des instruments chirurgicaux, et la maîtrise de l'espace opératoire. Une meilleure compréhension du geste chirurgical en laparoscopie permettrait d'améliorer les outils utilisés pour la formation des jeunes chirurgiens.

L'objectif de ce travail était de développer et valider une méthode visant à expliquer certains aspects clés de la pratique du geste chirurgical en termes cliniques, à partir d'une approche algorithmique. La compréhension du contexte clinique de cette thèse étant essentielle, un important travail d'explicitation et de formalisation des connaissances du chirurgien a été effectué. La deuxième partie de ce travail a consisté à développer une méthode algorithmique visant à prédire la qualité du geste chirurgical et le chirurgien pratiquant. Enfin à travers l'analyse de données décrivant la qualité et la pratique du geste chirurgical, nous avons étudié et validé la pertinence clinique de nouveaux éléments de connaissances cliniques.

Nous avons travaillé sur une cohorte de 30 patients opérés par gastrectomie longitudinale au sein du département de chirurgie digestive du CHU de Grenoble. Grâce à une réflexion commune avec notre partenaire chirurgien, nous avons pu formaliser les notions importantes de cette procédure chirurgicale. Pour chacune des chirurgies de la cohorte, nous avons effectué trois annotations distinctes : une annotation de la procédure et des actions des mains du chirurgien, une évaluation de la qualité d'exposition de la scène chirurgicale à chaque geste de dissection effectué par le chirurgien, et enfin la segmentation complète de l'image associée à chacun des gestes de dissection évalués. L'annotation de la procédure et la segmentation ont rendu possible l'extraction de métriques caractéristiques du geste et de la scène chirurgicale.

Ensuite, nous avons développé un algorithme dont l'objectif était la prédiction de la qualité d'exposition à partir des métriques. Nous avons également développé un environnement dédié à l'optimisation des hyper-paramètres de notre algorithme pour maximiser les performances en prédiction. L'intérêt de cet environnement était notamment de gérer les spécificités de notre jeu de données.

Dans un troisième temps, nous avons mis en place une méthode permettant de confronter l'analyse algorithmique quantitative de nos données à l'expertise clinique des chirurgiens ayant effectué les chirurgies. Pour ce faire, nous avons d'abord extrait les variables les plus importantes pour notre tâche de prédiction. Puis nous avons traduit l'information portée par ces variables sous forme d'énoncés présentant une signification clinique. Enfin nous avons extrait des échantillons vidéos représentatifs de chacun de ces énoncés. A partir de ces énoncés accompagnés de leurs échantillons vidéos, nous avons pu construire un questionnaire de validation, et le présenter à nos partenaires chirurgiens. Nous avons ainsi mené une validation clinique visant à recueillir leur avis quant à la pertinence clinique de notre approche.

Nous avons donc proposé une méthode d'analyse quantitative explicitant le lien entre

des observations visuelles et temporelles et des critères cliniques relatifs à des chirurgies laparoscopiques. Une meilleure compréhension de ces liens permettrait, à terme, de proposer des systèmes d'aide à la formation des chirurgiens sur cette pratique complexe.

Abstract

The practice of laparoscopy is getting more and more common in the daily routine of various surgical specialties. This technique offers great advantages for the patient, regarding potential complications as well as the ambulatory time. It is however a very challenging practice for the surgeon, with its own problematics and difficulties related to instruments manipulation and surgical space management for example. A better understanding of the surgical gestures in laparoscopy would bring new tools for the formation of residents and young surgeons.

In this work, our objective is to develop and validate a data science approach to explain key aspects of the practice of surgical gestures in a clinically meaningful way. An important work of knowledge elicitation and modelling was performed together with an expert surgeon to constitute a satisfying understanding of the clinical context. We then developed an algorithmic methodology to predict the quality of surgical gestures as well as the surgeon practicing. Finally, throughout the data analysis of the quality and profile of surgical gestures, we studied and validated clinically some new elements of clinical knowledge.

We worked on a cohort of 30 patients operated by Sleeve Gastrectomy in the department of digestive surgery of the Grenoble CHU. Sleeve Gastrectomy is a surgical treatment against morbid obesity. With the help of our surgeon partner, we formalized the important aspects of this technique. For each surgery of the cohort, we performed three video annotations : a procedural annotation of the surgeon hands' actions, an evaluation of the quality of exposure in the surgical scene for each activity of dissection performed by the surgeon, and a pixel-wise segmentation of each image associated with an evaluated dissection activity.

Based on the procedural annotation and the segmentation, we computed a set of metrics characterizing the surgeon's gestures and surgical scene. Then we developed an algorithm to predict the quality of exposure with these metrics as input data. We also developed a cross validation environment dedicated to the optimization of the hyper parameters of our algorithm, in order to maximize its prediction performances. This cross validation environment helped us to deal with the specificities of our dataset and its related bias.

Our third and final study focused on the conception of a methodology to confront our surgical data science approach with the clinical expertise of surgeons. To this end, we first extracted the most important variables for our prediction task, and we translated information resulting from these important variables into utterances with clinical meanings. We illustrated each of these statements with video clips extracted from videos in our dataset. Based on these statements and their associated clips, we constructed a validation survey and submitted it to surgeons. Results of these submissions expressing the surgeons' opinions allowed us to clinically validate our approach, and to assess its relevance.

To conclude, we proposed a methodology of surgical data analysis to explicit the relationships between visual and procedural information and some clinical criteria for some laparoscopic surgeries. A better understanding of these relationships would allow, in the long run, to propose some supporting tools for the formation of surgeons.

Table des matières

Table des figures

Liste des tableaux

1

Contexte clinique : l'évaluation et la formation en chirurgie

Sommaire

1.1 | Contexte générale de la chirurgie

1.1.1 | Changement de paradigme dans la pratique de la chirurgie

La chirurgie est une pratique en constante évolution, et les outils à disposition du chirurgien ne cessent de changer. Hier, le chirurgien travaillait avec son assistant et une ou deux infirmières de bloc. Les informations sur le patient et les examens complémentaires étant peu nombreux, le chirurgien s'en tenait aux quelques scénarios opératoires connus pour limiter les risques. Mais, du fait de ces connaissances réduites sur le patient, la possibilité de rencontrer une situation chirurgicale inconnue était grande. Le chirurgien ne pouvait alors se fier qu'à son expérience et son instinct pour gérer les risques et résoudre les difficultés. De plus, le chirurgien n'avait que quelques outils à sa disposition. Bref, il devait faire beaucoup avec très peu.

Aujourd'hui, le chirurgien travaille en équipe avec ses assistants, l'anesthésiste, les infirmières de bloc, et lors de certaines opérations complexes d'autres spécialistes prennent également part à l'opération. La communication est devenue centrale. Avant l'opération, le chirurgien a désormais accès à de nombreuses informations sur son patient grâce aux examens biologiques, aux différentes modalités d'imagerie médicale, et au suivi médical et chirurgical du patient en général. Il peut donc mieux planifier et anticiper le déroulement de l'opération. Pendant l'opération, les équipements aussi ont évolué. Les types de chirurgie eux-mêmes se sont multipliés : la chirurgie ouverte partage à présent le bloc opératoire avec la cœlioscopie, l'endoscopie, la chirurgie robotisée et la microchirurgie.

A mesure que les traitements chirurgicaux se perfectionnent et permettent de mieux soigner les patients, la charge de travail au bloc opératoire est distribuée entre de plus en plus de professionnels et de matériels. La coordination de toutes les composantes du processus chirurgical est à la charge du chirurgien qui doit garder une vision globale et anticiper les complications potentielles pour que la chirurgie se déroule au mieux. Aussi, le métier du chirurgien s'est transformé et complexifié. En plus de son savoir médical et de ses compétences de manipulation peropératoire, il a la charge de manager son équipe pendant l'opération, de communiquer, de gérer la surcharge cognitive, et de prioriser ses choix. C'est un environnement à risque où les erreurs potentielles prennent des formes très variées.

1.1.2 | Le cas particulier de la laparoscopie

La chirurgie endoscopique est une pratique chirurgicale de plus en plus généralisée. En endoscopie, on cherche à insérer les instruments chirurgicaux dans le corps du patient en passant au maximum par les voies naturelles (orale, anale, veineuse, ...) ou par de petites incisions. Le but est d'éviter de pratiquer de grandes incisions dans le corps du patient comme la laparotomie. Cette pratique s'impose dans de nombreuses spécialités chirurgicales du fait des avantages importants pour le patient comparé à la laparotomie (chirurgie ouverte) : le temps de cicatrisation, le temps d'hospitalisation, le nombre de complications post-opératoires et le taux de mortalité ont significativement diminué (cancer Laparoscopic or Open Resection Study Group, 2005; Faiz et al., 2009). Dans ces travaux, nous nous intéressons plus précisément au cas de la laparoscopie qui est la forme de chirurgie endo-

scopique dans laquelle on accède et on opère dans la cavité abdominale du patient sans ouvrir la cavité abdominale, et en ne pratiquant que les quelques incisions nécessaires pour insérer l'endoscope et les outils dans la cavité abdominale.

Si cette approche présente de grands avantages pour le patient, sa pratique présente ses propres difficultés pour le chirurgien. Avant tout, le chirurgien doit opérer avec une visibilité fortement réduite de l'environnement opératoire ainsi que de ses outils. Il a également moins de liberté pour manipuler ses outils, et comme de nombreuses zones ne sont pas couvertes par la vision de l'endoscope, il peut potentiellement léser des tissus. Le retour haptique des instruments et donc sa sensibilité aux tissus manipulés est transformé et diminué. On notera également que la position du chirurgien quand il pratique est contrainte et difficile à tenir, surtout sur de longues opérations.

Toutes ces difficultés propres à la laparoscopie ont un impact sur l'apprentissage de la pratique chirurgicale : la courbe d'apprentissage des chirurgiens en formation à la laparoscopie est plus longue que celle des internes en laparotomie (Hedrick et al., 2009). En effet, les internes en laparoscopie passent plus longtemps à acquérir et maîtriser les compétences de base de cette pratique, et lorsqu'ils arrivent en fin d'internat, ils n'ont pas encore acquis une maîtrise complète de leur pratique. Ainsi, les jeunes chirurgiens doivent continuer à se former afin d'atteindre une réelle autonomie.

Chez les chirurgiens experts, le risque d'erreur existe encore. Ce risque est faible, mais des évènements indésirables EI sont encore observés. L'étude de Francis et al. (2018) montre qu'en grande majorité, ces EIs n'ont pas d'impact (60.1%) ou un impact faible (37.1%) sur le déroulement opératoire et pas d'impact sur les suites post-opératoires du patient. Mais, si la pratique de l'expert peut encore être perfectionnée, la marge de progression des chirurgiens non-experts est plus conséquente et leur pratique en routine clinique est sujette à plus de risques.

Que ce soit en formation ou en routine clinique, la pratique de la chirurgie laparoscopique est très complexe et il est raisonnable de se questionner sur les outils existants pour maîtriser cette complexité et les risques associés.

1.2 | L'erreur humaine dans un environnement complexe

A l'instar du cockpit dans l'aviation civile, ou de la centrale nucléaire, la notion d'erreur humaine dans un environnement aussi complexe que le bloc opératoire est une notion très ambigüe, à manipuler avec de grandes précautions. Quand on est confronté à un incident impliquant des professionnels se trouvant « en première ligne » (Reason, 1990), Cook and Woods (1994) insistent sur l'importance de considérer et d'analyser le système dans son ensemble. Tel le chirurgien, le pilote d'aviation, l'astronaute, … le spécialiste en première ligne fait face, au quotidien, à des imprévus et des hasards liés à la très grande complexité du processus auquel il est confronté.

Face à une erreur qui, à première vue, est le fait du spécialiste en première ligne, les causes doivent en réalité être envisagées de manière beaucoup plus globales. Dans l'aviation civile, on utilise la classification « Human Factors Analysis and Classification » (HFACS) dans laquelle le processus menant à l'erreur se décompose comme suit :

— **L'acte risqué.** Ce peut être une erreur (par nature non-intentionnelle) ou une violation (intentionnelle).
— **Les préconditions à l'acte risqué.** Ces préconditions peuvent être liées à l'environnement physique et technologique, à la condition physique ou mentale de la personne, ou à la communication entre les personnes présentes.
— **La supervision à risque.** On regroupe ici les causes liées à un mauvais management des personnes ou une mauvaise planification, intentionnelles ou non-intentionnelles.
— **L'influence organisationnelle.** On regroupe ici l'influence de la gestion des ressources humaines, matérielles et financières, l'influence de l'ambiance globale au sein de l'organisation, et enfin l'influence des procédures, règlements et lois.

Cette classification s'applique parfaitement à l'environnement du bloc opératoire, vu au sein de l'hôpital, du système de santé et de l'appareil législatif d'un état. De surcroît, le spécialiste étant confronté quotidiennement à cette « première ligne », il développe des automatismes afin de prendre la meilleure décision face à la complexité des situations auxquelles il fait face. Mais il est loin d'être évident, voire impossible pour lui d'expliciter le raisonnement complet et exact qui l'a mené à sa décision du fait même de la complexité de son environnement, du stress et de l'urgence qui accompagnent ces incidents.

Dans le cadre d'une politique fort compréhensible de réduction du nombre d'accidents aériens, une méthodologie de gestion des risques liés aux personnes a été développée dans l'aviation civile à partir des années 80. Initialement pensée pour l'équipe du cockpit, la « Cockpit Resource Management » (CRM) a ensuite été étendue à l'ensemble du personnel impliqué dans la gestion de l'avion : la « Crew Resource Management ». Les trois piliers de cette approche de détection et de gestion des situations à risque sont la *communication*, la *définition des priorités*, et la *gestion effective de la charge de travail*. Cela se traduit en pratique par des protocoles de communication, des checklists, des procédures de sécurité, et des rapports d'activité. Un autre outil mis en place dans l'aviation civile pour réduire les risques d'erreurs est la formation continue et le simulateur de vol. Pendant leur formation les membres de l'équipe de de vol apprennent sur des simulateurs de vol bien sûr, mais pendant leur carrière aussi : le pilote de ligne est en formation continu et est amené à pratiquer dans des environnements simuler de façon régulière. Sa pratique est ainsi évaluée régulièrement.

Dans l'environnement chirurgical, ces outils sont de plus en plus utilisés. La CRM s'inscrit dans une vision à long terme afin d'accumuler une expérience commune et de permettre aux mécanismes de communication, checklists et rapports de rentrer dans les habitudes de travail. Cependant, l'environnement chirurgical étant moins connu et moins maîtrisé que le cockpit d'avion, la mise en place de ces outils formalisateurs et simplificateurs n'est pas évidente. Nous en parlerons plus en détail en section 1.4, mais le développement de simulateurs chirurgicaux présentent aujourd'hui encore de grandes limitations. La grande variabilité des procédures chirurgicales complexifie d'autant leur développement, et rend l'explicitation des mécanismes décisionnels au sein du bloc opératoire très difficile. Il existe également des blocages socio-culturels liés au milieu hospitalier, et à plus forte raison au milieu chirurgical, qui ralentissent la mise en place de ces outils. On notera que le corps des anesthésistes-réanimateurs a déjà très bien intégré la logique CRM, et ce, malgré la com-

plexité de leur spécialité.

Par ailleurs, quand un pilote de ligne n'a pas pratiqué depuis une longue période, il doit obligatoirement suivre un stage de remise à niveau afin de retrouver ses réflexes et son aisance. Ce principe commence à être appliqué chez les anesthésistes-réanimateurs mais absolument pas chez les chirurgiens. En tant que patient, il me semblerait pourtant rassurant de savoir qu'après une longue période d'inactivité mon chirurgien ne recommence pas à pratiquer directement dans une chirurgie réelle mais reprend la main dans un cadre contrôlé.

La mise en place de ces mécanismes de gestion des risques au bloc opératoire et d'amélioration de la formation à la chirurgie nécessite donc une meilleure connaissance et compréhension de l'environnement chirurgical, mais également d'axer la formation des chirurgiens autour de ces nouveaux outils. Face à ces risques et erreurs potentielles, le comportement du chirurgien « en première ligne » dépend grandement de son cursus d'apprentissage de la pratique chirurcale. La qualité de la formation suivie par le chirurgien est garante de la qualité de sa pratique future, et donc d'une réduction des risques pour le patient.

1.3 | Quelles transformations pour la formation du chirurgien ?

Au fur et à mesure que les connaissances chirurgicales s'affinent, la formation des chirurgiens évolue. L'entrée des notions de risque et d'erreur au bloc opératoire a transformé la façon dont les chirurgiens sont formés. Le modèle classique de formation par compagnonnage des internes en chirurgie sera d'abord introduit, avec ses forces et faiblesses. Puis nous décrirons deux exemples d'évolutions du milieu socio-professionnel chirurgical et plus précisément de l'environnement de formation des internes.

1.3.1 | Forces et faiblesses du tutorat : la relation tuteur-élève

Le modèle de formation du chirurgien dit d'Halsted, est basé sur le couple tuteur-élève. Le tuteur est un chirurgien confirmé qui prend en charge la formation d'un étudiant en médecine, son élève. Ce modèle a l'avantage que l'élève peut observer son tuteur, s'inspirer de sa pratique, et le moment venu, reproduire son comportement dans ses prises de décision, sa prise en charge du patient, et sa pratique lors de l'opération chirurgicale.

Entezami et al. (2012) ont mené un état de l'art des articles traitant de ce modèle de tutorat dans le cadre de la chirurgie. On notera avant tout que la plupart de ces articles traitent de l'expérience personnelle de professionnels, ou de prise de position, il n'y a que peu d'études proposant une approche scientifique rigoureuse visant à caractériser objectivement le tutorat. Cet état de l'art présente donc moins des résultats scientifiques que l'aspect consensuel ou non d'opinions sur le tutorat.

Tout d'abord Entezami et al. notent la faible formalisation du tutorat : tout au plus, certains articles décrivent un programme de mise en place du tutorat, et définissent les modalités d'attribution des couples tuteur-élève. Dans d'autres études cette formalisation n'est pas mentionnée du tout. Ainsi la relation tuteur-élève est globalement peu formalisée, et est

laissée essentiellement à l'appréciation et à l'expérience du tuteur. Les qualités attendues du tuteur sont en revanche bien discutées, il en ressort que le bon tuteur doit :

— être un modèle pour son élève sur le plan professionnel,
— s'investir en terme de temps et de travail avec l'élève,
— être compréhensif/bon/bienveillant et critique/exigeant/évaluateur,
— être un guide au bloc opératoire,
— pousser l'élève à se surpasser.

Les articles s'accordent également sur certaines difficultés rencontrées dans le cadre du tutorat. Le tuteur doit gérer une forte contrainte temporelle liée à son emploi du temps surchargé, et doit malgré tout réussir à consacrer suffisamment de temps à son élève. Il y a un clair manque de chirurgiens qualifiés pour le rôle de tuteur, ainsi qu'un manque de tutrices. De fait, la relation tuteur-élève rencontre des difficultés liées aux différences de sexe, mais également de culture et de génération. Concernant ce dernier point, il y a une différence claire entre la nouvelle génération qui assume la primauté de sa vie personnelle et de famille sur son travail, tandis que l'ancienne génération fait passer la perfection professionnelle avant son bien-être personnel.

Enfin, les trois quarts des articles analysaient les avantages de la relation tuteur-élève, et aucun ne considérait le tutorat comme non-nécessaire ou comme ayant un impact négatif sur l'élève. Ainsi, le tutorat est, et restera d'actualité : au regard de la complexité du travail du chirurgien, il est nécessaire que l'étudiant soit suivi par un expert. Des évolutions sont néanmoins à envisager, et des transformations sont déjà en cours.

1.3.2 | Une contrainte à la formation des internes : le nombre d'heures travaillées

En 2003 a été promulgué aux Etats-Unis une loi limitant à 80 heures le nombre d'heures travaillées par semaine des internes. Cette loi répond à un ensemble de problématiques liées au bien-être et à la qualité de vie des internes, mais aussi à la sécurité des patients. MacGregor and Sticca (2010) ont mené une étude dans laquelle ils ont envoyé un questionnaire aux internes en chirurgie générale, afin de recueillir leur avis sur l'impact que cette réglementation a sur leur travail quotidien. Les résultats montrent que dans une proportion non négligeable :

— Les internes sous-évaluent le nombre d'heures qu'ils effectuent.
— Leur hiérarchie ordonne aux internes de sous-évaluer leur nombre d'heures.
— Les internes pensent que leur co-internes font de même.
— Les jeunes internes sont plus en faveur d'une augmentation du nombre d'heures travaillées que les internes plus âgés.
— Les internes ont majoritairement sous-évalué leur nombre d'heures pour mieux prendre en charge un patient.
— Les internes pensent que cette restriction du nombre d'heures travaillées a un impact négatif sur leur formation de chirurgien.

Cette réforme a été engagée dans le but de réduire la charge de travail des internes, et de réduire les risques dans la prise en charge des patients. Les internes en chirurgie,

qui font partie des principaux bénéficiaires, présentent à travers leur avis des conséquences étonnantes et plutôt négatives de cette réforme. Ces résultats nécessitent donc d'être remis dans le contexte hospitalier pour être mieux compris.

Dans leur travail quotidien, les internes ont avant tout comme priorité de soigner leurs patients, puis dans une moindre mesure de se former. Ainsi, cette réduction du temps de travail impacte en premier lieu la formation, moins essentielle que la prise en charge des patients. Cependant, dans une optique à plus long terme, le temps passé à se former assure une prise en charge plus sûre des patients, et donc une réduction des risques et erreurs. On comprend donc que les internes ne soient pas favorables à cette nouvelle réforme qui dégrade leurs conditions de formation ainsi que leurs conditions salariales, alors même que, dans les faits, leur charge de travail reste la même.

Cette constatation soulève des questions, notamment sur l'absence de solutions proposées aux internes pour se former dans de bonnes conditions, tout en respectant la législation. Des éléments de réponse au niveau du management et de l'organisation des services de chirurgie pourraient permettre de mieux respecter cette restriction du nombre d'heures tout en permettant aux internes de bien se former. Mais l'interne étant une main d'œuvre « bon marché » permettant de faire face au flux de patients et de faire fonctionner l'hôpital, le remplacer ne pourrait aller sans un soutien financier et une augmentation de la masse salariale dans les hôpitaux. Les instances décisionnelles et régulatrices sont donc également impliquées dans la gestion de cette situation.

Finalement, on observe ici que, malgré son importance pour la sécurité des patients, la formation des internes en chirurgie est mise en difficulté. Comme le bon fonctionnement de l'hôpital repose lourdement sur le travail des internes, leur formation est reléguée au second plan. Une évolution du système de formation des internes pourrait permettre d'améliorer cette situation.

1.3.3 | Une évolution de la formation des internes : la réforme du 3ème cycle en France

Pour répondre à de telles difficultés, la réforme du 3ème cycle de médecine a été mise en place au niveau national en France en 2016. Cette réforme est le résultat d'un long travail de réorganisation et de refonte de l'internat et du post-internat. Dans le cadre de cette réforme, les spécialités médicales sont distinguées en Diplôme d'Etudes Spécialisées (DES) que les étudiants vont choisir suite à leur classement au concours des Epreuves Classantes Nationales (ECN) de 6^{me} année de l'externat. Les spécialités sont désormais séparées en filières, et un étudiant s'engage directement dans une de ces filières suite au concours alors qu'auparavant il suivait un tronc commun pendant 2 ou 3 ans avant de se spécialiser.

Cette réforme définit une formalisation du parcours du troisième cycle en trois phases :

— **La phase socle** a pour vocation d'introduire l'étudiant à sa spécialité, et d'acquérir les compétences de base essentielles à la profession de médecin (communication avec le patient et avec les autres professionnels de santé, éthique, ...). Elle dure un an et comporte un stage dans la spécialité en question permettant à l'étudiant d'être confronté à la réalité de son futur travail. Il possède de plus le droit au remord, et

peut donc changer d'avis et de spécialité.

— **La phase d'approfondissement** dans laquelle l'étudiant approfondit ses compétences au cours de stages, soit imposés, soit librement choisis pour satisfaire ses aspirations. Elle dure deux à trois ans.

— **La phase de mise en situation** dans laquelle l'étudiant a désormais acquis une certaine autonomie, et exerce comme un médecin. Il n'est cependant pas encore responsable, et peut à tout moment faire appel à un médecin en cas de doute dans sa pratique. Cette phase d'un ou deux ans se conclut avec la rédaction de la thèse clinique.

De plus, chaque filière de spécialité médicale doit définir un portfolio, c'est-à-dire un cursus avec des connaissances à acquérir et des compétences à valider en fonction de la phase de l'internat. On assiste donc ici à une réelle formalisation de la courbe d'apprentissage des internes. Cette formalisation du parcours de l'internat engagée au niveau national vient impacter la formation par tutorat et engage une réflexion dans chaque spécialité pour standardiser les connaissances enseignées par le tuteur à son élève.

En lien avec cette standardisation des connaissances, la réforme introduit un contrôle continu et une évaluation des compétences que l'interne doit acquérir tout au long de son cursus. Cette dimension d'évaluation et de validation du niveau de l'interne était inexistante auparavant. En chirurgie, la validation des compétences de l'interne était laissée à l'appréciation de son encadrement. L'évaluation formelle des compétences a donc été introduite en tant qu'outil pédagogique fort dans l'internat et implique la recherche de méthodes communes de suivi et d'évaluation des internes. Dans les filières chirurgicales, cette évaluation des compétences concerne tant les aspects relationnels, que les connaissances médicales et la pratique chirurgicale elle-même. Afin de faciliter cette évaluation des compétences, il est nécessaire de développer des outils de soutien à l'évaluation. Dans la prochaine section, nous présenterons une certaine forme d'outils d'évaluation : les simulateurs.

1.4 | La simulation dans la formation chirurgicale

Pour faire face au manque de temps dédié à la formation des internes, à la faible disponibilité des blocs opératoires, et aux contraintes éthiques protégeant les patients, la simulation est un outil d'une grande aide pour la formation des internes en chirurgie. La réforme du 3^{me} cycle en France a transformé le modèle de formation des internes, et établi un cadre favorable à l'utilisation de la simulation comme support pour l'évaluation des compétences chirurgicales. Palter and Grantcharov (2010) présentent un état de l'art sur la simulation comme outil dédié à la formation des internes en chirurgie.

Il y a deux intérêts majeurs à former l'interne dans le cadre de la simulation avant de le faire pratiquer au bloc opératoire. D'une part, la simulation est un environnement sûr dans lequel l'interne peut s'entraîner, avoir moins de contraintes de temps d'accès, faire des erreurs, et bénéficier de retours sur sa pratique. D'autre part, la simulation est un pré-entraînement : l'interne perfectionne son geste et sa compréhension de la chirurgie dans des conditions simplifiées sur simulateur. Il développe ainsi des compétences de base en prévision d'une pratique postérieure en conditions réelles dans la suite de sa formation. Je

n'ai parlé que d'un entraînement aux compétences techniques chirurgicales, mais des outils de simulation existent aussi pour la formation aux compétences non-techniques que sont la communication, le leadership et le travail en équipe.

En plus des modèles animaux, des cadavres et du patient comme supports d'entraînement classiques, la simulation moderne réunit de nouvelles modalités. Palter et Grantcharov distinguent différents simulateurs :

— Les simulateurs synthétiques qui reproduisent une tâche chirurgicale basique sur un support physique,
— Les tours d'entraînement généralement utilisées pour l'entraînement à la laparoscopie,
— Les simulateurs en réalité virtuelle,
— les salles d'opération simulées (réelles ou virtuelles).

Dans le cadre de l'apprentissage des compétences techniques, Palter and Grantcharov (2010) constatent que le simulateur est un bon moyen d'acquérir les compétences de base, et que le transfert vers le bloc opératoire est effectif. En revanche, il n'y a pas de clair avantage entre l'utilisation d'un modèle synthétique ou d'un simulateur en réalité virtuelle pour l'apprentissage, si ce n'est le coût plus important des simulateur en réalité virtuelle. Pour l'apprentissage des compétences non-techniques, les études présentées montrent une grande satisfaction des équipes de bloc mises en situation. Mais dans l'ensemble, les études analysées ne vont pas assez loin dans la validation de leurs outils. Et les auteurs fustigent l'absence d'une vision plus globale du parcours de formation de l'étudiant en chirurgie, dans lequel viendraient s'insérer des outils de simulation dédiés. Ils constatent également le manque de coordination des recherches qui empêche de bien déterminer la place et la pertinence des simulateurs tout au long de la courbe d'apprentissage du chirurgien.

La réforme du 3^{me} cycle engagée depuis 2016 en France laisse envisager une place de choix pour ces simulateurs dans le cadre de l'évaluation des compétences chirurgicales des internes. Plus généralement, l'ensemble des outils proposant une aide à la formation des chirurgiens ont désormais la possibilité d'être intégrés dans le cursus des internes en chirurgie et d'être utilisés dans le but d'améliorer le processus de formation à la chirurgie.

Dans une vision à plus long terme, ces outils pourraient également être utilisés pour l'évaluation du chirurgien dans sa pratique quotidienne. Une telle évaluation soulève néanmoins des questions très délicates, notamment en ce qui concerne la légitimité des institutions et personnes habilitées à évaluer des chirurgiens. On observe une évolution lente mais réelle des mentalités en faveur d'une évaluation continue du chirurgien tout au long de sa carrière, dans une dynamique de réduction des risques de complications per- et post-opératoires. Mais l'évaluation du chirurgien en exercice, ne serait-ce que par ses pairs, est aujourd'hui encore difficilement acceptée et envisagée par la communauté des chirurgiens. Dès lors, sans même parler des contraintes techniques actuelles, l'acceptation d'outils automatique développés par des non-chirurgiens, non-experts pour l'évaluation des chirurgiens n'est actuellement absolument pas à l'ordre du jour.

Ainsi, la suite de notre réflexion se focalisera sur l'état des connaissances et des techniques récentes portant sur l'évaluation et l'analyse de la pratique chirurgicale.

En résumé

Le chirurgien travaille au quotidien dans un environnement complexe propice à l'apparition d'évènements risqués et d'incidents. La maîtrise de cet environnement complexe passe, pour le chirurgien, par la mise en place de mécanismes de contrôle et de gestion des risques dans sa pratique quotidienne d'une part, mais également par une formation extrêmement exigeante. L'interne en chirurgie est traditionnellement accompagné pendant sa formation par un tuteur qui lui sert d'exemple et de guide, mais ce modèle de formation évolue. Dernièrement en France, cette formation a été formalisée à travers la réforme du troisième cycle qui donne une place prépondérante au contrôle continu et à l'évaluation de l'interne. Dans ce cadre, l'évaluation des compétences chirurgicales peut donner lieu à l'utilisation d'outils prévus à cet effet, tels que la simulation pour l'entraînement à la chirurgie. On se trouve donc aujourd'hui dans un environnement très favorable au développement d'outils d'aide à l'évaluation et à l'analyse de la pratique chirurgicale.

2

Etat de l'art sur l'analyse et l'évaluation de la pratique chirurgicale

Préambule

Comme nous l'avons vu dans le chapitre précédent, la pratique chirurgicale est un phénomène complexe aux multiples facettes, notamment la pratique de la laparoscopie. La qualité de cette pratique, la question de l'erreur en chirurgie, et le rapport du chirurgien à cette erreur sont des problématiques éminemment actuelles. Ici, mon champ d'étude ne traitera que d'un aspect précis et limité de la pratique chirurgicale. Tout d'abord je ne m'intéresserai pas aux compétences non techniques du chirurgien, et traiterai uniquement des compétences techniques du chirurgien au bloc opératoire. Cela englobe la maîtrise des outils et des équipements du bloc, la connaissance de la procédure chirurgicale, et de manière générale, tout ce qui, indépendamment des autres membres de l'équipe de soin, limitera pour le chirurgien les risques de complications per-opératoires. Dans l'état de l'art qui suit je présenterai des études traitant de la chirurgie ouverte, de la laparoscopie et de la chirurgie robotisée.

Je présenterai tout d'abord la méthode d'état de l'art qui m'a permis de constituer le corpus d'études en section 2.1. Je proposerai ensuite un prisme de lecture générale de ces études en section 2.2. Puis je m'intéresserai en section 2.3 aux grandes catégories d'études qui m'ont semblées le mieux rendre compte des tendances actuelles du domaine, ainsi qu'aux conclusions et aux connaissances qu'elles nous apportent sur la pratique chirurgicale. Je conclurai en positionnant mon travail de thèse et mes objectifs en section 2.4.

Sommaire

2.1 | Méthodologie d'état de l'art

Pour cet état de l'art, nous avons mené une recherche systématique à partir des moteurs de recherche Google Scholar et PubMed en utilisant la combinaison des mots clés suivants :

> *("surgical" OR "surgeon") AND ("practice" OR "skill") AND "automatic" AND ("evaluation" OR "assessment" OR "estimation")*

En plus des références obtenues ainsi, nous avons recueilli les références citées dans ces articles, et notamment dans les articles d'état de l'art. Les articles sélectionnés devaient être soumis en anglais entre 2010 et Juillet 2019 dans des journaux ou des conférences. La recherche initiale nous a donné 95 articles via Google Scholar et 100 articles via PubMed. 7 de ces articles étaient des états de l'art à partir desquels nous avons tiré 372 références d'articles. Après suppression des doublons, des articles non-accessibles, des articles d'état de l'art, et des articles rédigés dans une autre langue que l'anglais, nous avons réuni 493 articles.

Par une lecture des titres d'articles nous avons ensuite regroupé les articles dans les catégories suivantes :

— Analyse de compétences techniques : 79 articles
— Analyse de compétences non-techniques : 42 articles
— Test d'un banc d'entrainement : 39 articles
— Détection d'outils : 35 articles
— Reconnaissance du contexte : 80 articles
— Etude du résultat patient : 8 articles
— Hors-sujet : 46 articles

Nous n'avons conservé que les 79 articles traitant de l'« Analyse de compétences techniques ». Dans la suite de cette étude, nous avons tout d'abord caractérisé des aspects généraux et communs à l'ensemble de ces articles en section 2.2. Nous nous sommes ensuite focalisés sur différentes catégories d'articles définies en considérant le type de données utilisé pour étudier la pratique chirurgicale en section 2.3. Dans cette section nous nous sommes attachés à souligner les conclusions à portée clinique dans ces articles.

2.2 | Quelles formes prennent les études portant sur l'analyse et l'évaluation de la pratique chirurgicale ?

Les études auxquelles je m'intéresse traitent toutes de la pratique chirurgicale, de la façon dont on peut mieux la prédire ou mieux la comprendre. Mais elles diffèrent toutes par leurs approches méthodologiques. Je présenterai dans ce qui suit différents aspects méthodologiques permettant de mieux saisir cette diversité.

2.2.1 | Le lien entre l'objectif global de l'étude et la pratique chirurgicale

Positionnement	Id	Description
Validation	1	L'étude s'intéresse uniquement à la validation d'un outil, dans un cadre où il n'a jamais été validé
	2	L'étude présente la validation d'un outil, dans un cadre où l'outil a déjà été présenté
	3	L'étude présente un outil d'évaluation, et traite également d'un aspect de validation
Evaluation	4	L'étude traite uniquement d'un outil permettant l'évaluation de la pratique chirurgicale
	5	L'étude présente un outil d'évaluation, apportant également des éléments d'informations sur la pratique chirurgicale
	6	L'étude présente une analyse de la pratique chirurgicale pour laquelle un outil d'évaluation a été utilisé
Analyse	7	L'étude a pour objectif d'analyser la pratique chirurgicale, et n'utilise aucun outil de prédiction dans cet objectif

Tableau 2.1 – Positionnement de l'objectif de l'étude vis-à-vis de la notion de pratique chirurgicale

Dans mon analyse, la notion centrale est la pratique chirurgicale. Dans le corpus d'études analysées, je questionne la place de cette notion dans les approches méthodologiques proposées. Autrement dit, je m'intéresse dans chacune de ces études à la façon dont l'objectif gravite autour de la notion de pratique chirurgicale. Je discerne ainsi trois catégories d'études :

— *Evaluation :*

La catégorie la plus commune regroupe les études dont l'objectif est l'évaluation de la pratique chirurgicale. Cette évaluation peut être une prédiction automatisée ou manuelle. Dans ces études, la pratique chirurgicale est décrite par un indicateur, tel que le niveau d'expérience du chirurgien, ou un score dédié à l'évaluation de sa performance.

— *Validation :*

Les études de la deuxième catégorie ont pour objectif d'évaluer la validité de construction et la fiabilité d'un outil d'évaluation de la pratique chirurgicale. La validité de construction concerne la capacité de l'outil à réellement mesurer ce pour quoi il a été créé. La fiabilité concerne la capacité de l'outil à mesurer avec précision et robustesse ce pour quoi il a été développé.

— *Analyse :*

Les études de la troisième catégorie analysent plus en profondeur le fonctionnement et le comportement d'un outil d'évaluation de la pratique chirurgicale. L'objectif étant, à travers cette analyse, d'extraire de nouveaux éléments de connaissance clinique sur la pratique chirurgicale.

Les objectifs de validation et d'analyse d'un outil d'évaluation s'articulent autour de l'objectif d'évaluation. En effet, ce n'est qu'une fois que l'outil d'évaluation a été développé, qu'on peut envisager de le valider et/ou d'en tirer des analyses cliniques. Ce sont donc deux développements distincts qui se basent sur un outil d'évaluation pré-existant.

Une lecture plus fine des études rend cette catégorisation plus nuancée. Pour une étude donnée, les résultats portant principalement sur l'évaluation de la pratique chirurgicale peuvent également comprendre des éléments de compréhension de ladite pratique. A l'inverse, une étude dont l'objectif principal était d'apporter de nouvelles connaissances sur la pratique chirurgicale prouvera également la pertinence de son analyse en s'accompagnant d'un résultat de prédiction. Finalement, on peut représenter le positionnement de l'objectif de l'étude sur une échelle à 7 niveaux (voir tableau 2.1).

Dans la figure 2.1, je décris la répartition des études selon le positionnement de leur objectif vis-à-vis de la pratique chirurgicale. On observe une prédominance attendue des études traitant de la tâche d'évaluation. Comme expliqué précédemment, cette tâche est un prérequis aux deux autres, il est donc normal qu'elle soit traitée en priorité. Par ailleurs on a plus d'études qui tendent à analyser la pratique du chirurgien, que d'études de validation.

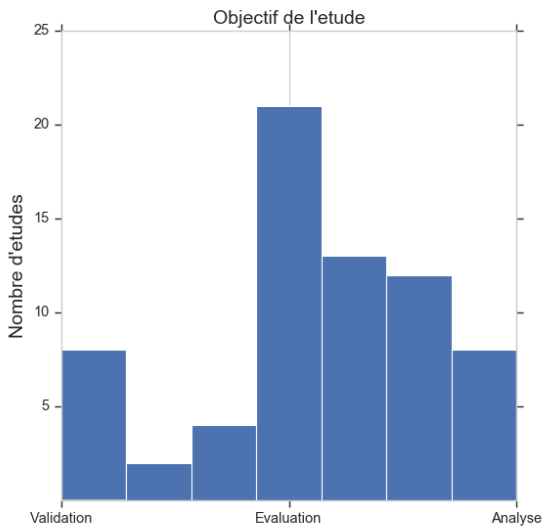


FIGURE 2.1 – Quelle est la position de la pratique chirurgicale dans l'objectif global de l'étude ?

Comme nous le verrons plus tard (section 2.3.1), le développement des premiers scores dédiés à l'évaluation des compétences chirurgicales remonte à une vingtaine d'années, et les outils d'évaluation automatiques pour cet usage sont encore plus récents. De plus, ces outils sont encore rarement utilisés en environnement clinique en raison de leur manque de maturité. Avant donc de chercher à valider des outils qui manquent encore de pertinence, il

est logique de chercher à les perfectionner. Ces observations sont néanmoins cohérentes avec le constat de Palter and Grantcharov (2010) sur le manque certain d'études de validation des outils de simulation chirurgicale.

Dans notre corpus d'étude, une grande majorité des articles s'attachent à prédire, à analyser ou à valider le même aspect de la pratique chirurgicale : le niveau d'expertise du chirurgien. En général les chirurgiens sont séparés soit en deux catégories, le novice et l'expert, soit en trois catégories, le novice, l'intermédiaire et l'expert :

1. Le chirurgien expert ou confirmé qui a effectué plus de 50 chirurgies.
2. Le chirurgien de niveau intermédiaire qui effectué entre 5 et 50 chirurgies. Il s'agit en général d'internes séniors ou de jeunes chirurgiens.
3. Le novice en chirurgie. Cette catégorie prend des formes variées, le novice peut-être un non-chirurgien, un externe, un interne jeune ou sénior, en allant parfois jusqu'au chirurgien junior. La catégorie novice peut aussi regrouper un mélange de ces différents profils.

On notera que le nombre de chirurgies nécessaires pour être reconnu comme expert ne fait pas consensus, certaines études ne le précisent pas tandis que d'autres fixent la barre à 100 chirurgies. Il serait intéressant de comparer les raisons de ces différents choix. Par ailleurs, lorsque des novices interviennent dans l'étude, la procédure étudiée est une simulation. En effet, il est inenvisageable d'étudier le geste d'un novice lors d'une opération sur un patient. Exception faite de certaines études dans lesquelles on désigne comme « novice » un jeune chirurgien se formant à une spécialité chirurgicale.

Selon moi, cette focalisation des travaux sur le niveau d'expertise s'explique par le besoin de comparer ses résultats à l'état de l'art, même si on a vu le manque de fiabilité de la définition de cette notion d'expertise. L'expertise des chirurgiens impliqués dans une étude est de plus une notion facilement accessible.

Nous avons néanmoins rencontré quelques études prenant le risque de s'intéresser à d'autres facettes de la pratique chirurgicale. (El Ahmadieh et al., 2014; Khan et al., 2015; Malpani et al., 2015; Matsuda et al., 2014) ont étudié la qualité de pratique des chirurgiens que nous distinguons de l'expertise, en effet un chirurgien expert n'a pas la même qualité de pratique tous les jours, et un interne junior peut très bien mieux réussir une certaine tâche d'entraînement qu'un interne sénior. (Kumar et al., 2012; Menhadji et al., 2012; Mitchell et al., 2011; Nugent, 2012) se sont intéressés à la courbe d'apprentissage pendant la formation des chirurgiens, tandis que (Ershad et al., 2016; Huaulmé et al., 2018) ont étudié le profil de pratique des chirurgiens, et que (Jiang et al., 2013) ont étudié la complexité de tâches d'entrainement. Les approches alternatives proposées par ces études présentent une diversité qui nous laisse envisager les possibilités d'étude autour de cette pratique chirurgicale.

2.2.2 | L'application clinique

Indépendamment de son objectif, l'expérimentation clinique menée au sein d'une étude permet de situer cette dernière du point de vue de son application, mais également de l'impact de l'étude en elle-même. Deux types de paramètres me permettent de classifier le

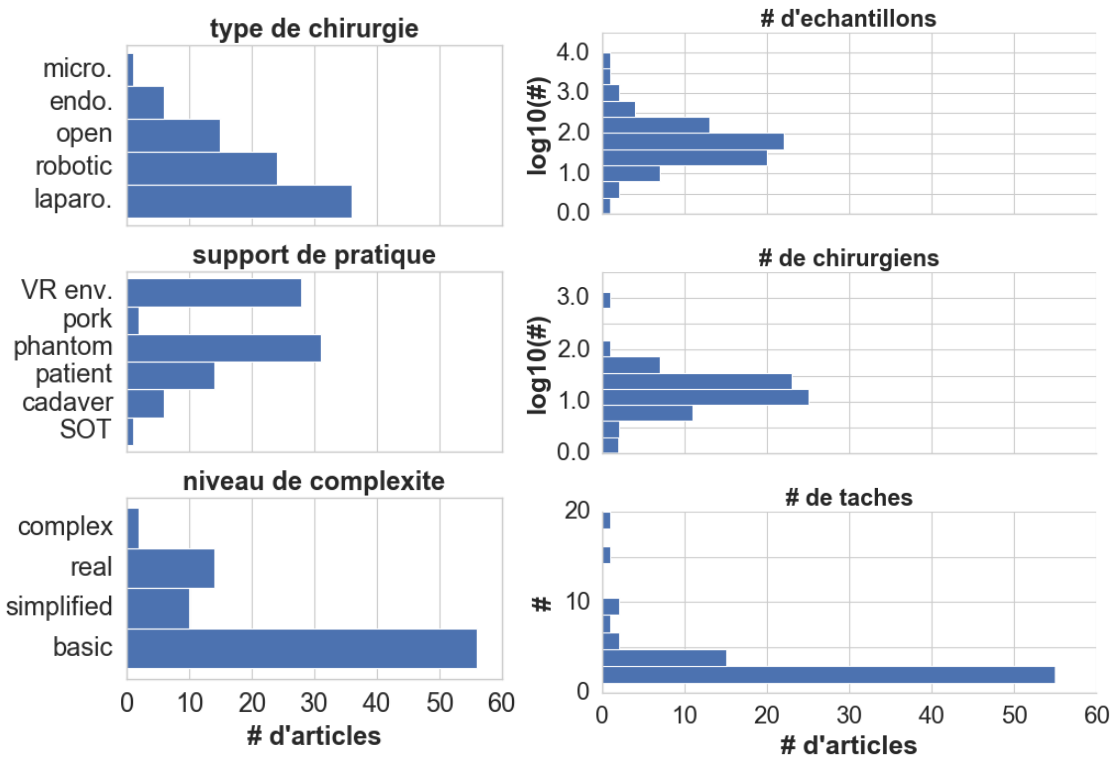
Nom	Description
	Paramètres qualitatifs
Le type de chirurgie	Il peut s'agir d'une chirurgie ouverte, laparoscopique, ou robotisée.
La spécialité chirurgicale	Certaines études s'appliquent spécifiquement à une spécialité (digestive, urologie, ophtalmologique, orthopédique, ...). D'autres sont indépendantes de la spécialité, notamment lorsque l'étude se fait dans le cadre d'une simulation d'entraînement.
Le support de pratique	Dans le cadre d'une chirurgie réelle, le support est le patient. S'il s'agit d'une simulation, les supports sont variés : cadavre, modèle porcin, fantôme, banc d'essai, environnement virtuel, ...
La complexité de la tâche	La complexité de la tâche dépend du niveau d'expertise étudié. Elle peut être basique pour des novices. Il peut s'agir d'une procédure simplifiée inspirée d'une procédure réelle modélisée dans un environnement d'entraînement. Ou bien il peut s'agir de procédure réelle classique, voire complexe.
	Paramètres quantitatifs
# de catégories de niv. d'expertise	En général, on a deux niveaux d'expertise (novice et expert) ou trois (novice, intermédiaire et expert).
# de chirurgiens	Ce sont les chirurgiens ayant effectué au moins une tâche dans le cadre de l'expérimentation.
# de tâches	Selon les études, on s'intéresse à une ou à différentes tâches.
# d'échantillons	En général il s'agit du nombre de chirurgiens ayant participé multiplié par le nombre de tâches, mais certaines études proposent également d'autres configurations.

Tableau 2.2 – Description des paramètres qualitatifs et quantitatifs utilisés pour caractériser l'expérimentation clinique d'une étude

contexte clinique des études : les paramètres qualitatifs et les paramètres quantitatifs (voir tableau 2.2).

Une caractérisation qualitative permet de situer ces expérimentations cliniques dans le domaine chirurgical. La figure 2.2a présente la répartition des études selon trois critères : le type de chirurgie, le support de pratique et le niveau de complexité de la tâche chirurgicale. Les types de chirurgie les plus représentés sont la chirurgie laparoscopique et robotisée, la chirurgie ouverte étant également étudiée dans une quinzaine d'études. Les supports de pratique prédominants sont les environnements de réalité virtuelle et les fantômes, qui comprennent également les bancs d'essais. On observe également qu'une très large majorité d'études s'intéressent à des tâches chirurgicales basiques. Nous définissons une tâche simplifiée comme une tâche d'entraînement cherchant à reproduire avec fidélité la réalité du bloc opératoire. A noter que nous différencions les chirurgies réelles de difficulté normale nécessitant une expérience de la chirurgie en général, des chirurgies complexes, qui sont plutôt maîtrisées par des experts d'une spécialité chirurgicale précise.

La caractérisation quantitative rend plus compte de l'impact de l'étude. Trois critères sont considérés : le nombre d'échantillons, le nombre de chirurgiens impliqués et le nombre de tâches chirurgicales. Ici, on définit une tâche chirurgicale comme étant un certain exercice d'entraînement ou une certaine technique chirurgicale, une tâche chirurgicale peut donc être tant un exercice de suture sur banc d'essai ou en environnement virtuel, qu'une procédure réelle menée sur patient. Un échantillon est une instance de tâche chirurgicale effectuée par



(a) (b)

FIGURE 2.2 – Distribution des critères qualitatifs et quantitatifs caractérisant les expérimentations cliniques menées dans les études présentées - (a) Répartition des études selon leurs conditions d'expérimentation clinique - *endo.* pour endoscopic - *micro.* pour microsurgery - *SOT* pour Simulated Operating Theatre - *VR env.* pour Virtual Reality environment - (b) Répartition des études selon leur nombre de participants, de tâches distinctes et d'échantillons - la répartition est donnée selon une échelle logarithmique en base 10 pour le nombre d'échantillons et de chirurgiens.

un chirurgien dans le cadre de l'expérimentation clinique d'une étude.

Plus l'étude compte de chirurgiens impliqués, de tâches chirurgicales distinctes effectuées, et d'échantillons, plus les résultats obtenus seront impactant et robuste d'un point de vue clinique. La force des résultats peut aussi venir des catégories qualitatives si, par exemple, plusieurs types de chirurgie, plusieurs supports de pratique, ou plusieurs difficultés de tâches sont comparés. La figure 2.2b présente la distribution des études selon leur nombre de chirurgiens participants, leur nombre de tâches différentes effectuées, et leur nombre total d'échantillons. On observe qu'en moyenne, quelques dizaines de chirurgiens participent à l'étude. Une étude a fait participer presque 80 chirurgiens (Ghani et al., 2016), et une autre presque mille (Matsuda et al., 2014). La majorité des études s'intéressent à une seule tâche. En moyenne, les études travaillent sur une centaine d'échantillons, avec certaines études n'utilisant que quelques échantillons, tandis que d'autres ont plusieurs milliers d'échantillons.

La figure 2.3 associe chacun de ces critères deux à deux pour mieux rendre compte des conditions expérimentales ayant réellement été mises en place. Sur le corpus d'études que nous présentons ici, on peut faire les observations suivantes :

— Fig. 2.3a : La chirurgie robotisée a essentiellement été étudiée en environnement vir-

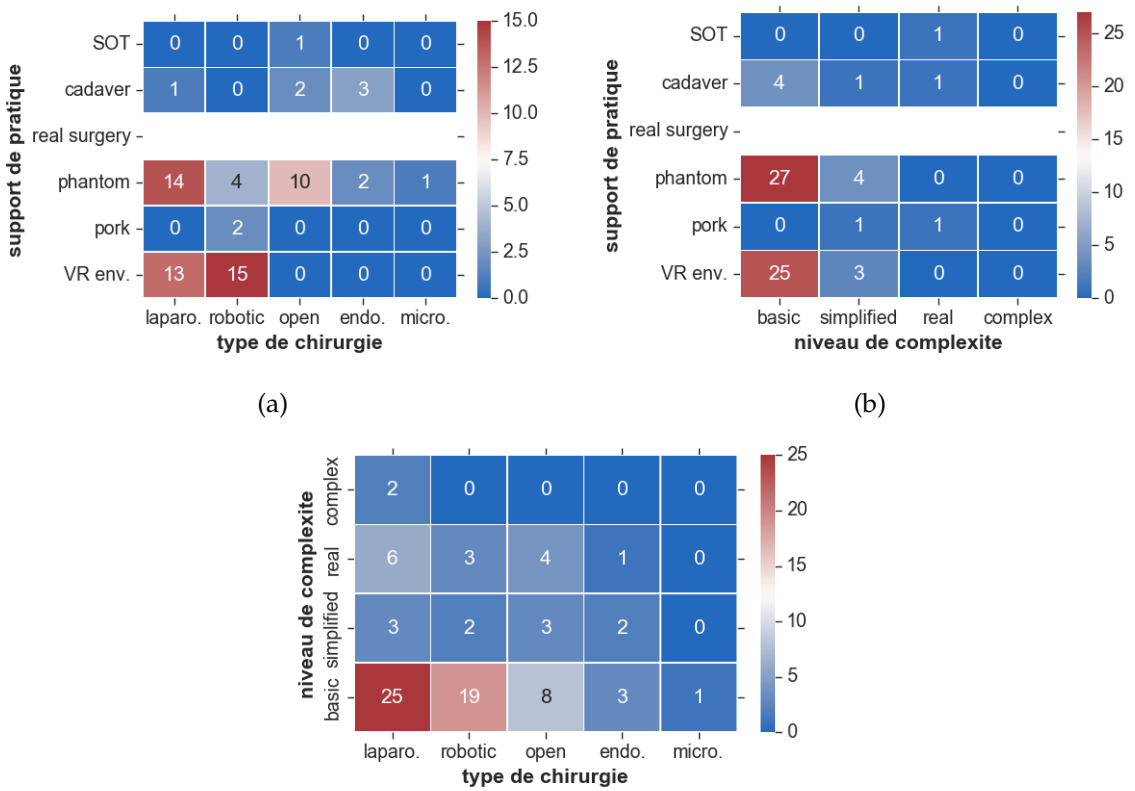


FIGURE 2.3 – Répartition des études selon leurs conditions d'expérimentation clinique, la mise en relation des critères dans ces cartes de chaleurs permet de mieux comprendre les conditions expérimentales ayant réellement lieu dans ces études - endo. pour endoscopic - micro. pour microsurgery - SOT pour Simulated Operating Theatre - VR env. pour Virtual Reality environment

tuel.

- Fig. 2.3a : La chirurgie laparoscopique a été étudiée à part égale en environnement virtuel et sur fantôme, mais également 8 fois en condition réelle.
- Fig. 2.3a : La chirurgie ouverte a essentiellement été étudiée sur fantôme.
- Fig. 2.3a : La chirurgie endoscopique, la microchirurgie, le cadavre et le théâtre d'opération simulé ou « simulated operating theatre » SOt n'ont que peu été étudiés.
- Fig. 2.3b : Les tâches chirurgicales basiques ont soit été étudiées en environnement virtuel, soit sur fantôme.
- Fig. 2.3b : La chirurgie sur patient réel s'est évidemment effectuée dans des conditions de difficulté réelle à complexe.
- Fig. 2.3b : Les procédures simplifiées ont été étudiées sur une grande variété de supports.
- Fig. 2.3c : La chirurgie laparoscopique a été étudiée pour tous les niveaux de difficulté, mais surtout sur des tâches basiques.
- Fig. 2.3c : La chirurgie robotisée a été étudiée surtout sur des tâches basiques mais aussi sur des procédures simplifiées et réelles.
- Fig. 2.3c : Les tâches chirurgicales basiques ont été étudiées sur tous les types de chirurgies, mais surtout en chirurgie laparoscopique ou robotisée.

Il existe une grande diversité d'expérimentations cliniques possibles lorsqu'on considère les combinaisons possibles de support de pratique, de type de chirurgie, et de niveau de complexité. Dans la réalité, une fraction de ces possibilités est exploitée, d'une part parce que certaines combinaisons sont aberrantes (telles qu'une procédure complexe sur un fantôme d'entraînement, ou une procédure basique sur un patient réel). D'autre part, certaines conditions cliniques sont beaucoup plus simples à mettre en place que d'autres, permettent de faire intervenir un nombre plus important de chirurgiens, ou permettent de récupérer plus d'échantillons d'étude.

Finalement, l'étude de la pratique chirurgicale peut être menée dans des conditions chirurgicales variées. Et les études de la pratique chirurgicale s'appliquent tant à la chirurgie ouverte, qu'à la laparoscopie ou qu'à la chirurgie robotisée, sur des niveaux d'expertise différents, et sur des supports de pratique différents. L'étude de la pratique chirurgicale constitue donc un soutien potentiel pour l'ensemble des pratiques chirurgicales.

2.2.3 | La forme des données traitées

Tandis que le chirurgien pratique la tâche chirurgicale dans le cadre de l'expérimentation, un flux de données est enregistré pour permettre un traitement futur. Ces données sont issues de différentes sources pouvant potentiellement être combinées pour obtenir une information plus complète sur le déroulement de la tâche chirurgicale étudiée. Voici les différentes sources de données, et les types de données associées qui peuvent être enregistrées lors d'une procédure :

— Les gyroscopes, accéléromètres, et capteurs électromagnétiques, à partir desquels on obtient un signal décrivant la trajectoire de l'objet auquel ils sont attachés. Cette trajectoire est en trois dimensions si on s'intéresse uniquement aux translations, ou en 6 dimensions si on rajoute les rotations.
— Une caméra infrarouge accompagnée de marqueurs infrarouges. Ce système permet également de récupérer une trajectoire de l'objet auquel les capteurs ont été attachés.
— L'interface robotique sur laquelle le chirurgien effectue les gestes chirurgicaux. Ici, et selon la complexité de l'interface, on peut récupérer tout une variété de signaux cinématiques liés aux mouvements des mains du chirurgien, des articulations du robot et des instruments chirurgicaux.
— Les capteurs de forces, qui sont placés sur les outils ou sur des bancs d'essai dédiés. Dans les études présentées ici, ces capteurs permettent de récupérer les composantes de forces en trois dimensions.
— Une ou des caméra(s) externe(s) qui sont placées dans l'environnement opératoire et enregistre(nt) le flux vidéo en deux ou trois dimensions.
— Une caméra endoscopique qui enregistre le déroulement opératoire de l'intérieur. Ce système enregistre les mouvements des outils et des cibles chirurgicales. On retrouve des caméras endoscopiques dans des chirurgies laparoscopiques, endoscopiques, et robotisées. A la différence des caméras externes, la caméra endoscopique est manipulée par le chirurgien ou son assistant, son point de vue n'est donc pas fixe. Ainsi la qualité même du signal enregistré est beaucoup plus variable et moins maîtrisée que

les autres modalités présentées ci-dessus. Ici aussi on enregistre le flux vidéo en deux ou en trois dimensions.

Ces sources de données primaires peuvent ensuite être traitées une première fois pour obtenir d'autres types de données. Trois cas se présentent :

— Les signaux mesurés sont traités afin d'obtenir un ensemble de métriques représentatives de l'état des données traitées et donc de la tâche chirurgicale.
— La vidéo enregistrée est traitée afin d'obtenir les trajectoires des mains du chirurgien ou des instruments visibles dans cette vidéo. Ces trajectoires peuvent ensuite être elles-mêmes traitées.
— La vidéo enregistrée peut être manuellement annotée et on distingue trois types d'annotation :
  — L'évaluation de la pratique chirurgicale qui est en général effectuée par un chirurgien expert, mais pas seulement.
  — L'annotation de la procédure chirurgicale, qui est aussi effectuée par un chirurgien expert.
  — La détection, le suivi, ou la segmentation d'objets dans les images (le plus souvent les instruments). Dans certains cas, ces annotations sont effectuées au cours de la tâche elle-même.

Le recueil de ces données nécessite au préalable l'installation des capteurs dans l'environnement de la tâche étudiée. S'il s'agit d'une procédure réelle, des problématiques de stérilisation, de sécurité du patient et d'encombrement du bloc opératoire se posent, et viennent énormément complexifié la mise en place de ces capteurs. De fait, les capteurs les plus simples à utiliser sont encore ceux déjà présents au bloc opératoire, c'est à dire la vidéo endoscopique et dans certains cas l'interface robotisée.

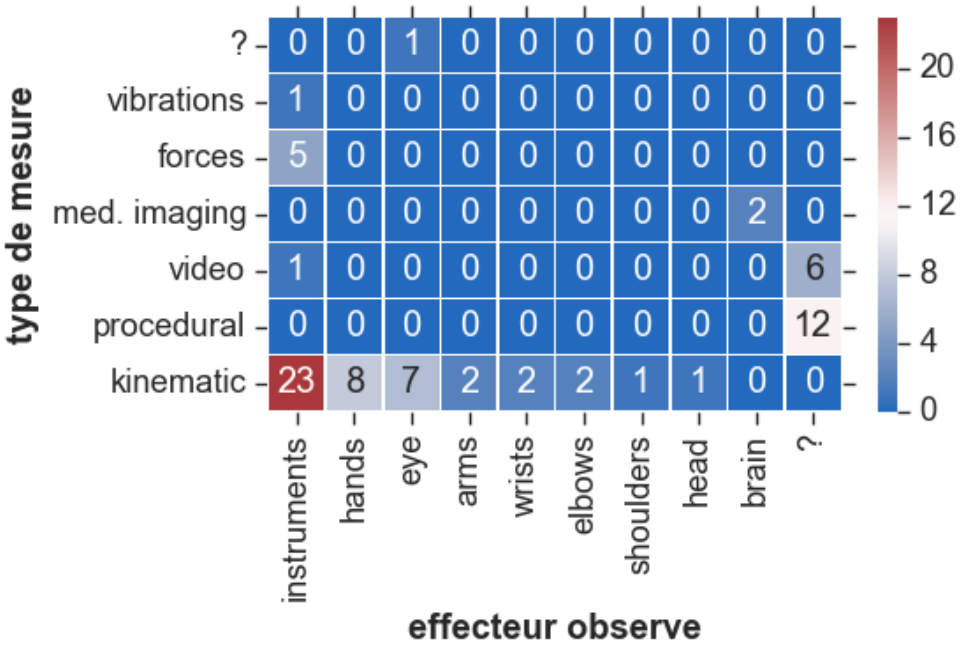


FIGURE 2.4 – Répartition des études en fonction du type de données mesurées, et de l'effecteur observé

La figure 2.4 présente la répartition des études en fonction du type de données mesurée et de l'effecteur observé. Cette carte de chaleur montre avant tout que la tendance majoritaire est l'étude de données cinématiques, principalement pour les instruments et les mains du chirurgien, mais aussi pour d'autres parties du corps du chirurgien. Les yeux du chirurgien notamment, ou plus exactement les mouvements de son regard et les variations de taille de ses pupilles ont aussi été étudiés dans 8 études. 5 études se sont intéressées au lien entre les forces appliquées lors de la manipulation des outils et la pratique du chirurgien. 12 études ont étudié l'aspect procédural de la pratique du chirurgien. Dans la section suivante 2.3, je présenterai chacun de ces groupes d'études plus en détail.

2.2.4 | La méthodologie de traitement des données

Jusqu'ici dans cet état de l'art, j'ai introduit des études présentant une grande variété, tant dans leurs objectifs, que dans leurs conditions d'expérimentation clinique, ou dans les sources de données exploitées. Cette variété s'accentue encore lorsqu'on considère les méthodologies d'analyse statistique et algorithmique. Comme mon intérêt porte avant tout sur la façon dont ces études s'articulent autour de la notion de pratique chirurgicale, je considère qu'il serait trop complexe et laborieux d'aborder cet aspect méthodologique. Ainsi je n'aborderai pas dans mon état de l'art les méthodologies de traitement de données.

En résumé

— La majorité des études ont pour objectif d'évaluer la pratique chirurgicale, le deuxième type d'objectif concerne l'analyse de la pratique chirurgicale, et le reste des études, moins nombreuses, cherchent à valider un outil d'évaluation de la pratique chirurgicale.
— Sur l'ensemble des études, les expérimentations cliniques se font en majorité sur fantôme ou sous environnement virtuel, et simulent des tâches basiques laparoscopiques, robotisées ou non. D'autres types de chirurgie, sur d'autres types de support et avec des niveaux de complexité plus élevés ont également été étudiés, mais en moins grand nombre.
— Les données exploitées dans ces études sont en très grande majorité des données cinématiques des instruments chirurgicaux. La cinématique des mains du chirurgien et de ses yeux, ainsi que les forces appliquées par les instruments aux tissus ont aussi été traitées, mais dans une moindre mesure.

2.3 | Les principaux sujets de recherche

En considérant la variabilité des approches et des angles d'attaque proposés pour l'évaluation et l'analyse de la pratique chirurgicale, il m'a semblé pertinent de ne pas me focaliser sur les approches de traitement statistique et algorithmique mais plutôt sur les résultats des études. Plus précisément, je me suis focalisé sur les nouvelles connaissances apportées par ces études sur la pratique et le geste chirurgical, et à la façon dont cette pratique peut être mise en lien avec le comportement du chirurgien, tant dans ses mouvements, que dans ses prises de décisions, ou dans son attitude.

Pour structurer cette analyse, j'ai d'abord introduit les approches par scores structurés, puis j'ai étudié différentes sources de données utilisées pour analyser la pratique chirurgicale.

2.3.1 | L'évaluation par scores structurés

Dans le cadre de l'entraînement et de la formation des chirurgiens, l'évaluation des compétences techniques nécessite la présence d'un chirurgien confirmé. Depuis une vingtaine d'années, les scores structurés sont apparus comme une modalité valide, pertinente et simple pour l'évaluation des chirurgiens en formation, lors de leur entraînement à la pra-

	1	2	3	4	5
Respect for tissue	Frequently used unnecessary force on tissue or caused damage by inappropriate use of instruments		Careful handling of tissue but occasionally caused inadvertent damage		Consistently handled tissues appropriately with minimal damage
Time and motion	Many unnecessary moves		Efficient time/motion but some unnecessary moves		Economy of movement and maximum efficiency
Instrument handling	Repeatedly makes tentative or awkward moves with instruments		Competent use of instruments although occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness
Knowledge of instruments	Frequently asked for the wrong instrument or used an inappropriate instrument		Knew the names of most instruments and used appropriate instrument for the task		Obviously familiar with the instruments required and their names
Use of assistants	Consistently placed assistants poorly or failed to use assistants		Good use of assistants most of the time		Strategically used assistant to the best advantage at all times
Flow of operation and forward planning	Frequently stopped operation or needed to discuss next move		Demonstrated ability for forward planning with steady progression of operative procedure		Obviously planned course of operation with effortless flow from one move to the next
Knowledge of specific procedure	Deficient knowledge. Needed specific instruction at most operative steps		Knew all important aspects of the operation		Demonstrated familiarity with all aspects of the operation

Tableau 2.3 – Echelle d'evaluation OSATS (Martin et al., 1997)

tique chirurgicale. Ces scores présentent notamment l'avantage d'être indépendants de la procédure et permettent de maîtriser la variabilité inter- et intra-évaluateur.

Trois scores structurés existent : OSATS pour la chirurgie ouverte, GOALS pour la chirurgie laparoscopique et GEARS pour la chirurgie laparoscopique robotisée. Ces scores évaluent de manière distincte différents aspects de la pratique à l'aide d'une échelle de Likert à 5 niveaux. Je présente ici chacun de ces scores, ainsi qu'un autre score proposant une approche plus orientée sur le respect de la procédure et l'analyse des erreurs (méthode OCHRA). Hormis le GEARS, ces scores et leur validation ont déjà été présentés par Wolf (2013).

2.3.1.1 | Objective Structured Assessment of Technical Skills (OSATS)

Le score OSATS a été proposé par Martin et al. (1997) pour l'évaluation des compétences techniques des internes en chirurgie ouverte. Ce score capture 7 aspects différents des compétences techniques du chirurgien, et décrit chacun de ces aspects à l'aide d'une échelle de Likert à 5 niveaux. Les niveaux 1, 3 et 5 sont décrits par un énoncé précisant les conditions liées à ce niveau de compétence. Le tableau 2.3 présente la grille d'évaluation du score OSATS.

La revue de littérature menée par Hatala et al. (2015) établit la validité de ce score pour l'évaluation dans le cadre de chirurgies simulées, et dans le cadre de la formation. En revanche il n'y a pas suffisamment d'arguments en faveur de l'utilisation du score OSATS pour certifier d'un certain niveau de pratique ni pour l'évaluation autonome.

2.3.1.2 | Global Operative Assessment of Laparoscopic Skills (GOALS)

Le score GOALS a été proposé par Vassiliou et al. (2005). L'objectif de cette méthode est de proposer une mesure objective, fiable, et valide des performances intra-opératives en laparoscopie. L'étude a montré que le score GOALS vérifiait ces conditions d'une tâche de dissection de vésicule biliaire.

Ce score se présente sous la même forme que le score OSATS, les 5 aspects de la pratique chirurgicale en laparoscopie qu'il décrit sont présentés dans le tableau 2.4.

2.3.1.3 | Global Evaluative Assessment of Robotic Skills (GEARS)

Le score GEARS a été proposé par Goh Alvin C et al. (2012). Ce score a été développé afin d'évaluer les performances intra-opératives lors de chirurgie en laparoscopie robotisée. Inspiré du score GOALS, il se présente sous la même forme et traite de 6 aspects de la pratique chirurgicale en laparoscopie robotisée (voir tableau 2.5). Comme ce score est encore récent, diverses études s'intéressent à sa validation (Aghazadeh et al., 2015; Sánchez et al., 2016), mais aucune n'a encore établi sa validité de façon globale.

2.3.1.4 | Observational Clinical Human Reliability Assessment (OCHRA)

L'approche OCHRA a été développée par Joice et al. (1998). A la différence des scores décrits ci-dessus, il ne s'agit pas d'une grille de notation. L'idée ici est d'analyser les différentes étapes de la procédure, et d'y détecter les erreurs potentielles. Cette méthodologie

	1	2	3	4	5
Depth perception	Consistently overshoots target, wide swings, slow to correct		Some overshooting or missing of target, but quick to correct		Accurately directs instruments in the correct plane to target
Bimanual dexterity	Uses only one hand, ignores non dominant hand, poor coordination between hands		Uses both hands, but does not optimize interaction between hands		Expertly uses both hands in a complementary manner to provide optimal exposure
Efficiency	Uncertain, inefficient efforts, many tentative movements, constantly changing focus or persisting without progress		Slow, but planned movements are reasonably organized		Confident, efficient and safe conduct, maintains focus on task until it is better performed by way of an alternative approach
Tissue handling	Rough movements, tears tissue, injures adjacent structures, poor grasper control, grasper frequently slips		Handles tissues reasonably well, minor trauma to adjacent tissue (ie. occasionnal unnecessary bleeding or slipping of the grasper		Handles tissues well, applies appropriate traction, negligeable injury to adjacent structures
Autonomy	Unable to complete entire task, even with verbal guidance		Able to complete task safely with moderate guidance		Able to complete task independently without prompting

Tableau 2.4 – Echelle d'evaluation GOALS(Vassiliou et al., 2005)

	1	2	3	4	5
Depth Perception	Constantly overshoots target, wide swings, slow to correct		Some overshooting or missing target, but quick to correct		Accurately directs instruments in the correct plane to target
Bimanual Dexterity	Uses only one hand, ignores non dominant hand, poor coordination		Users both hands, but does not optimize interaction between hands		Expertly uses both hands in a complementary way to provide optimal exposure
Efficiency	Inefficient efforts, many tentative movements, constantly changing focus or persisting without progress		Slow, but planned movements are reasonably organized		Confident, efficient and safe conduct, maintains focus on task, fluid progression
Force Sensitivity	Rough moves, tears tissue, injures nearby structures, poor control, frequent suture breakage		Handles tissue reasonably well, minor trauma to adjacent tissue, rare suture breakage		Applies appropriate tension, negligible injury to adjacent structures, no suture breakage
Robotic Control	Consistently does not optimize view, hand position, or repeated collisions even with guidance		View is sometimes not optimal. Occasionally needs to relocate arms. Occasional collisions and obstruction of assistant		Controls camera and hand position optimally and independently. Minimal collisions or obstruction of assistant
Autonomy	Unable to complete entire task, even with verbal guidance		Able to complete task safely with moderate verbal guidance		Able to complete task independently without verbal prompting

Tableau 2.5 – Echelle d'evaluation GEARS (Goh Alvin C et al., 2012)

nécessite d'établir a priori une décomposition de la procédure et une taxonomie des erreurs envisageables à chaque étape de la procédure.

C'est donc une approche beaucoup moins universelle que les scores précédents, car un panel d'experts doit établir cette décomposition et cette taxonomie pour chaque technique chirurgicale. En revanche, elle présente l'avantage d'être adaptée à un usage en routine clinique avec un retour possible, y compris pour des chirurgiens confirmés. De plus, l'analyse de la procédure menée par le panel d'experts établit, pour une chirurgie donnée, un état de leur connaissance de la procédure, des difficultés, et des erreurs, avec leur cause, leur mode, et leur portée.

Les trois scores (OSATS, GOALS et GEARS) présentés ici sont des outils qui tendent à un usage global dans le cadre de l'entraînement à la chirurgie. Chaque score est adapté à un type de chirurgie. Ces scores présentent l'avantage d'être simples d'utilisation pour l'évaluateur, tout en proposant un retour sur expérience clair à l'étudiant. Ils nous informent sur les aspects essentiels et critiques de chaque pratique, sur la bonne façon de pratiquer, mais aussi sur l'évolution attendue des compétences de l'étudiant au cours de sa formation. En revanche, au-delà d'un certain niveau d'expertise, ces scores ne permettent plus de distinguer les performances des chirurgiens, et ne sont donc pas adaptés pour l'évaluation des performances de chirurgiens confirmés lors de chirurgies réelles. Par ailleurs, ces scores présentent un retour sur une performance dans son ensemble, et ne permettent pas de distinguer l'évolution de la qualité de pratique au cours de la procédure.

L'approche OCHRA est plus adaptée à une utilisation lors de chirurgies réelles, du fait de sa construction. Même pour des chirurgiens confirmés, on pourra observer sur plusieurs échantillons de procédure les erreurs faites, et donc comparer les pratiques des chirurgiens. En revanche, la méthodologie OCHRA nécessite un travail préalable important, et ce pour chaque technique chirurgicale. Cette approche étant construite à partir de l'état de connaissance sur une procédure donnée à un instant donné, elle doit donc être maintenue pour ne pas devenir obsolète, du fait de l'évolution des connaissances sur la procédure elle-même.

Dans l'ensemble, ces outils sont pertinents pour l'évaluation des chirurgiens en formation, ou en routine clinique. Ils se basent néanmoins sur l'expertise et l'implication directe de chirurgiens confirmés. Nous discuterons donc par la suite d'autres approches qui cherchent à s'affranchir de la subjectivité, de l'expérience, et du temps du chirurgien.

2.3.2 | L'évaluation par la foule

Cette approche plutôt atypique du problème de l'évaluation de la pratique chirurgicale a été proposée en premier par Chen et al. (2014). L'idée de recueillir l'avis d'une large population de profanes a déjà été appliquée dans plusieurs domaines, tels que la science du langage [1], l'épigénétique [2], ou le diagnostic médical [3]. En simplifiant suffisamment la tâche à effectuer, grâce à la force du plus grand nombre, et à l'aide d'un post-traitement adapté, on peut obtenir des résultats similaires à ceux de quelques experts.

1. https://vizwiz.org/
2. https://foldit.fandom.com/wiki/Foldit_Wiki
3. https://www.crowdmed.com/

L'étude menée par Chen et al. (2014) consistait à faire évaluer un unique échantillon vidéo d'une tâche chirurgicale sur banc d'entraînement par une population de chirurgiens experts d'une part, et par une foule anonyme grâce à l'outil en ligne « Amazon Mechanical Turk »[4]. La conclusion principale de cette étude est que l'Evaluation par la Foule (EF) est un complément, voire une alternative valide à l'évaluation par les pairs, dans le cas de tâches chirurgicales basiques en environnement virtuel de laparoscopie robotisée.

Dans toutes les études d'EF qui ont été menées par la suite, plusieurs recommandations méthodologiques sont systématiquement respectées sur la façon dont les évaluateurs anonymes sont intégrés dans l'étude :

— Les données vidéos sont anonymisées.
— L'évaluateur doit déjà avoir un certain nombre de tâches à son actif sur la plateforme « Mechanical Turk » (classiquement 100).
— L'évaluateur doit satisfaire un certain niveau de qualité dans ces opérations.
— L'évaluateur doit être majeur.
— Un test d'attention et de compréhension est présenté à l'évaluateur avant de lui présenter les échantillons réels. On peut ainsi rejeter les évaluateurs qui ne passent pas ce test.

Dans l'ensemble, l'EF pour le niveau de pratique chirurgicale permet d'obtenir des évaluations moins biaisées, et plus rapidement que l'évaluation par les pairs. En effet, la fiabilité inter-utilisateur et inter-institution est beaucoup plus grande pour une foule d'évaluateurs aux origines diverses que sur les échantillons classiques de quelques chirurgiens experts d'un même centre (Deal et al., 2017). Finalement, grâce à l'EF, on permet au chirurgien de ne plus consacrer son temps à cette tâche d'évaluation, c'est donc une méthode beaucoup moins coûteuse.

A noter que pour toutes ces études portant sur l'évaluation par la foule, la comparaison de la foule et d'experts se fait sur des scores cliniques établis, à savoir le score OSATS pour la chirurgie ouverte, le GOALS pour la chirurgie laparoscopique, et le GEARS pour la chirurgie robotisée.

Ces différents avantages, ainsi que les questions suivantes portant sur les possibilités liées à l'EF, ont été étudiés par la suite dans d'autres études :

— L'EF peut-elle évaluer des tâches chirurgicales plus complexes, et allant jusqu'à la procédure réelle ?
— L'EF peut-elle discriminer des chirurgiens selon leur niveau de pratique ?
— L'EF peut-elle traiter d'autres types de chirurgie (ouverte, laparoscopique, ...) ?

2.3.2.1 | validation sur tâche basique

Les possibilités ouvertes par Chen et al. (2014) sur l'évaluation de la pratique chirurgicale par la foule ont été étudiées extensivement par de nombreux travaux. Dans toutes les études, l'objectif est d'obtenir les mêmes résultats d'évaluation par la foule que par les pairs. Ainsi, et de façon incrémentale, de nombreuses réponses ont été données quant au potentiel de l'EF.

4. https://www.mturk.com/

Tout un groupe d'études s'est intéressé spécifiquement au cas des tâches d'entraînement basiques. Leur objectif commun était d'étudier la capacité de l'EF à évaluer différents niveaux de pratique d'un groupe de chirurgien donnés. Cet objectif a été évalué dans différentes configurations (voir tableau 2.6). Holst et al. (2014) l'ont validé dans sa configuration la plus simple en étudiant un groupe restreint de chirurgiens de niveaux variés pour la chirurgie robotisée. Toujours pour la chirurgie robotisée, White et al. (2015) ont validé ce même objectif sur un groupe de 49 chirurgiens et pour deux tâches distinctes. Kowalewski et al. (2016) et Deal et al. (2016) ont mené des travaux similaires, mais pour le cas de la chirurgie laparoscopique. Enfin, Vernez et al. (2016) ont validé la capacité de l'EF à évaluer différents niveaux de pratique indifféremment pour la chirurgie ouverte, robotisée, et laparoscopique.

Réference	Type de chirurgie	# de praticiens	# de tâches	# d'échantillons
Chen et al. (2014)	robotique	1	1	1
Holst et al. (2014)	robotique	5	1	5
White et al. (2015)	robotique	49	2	96
Kowalewski et al. (2016)	laparoscopique	?	2	24
Deal et al. (2016)	laparoscopique	7	3	21
Vernez et al. (2016)	robotique, laparoscopique, ouverte	25	4	100

Tableau 2.6 – Description des configurations cliniques mises en place dans les études traitant de l'évaluation par la foule pour des tâches chirurgicales basiques

Finalement, les résultats de validation positifs de ces différents travaux ont permis de montrer la fiabilité et l'efficacité de l'évaluation de la pratique chirurgicale par la foule pour des tâches simples d'entraînement chirurgical. Et ce, pour des groupes de plusieurs dizaines de chirurgiens de niveaux variés, pour des tâches diverses, et indifféremment pour l'entraînement aux chirurgies ouverte, robotisée, et laparoscopique.

De tels résultats laissent envisager une utilisation réelle de l'EF dans le cadre de l'entraînement chirurgical (Vernez et al., 2016). En effet la foule pourvoit en échantillons d'évaluation de haute qualité pour des tâches chirurgicales simples (White et al., 2015). Néanmoins, cette méthode n'a pas pour objectif de remplacer le retour d'expérience du maître à l'élève, mais plutôt d'améliorer les méthodes d'entraînement actuelles (Aghdasi et al., 2015).

De plus ces études ont soulevé de nouveaux questionnements. Sur la plateforme « Mechanical Turk » utilisée dans toutes ces études, l'évaluateur anonyme est rémunéré :

— Quel impact la rémunération des évaluateurs a-t-elle sur la qualité des échantillons d'évaluation (White et al., 2015) ?
— Quel impact la distribution des niveaux de pratique dans le groupe de chirurgiens a-t-elle sur la qualité d'évaluation (White et al., 2015) ?
— Comment des niveaux de pratique plus rapprochés ou globalement plus élevés seraient-ils évalués par l'EF ?

2.3.2.2 | validation dans conditions plus complexe

Après avoir constaté le potentiel de l'EF pour des tâches d'entraînement simples et dans des conditions variées, les recherches ont été menées plus avant dans l'objectif d'évaluer des procédures réelles. L'EF s'est donc confrontée à une variété de jeux de données (voir tableau 2.7), et les résultats ont été, encore une fois, étonnamment satisfaisants.

Réference	type de chirurgie	procédure	# de tâches
Aghdasi et al. (2015)	ouverte	cricothyrotomie sur fantôme	26
Holst et al. (2015)	robotique	fermeture de vessie	12
Powers et al. (2015)	robotique	parial nephrectomy	14
Ghani et al. (2016)	robotique	prostatectomie radicale	76
Deal et al. (2017)	laparoscopie	cholecystectomie	160

Tableau 2.7 – Description des configurations cliniques mises en place dans les études traitant de l'évaluation par la foule pour des tâches chirurgicales complexes.

Toujours de façon incrémentale, Aghdasi et al. (2015) ont validé l'EF pour une procédure simplifiée sur fantôme de chirurgie ouverte en rencontrant toutefois des difficultés dans la distinction entre les praticiens de niveau moyen et les novices. Holst et al. (2015) ont validé avec succès l'EF pour des procédures d'entraînement complètes sur modèle porcin en chirurgie robotisée. Powers et al. (2015) ont porté avec succès l'EF sur des extraits courts de chirurgies robotisées réelles. Ghani et al. (2016) se sont intéressés à un grand nombre d'échantillons d'étapes chirurgicales d'une procédure robotisée réelle. Pour ces échantillons l'EF permet de reconnaître les moins bonnes performances, mais pas de distinctions plus complexes. Cela s'explique notamment par le fait que le score utilisé ici, le GEARS, est discriminant pour des chirurgiens en cours de formation, et pas pour des chirurgiens confirmés, comme c'est le cas dans cette étude.

Deal et al. (2017) sont allés encore plus loin, en s'interrogeant sur la capacité de la foule à évaluer une étape précise d'une procédure réelle. Pour cela il a introduit au début de la procédure d'évaluation en ligne une vidéo de tutoriel, et des instructions accompagnées d'images décrivant le déroulement de l'étape en question et l'échelle d'évaluation associée. L'EF est encore une fois validée pour cette tâche, hormis dans la distinction entre les bonnes et les très bonnes performances.

On constate donc la puissance de cette approche qui, en 2-3 ans, a été validée sur des chirurgies en conditions réelles, et qui ne classe pas seulement les chirurgiens par niveau d'expertise, mais les évaluent avec des scores reconnus et établis. Il reste encore néanmoins des questions ouvertes :

— Toutes les études traitent d'extraits vidéos courts n'excédant pas quelques minutes, afin de rester sur des tâches simples à évaluer pour la foule (Holst et al., 2015). Est-il possible par une approche combinée d'évaluer par la foule des procédures réelles complètes ?

— Initialement, le faible coût de l'EF comparé au l'évaluation par les pairs l'a rendue particulièrement attractive. Cependant, Deal et al. (2017) questionnent le coût réel de cette approche dans le cas de chirurgies réelles nécessitant une formation a priori

des évaluateurs anonymes, et peut-être une plus haute rémunération du fait de la complexité de la tâche.
— Quels impacts ont le niveau de rémunération, le niveau d'expérience, et plus généralement le profil des évaluateurs anonymes sur l'EF ? (White et al., 2015)
— Une meilleure compréhension des dynamiques propre à l'évaluation par la foule permettrait d'optimiser la précision, le coût et la vitesse d'évaluation. (White et al., 2015)
— Dans une dynamique globale de protection des données, le fait de mettre à disposition des données patients, même anonymisées, sur une plateforme privée est un frein à l'utilisation d'une telle approche.

2.3.2.3 | Approches alternatives de l'évaluation par la foule

Deux approches alternatives m'ont intéressée du fait des nouvelles modalités d'évaluation qu'elles proposaient. Malpani et al. (2015) traitent des tâche basiques d'entraînement chirurgical. Lors d'une évaluation, deux extraits vidéos sont proposés, et après avoir regardé les deux échantillons, l'évaluateur anonyme doit simplement sélectionner celui présentant la meilleure qualité de pratique. Cette approche relativiste donne, au global, des résultats satisfaisants.

Ershad et al. (2019) traitent également des tâches d'entraînement, et se basent sur le score GEARS. Il a construit un lexique de paires d'adjectifs contraires associés à partir des catégories du score GEARS (par exemple : net/tremblant, calme/stressé). Ces adjectifs qualifient donc différentes composantes de la pratique chirurgicale, ce qu'il appelle le « style » du chirurgien. En décrivant à l'aide d'un vocabulaire facilement intelligible la pratique du chirurgien, on lui rend accessible l'évaluation qui est faite de sa pratique. Cette approche peut aussi faciliter le processus d'annotation, toujours grâce à l'intelligibilité de ce vocabulaire. Cette approche présente de très bons résultats de corrélation entre la foule et les chirurgiens évaluateurs.

En conclusion, l'évaluation de la pratique chirurgicale par la foule est une approche très pertinente pour compléter et améliorer les outils d'entraînement chirurgicaux, tant pour la chirurgie ouverte, que pour la chirurgie robotisée ou la laparoscopie. La foule est également capable de noter des extraits de procédures réelles, avec néanmoins plus de difficultés à distinguer les niveaux de pratique dans une population de chirurgiens confirmés uniquement.

Une limite intrinsèque de cette méthode, face à un phénomène complexe telle que la chirurgie, est de proposer à l'évaluateur profane des tâches simples. Le fait qu'une population de profanes soit capable d'évaluer des extraits de chirurgies réelles est déjà un résultat extrêmement satisfaisant, et il me semble compliqué pour une foule d'évaluer avec succès des procédures complètes. Ou bien cela viendra par une approche détournée ou combinée à d'autres technologies. Enfin plusieurs questions restent ouvertes sur des sujets tels que la place de la rémunération de l'évaluateur dans cette approche, l'impact du profil des évaluateurs, ou les dynamiques propres à l'évaluation par la foule.

2.3.3 | L'analyse de trajectoires

Plus classiquement, l'évaluation du niveau d'expertise peut se faire à partir de données extraites au cours de la procédure et à l'aide d'approches algorithmiques, comme nous l'avons introduit précédemment (voir section 2.2.3). Il est notamment possible d'étudier différentes trajectoires d'objets évoluant dans l'environnement chirurgical à partir de données cinématiques. Ces données cinématiques proviennent en général de capteurs faisant partie du robot chirurgical, ou bien installés en plus dans l'environnement chirurgical. Dans le deuxième cas, les capteurs peuvent être des capteurs électromagnétiques, des accéléromètres, ou des réflecteurs infra-rouges associés à une caméra infra-rouge. Ces trajectoires peuvent également provenir d'un traitement algorithmique de la vidéo chirurgicale. A partir de ces données cinématiques, il est possible de prédire et/ou d'étudier la pratique du chirurgien. Deux catégories de trajectoires sont étudiées séparément ou de façon conjointe : les trajectoires des instruments dans l'espace chirurgical, et les trajectoires de certaines parties du corps du chirurgien que sont les mains, les bras, et la tête, et les articulations associées (poignets, coudes et épaules).

Dans cet état de l'art, j'ai rencontré des approches très diverses, utilisant une source de données bien définie, tout comme des combinaisons de différentes sources, allant même jusqu'à des approches orientées fusion globale de données capteurs. Mon objectif n'est pas de décrire les types de capteurs, les configurations expérimentales, ou les combinaisons de données traitées dans ces études. Cela s'apparenterait à un travail de catalogage peu lisible, dont les conclusions manqueraient de pertinence au vu de mon objectif. Dans la suite je continuerai donc à me focaliser sur l'analyse et l'évaluation de la pratique chirurgicale. J'ai donc regroupé ces études selon le type de données traitées et selon la pertinence de leurs conclusions en lien avec la pratique chirurgicale.

2.3.3.1 | L'analyse des trajectoires d'outils

La prédiction du niveau d'expertise chirurgicale a été extensivement étudiée à partir d'informations sur les trajectoires et mouvements des instruments chirurgicaux, ainsi que de diverses parties du corps (voir tableau 2.8). Ces données sont récupérées à partir de différents types de capteurs : accéléromètres [56; 57; 109], gyroscope [57], capteurs électromagnétiques [62; 91; 95] et infra-rouge [95]. Ces différents capteurs peuvent être fixés soit sur les instruments, soit sur les parties du corps étudiées que sont les mains[20; 25; 59; 95; 109], les poignets et les coudes [20; 57], et les épaules [20] du chirurgien. Dans le cas de la chirurgie robotisée, ces données sont directement récupérées sur le robot, et donnent accès aux trajectoires des instruments [6; 20; 25; 59] ou des mains [20; 59] du chirurgien.

Une autre approche possible est de travailler à partir d'un enregistrement vidéo. Cet enregistrement peut être endoscopique et ne capturer que les mouvements des instruments et objets manipulés [31; 33; 64; 96; 114; 115], ou bien externe et capturer les mouvements du chirurgien lui-même [9; 38; 90]. Ces vidéos sont ensuite traitées pour en extraire les mouvements des instruments [33; 90; 96] ou des mains [9; 90], la texture du mouvement [90], des métriques statiques ou dynamiques [64]. Ces données sont elles-mêmes utilisées pour

Référence	Type de capteur	Suivi des instruments	Suivi des mains	Suivi des poignets	Suivi des coudes	Suivi des épaules
Loukas and Georgiou (2011)	électromagnétique	√	×	×	×	×
Zhang and Li (2011)	vidéo laparoscopique	×	×	×	×	×
Gray et al. (2012)	vidéo 3D externe	×	×	×	×	×
Kumar et al. (2012)	cinématiques	√	√	×	×	×
Sonnadara et al. (2012)	infrarouge, électromagnétique et vidéo	×	√	×	×	×
Sharma et al. (2014)	vidéo 2D externe	√	√	×	×	×
Watson (2014)	accéléromètre	×	√	×	×	×
Khan et al. (2015)	accéléromètre	√	×	×	×	×
Kirby et al. (2015)	accéléromètre et gyroscope	×	×	√	√	×
Suzuki et al. (2015)	vidéo laparoscopique	√	×	×	×	×
Zhang and Li (2015)	vidéo laparoscopique	×	×	×	×	×
Ahmmad et al. (2016)	cinématiques	√	×	×	×	×
Loukas and Georgiou (2016)	vidéo laparoscopique	×	×	×	×	×
Sharon et al. (2017)	cinématiques et électromagnétique	√	×	×	×	×
Fawaz et al. (2018)	cinématiques	√	√	×	×	×
Ganni et al. (2018)	vidéo laparoscopique	√	×	×	×	×
Azari et al. (2019)	video externe	×	√	×	×	×
Ershad et al. (2019)	cinématique	√	√	√	√	√
Funke et al. (2019)	vidéo laparoscopique	×	×	×	×	×

Tableau 2.8 – Caractérisation des données utilisées par les articles étudiant les trajectoires et mouvement pour la prédiction du niveau d'expertise - √ = oui - × = non.

la prédiction du niveau d'expertise. Dans certains cas les données vidéos sont directement utilisées pour prédire le niveau d'expertise [31; 38; 114; 115].

A noter qu'à partir du moment où des capteurs supplémentaires sont utilisés, il faut les installer dans l'environnement opératoire. Cela pose des problématiques de stérilisation, et de modification du contexte opératoire qui peuvent avoir un impact sur le déroulement normal de l'opération. Bien sûr ces problématiques sont inexistantes dans le cadre de l'entraînement chirurgical, mais bien pregnantes dans des chirurgies réelles. Néanmoins, dans l'optique de simplifier le transfert vers l'environnement clinique, les études n'utilisant que des données vidéos nous paraissent plus pertinentes, a fortiori lorsqu'il s'agit de vidéos endoscopiques ne nécessitant aucune modification de l'environnement opératoire. Bien sûr la contrepartie de ce choix est la complexité de l'extraction d'informations pertinentes à partir de ces données vidéos bruitées, extrêmement variables et très sujettes aux artefacts vidéos [63].

Une exception à la remarque précédente concerne la chirurgie robotisée, et le cas particulier du robot « Da Vinci » développé par l'entreprise Intuitive Surgical [5]. Dans les salles d'opération équipées avec ce type d'installation, on a un accès simple et direct aux données cinématiques des instruments et des mains du chirurgien, dans la mesure où ces données sont librement mises à disposition. Je ne discuterai pas la complexité, l'effectivité ni les coûts liés à l'installation de ce robot.

Dans le tableau 2.9 sont décrites les configurations d'étude clinique des articles. L'ensemble des articles présentés ici montrent une grande diversité dans les types de chirurgie étudiés avec des chirurgies ouvertes [9; 56; 57; 90; 95; 109], laparoscopiques [6; 33; 38; 62; 64; 96; 114; 115] et robotisées [20; 25; 31; 59; 91]. Il serait intéressant d'analyser plus en détail ce en quoi les mouvements et trajectoires diffèrent pour ces différents types de chirurgie.

Réference	Type de chirurgie	Support de pratique	# de tâches	Complexité des tâches	Niveau(x) de pratique étudié(s)
Loukas and Georgiou (2011)	laparoscopie	réalité virtuelle	1	basique	IntJ et IntS
Zhang and Li (2011)	laparoscopie	réalité virtuelle	1	basique	IntJ et IntS
Gray et al. (2012)	laparoscopie	banc d'essai	1	basique	PGY1 – PGY3, PGY5 ou fellow
Kumar et al. (2012)	robotisée	banc d'essai	4	basique	Int, fellow et E
Sonnadara et al. (2012)	ouverte	banc d'essai	1	basique	Ext et E
Sharma et al. (2014)	ouverte	banc d'essai	1	basique	Int et E
Watson (2014)	ouverte	banc d'essai	1	basique	IntJ et IntS
Khan et al. (2015)	ouverte	banc d'essai	1	basique	Int
Kirby et al. (2015)	arthroscopie	fantôme	3	basique	N, I et E
Suzuki et al. (2015)	laparoscopie	banc d'essai	1	basique	N et E
Zhang and Li (2015)	laparoscopie	banc d'essai	1	basique	Int et E
Ahmmad et al. (2016)	laparoscopie	réalité virtuelle	1	basique	non-chirurgien et chirurgien
Loukas and Georgiou (2016)	laparoscopie	réalité virtuelle	2	basique	N et E
Sharon et al. (2017)	robotisée et ouverte	réalité virtuelle et banc d'essai	2	basique	non-chirurgien et E
Fawaz et al. (2018)	robotisée	réalité virtuelle	3	basique	N, I et E
Ganni et al. (2018)	cholécystectomi	réelle	1	réelle	Int et E
Azari et al. (2019)	ouverte	réelle	1	réelle	Int senior et E
Ershad et al. (2019)	robotisée	réalité virtuelle	2	basique	N, Int, I et E
Funke et al. (2019)	robotisée	réalité virtuelle	3	basique	N, I et E

Tableau 2.9 – Description des configurations cliniques mises en place dans les articles étudiant des relations simples entre les trajectoires et le niveau d'expertise - N = novice - IntJ = interne junior, IntS = interne sénior - I = intermédiaire - E = Expert - PGY = Post-Graduate Year.

5. https://www.intuitive.com/

A noter que la grande majorité des études ont prédit le niveau d'expertise sur des tâches d'entraînement basiques.

Enfin, les études les plus récentes dans ce domaine présentent une capacité à distinguer trois niveaux d'expertise lors de chirurgies réelles. Les études [9] et [20] ont mené cette tâche de prédiction sur des données de chirurgies réelles, et ont donc réussi à analyser l'environnement chirurgical réel et complexe à partir de données de trajectoires. Par ailleurs l'approche développée par [31] est tout particulièrement intéressante : à l'aide de réseaux neuronaux profonds, des données vidéos ont été traitées directement sans extraire d'autres formes de données. Cette approche permet d'éviter de faire des hypothèses de travail et de poser des a priori propres aux traitements de vidéos laparoscopiques. A partir d'approches algorithmiques basées sur des données issues de procédures réelles, on est désormais capable de distinguer les niveaux novice, intermédiaire et expert en chirurgie.

Nous présentons ici des constatations extraites des articles précédents. Ces observations sont organisées de la plus générale et la plus validée, à la plus spécifique et la plus novatrice :

— Le niveau de pratique chirurgicale est lié aux trajectoires et mouvements d'outils [6; 33; 38; 56; 59; 64; 90; 91; 96; 115].
— Le niveau de pratique chirurgicale est lié aux trajectoires et mouvements des mains [9; 20; 25; 59; 90; 95; 109], et des articulations des bras [20; 57].
— Le niveau d'expertise est corrélé à la durée de réalisation de la tâche [33; 91; 96; 115].
— La synchronisation des mains de l'élève s'améliore au cours de son entraînement [62].
— Le niveau d'expertise est corrélé à la longueur de chemin parcouru par les outils [96].
— La fréquence de changement d'orientation des outils montre une différence statistiquement significative entre les experts et les novices [91].
— Les métriques liées à la dynamique d'orientation des outils pourraient aider à mieux estimer le niveau de pratique chirurgicale [91].
— Les mouvements des experts sont plus efficients [33].
— Il est possible, en utilisant des données brutes de vidéos laparoscopiques de prédire le niveau d'expertise chirurgicale [31].

L'étude extensive depuis presque dix ans des données de trajectoire extraites de l'environnement chirurgical présente désormais des conclusions observées dans des conditions d'études variées. Ces premiers éléments de réponse sur la pratique chirurgicale et le niveau d'expertise laissent envisager des possibilités de généralisation de ces résultats, ainsi que de production de nouvelles connaissances, décrites dans la section qui suit.

2.3.3.2 | Etude d'analyse de la pratique chirurgicale

Les études présentées ici traitent différents aspects de la pratique chirurgicale et ont des objectifs variés présentés dans le tableau 2.10. Elles présentent surtout des conclusions analytiques portant sur la pratique chirurgicale. Ces conclusions nous apportent des pistes de réflexion intéressantes. Comme les analyses de la pratique chirurgicale proposées dans les études présentées ici vont plus loin que celles de la section précédente, les résultats nous apportent des informations réellement novatrices sur divers aspects de cette pratique chirurgicale, et nourrissent donc notre objectif d'étudier et de comprendre la notion de pratique chirurgicale.

Réference	Objectif de l'étude
Nugent et al. (2012)	Evaluer l'impact des aptitudes visuo-spatiales et psychomotrices sur les performances lors de colectomies réelles.
Lin et al. (2013)	Valider l'utilisabilité et l'efficacité d'un système évaluant les mouvements des bras du chirurgien dans le cadre d'entrainement à la chirurgie.
Loukas et al. (2013)	Evaluer l'impact de la coordination des mains sur les performances lors de cholécystectomies simulées.
Uemura et al. (2014)	Discriminer le niveau d'expertise de chirurgiens experts et novices à l'aide du mouvement de leurs mains
Ahmidi et al. (2015)	Evaluer le niveau d'expertise de chirurgiens experts et novices et apporter un retour d'expérience suite à des chirurgies réelles.
Forestier et al. (2018)	Identifier des schémas de mouvements discriminants pour le niveau d'expertise et cliniquement interprétables.
Fard et al. (2018)	Discriminer le niveau d'expertise de chirurgiens experts et novices à partir de données cinématiques des instruments chirurgicaux.
Jin et al. (2018)	Evaluer le score GOALS à partir du suivi automatique des outils dans les vidéos laparoscopiques.

Tableau 2.10 – Description de l'objectif des études étudiant des relations complexes entre les trajectoires et le niveau d'expertise.

A noter que ces conclusions sont spécifiques aux conditions expérimentales de chaque étude (voir le tableau 2.11). Dans ce qui suit, nous présenterons un discours cohérent sur les connaissances que l'étude des trajectoires peut apporter sur la pratique du geste chirurgical. A travers ce discours, nous faisons l'hypothèse que les conclusions apportées par les différentes études sont généralisables à l'ensemble des pratiques chirurgicales, ce qui n'est rien moins que sûr. Un énorme travail de généralisation serait nécessaire pour valider chacune de nos affirmations à venir.

Réference	Type de chirurgie	Support de pratique	# de tâches	Complexité des tâches	Niveau(x) de pratique étudié(s)
Nugent et al. (2012)	Colectomie	réalité virtuelle	1	simplifiée	N
Lin et al. (2013)	laparoscopique	banc d'essai	2	basique	N et E
Loukas et al. (2013)	Cholécystectomie	réalité virtuelle	1	simplifiée	IntJ et IntS
Uemura et al. (2014)	ouverte	fantôme	1	basique	N et E
Ahmidi et al. (2015)	Septoplastie	réelle	1	réelle	N et E
Forestier et al. (2018)	robotisée	réalité virtuelle	6	basique	N, I et E
Fard et al. (2018)	robotisée	réalité virtuelle	1	basique	N et E
Jin et al. (2018)	Cholécystectomi	réelle	1	réelle	E

Tableau 2.11 – Description des configurations cliniques mises en place dans les articles étudiant des relations complexes entre les trajectoires et le niveau d'expertise - N = novice - IntJ = interne junior, IntS = interne sénior - I = intermédiaire - E = Expert.

La coordination des mains est un indicateur fort du niveau d'expertise : les experts ont des mouvements plus fluides (Loukas et al., 2013) et plus stables (Ahmidi et al., 2015; Fard et al., 2018) que les novices qui montrent des enchaînements plus saccadés (Loukas et al., 2013). Sur une population de chirurgiens confirmés pratiquant une procédure réelle, Jin et al.

(2018) ont également montré que de meilleurs chirurgiens manipuleront généralement leurs outils dans une zone plus restreinte du champ opératoire, avec une plus grande précision et une plus grande économie de mouvement. Ce dernier point est cohérent avec l'étude de Loukas et al. (2013) qui a observé de plus longs enchaînements de gestes chez les experts que chez les novices. Par ailleurs, Fard et al. (2018) ont constaté, en accord avec la littérature clinique (Pouliquen, 2009), l'importance voire la prépondérance des métriques décrivant le comportement de la main auxiliaire vis-à-vis de la main principale.

L'étude de Lin et al. (2013) vient compléter cette analyse en s'intéressant notamment aux trajectoires des bras lors de procédures d'entraînement laparoscopique. Dans cette étude, l'analyse des seuls mouvements des bras ne permet pas l'évaluation du niveau d'expertise mais apporte des informations intéressantes sur la pratique du geste chirurgical. Les experts ont tendance à plus déplacer leurs coudes et poignets que leurs épaules. En comparaison des novices, les experts déplacent moins leur épaule principale, et plus leur épaule auxiliaire, ce qui vient confirmer le résultat de Fard et al. (2018) sur l'importance de la main auxiliaire. Enfin, les experts pratiquent les gestes chirurgicaux en variant plus l'orientation de leurs bras : Fard et al. (2018) constatent que les mouvements de rotation des bras sont plus rapides et plus amples chez les experts.

Les capacités à se projeter dans les mouvements à venir et à planifier sont discriminantes : les experts suivent des schémas de gestes reconnaissables (Loukas et al., 2013), prédictibles (Ahmidi et al., 2015; Fard et al., 2018), et cohérents (Ahmidi et al., 2015). Dans leur façon de gérer la procédure, les experts font aussi preuve d'adaptabilité et de flexibilité (Uemura et al., 2014), et ont des mouvements d'instruments plus décisifs (Ahmidi et al., 2015). A l'inverse, Loukas et al. (2013) ont observé une tendance à improviser chez les novices, ce qui expliquerait en grande partie tous les défauts de manipulation décrit ci-dessus.

Pour aider l'élève à se projeter dans sa pratique et à mieux comprendre ses défauts, les travaux de Forestier et al. (2018) proposent une approche intéressante. En extrayant des schémas de mouvements des trajectoires d'instruments, ils sont capables de déterminer le niveau de pratique, tout en proposant un retour d'expérience au cas par cas. Les segments de trajectoire fortement associés au niveau de pratique sont mis en valeur, mais du fait de leur variabilité aucune conclusion clinique pertinente n'a pu être proposée.

Dans leur étude, Nugent et al. (2012) étudient l'influence des capacités visuo-spatiales et psychomotrices sur les performances des internes. Les capacités visuo-spatiales caractérisent les relations entre la vision et la représentation dans l'espace. Ils constatent que ces capacités sont corrélées à des mouvements plus fluides, et plus rapides. Et il semblerait que, même avec un niveau d'expertise plus avancée, lors de l'apprentissage d'une nouvelle tâche ces aptitudes influencent la capacité à bien réussir la tâche demandée.

2.3.4 | L'étude des yeux et du regard

De même que pour l'étude des trajectoires, nous nous intéressons ici à l'évaluation, la validité et l'analyse des relations entre le niveau de pratique et les yeux du chirurgien. Dans de nombreux secteurs, l'observation des yeux et du regard du sujet permet de caractériser son comportement. On peut ainsi s'intéresser à la cible de l'attention, à la prise de décision, et on peut mesurer d'autre aspects cognitifs comme le niveau d'attention, la fatigue ou le

stress du sujet. Ces informations ont été étudiées pour le chirurgien dans plusieurs études. Le tableau 2.12 présente les expérimentations cliniques de travaux d'intéressant aux yeux et au regard du chirurgien.

Réference	Type de chirurgie	Support de pratique	# de tâches	Complexité des tâches	Niveau(x) de pratique étudié(s)	Données analysées
Ahmidi et al. (2010)	Endoscopie sinus	cadavre	7	basique	novice et expert	cinématique
Richstone et al. (2010)	Laparoscopie	réalité virtuelle	4	variable	non-expert et expert	métriques dérivées de cinématique
Ahmidi et al. (2012)	Endoscopie sinus	cadavre	9	basique	novice et expert	cinématique
Jiang et al. (2013)	laparoscopie	réalité virtuelle	9	basique	non-chirurgien et chirurgien	métriques dérivées d'une vidéo de la pupille
Harvey et al. (2014)	thyroïde lobectomie ouverte	cadavre	1	réelle	interne seniors et experts	cinématique
Snaineh and Seales (2015)	laparoscopie	banc d'essai	1	basique	novice et expert	métriques dérivées de cinématique de vidéo de l'oeil
Tien et al. (2015)	réparation d'hernie inguinale	réelle	1	réelle	junior et sénior	cinématique et métriques dérivées de vidéo de l'oeil

Tableau 2.12 – Description des configurations cliniques mises en place par les articles étudiant l'oeil du chirurgien.

Trois études se sont d'abord intéressées à la pertinence de cette approche. Elles se sont questionnées sur la possibilité de prédire le niveau d'expertise en utilisant des données cinématiques décrivant uniquement le chemin du regard (Richstone et al., 2010), ou accompagnées des données cinématiques des instruments (Ahmidi et al., 2010, 2012). Ces trois études portent soit sur des tâches d'entraînement basiques en endoscopie des sinus (Ahmidi et al., 2010, 2012), soit sur des tâches de difficulté croissante en laparoscopie simulée en réalité virtuelle (Richstone et al., 2010). Ces études montrent que l'analyse du chemin du regard permet une très bonne distinction entre novices et experts (Ahmidi et al., 2010, 2012), voire entre chirurgiens juniors et séniors (Richstone et al., 2010). Ainsi l'analyse du chemin parcouru par le regard du chirurgien est une autre façon d'évaluer le niveau d'expertise chirurgicale.

D'autres études se sont intéressées aux données décrivant le regard (Harvey et al., 2014; Snaineh and Seales, 2015; Tien et al., 2015), et les yeux eux-mêmes (Jiang et al., 2013; Tien et al., 2015). Plus précisément, l'analyse de yeux passe par la mesure des battements de paupières, de la taille et de la variation de taille des pupilles. Ce sont des indicateurs de

l'activité cognitive (Richstone et al., 2010) et du niveau d'attention (Tien et al., 2015) du sujet.

L'étude de ces données mène à des conclusions sur les liens entre le regard et les yeux d'une part, et la pratique du chirurgien d'autre part. Tout d'abord, Richstone et al. (2010) calculent l'Index d'Activité Cognitive, et observent des valeurs plus grandes et donc une activité cognitive plus importante chez les non-experts que chez les experts. Jiang et al. (2013) ont également observé que lorsque la complexité de la tâche augmente, la taille de la pupille est significativement plus grande. Ainsi, lorsque la complexité de la tâche chirurgicale augmente, la charge cognitive du chirurgien augmente, et ce d'autant plus que son niveau d'expertise est bas.

De plus, Tien et al. (2015) ont observé que la taille des pupilles a une plus grande entropie chez les novices. Ce désordre a également été observé dans les objets fixés du regard par les novices, ils passent la moitié du temps de la tâche à regarder des tissus non-identifiés et non-pertinents (Harvey et al., 2014). A l'inverse les experts fixent plus longtemps du regard la bonne cible chirurgicale, et changent moins souvent de cible (Harvey et al., 2014; Tien et al., 2015). L'expert garde plus facilement un haut niveau d'attention, et sait mieux où fixer son regard. En plus d'avoir de meilleures capacités d'attention, les experts savent également mieux anticiper le déroulement de la procédure. Par exemple, lorsqu'il doit demander un nouvel outil à l'infirmière de bloc, l'expert saura garder les yeux fixés sur l'écran, tandis que le chirurgien junior cherchera sur la table de quel outil il a besoin, et arrêtera donc de regarder l'écran (Tien et al., 2015).

Le désordre caractéristique des novices est aussi visible lors d'exercices d'entraînement à la cœlioscopie. Au moment où le novice arrête de regarder l'écran, il continue à bouger ses outils, tandis que l'expert les gardera immobile, évitant ainsi des mouvements à risque (Snaineh and Seales, 2015). Ce manquement peut être lié à un manque de concentration du novice, mais aussi à une mauvaise coordination visuelle et psychomotrice. Snaineh and Seales (2015) ont aussi observé que les novices regardent souvent leurs mains plutôt que l'écran où la tâche est réellement visible. Cela montre la complexité des tâches de laparoscopie en général, et l'importance des capacités visuelles et psychomotrices dans ce genre de tâche.

Ainsi, le manque d'expérience chirurgicale se traduirait par un manque d'aisance et de coordination visuelle et psychomotrice, et par un déficit cognitif d'attention et de prise de décision.

2.3.5 | l'étude des forces

Une autre approche possible de la pratique chirurgicale, et encore une fois de la prédiction du niveau d'expertise, est la mesure des forces appliquées. Différentes configurations expérimentales sont proposées (voir tableau 2.13) : les forces peuvent être mesurée sur une plateforme équipée de capteurs de force sur laquelle on vient effectuer la tâche demandée [37; 44; 45; 46; 83]. [85] ont installé des capteurs de forces directement sur les instruments. [92] proposent une plateforme haptique. [37] ont également complété leur approche en équipant les instruments d'accéléromètres à haute-fréquence afin d'enregistrer les vibrations de

ces instruments. Ces vibrations sont mesurées dès que l'instrument entre en contact avec un autre objet.

Réference	Type de chirurgie	Support de pratique	# de tâches	Complexité des tâches	Niveau(x) de pratique étudié(s)	Capteurs utilisés	interaction outil/tissus étudiée
[85]	fundoplication	modèle porcin	1	réelle	N et E	3DOF capteurs de force sur instrument	manipulation et dissection
[46]	laparoscopie	banc d'essai	1	basique	N et E	3DOF forces plateforme	suture
[44]	laparoscopie	banc d'essai	1	basique	N et E	3DOF forces plateforme	suture
[45]	laparoscopie	banc d'essai	2	basique	Ext, Int et E	3DOF forces plateforme	manipulation et déplacement
[83]	canulation d'artère	fantôme	3	simplifiée	ChirJ et ChirS	3DOF forces plateforme	cathétérisme
[92]	robotisée	réalité virtuelle	20	basique	N et E	interface haptique	pénétration
[37]	robotisée	réalité virtuelle	3	basique		3DOF forces plateforme et HF accéléromètres sur instruments	manipulation et suture

Tableau 2.13 – Description des configurations cliniques mises en place dans les articles étudiant les forces appliquées par le chirurgien - N = novice - IntJ = interne junior, IntS = interne sénior - I = intermédiaire - E = Expert - ChirJ = Chirurgien Junior - ChirS = Chirurgien Sénior - 3DOF = 3 Degrees of Freedom - HF = High Frequence.

Dans plusieurs études, ces forces sont un complément utile aux données de trajectoire [44] et de durée [45] pour la prédiction du niveau d'expertise. Utilisées seules, les forces donnent des résultats de prédiction encourageants dans le cas d'une tâche simulée de pose de cathéter dans des artères [83]. L'étude de [37] montre que des métriques basées sur les mesures de vibrations permettent une évaluation objective et valide de construction du niveau d'expertise.

La bonne intensité de force à appliquer, ou la direction dans laquelle cette force doit être appliquée diffèrent selon l'expérimentation clinique et la tâche effectuée. Dans le cas d'une tâche complexe sur modèle porcin, et donc sur tissus vivants, [85] ont observé que les experts appliquaient de plus grandes forces que les novices lors de tâches de dissection, mais des forces moindres lors de tâches de manipulation de tissus. Cela s'expliquerait par une plus grande assurance et connaissance des forces à appliquer lors de la dissection, mais de plus grandes précautions lors de la manipulation des tissus.

Lors de tâches de manipulations impliquant de déplacer les tissus sans les abîmer, plusieurs études s'accordent à dire que les experts appliquent des forces moindres sur les tissus que les novices [37; 85]. A l'inverse, lors de tâches impliquant de disséquer des tissus [85] ou de transpercer des tissus [92], les experts appliquent de plus grandes forces que les novices. A noter que l'étude de [85] s'est faite sur des tissus vivants, alors que l'étude de [92]

s'intéresse à la construction du ressenti haptique du chirurgien sur une interface dédiée. Ces deux résultats ont donc été observés dans des conditions expérimentales complètement différentes. On ne peut donc conclure et des études complémentaires sont nécessaires pour mieux caractériser les profils de forces d'experts et de novices sur des tâches de dissection.

Les études [44; 46] portant sur la suture concluent que cette tâche est menée par les experts en appliquant des forces moindres que les novices, tandis que l'étude de [37] ne permet pas de conclure de tendance significative. A mon sens, les tâches de suture et de dissection se ressemblent en terme de forces appliquées aux tissus, car il s'agit dans les deux cas de venir appliquer une force suffisante en un point précis du tissu, soit pour les transpercer, soit pour les séparer. Or les conclusions proposées par [45] contredisent celles de [85]. Il nous est donc ici aussi impossible de conclure sur les différences entre les profils de forces d'experts et les profils de forces de novices.

En revanche, plusieurs études [37; 44; 92] s'accordent sur le fait que les experts ont des profils de forces présentant une variance moindre que ceux des novices. En effet, il semble logique que les experts ressentent et maîtrisent mieux les forces minimale et maximale qu'ils appliquent aux tissus, alors que, par inexpérience, les novices testent une plus grande amplitude de force, et ont une sensibilité moins fine que les experts.

Finalement, l'étude des forces appliquées lors de tâches chirurgicales est un champ de recherche présentant des conclusions qui doivent encore être confirmées. Il semble que les conditions expérimentales aient un impact important sur les résultats des études. On notera notamment que les bancs d'essai et plateformes dédiées à la mesure de forces ne sont pas tous basés sur les mêmes techniques d'acquisition. Il est donc peu probable que les données de forces extraites à partir de ces différents appareils soient homogènes ou comparables. Cela pourrait expliquer en partie la diversité, voire les contradictions entre les résultats discutés précédemment. La comparaison de ces différentes solutions de mesure, voire le développement d'un environnement performant et robuste serait complètement pertinents.

Le choix de la tâche chirurgicale étudiée est également important. Il semble normal que les profils de forces optimaux varient d'un type de tâche à l'autre. Encore plus que pour l'étude des trajectoires ou des yeux, il semble que l'étude des forces soit sensible au support de pratique. Les profils de forces obtenus sur un fantôme constitué de tissus synthétiques, sur un modèle porcin, un cadavre ou un vrai patient auront certainement des différences marquées.

Enfin, l'entraînement à la chirurgie n'est aujourd'hui pas conçu pour apprendre à maîtriser ces forces appliquées. Les tâches d'entraînement dédiées à la maîtrise des forces appliquées proposées par [45], ou l'étude du ressenti haptique de [92] vont dans le sens du développement d'outils dédiés. Ces outils permettraient aux internes d'axer leur formation sur la maîtrise des forces, mais permettraient également d'étudier plus avant cet aspect de la pratique chirurgicale.

2.3.6 | L'analyse procédurale

Plusieurs états de l'art ont déjà traité extensivement de la modélisation et de la reconnaissance de la procédure chirurgicale (Lalys and Jannin, 2014; Pernek and Ferscha, 2017). La prédiction du niveau d'expertise y sert pour évaluer la validité de construction de l'ap-

proche proposée. La pertinence de l'étude procédurale pour analyser le niveau d'expertise et la pratique chirurgicale n'est donc plus à prouver. Nous nous intéresserons ici aux aspects procéduraux de l'expertise et de la pratique chirurgicale. Le tableau 2.14 présente les configurations cliniques des travaux étudiant ces aspects procéduraux.

Réference	Type de chirurgie	Support de pratique	# de tâches	Niveau(x) de pratique étudié(s)	granularité procédurale
Tao et al. (2012)	robotisée	réalité virtuelle	4	N, I et E	manoeuvre
Uemura et al. (2016)	laparoscopie	banc d'essai	1	N, I et E	activité
Vedula et al. (2016)	robotisée	réalité virtuelle	1	N et E	activité, manoeuvre et geste
Huaulmé et al. (2018)	microchirurgie robotisée	fantôme	1	N et E	activité
Wang and Fey (2018)	robotisée	réalité virtuelle	3	N, I et E	activité
Zia and Essa (2018)	robotisée	réalité virtuelle	3	N, I et E	activité

Tableau 2.14 – Description des configurations cliniques mises en place dans les articles étudiant l'aspect procédural de la chirurgie - N = novice - IntJ = interne junior, IntS = interne sénior - I = intermédiaire - E = Expert.

Dans la section 2.3.3, nous avions déjà observé la corrélation entre la durée opératoire et le niveau d'expertise. L'analyse procédurale confirme cette observation : les experts sont plus rapides que les novices. Cela est dû à la meilleure planification et anticipation des experts pendant la procédure en elle-même, notamment lors des temps de transition entre mouvements pendant lesquels le chirurgien pense au mouvement futur. Ces temps de transition sont plus courts chez l'expert (Uemura et al., 2016; Vedula et al., 2016). La capacité de planification a également été observée par Huaulmé et al. (2018) à travers ce qu'ils nomment les « séquences d'activités signatures ». Ces enchaînements d'activités reconnaissables et caractéristiques sont plus longs chez l'expert, montrant sa capacité à mieux anticiper le déroulement de la procédure. Ces séquences d'activités signatures apparaissent aussi plus souvent chez l'expert : sa pratique est plus prédictible, moins chaotique que celle du novice. Nous retrouvons ici une autre observation de la section 2.3.3.

En comparant les procédures d'internes à différents stades de leur entraînement Vedula et al. (2016), s'intéressent à la courbe d'apprentissage des chirurgiens en formation. Ils observent que dans un même exercice d'entraînement, ici une suture, les mouvements n'ont pas tous la même difficulté, et que les internes apprennent à maitriser les différents aspects de cet exercice à différents stades de leur entraînement. Ainsi l'activité « faire un noeud » pose problème aux internes débutants, tandis que les internes plus avancés dans leur entraînement font aussi bien que les experts. Et l'activité « passer le fil dans l'ouverture » pose autant de problèmes aux internes confirmés qu'aux internes débutants, qui sont significativement moins efficients que les experts. Il est donc important d'adapter l'entraînement au niveau de l'étudiant, et de lui prodiguer des conseils en conséquence.

En résumé

— L'évaluation de niveaux d'expertise novice, intermédiaire et expert est aujourd'hui possible à partir du traitement algorithmique de données issues de procédures chirurgicales réelles. En menant une évaluation par la foule, on est aujourd'hui capable de reproduire une évaluation comparable à celle d'un groupe d'experts sur score structuré pour des extraits de chirurgies réelles.

— l'étude avancée des trajectoires dans différentes tâches chirurgicales nous apprend que, comparés aux novices, les chirurgiens experts ont une pratique plus fluide, décisive et économe en mouvements, ainsi qu'une meilleure capacité d'anticipation et moins d'hésitations.

— l'étude des yeux et du regard du chirurgien nous permet de complémenter ces connaissances. On constate que, comparés aux novices, les chirurgiens experts ont une moins grande charge cognitive à gérer qui traduit leur experience et leur habitude des situations rencontrées. Le regard et les yeux des experts traduisent une meilleure focalisation sur des cibles pertinentes, et une pratique plus ordonnnée exempte de mouvements inutiles.

— l'étude des forces appliquées par les instruments du chirurgien sur les tissus montre une meilleure maîtrise de ces forces chez les experts, ainsi qu'une plus grande sensibilité. Le manque de validation et de robustesse des techniques d'acquisition de ces forces impacte la cohérence des analyses en découlant. L'étude de ces forces est néanmoins très prometteuse, et serait un apport pertinent à la formation au geste chirurgical.

— l'analyse procédurale vient confirmer certains résultats tels que la plus grande efficacité des experts, ainsi que leur meilleure capacité d'anticipation. Cet axe de recherche montre un potentiel d'analyse très encourageant et très riche pour mieux comprendre les ressorts de la pratique chirurgicale.

2.4 | Positionnement et objectifs de la thèse

2.4.1 | Positionnement

L'environnement chirurgical, et plus précisément celui de la formation à la chirurgie connait aujourd'hui des transformations notamment sur la place donnée à l'évaluation des compétences chirurgicales. Cet environnement est propice au développement d'outils de simulation des conditions opératoires, ainsi que d'outils d'évaluation et d'analyse de différentes facettes de la pratique chirurgicale.

Nous avons présenté une taxonomie du positionnement des études et de leurs objectifs vis-à-vis de la notion de pratique chirurgicale, entre « prédiction », « analyse » et « validation » de la pratique chirurgicale. Nous avons vu que la prédiction était étudiée en priorité, car c'est la tâche primaire à étudier avant d'envisager les deux autres. L'analyse de la pratique chirurgicale est moins étudiée, et la validation encore moins. Le domaine de l'analyse de données chirurgicales propose actuellement plusieurs défis qui axent nos recherches. D'une part, de nombreuses études cherchent à automatiser l'extraction de données complexes à partir de flux de données brutes issues de l'environnement chirurgicales, mais du fait de la variabilité et des difficultés inhérentes à cet environnement, cet objectif n'a pas encore été atteint dans des conditions chirurgicales réalistes. L'annotation manuelle de ces données est donc une solution logique pour envisager une analyse du phénomène chirurgical.

D'autre part, les analyses proposées étudient majoritairement le niveau d'expertise chirurgicale, or cet indicateur, pertinent certes, n'est qu'un aspect de la pratique chirurgicale parmi tant d'autres. Par exemple, on a vu certaines études s'intéresser à la qualité de pratique du chirurgien (El Ahmadieh et al., 2014; Khan et al., 2015; Malpani et al., 2015; Matsuda et al., 2014) ou au profil de pratique du chirurgien (Ershad et al., 2016; Huaulmé et al., 2018). Nous avons donc axé nos travaux sur ces deux aspects de la pratique chirurgicale.

Un autre apport potentiel de ces travaux est d'aider à l'amélioration des connaissances sur la pratique du chirurgien. De fait dans cet état de l'art, j'ai mené un travail de synthèse des connaissances obtenues par analyse algorithmique sur des données extraites de tâches chirurgicales. Ces analyses portaient surtout sur le niveau d'expertise, mais également sur les scores structurés, la courbe d'apprentissage des internes, la qualité de pratique ou le profil de pratique du chirurgien, et la complexité de la tâche chirurgicale. J'ai fait apparaître la cohérence des connaissances recuillies dans ces différents travaux ainsi que la variété des conclusions apportées par ces analyses de la pratique chirurgicale. Nous nous attacherons à proposer de telles conclusions sur la pratique chirurgicale.

L'analyse de la pratique chirurgicale à partir de données issues de tâches chirurgicales est donc un domaine plein d'opportunités. Que ce soit le phénomène complexe de la pratique chirurgicale, dont de nombreux aspects n'ont pas encore été étudiés. Les données chirurgicales utilisées laissent également envisager des possibilités dans le niveau de détail ou les types de données exploités. Comme nous l'avons vu en section 2.2.1, la validation des approches proposées est également une problématique d'avenir.

2.4.2 | Objectifs

Nos travaux sont découpés en trois études qui traiteront de l'annotation, de la prédiction, et enfin de l'analyse de la pratique chirurgicale. En parallèle, nous nous attacherons à valider différents aspects de notre méthodologie, et nous finirons notamment en proposant une méthodologie de validation clinique. Pour chaque étude, nous présentons dans ce qui suit un objectif principal et plusieurs objectifs secondaires.

2.4.2.1 | Création d'un jeu de données annotées

Objectif principal :

Création d'un jeu de données annotées multiforme à partir de chirurgies réelles.

Objectifs secondaires :

i Définition du protocole d'étude clinique.

ii Création de la cohorte de patients.

iii Modélisation du cadre sémantique permettant de conceptualiser l'annotation future, et notamment de définir des descripteurs de la pratique chirurgicale.

iv Conception d'une interface dédiée à la segmentation d'images.

v Annotation manuelle des aspect procédural et visuel de la vidéo cœlioscopique, ainsi que des critères décrivant la pratique chirurgicale.

vi Etude de la variabilité inter- et intra-opérateur de ces processus d'annotation.

vii Comparaison de notre jeu de données annotées avec les autres jeux de données similaires en libre accès du domaine.

2.4.2.2 | Prédiction des critères de la pratique chirurgicale.

Objectif principal :

Optimisation d'un algorithme d'apprentissage dédié à prédiction de la qualité d'exposition de la scène chirurgicale à partir de variables issues des données annotées.

Objectifs secondaires :

i Définition de métriques interprétables à partir des différentes données annotées.

ii Extraction d'une matrice d'entrée et d'un vecteur de sortie pour le processus algorithmique.

iii Conception d'un processus algorithmique dédié à la prédiction du vecteur de sortie à partir des données d'entrée.

iv Implémentation d'un environnement d'optimisation de notre algorithme répondant aux contraintes spécifiques à nos données.

2.4.2.3 | Analyse et interprétation clinique à partir de notre algorithme de prédiction

Objectif principal :

Proposer et valider des interprétations compréhensible cliniquement à partir de nos données annotées traitées dans notre algorithme prédictif.

Objectifs secondaires :

i Définir une méthodologie permettant de sélectionner les variables les plus importantes pour la tâche de prédiction.

ii Définition d'une métrique caractérisant la capacité d'une variable en entrée de l'algorithme à prédire le score en sortie.

iii Traduire ces résultats d'importance sous une forme compréhensible cliniquement.

iv Définir un cadre expérimental permettant d'analyser différentes configurations de données.

v Evaluer la validité clinique de ces résultats auprès de chirurgiens.

3

Étude 1 : création du jeu de données annotées issues de vidéos de chirurgies laparoscopiques

Préambule

Mon travail de thèse a nécessité la construction d'un jeu de données exploitable dans le cadre de traitements algorithmiques. Cette construction se base sur les données issues d'une étude clinique présentée en section 3.1. Afin de rendre ces données exploitables, un formalisme a été défini en section 3.2 et une méthodologie d'annotation a été suivie pour obtenir un jeu de données annotées (voir section 3.3). Le jeu de données résultant de cette annotation a ensuite été présenté en section 3.4, puis a été comparé aux autres jeux de données annotées existants en section 3.5, et a été discuté en section 3.6. Enfin nous avons conclus sur la création de ce jeu de données annotées en section 3.7.

Sommaire

3.1 | La création du jeu de données cliniques

Mon travail de thèse est basé sur un cas d'étude clinique : la gastrectomie longitudinale, et un jeu de données cliniques obtenu dans des conditions opératoires réelles sur une cohorte de patients. Ici, je décrirai le contexte clinique et la technique opératoire étudiée. Je préciserai ensuite le protocole d'étude clinique au sein duquel la cohorte de patient a été constituée. Et je terminerai en décrivant les données recueillies à partir de cette cohorte.

3.1.1 | Cas d'étude clinique : La gastrectomie longitudinale

La gastrectomie longitudinale, aussi appelée « sleeve » pour « Sleeve Gastrectomy » en anglais, est une technique chirurgicale bariatrique. La chirurgie bariatrique regroupe toutes les techniques chirurgicales appliquées au système digestif dans le but de traiter l'obésité. La chirurgie bariatrique s'effectue le plus souvent en cœlioscopie (chirurgie minimale invasive), et rarement en laparotomie (chirurgie ouverte).

3.1.1.1 | L'obésité, un problème de santé publique majeur

Selon l'Organisation Mondiale de la Santé (OMS) en 2018 :

> « A l'échelle mondiale, le nombre de cas d'obésité a presque triplé depuis 1975. »
>
> « En 2016, plus de 1,9 milliard d'adultes étaient en surpoids. Sur ce total, plus de 650 millions étaient obèses. »
>
> « En 2016, plus de 340 millions d'enfants et d'adolescents âgés de 5 à 19 ans étaient en surpoids ou obèses. »

Aujourd'hui, l'obésité a donc le statut d'épidémie mondiale (on Obesity, 2003), notamment dans les pays développés qui sont les plus touchés.

Chez l'adulte, le surpoids se définit par un Indice de Masse Corporelle (*IMC*) supérieur à $25 kg/m^2$, l'obésité par un *IMC* supérieur à $30 kg/m^2$, et l'obésité morbide par un *IMC* supérieur à $40 kg/m^2$ (on Obesity, 2003). Pour rappel, l'*IMC* correspond au rapport du poids (en kg) sur le carré de la taille (en m).

Par le passé, cette maladie était avant tout un signe d'opulence et de prospérité. Dans un monde où l'on craignait et subissait les famines, la prise de poids et les réserves de graisses étaient des signes de bonne santé. Aujourd'hui, dans les pays développés et en développement, l'obésité est désormais considérée comme une maladie chronique, un handicap, voire les deux. Dans nos sociétés modernes, l'obésité s'explique surtout par la surabondance de nourriture, ainsi que par la sédentarité des modes de vie (on Obesity, 2003).

Les solutions proposées pour s'attaquer à ce problème de santé publique sont nombreuses et variées. Il s'agit tant de prévenir cette maladie auprès de la population, que de prendre en charge les personnes souffrant de surpoids et d'obésité, et d'établir des partenariats entre les différents secteurs responsables (pouvoirs publics, consommateurs, industries, médias). La prise en charge du patient obèse est complexe, car elle implique son mode de vie, avec ses habitudes alimentaires, son activité physique, sa psychologie, et son environnement social. Le traitement standard de l'obésité est donc avant tout un traitement diététique s'accompagnant d'une thérapie comportementale. Divers traitements médicamenteux

existent, mais leurs effets secondaires n'en font pas, aujourd'hui, des traitements viables. La chirurgie bariatrique est la seule modalité de traitement alternative viable présentant de bons résultats en terme de perte de poids, et de maintien de cette perte de poids dans la durée (Thereaux et al., 2010).

3.1.1.2 | L'apport de la chirurgie bariatrique au traitement de l'obésité

En chirurgie bariatrique, l'objectif est de modifier le parcours des aliments au niveau de l'estomac et/ou de l'intestin grêle, afin d'induire une réduction de la quantité de nourriture absorbée. Pour ce faire, deux leviers existent sur lesquels on peut influer (Thereaux et al., 2010) :

— L'approche « restrictive » par laquelle on limite physiquement la quantité de nourriture absorbée. Ce sont notamment les techniques de gastroplastie par anneau, et de gastrectomie longitudinale.
— L'approche « malabsorptive » dans laquelle on court-circuite le trajet des aliments et on limite donc leur absorption par l'organisme. Ce sont notamment les techniques de court-ciruit gastrique ou « Bypass » et de diversion biliopancréatique.

A noter que ces approches chirurgicales ne sont proposées qu'après qu'une thérapie standard ait échoué à induire une perte de poids suffisante sur une période de 6 à 12 mois, et à maintenir cette perte de poids. De plus, ce type d'intervention nécessite par la suite un suivi médical et chirurgical à long terme. Le chirurgien propose une approche chirurgicale aux patients ayant un $IMC \geq 40 kg/m^2$ ou bien un $IMC \geq 35 kg/m^2$ associé à une ou plusieurs comorbidités (Thereaux et al., 2010).

Depuis que ces techniques ont été adaptées en cœlioscopie, elles induisent nettement moins de complications, une réduction de la mortalité post-opératoire et permettent des temps ambulatoires raccourcis. En France, le succès des approches bariatriques est clair : entre 2006 et 2013, le nombre d'opérations a triplé, passant de moins de 15 000 opérations à plus de 42 000. La gastrectomie longitudinale enregistre l'évolution la plus spectaculaire car sur cette même période, le nombre d'opérations a été multiplié par 24, et en 2013 56% des opérations bariatriques sont des gastrectomies longitudinales (Schaaf et al., 2015).

3.1.1.3 | La gastrectomie longitudinale

Du fait de l'effectivité du traitement chirurgical de l'obésité, et du fort engouement actuel autour de la gastrectomie longitudinale, cette technique nous a paru être un sujet d'étude pertinent. De plus, c'est une technique relativement récente car elle a été décrite pour la première fois en 1993 (Marceau et al., 1993), comme intervention préparatoire à un court-circuit gastrique : l'idée initiale était de pratiquer la gastrectomie longitudinale dans des cas d'obésités morbides nécessitant une intervention préparatoire induisant une perte de poids suffisante pour permettre, ensuite, de pratiquer un court-circuit gastrique comme opération principale. La gastrectomie longitudinale a ensuite été proposée de façon autonome comme traitement de l'obésité en raison de ses avantages vis-à-vis des autres chirurgies : c'est une technique moins complexe pour le chirurgien, relativement au court-circuit gastrique notamment, avec des complications post-opératoires moins nombreuses et moins graves, et

induisant une perte de poids conséquente. Pour décrire cette technique, nous nous basons sur les travaux de Verhaeghe et al. (2011).

Présentation de la technique chirurgicale Les termes anatomiques utilisés par la suite sont illustrés dans la figure 3.1. Le principe général de la gastrectomie est assez simple : on réduit le volume de l'estomac en enlevant l'antre gastrique (le corps de l'estomac) sur toute sa longueur. L'estomac, qui a normalement la forme d'une outre, est transformé en tube (voir Figure 3.2b). Cette transformation permet avant tout une réduction du volume d'aliments que l'estomac peut contenir, et ainsi une réduction des apports alimentaires. Un autre effet de la gastrectomie longitudinale encore mal compris est l'action anorexigénique. Cette action est liée à la forte réduction de production de l'hormone orexigénique dans le fundus (partie supérieur de l'estomac (voir figure 3.1). En retirant cette zone de l'estomac, on coupe la production de cette hormone responsable de la sensation de faim, ce qui induit également une diminution des apports alimentaires et amplifie d'autant l'efficacité de cette opération.

Cette technique, aujourd'hui menée sous cœlioscopie, consiste en 6 phases : l'approche chirurgicale, l'exposition de l'estomac, la dissection de la grande courbure de l'estomac, la résection de l'antre gastrique, le nettoyage de la cavité abdominale, et la fermeture de la cavité abdominale. Les deux phases clés sont la dissection et la résection, que nous allons décrire plus en détail.

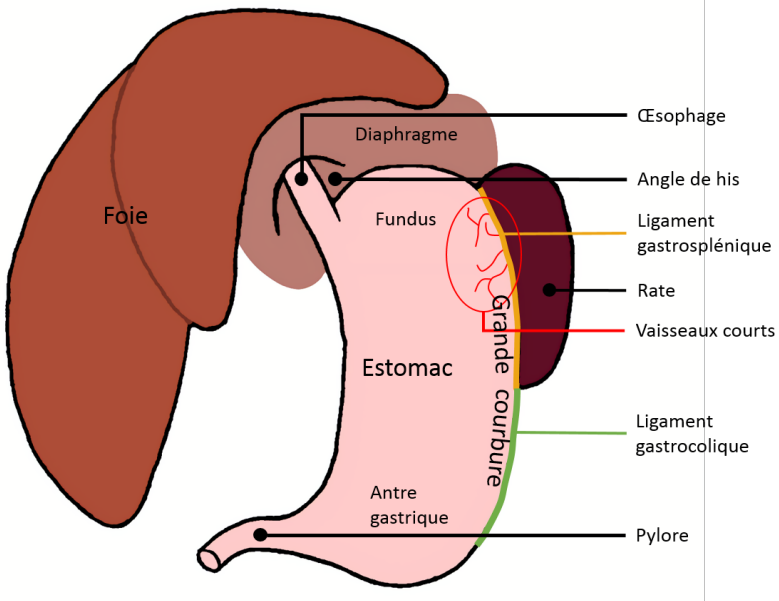
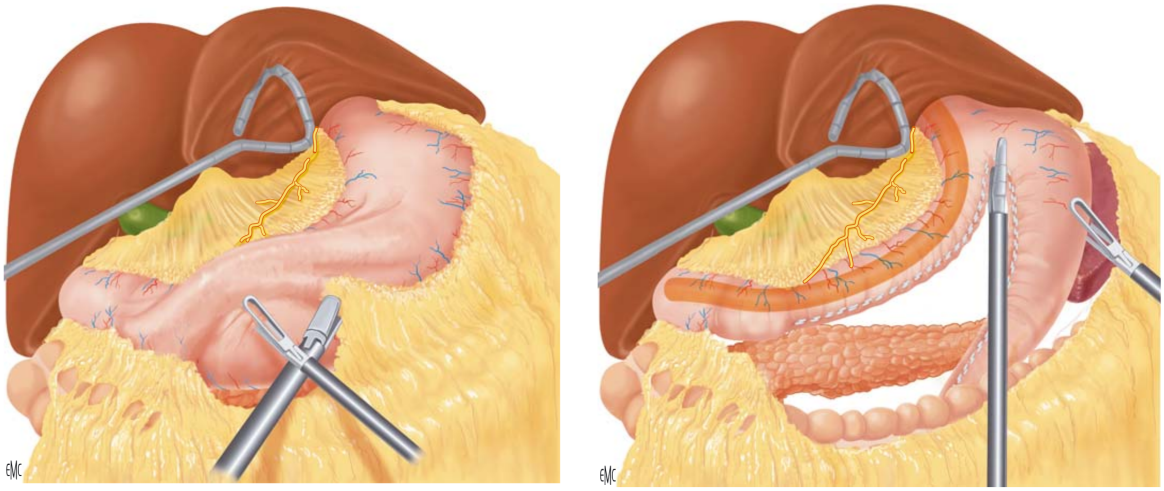


FIGURE 3.1 – Schéma anatomique de l'estomac avec annotation des principaux repères anatomiques.



(a) (b)

FIGURE 3.2 – Illustrations de deux phases critiques de la gastrectomie longitudinale : (a) la dissection de l'estomac en libérant la grande courbure des tissus adipeux qui l'entourent et (b) la résection du corps de l'estomac à l'aide d'une pince agrapheuse (Verhaeghe et al., 2011).

L'étape de dissection Lors de la dissection, l'idée est de libérer sur toute sa longueur la grande courbure de l'estomac des tissus adipeux qui l'entourent (voir figure 3.2a). Ainsi, lorsqu'on viendra ensuite réséquer l'antre gastrique, celle-ci ne sera plus reliée à aucun tissu. Tout le long de la grande courbure, l'estomac est relié aux tissus adipeux par un ligament, appelé ligament gastrocolique pour la partie inférieure de l'estomac, et ligament gastrosplénique pour sa partie supérieure (à proximité de la rate). Ce sont précisément ces ligaments qui sont disséqués.

Classiquement, on commence la dissection par la partie inférieure de l'estomac, en partant à une distance de 4 à 6 cm du pylore. On dissèque ensuite en remontant tout le long de la grande courbure. Enfin, on dissèque le fundus du ligament gastrocolique, du vaisseau court, puis du ligament gastrocolique et on est bien attentif à séparer le fundus du muscle du diaphragme. La dissection du fundus est terminée une fois qu'on peut voir l'angle de His (marquant la séparation entre l'estomac et l'œsophage). On finit la phase de dissection en vérifiant que l'estomac a bien été libéré du méso-gastro postérieur ainsi que des adhérences congénitales potentielles sur sa face postérieure. La bonne dissection du fundus et l'exposition de l'angle de His sont importantes pour réduire les risques de complications.

Lors de cette étape, l'assistant du chirurgien se sert d'un écarteur de foie pour bien exposer l'estomac en soulevant le foie (voir Fig. 3.3a). Le chirurgien, lui, utilise deux outils :

— *La pince électro-thermale :* Cet outil permet le geste de dissection, il fonctionne par effet électrothermique, il découpe et cautérise le ligament en un geste. Cet outil est traumatique et ne doit pas manipuler directement l'estomac (voir Fig. 3.3b).
— *La pince atraumatique ou « flat grasper » :* cet outil permet de manipuler l'estomac en exposant correctement la bordure disséquée de l'estomac afin d'assister l'autre main (voir Fig. 3.3c).

(a) (b) (c) (d)

FIGURE 3.3 – Illustrations des principaux outils utilisés lors de la gastrectomie longitudinale : (a) l'écarteur de foie[1], (b) la pince électro-thermale[2], (c) la pince atraumatique[3] et (d) la pince agrafeuse[4]

En soutien de ces outils, le chirurgien peut également utiliser une compresse pour « caler » les tissus adipeux qui l'empêchent de bien exposer le ligament à disséquer.

L'étape de résection Suite à la dissection de l'estomac, le chirurgien résèque le corps de l'estomac, c'est-à-dire qu'il enlève cette partie de l'estomac en la découpant (voir figure 3.2b). Au préalable, le chirurgien introduit par voie orale une bougie de diamètre *34 french* (unité de mesure) jusqu'au duodénum. Cette bougie sert de « patron » au chirurgien qui restera proche de cette bougie lors de la découpe, afin de donner à l'estomac transformé la forme d'un tube de diamètre bien défini. Pour cette résection, le chirurgien échange la pince électro-thermale contre une pince agrafeuse (voir Fig. 3.3d). La résection commence par la partie inférieure de l'estomac, et la pince agrafeuse s'utilise de la façon suivante :

1. Le chirurgien insère la pince agrafeuse dans la cavité abdominale,
2. il vient attraper l'estomac avec la pince agrafeuse, et la serre contre la bougie,
3. il ferme la pince et la verrouille en position fermée,
4. il enclenche la lame de découpe sur tout la longueur de la pince,
5. il enclenche l'agrafage,
6. il relâche les tissus de l'estomac,
7. il extrait la pince agrafeuse de la cavité abdominale.

Cet enchaînement est répété 5 à 6 fois avec une nouvelle pince agrafeuse à chaque fois. Le dernier geste de résection par lequel on vient entièrement séparer le fundus du corps de l'estomac est critique. En effet lors de cette dernière découpe, il faut bien réséquer le fundus en entier pour contrer au maximum l'action orexigénique, mais également faire attention à limiter la tension dans les tissus au niveau de l'angle de His qui fait la liaison entre l'estomac et l'œsophage. Si cette tension est trop importante, on peut craindre des complications post-opératoires telles que le reflux gastroœsophagien ou la fistule gastrique.

1. https ://www.bariatric-solutions.com/images/products-cobra-liver-retractor.png
2. https ://cdn2.bigcommerce.com/n-nr1m3w/22fhhbs/products/225/images/440/ls1020_covidien__08765.1441719235.128
3. https ://catalog1.bd.com/media/catalog/product/2/5/256.01010U-0003.jpg
4. https ://www.medtronic.com/content/dam/covidien/library/global/en/product/surgical-stapling/endo-gia-black-reload-b.jpg

3.1.2 | Méthodologie : Le protocole d'étude clinique

Le bon déroulement de ma thèse a nécessité de constituer un jeu de données à partir duquel j'ai mené plusieurs études. Nous avons donc travaillé sur des données cliniques afin d'étudier une notion clinique : la pratique du geste chirurgical. Ces données cliniques ont été recueillies en nous basant sur la cohorte de patients créée dans le cadre d'un protocole d'étude clinique mis en place au CHU de Grenoble.

Le processus d'approbation de notre étude clinique a été le suivant :

1. Un dossier a été constitué afin de décrire l'étude, avec l'aide du Centre d'Investigation Clinique – Innovation Technologique (*CIC-IT*) du CHU de Grenoble.
2. Le dossier complet a été soumis au Comité d'Ethique des Centres d'Investigation Clinique de l'inter-région Rhône-Alpes-Auvergne, afin d'obtenir un avis sur la faisabilité de l'étude.
3. Une fois qu'un avis favorable a été obtenu auprès du comité d'éthique, les données ont pu être recueillies au sein du CHU-Grenoble.

Le dossier devait détailler les objectifs de la recherche, la population concernée par l'étude, la méthode de recueil de données, et les types de données recueillies.

3.1.3 | Résultat : L'étude clinique LapEx

3.1.3.1 | Approbation de l'étude clinique

Notre étude et le jeu de données associé s'appellent *LapEx* (pour LAParoscopic EXpertise). C'est une étude mono-centrique n'impliquant pas la personne humaine, menée sur des données rétrospectives. En effet, le jeu de données a été recueilli avant que l'étude n'ait été approuvée par le comité éthique, ou que la déclaration à la CNIL n'ait été faite. Le responsable scientifique de l'étude est le docteur Fabian Reche, chirurgien au sein du service de chirurgie digestive et de l'urgence au CHU-Grenoble. Le docteur Fabian Reche a été impliqué tout au long de mon travail de thèse pour son expertise en cœlioscopie et en chirurgie bariatrique.

L'objectif de cette étude recoupe ceux de ma thèse : en étudiant la gastrectomie longitudinale à partir de données vidéos et cliniques, nous cherchions à améliorer les pratiques professionnelles et l'enseignement de cette procédure.

La population cible devait être éligible vis-à-vis de critères d'inclusion/non-inclusion portant notamment sur l'âge, l'IMC, les maladies chroniques, l'historique chirurgical, ... du patient. Le patient devait aussi remplir et signer un formulaire d'information et de non-opposition à l'utilisation de ses données dans le cadre défini par l'étude. Le patient a un droit de rétractation.

Les données ont été recueillies depuis l'inclusion du patient dans l'étude jusqu'à la fin de son suivi post-opératoire. Le recueil des données concernait également les phases pré- et peropératoire. Nous avons surtout enregistré la vidéo endoscopique tout au long de l'opération chirurgicale.

Le jeu de données est géré par le CIC-IT qui a assuré l'anonymisation des données, avant de les mettre à disposition des deux laboratoires partenaires : le TIMC, et le LTSI.

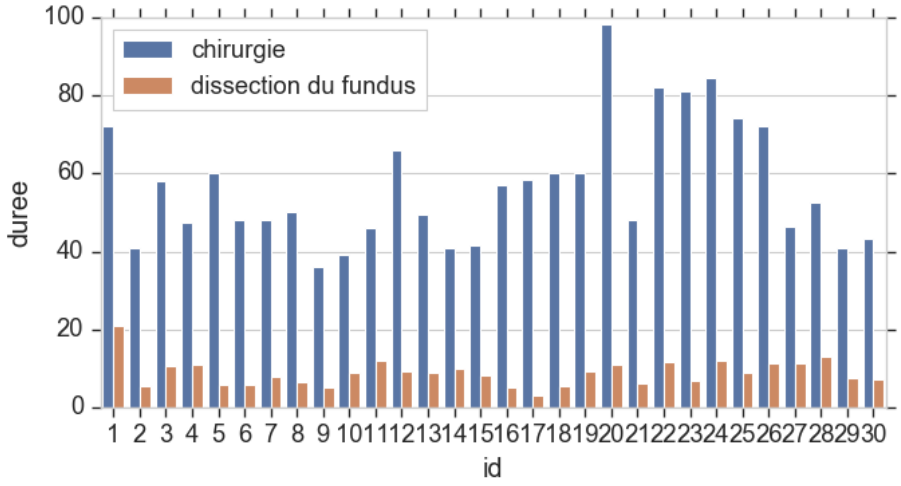


FIGURE 3.4 – Durée totale enregistrée de chirurgie, durée de l'étape de dissection du fundus pour chaque opération de la cohorte.

	Réduction d'IMC (en kg/m^2)	Perte de Poids (en kg)
maximum	26	62
moyenne/ecart-type	13 ± 5.1	35.7 ± 14.1
minimum	2	6

Tableau 3.1 – Statistiques sur l'évolution du poids et de l'IMC des 30 patients de l'étude entre la première visite médicale, et la visite de fin de suivi, 6 mois plus tard.

Un avis éthique consultatif favorable a été obtenu le 24/05/2018 (CECIC Rhône-Alpes-Auvergne, Clermont-Ferrand, IRB 5891).

3.1.3.2 | Le jeu de données cliniques

Le jeu de données *LapEx* est composé des données patients et des vidéos endoscopiques enregistrées lors des opérations.

Les données patients ont été recueillies au cours des visites patients pré- et postopératoires, et incluent des informations sur la démographie, les antécédents médicaux et chirurgicaux du patient, les évènements indésirables lors de l'opération, les complications per- et postopératoires, et l'évolution du poids, de l'IMC et des comorbidités du patient.

Au total, 30 patients, 22 femmes et 8 hommes, ont pris part à cette étude, et ont été opérés. Ils sont âgés de 44.1 ± 11.2 ans, et ont un IMC de 45.7 ± 7.1. Le suivi patient a été effectué par le docteur Reche tout au long de l'étude. Les opérations ont été menées pour 15 d'entre elles par un chirurgien spécialisé en cœlioscopie, et pour les 15 autres par un chirurgien spécialisé en laparotomie. Les opérations se sont déroulées sur une période de 15 mois, du 17/12/2015 au 28/03/2017. La vidéo endoscopique enregistre l'opération dès l'instant où l'endoscope est introduit dans la cavité abdominale, et jusqu'à l'instant où l'endoscope est définitivement extrait de la cavité abdominale. Ces vidéos ont une durée de $56.7 \pm 15.7 min$. La chirurgie la plus longue dure $1h38min$, la plus courte $36min$ (voir figure 3.4).

Sur l'ensemble de la cohorte, un seul évènement indésirable accompagné de saignement a eu lieu : une plaie de la rate. Cet évènement indésirable n'a pas eu d'impact post-opératoire. Dans le tableau 3.1, on décrit les statistiques d'évolution du poids et de l'IMC liées à l'opération, avec notamment une perte de poids moyenne de 35.7 kg et une réduction d'IMC de 13 kg/m^2.

Comme nous l'avons expliqué en section 3.1.1.3, les deux étapes critiques de la gastrectomie longitudinale la dissection et la résection du fundus. Dans la suite de mes travaux, je me suis uniquement intéressé à l'étape de dissection du fundus. Ce choix a été fait avec notre partenaire clinique, qui a considéré que le geste de dissection effectué de façon répétée dans cette étape est à la fois très classique et également un geste présentant une grande disparité selon les niveaux de pratique en cœlioscopie. C'est donc un bon sujet d'étude quand on s'intéresse à la formation et à la pratique du geste chirurgical. Ainsi, les données auxquelles je me suis intéressé sont uniquement les extraits vidéos correspondant à l'étape de dissection du fundus pour chacune des 30 opérations. Cette étape dure $8min57sec \pm 3min26sec$, avec l'extrait le plus long durant $20min55sec$ et le plus court $3min19sec$ (voir figure 3.4).

En résumé

- L'obésité est aujourd'hui un problème de santé publique à l'échelle mondiale.
- La chirurgie bariatrique, et a fortiori la gastrectomie longitudinale, ont des résultats effectifs notamment en terme de perte pondérale.
- La gastrectomie longitudinale est de plus en plus pratiquée du fait de sa reproductibilité, du faible taux de complications post-opératoires, et des bons résultats en terme de perte pondérale.
- L'idée générale de la gastrectomie longitudinale est de réduire le volume de l'estomac en le transformant en « tube ».
- Dans mes travaux, je me suis focalisé sur une étape critique de cette procédure : la dissection du fundus.
- L'objectif principal de cette étape est de séparer la partie supérieure de l'estomac (le fundus) des tissus adipeux qui l'entourent.
- Un protocole d'étude clinique a été mené et approuvé pour une cohorte de patients.
- Cette cohorte est constituée de 30 patients atteints d'obésité et opérés au CHU de Grenoble par deux chirurgiens séniors.
- Le jeu de données *LapEx* est constitué de 30 vidéos enregistrées lors de l'étape du fundus des 30 procédures menées lors de cette étude clinique.

3.2 | Modélisation de la chirurgie

A ce point de mon travail, j'avais à ma disposition un jeu de données associées aux opérations et profils de 30 patients opérés au CHU de Grenoble. Au vu des données traitées dans les études de notre état de l'art (voir section 2.2.3), je me focaliserai par la suite uniquement sur les données per-opératoires, c'est-à-dire les vidéos. Dans cette partie, je présente une étape préalable à l'analyse quantitative des vidéos : la modélisation de la procédure chirurgicale.

3.2.1 | Pourquoi modéliser la chirurgie ?

Le but de cette modélisation était de poser les bases de connaissances concernant la cœlioscopie, la gastrectomie longitudinale, et a fortiori l'étape de dissection du Fundus sur laquelle s'est focalisé mon travail de thèse. Plus précisément le but commun du chirurgien et du scientifique était de créer un terrain d'entente et de compréhension pour pouvoir travailler ensemble.

Pour le scientifique, la découverte de ce contexte clinique complexe, et le dialogue avec le chirurgien expert, ont permis de mieux comprendre les problématiques propres à ce domaine, et de faire émerger les attentes du clinicien quant aux apports du travail scientifique. De plus, le formalisme utilisé dans la modélisation et le modèle résultant ont été la base sur laquelle s'est définie la méthodologie d'annotation présenté en section 3.3.

Pour le chirurgien, l'intérêt de cette démarche est moins évident : définir le modèle chirurgical nécessaire à la suite de ce travail a demandé au chirurgien de mettre en mots des concepts et des connaissances qu'il maîtrisait en pratique et par expérience, mais qu'il avait rarement l'occasion d'expliciter, surtout à un néophyte en chirurgie (ce qu'est le scientifique).

3.2.2 | Comment modéliser la chirurgie ?

La modélisation a été menée pour deux aspects de la chirurgie : la procédure et la qualité. Nous avons suivi une méthodologie de type top-down pour modéliser la procédure chirurgicale, et une méthodologie de type bottom-up pour modéliser la qualité chirurgicale.

3.2.2.1 | Méthode de modélisation de la procédure

Comme nous le présenterons dans la section suivante 3.2.3, la formalisation de la procédure chirurgicale est un domaine de recherche bien établi, et des outils existent pour structurer ce processus. En nous basant sur ces outils établis, nous avons appliqué une approche top-down qui nous a permis d'inscrire notre étude dans un cadre pré-établi.

Ces outils sont :

— Le formalisme de *Modélisation de la Procédure Chirurgicale* (*SPM* pour « Surgical Process Model ») et les outils ontologiques associés.
— Le corpus des études cliniques décrivant les protocoles et règles de bonnes pratiques pour la gastrectomie longitudinale.
— L'expérience du chirurgien expert en chirurgie bariatrique.

3.2.2.2 | Méthode de modélisation de la qualité chirurgicale

La modélisation de la qualité chirurgicale est un domaine de recherche qui a émergé dans les années 90 avec le score OSATS (voir section 2.3.1). Les approches proposées sont des travaux cliniques proposant des scores d'évaluation génériques pour noter la qualité de chirurgies complètes. Ici, nous nous sommes plutôt intéressés à l'analyse d'aspects spécifiques de la pratique chirurgicale qui présentent des perspectives intéressantes en termes de compréhension du geste chirurgical. Et dans ce domaine, peu, voire aucune étude ne proposent de méthodologie de formalisation.

De fait, nous avons appliqué une approche bottom-up, en nous basant sur du contenu issu de procédures de gastrectomie longitudinale et sur l'expérience du chirurgien expert en chirurgie bariatrique. Nous nous sommes également basés sur la littérature clinique établissant les bonnes pratiques et les protocoles chirurgicaux pour la gastrectomie longitudinale.

En pratique, nous avons mené un certain nombre d'entretiens avec le chirurgien expert. Pendant les entretiens, nous avons visionné les chirurgies du jeu de données *LapEx*, et questionné le chirurgien sur sa pratique. Au fil des entretiens et dès lors que le projet a gagné en maturité, des idées de plus en plus claires et concrètes ont émergé. Afin d'étayer ces entretiens, nous avons également étudié plusieurs articles établissant les standards pour cette technique chirurgicale. Nous présenterons en section 3.2.4 les résultats de cette réflexion.

3.2.3 | Modélisation de la procédure chirurgicale

La modélisation de la procédure chirurgicale consiste à décrire dans un certain formalisme les évènements ayant lieu tout au long de la chirurgie. Dans les années 2000, une première période de recherche s'est focalisée sur la reconnaissance du flux d'évènements se rapportant strictement à la procédure chirurgicale (Lalys and Jannin, 2014). Plus récemment, le domaine de recherche s'est élargi, et inclut désormais l'analyse cognitive de l'état du chirurgien et la reconnaissance du contexte de bloc opératoire au sens large (Pernek and Ferscha, 2017).

3.2.3.1 | Formalisme de description de la procédure

La modélisation de la procédure chirurgicale est avant tout une modélisation temporelle. A cet égard, il convient de définir selon quelle échelle temporelle les évènements observés sont modélisés. Cette notion d'échelle est appelée la *granularité*. Si on observe de haut et dans sa globalité la procédure chirurgicale, on aura une *granularité* grossière. Plus on considère des évènements précis et courts, plus la *granularité* s'affine.

4 niveaux de granularité sont communément acceptés pour décrire le déroulement de la procédure chirurgicale (Lalys and Jannin, 2014). De la granularité la plus grossière à la plus fine :

— **Phase.** La phase décrit les épisodes généraux de la chirurgie. Ils peuvent être communs à plusieurs techniques chirurgicales distinctes. La phase est composée d'une séquence d'étapes.
Exemple : la phase de dissection.

— **Etape.** L'étape est associée à un objectif chirurgical visant une cible anatomique précise. Pour atteindre cet objectif, le chirurgien effectuera une succession d'activités.
Exemple : l'étape de dissection du fundus.
— **Activité.** L'activité, ou action, décrit un mouvement effectué par le chirurgien afin de remplir un objectif chirurgical simple/atomique. L'activité peut elle-même être décomposée en mouvements.
Exemple : un geste de dissection du ligament gastrocolique effectué par la pince électro-thermale.
— **Mouvement.** Le mouvement, ou « surgème » (combinaison de « surgical » et « phoneme »), n'a pas de signification chirurgicale, il se définit par le déplacement ou la transformation physique qu'il induit.
Exemple : un mouvement de la pince électro-thermale en direction de l'estomac.

Finalement, la modélisation d'une opération réelle peut se faire à chacun de ces niveaux de granularité sous la forme d'une séquence d'évènements. Et les modélisations d'une même séquence à différents niveaux de granularité doivent être cohérentes. C'est à dire que pour une procédure donnée, la séquence de phases la modélisant couvrent également la séquence d'étapes la modélisant, et ainsi de suite.

De plus, l'activité est communément formalisée par une association de trois termes appelée *triplet*. Ce *triplet* est constitué d'un verbe, d'un instrument chirurgical, et d'une cible chirurgicale. Si on reprend l'exemple proposé pour l'activité ci-dessus, sa formalisation sous forme de triplet $< verbe; instrument; cible >$ sera :

<*disséquer ; pince électro-thermale ; ligament gastrocolique*>

Pour ces trois éléments, nous avons pu établir un vocabulaire exhaustif décrivant quels instruments chirurgicaux, quels verbes, et quelles cibles intervenaient pendant la gastrectomie longitudinale, mais surtout pendant l'étape de dissection du fundus. Ce vocabulaire a été établi grâce à un travail commun avec notre partenaire chirurgien en nous basant sur son expérience, sur une analyse d'articles établissant des standards opératoires pour la gastrectomie longitudinale (Gagner et al., 2016; Iossa et al., 2016; Rosenthal, 2012; Verhaeghe et al., 2011), et sur une ontologie dédiée à la cœlioscopie (Katić et al., 2015). Le vocabulaire établi pour l'étape de dissection du fundus est décrit ci-dessous (voir figure 3.1 pour les termes anatomiques) :

— **instrument** ∈ [écarteur de foie ; pince atraumatique ; pince électro-thermale]
— **verbe** ∈ [coaguler un tissu ; écarter un objet d'un endroit ; pousser un objet ; tenir un objet ; tirer un objet ; verrouiller et disséquer]
— **cible** ∈ [adhérence ; compresse ; diaphragme ; estomac ; foie ; ligament gastrocolique ; ligament gastrosplenique ; lipome de his ; meso-gastro posterieur ; paroi abdominale ; rate ; tissus adipeux ; vaisseau court gastrique]

3.2.4 | Modélisation de la pratique chirurgicale

Comme nous l'avons présenté dans notre état de l'art (voir chapitre 1), la notion de qualité en chirurgie doit être abordée avec précaution : c'est une notion complexe aux multiples

facettes, présentant différents niveaux de lecture (voir section 1.2). Ici, je me suis focalisé sur la pratique du chirurgien, indépendamment des autres membres de l'équipe opératoire, de l'environnement du bloc opératoire, et du contexte organisationnel. Plus précisément j'ai traité un aspect de la qualité de pratique du chirurgien : l'exposition dans la scène chirurgicale, je me suis aussi intéressé au profil de pratique du chirurgien.

3.2.4.1 | L'exposition dans la scène chirurgicale

La notion d'exposition dans la scène chirurgicale est décrite dans des manuels de technique chirurgicale (Pouliquen, 2009), mais de façon générale pour la cœlioscopie. Selon les procédures, et donc les organes et structures anatomiques traités, cette notion présente des spécificités et peut être décrite de différentes façons. Nous nous sommes donc intéressés ici à la notion d'exposition dans la scène chirurgicale lors de la gastrectomie longitudinale, car cette notion est considérée comme cruciale pour le bon déroulement de la procédure par notre partenaire clinique. Avec l'aide de notre partenaire clinique et de son expertise en cœlioscopie, nous avons donc proposé une formalisation de cette notion dans le cadre de la gastrectomie longitudinale.

En chirurgie laparoscopique, la vision du chirurgien se base exclusivement sur la vidéo fournie par l'endoscope qui ne donne qu'un champ de vision réduit de l'espace chirurgical. Les instruments chirurgicaux étant insérés par des trocarts, ils ont de plus des possibilités réduites de mouvements. La problématique de l'exposition consiste à gérer au mieux ces contraintes de vision et de mouvement en modifiant, à l'aide des outils chirurgicaux, la configuration spatiale de l'espace chirurgical. Plus précisément, l'objectif est de rendre la visibilité et l'accès à la cible chirurgicale courante les meilleurs possible (Pouliquen, 2009).

Avec l'expert chirurgien, nous avons donc défini la *qualité d'exposition d'une cible chirurgicale* comme l'état d'exposition de cette cible dans le contexte du déroulement chirurgical et vis-à-vis de l'objectif recherché pour cette cible. Lorsque l'exposition de la cible chirurgicale est bonne, le chirurgien peut se focaliser sur les objectifs de la procédure et n'a besoin de faire que quelques modifications mineures de l'environnement chirurgical. Plus l'exposition est mauvaise, plus le chirurgien devra modifier les positions relatives des organes et des outils pour obtenir une meilleure exposition. Autrement dit, la gestion de l'exposition est un préalable aux objectifs chirurgicaux de l'opération.

Dès lors que la cible chirurgicale change et se déplace au cours de la procédure chirurgicale, l'exposition évolue et alterne entre de bons et de mauvais états. Dans une chirurgie bien menée par un chirurgien expérimenté, ces deux états se distinguent facilement : soit l'exposition de la cible chirurgicale est bonne et le chirurgien peut remplir les objectifs chirurgicaux, soit cette exposition est dégradée, et le chirurgien rétabli une exposition correcte de la cible chirurgicale. Si la chirurgie est plus compliquée, ces états deviennent indistincts, et il devient difficile de savoir si le chirurgien est en train de modifier l'exposition, ou s'il cherche à remplir les objectifs de la chirurgie.

Ainsi, je définis la *qualité de gestion de l'exposition de la cible chirurgicale* comme la capacité du chirurgien à connaître, à un instant donné, l'état de l'exposition de la cible ; à savoir s'il doit modifier ou non cette exposition pour atteindre son objectif ; et, le cas échéant, à transformer le plus efficacement possible cette exposition. Pour alléger notre discours, nous

parlerons par la suite de la *qualité d'exposition* au lieu de la *qualité de gestion de l'exposition de la cible chirurgicale*.

3.2.4.2 | Le profil de pratique chirurgicale

J'ai également étudié le profil de pratique chirurgicale. Au sein d'un même pôle chirurgical, il existe des différences entre chirurgiens. Lorsqu'un chirurgien pratique, ses collègues chirurgiens sauront le reconnaître simplement en visionnant la vidéo d'une de ses chirurgies. Ils reconnaitront ses habitudes de pratique, la décision qu'il prendra plutôt qu'une autre, voire la marque de l'enseignement de son mentor à travers sa façon d'effectuer un geste spécifique par exemple.

Ces différences sont encore plus prégnantes entre des chirurgiens d'hôpitaux différents, et peuvent s'accentuer encore entre pays et entre continents. Questionner les différences de pratique entre chirurgiens permet de mieux comprendre ce qui les distingue, notamment au regard de la qualité de leurs pratiques.

Je définis le *profil de pratique du chirurgien* comme l'ensemble des spécificités de pratique, aussi bien dans les choix faits par le chirurgien tout au long de la procédure, que dans ses habitudes gestuelles, permettant de l'identifier.

Cette notion nous a intéressé car deux chirurgiens ont opéré les patients de notre cohorte, et ils sont tout deux chirurgiens séniors. En étudiant leurs profils de pratique, nous avons comparé les différences et les points communs de leurs pratiques respectives. Cette approche est donc similaire à l'étude du niveau d'expertise, mais plutôt que de comparer des populations de chirurgiens de niveaux différents, nous n'avons comparé que deux chirurgiens présentant un niveau d'expertise élevé. Notre analyse n'a donc pas porté sur leur niveau d'expertise mais sur ce qui distinguait réellement leurs pratiques respectives.

En résumé

- La modélisation de la chirurgie rend possible la communication et l'échange entre le chirurgien et le scientifique.
- Selon le sujet de la modélisation, le processus de modélisation peut prendre une forme très structurée pour l'aspect procédural, ou moins formalisée pour l'aspect qualitatif.
- La procédure chirurgicale est décrite à travers les notions imbriquées de phase, d'étape, d'activité et de mouvement.
- l'activité chirurgicale est représenté par un triplet $< verbe; instrument; cible >$, et un vocabulaire exhaustif pour les trois éléments du triplet.
- Deux aspects de la pratique chirurgicale ont été étudiés : la gestion de l'exposition de la cible chirurgicale et le profil de pratique chirurgicale.

3.3 | Méthode d'annotation de données vidéos

En section 3.1.3, nous avons décrit la cohorte de patients ainsi que les données cliniques recueillies sous la forme de vidéos endoscopiques notamment. En section 3.2, nous avons établi un modèle décrivant la procédure chirurgicale de façon formelle au cours de l'étape de dissection du fundus de la gastrectomie longitudinale, de même que deux aspects de la qualité de pratique chirurgicale. Désormais munis de ces deux préalables, nous pouvons considérer une annotation de ces vidéos endoscopiques à partir de cette modélisation.

Sous leur forme initiale brute, ces vidéos ne nous permettaient pas d'envisager un traitement algorithmique proposant des résultats sur la pratique chirurgicale. Un prétraitement a donc été nécessaire pour extraire de ces vidéos des donnés plus exploitables. Ce prétraitement a pris la forme d'un ensemble de tâches d'annotation permettant de sélectionner des informations dans le flux vidéo. Etant donné la modélisation présentée dans la section précédente, nous présentons ici trois tâches d'annotation effectuées manuellement et successivement sur les vidéos laparoscopiques. Nous avons commencé par une annotation procédurale, suivie d'une annotation de la qualité d'exposition de la cible chirurgicale, puis nous avons fini par une tâche de segmentation d'images extraites des vidéos.

Je commencerai par argumenter notre choix d'une annotation manuelle plutôt qu'automatique, puis je décrirai la méthodologie d'annotation que nous avons suivi pour chaque tâche.

3.3.1 | Pourquoi annoter manuellement des données vidéos ?

Dans une vidéo, l'essentiel de l'information est contenu dans les images et leur succession. Une communauté scientifique très active travaille à développer des méthodes de reconnaissance automatique du contenu de vidéos laparoscopiques. Ces méthodes concernent la reconnaissance de la procédure chirurgicale, la reconnaissance d'objets et notamment d'instruments chirurgicaux dans les images, et parfois aussi la qualité et les performances chirurgicales.

Pour l'annotation de la procédure à partir de données vidéos, la tâche consiste à reconnaître les évènements ayant lieu au cours de la procédure, et à obtenir une séquence de ces évènements. La difficulté de la tâche varie en affinant le niveau de granularité temporelle observée. Ainsi la difficulté va croissante quand on considère d'abord les phases, les étapes, puis les activités. Dans le cas d'une annotation fine au niveau des activités par exemple, on peut complexifier l'annotation procédurale en annotant plusieurs acteurs en même temps. On doit alors labelliser l'action effectuée par chaque acteur à chaque instant.

Pour l'annotation visuelle basée sur des données vidéos, je définis les trois tâches suivantes par ordre de difficulté croissante :

1. L'*étiquetage* consiste à reconnaître le ou les objets présents dans les images de la vidéo.
2. Le *suivi* consiste à connaître la position d'un ou plusieurs objets dans les images de la vidéo.
3. La « bounding box » ou rectangle de localisation consiste à placer un rectangle englobant entièrement l'objet ou les objets dans les images de la vidéo.

4. La *segmentation* consiste à savoir exactement quels pixels de l'image en cours appartiennent à l'objet ou aux objets dans les images de la vidéo.

Un autre moyen, là aussi, de faire varier la difficulté de la tâche, est d'augmenter le nombre d'objets à traiter pour l'algorithme. Il existe aussi une différence de difficulté importante entre le traitement des instruments chirurgicaux sur lesquels nombre d'études existent (Bouget et al., 2017), et le traitement des organes et structures anatomiques qui n'est que peu, voire pas étudié.

De plus, l'environnement visuel laparoscopique étant très complexe, les méthodes proposées doivent être robustes afin de gérer des artefacts visuels tels que les occultations, les illuminations, la fumée, le sang, ou la buée. Ainsi, les méthodes proposées ont d'abord été testées dans des environnements visuels présentant un faible niveau de complexité, tels que des vidéos sur banc d'essai, sous environnement virtuel, ou dans des situations chirurgicales bien précises ne présentant aucun artefact.

Finalement, nous avons décidé d'annoter les données vidéos correspondant à l'étape de dissection du fundus de chacune des 30 procédures de notre jeu de données. Ces vidéos montrent un certain niveau de complexité car ce sont des vidéos de chirurgies réelles présentant tous les artefacts mentionnés ci-dessus. D'autre part, notre annotation procédurale a consisté à reconnaître les activités de tous les acteurs manipulant des outils visibles dans les vidéos, c'est-à-dire les mains du chirurgien et les mains de l'assistant. Nous avons également segmenté entièrement certaines images des vidéos (instruments et entités anatomiques visibles). Enfin nous avons annoté la qualité de gestion de l'exposition au cours de ces procédures. Actuellement, les algorithmes proposés par la communauté ne permettent pas d'envisager de remplir automatiquement une de ces trois tâches sur des vidéos présentant ce niveau de complexité.

Par ailleurs, nous avions présenté dans notre état de l'art (voir section 2.3.2) une méthode d'annotation par la foule qui aurait potentiellement permis de grandement accélérer le processus d'annotation. Cependant, l'une des limitations de l'annotation par la foule est qu'elle ne s'applique qu'à des tâches élémentaires, pour lesquelles la foule annote aussi bien, voire mieux, qu'une personne avertie. Or dans notre cas, les tâches d'annotation procédurale et visuelle présentent un niveau de complexité trop élevé pour que nous envisagions l'annotation par la foule. Une simplification des tâches de segmentation aurait néanmoins été envisageable, en ne demandant aux opérateurs de n'annoter par exemple que les outils, ou qu'un organe bien spécifique.

A noter qu'une problématique de confidentialité des données et images patients se posait aussi : le protocole clinique présenté en section 3.1.2 spécifiait clairement que l'accès aux données de l'étude n'était permis que pour les personnes membres du projet. De fait, l'annotation par la foule ne nous a pas semblée envisageable.

Pour toutes ces raisons, l'annotation manuelle s'est présentée comme la seule possibilité réaliste, au vu de nos objectifs d'annotation.

3.3.2 | L'annotation procédurale

L'annotation procédurale a consisté à décrire les activités des deux mains du chirurgien, et des deux mains de l'assistant. Sachant que nous traitions uniquement de l'étape de dissection du fundus, et au vu du formalisme défini en section 3.2.3, nous avons annoté les activités se déroulant lors de cette étape sous la forme du triplet $< instrument; verbe; cible >$.

3.3.2.1 | Le protocole expérimental

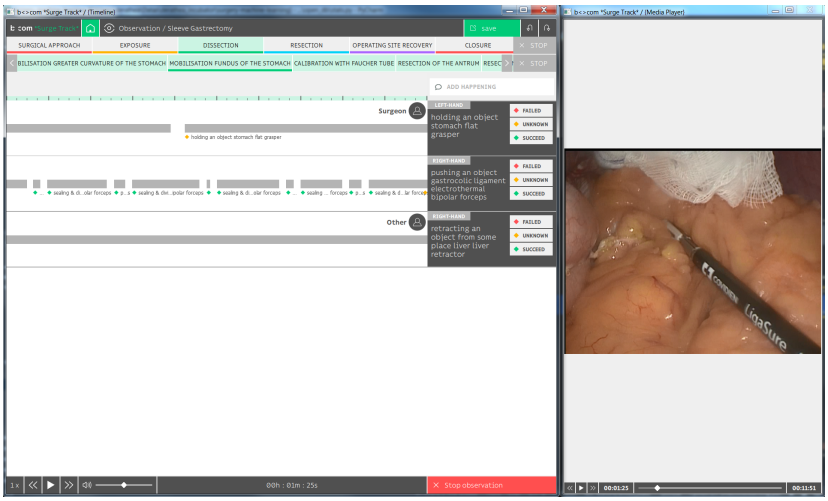


FIGURE 3.5 – Aperçu de l'interface d'annotation de la procédure SurgeTrack développé par la société b<>com

L'annotation a été menée conjointement par un chirurgien expert en chirurgien bariatrique, et par un scientifique sur la plateforme logicielle « Annotate »[1] (voir aperçu de l'interface d'annotation en figure 3.5). Pour chacune des 30 vidéos chirurgicales, ils ont regardé l'extrait correspondant à la dissection du fundus.

Ensemble, ils ont annoté les actions effectuées par 4 acteurs : les deux mains du chirurgien, la main de l'assistant qui tient l'écarteur de foie, et la seconde main de l'assistant dans quelques procédures seulement. Comme ces 4 acteurs agissent simultanément, l'annotation résultante prend la forme d'une séquence multidimensionnelle avec un label par acteur défini à chaque instant.

3.3.2.2 | Le formalisme des données

Chaque instance d'action annotée a été décrite par les informations suivantes :

— l'acteur,
— le triplet <verbe ; instrument ; cible>,
— la date de début de l'action,
— la date de fin de l'action,

Au final, pour chaque extrait vidéo de la dissection du fundus, on a obtenu un fichier au format XML contenant toutes les annotations des activités effectuées par les 4 acteurs.

1. https ://b-com.com/en/bcom-surgery-workflow-toolbox-annotate

3.3.2.3 | Méthode de validation

Cette tâche d'annotation de la procédure a été menée dans sa totalité, et conjointement, par le chirurgien et le scientifique. Ils sont les seuls à avoir annoté la procédure dans ces vidéos.

Il est important de noter l'expertise du chirurgien dans le domaine de la chirurgie bariatrique, et donc pour la gastrectomie longitudinale. Le fait qu'il ait été présent tout au long de la tâche d'annotation est un indicateur fort de la qualité de cette annotation.

Afin de valider quantitativement cette annotation, le scientifique a ré-annoté seul, une vidéo de dissection du fundus. Les métriques suivantes ont été calculées pour les deux versions afin de les comparer :

— le nombre d'activités annotées,
— le nombre de types d'activité annotés,
— la durée d'activité

3.3.3 | Annotation de la pratique chirurgicale

Je présente ici les annotations des deux aspects de la pratique chirurgicale présentés précédemment en section 3.2.4.

3.3.3.1 | La qualité d'exposition

La qualité d'exposition de la cible chirurgicale est un phénomène observable tout au long de l'opération, de même que la qualité de gestion de l'exposition de la cible par le chirurgien. Nous avons décidé d'évaluer la qualité de cette gestion à certains instants spécifiques de la procédure.

Pendant la dissection du fundus, l'objectif principal est de disséquer l'estomac du ligament gastrocolique, du ligament gastrosplénique, du méso-gastro posterior, puis des potentielles adhérences (voir section 3.1.1.3). Ainsi, chaque fois qu'un geste de dissection est effectué, on se rapproche un peu plus de l'objectif principal. Pour le bon déroulement de la dissection, la bonne gestion de l'exposition de la cible chirurgicale (le point d'attache entre l'estomac et le ligament gastrocolique par exemple) est essentielle. En évaluant la gestion de l'exposition de la cible chirurgicale pour chacun de ces gestes de dissection, nous analysons la qualité de la dissection tout au long de cette étape chirurgicale.

Le protocole expérimental L'annotation de l'exposition a été menée en même temps que l'annotation procédurale, toujours conjointement par le chirurgien expert et le scientifique et sur la même plateforme logicielle « Annotate ». En se basant sur l'annotation procédurale, nous avons observé un type d'activité caractérisé par son verbe : « verrouiller et séparer » ou « coaguler un tissu » (« sealing and dividing » ou « coagulating a tissue »). Ce type d'activité correspond à l'ensemble des activités de dissection. A chaque occurrence d'une telle activité, la gestion de l'exposition de la cible chirurgicale a été évaluée par le chirurgien expert.

Pour chacune des 30 vidéos, nous avons annoté la qualité d'exposition de la cible chirurgicale chaque fois que nous avons rencontré une activité de dissection.

Le formalisme des données L'expert a évalué la qualité d'exposition de la cible chirurgicale selon une échelle à trois niveaux :

— 1 = bonne qualité
— 2 = qualité suffisante mais pouvant être améliorée
— 3 = qualité insuffisante

Ces annotations ont été stockées dans le même fichier XML que l'annotation procédurale de la vidéo correspondante. Pour chaque chirurgie, on a obtenu autant d'annotations de la qualité d'exposition, qu'il y a d'activités dont le verbe est soit « verrouiller et séparer », soit « coaguler un tissu ». Il n'y a aucun a priori sur le nombre d'échantillons annotés dans une chirurgie, et ce nombre est variable d'une chirurgie à l'autre.

Méthode de validation L'annotation de la qualité d'exposition a été faite simultanément avec l'annotation procédurale. Le chirurgien et le scientifique sont donc également les deux seules personnes à avoir annoté ces vidéos. De même, l'expertise du chirurgien était l'argument fort en faveur d'une validation qualitative de cette tâche.

Aucune méthode n'a été proposée pour valider quantitativement la tâche d'évaluation de l'exposition.

3.3.3.2 | Le profil de pratique chirurgicale

Le profil de pratique chirurgicale associé à une chirurgie est celui du chirurgien ayant effectué la chirurgie. Notre jeu de données a été constituée à partir des chirurgies effectuées par deux chirurgiens experts. Ces annotations prennent donc la forme d'une valeur binaire selon le chirurgien ayant effectué la chirurgie (0 pour chirurgien 0 et 1 pour chirurgien 1).

3.3.4 | La segmentation d'images

L'annotation visuelle a consisté à segmenter entièrement chaque image correspondant à une instance d'annotation de la qualité d'exposition.

3.3.4.1 | Le protocole expérimental

Nous avons défini une liste exhaustive des objets pouvant être segmentés, en accord avec le vocabulaire défini en section 3.2.3 qui décrivait quels instruments chirurgicaux et structures anatomiques pouvaient apparaître pendant l'étape de dissection du fundus. Pour chaque image, tous les objets visibles de la liste ont été labellisés et segmentés. Nous avons défini des contraintes sur la segmentation de chaque image :

— Un pixel ne peut pas avoir deux labels différents.
— Plusieurs formes segmentées peuvent avoir le même label.
— L'image doit être entièrement segmentée, i.e. chaque pixel de l'image doit avoir un label.

Du fait de l'importante durée nécessaire pour mener à bien cette tâche de segmentation, nous avons restreint le nombre d'images à annoter. Nous nous sommes limités à annoter les images correspondant aux instants auxquels l'exposition de la cible chirurgicale avait été

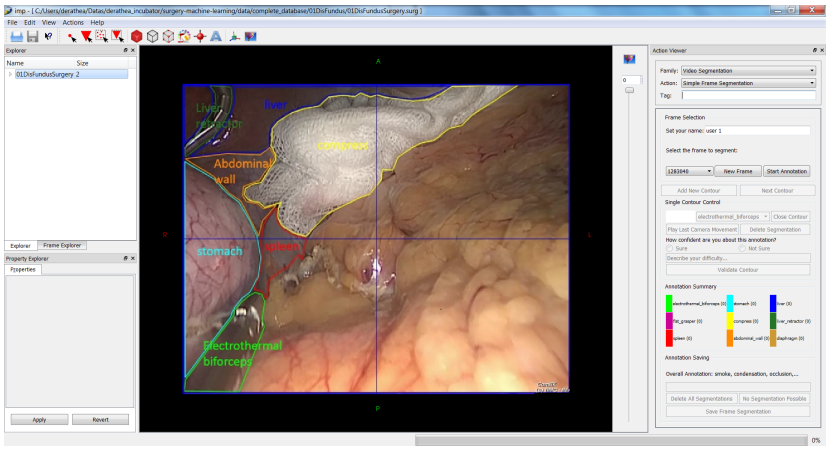


FIGURE 3.6 – Aperçu de l'interface de segmentation d'image développée à partir de l'interface de programmation d'application CamiTK (Fouard et al., 2012).

annotée. Ainsi, chaque fois que nous avons évalué l'exposition, l'image de la vidéo correspondante a été segmentée et labellisée.

Cette tâche de segmentation a été effectuée par trois scientifiques sur une application dédiée et développée à partir de l'interface de programmation d'application C++ CamiTK (Fouard et al., 2012) (voir aperçu de l'interface d'annotation en figure 3.6). En amont de l'annotation elle-même, les opérateurs ont été formés afin de s'approprier et le contexte clinique et l'outil de segmentation. Suite à cette formation, la charge de travail de segmentation a été distribuée petit à petit entre les trois scientifiques, en fonction de leur disponibilité, et jusqu'à segmentation complète de toutes les images. Chaque image n'a été segmentée que par un opérateur.

3.3.4.2 | Le formalisme des données

Chaque image segmentée se présente sous la forme d'une image en nuance de gris, dans laquelle chaque label est identifié par un niveau de gris unique. Et pour chaque annotation de la qualité d'exposition, on a une image segmentée.

3.3.4.3 | Méthode de validation

Avant d'entamer la tâche de segmentation principale, les opérateurs ont segmenté un même groupe de 40 images extraites des 30 vidéos, et représentatives de la variabilité de ces vidéos. Cela a permis une évaluation quantitative de la variabilité inter-opérateur.

Nous utilisons pour cette évaluation une métrique de mesure de similarité entre deux formes segmentées F_0 et F_1 appelée la *similarité régionale* ou l'intersection sur l'union IoU. Cette métrique est utilisée comme métrique d'évaluation dans des challenges de traitement d'image (Everingham et al., 2010; Zhou et al., 2017). La similarité régionale est définie comme le rapport de l'aire d'intersection des pixels de ces deux formes sur l'aire de l'union de leurs pixels :

$$IoU = \frac{|F_1 \bigcap F_2|}{|F_1 \bigcup F_2|} \tag{3.1}$$

Plus IoU est proche de 1, plus les deux segmentations sont similaires, plus IoU est proche de 0, moins elles ont de pixels en commun.

Ainsi nous pourrons comparer la cohérence de l'annotation pour les 3 opérateurs ayant segmenté une part du jeu de données, sur chacun des objets pouvant être segmentés.

En résumé

- Pour *LapEx*, l'annotation procédurale a été menée conjointement par un chirurgien expert en chirurgie bariatrique et un scientifique.
- Cette annotation procédurale a consisté à décrire l'enchaînement d'activités des mains du chirurgien et de son assistant pendant l'étape de dissection du fundus.
- Pour *LapEx*, l'annotation de la pratique chirurgicale a consisté à faire évaluer la qualité d'exposition lors de chaque activité de dissection et à identifier le chirurgien qui opère.
- Cette annotation de la pratique a été menée conjointement par un chirurgien expert et un scientifique.
- Pour *LapEx*, l'annotation visuelle a consisté à segmenter entièrement les objets présents dans chaque image associée à un échantillon d'annotation de la qualité d'exposition.
- Cette segmentation a été menée par trois scientifiques ayant été préalablement formés à cette tâche.

3.4 | Résultats de l'annotation

Comme nous l'avons vu dans la section précédente, l'annotation peut prendre différentes formes. Nous allons ici analyser successivement les annotations de la procédure, de la pratique, et des images chirurgicales et décrire en détail les données résultantes de ces annotations.

3.4.1 | L'annotation des données procédurales

3.4.1.1 | Contenu de l'annotation

Dans le tableau 3.2, nous présentons des statistiques sur les activités chirurgicales annotées sur l'ensemble du jeu de données *LapEx*. Pour chaque acteur nous avons récupéré le label de l'activité, la moyenne et écart-type du nombre d'activités et la moyenne et écart-type de la durée des échantillons d'activité.

	# de	# d'activités		durée des activités (en s)	
	labels	moyenne	écart-type	moyenne	écart-type
Main gauche de l'assistant	3	5.00	3.65	82.44	170.91
Main droite de l'assistant	4	1.77	1.74	285.05	197.89
Main gauche du chirurgien	14	13.57	8.48	34.01	39.73
Main droite du chirurgien	24	58.53	20.34	5.97	4.96

Tableau 3.2 – Statistiques des activités effectuées par chirurgie pour les différents acteurs du set de données LapEx

On observe que, comparé aux mains du chirurgien (gauche : 14, droite : 24), les mains de l'assistant ont peu de labels différents (gauche : 3, droite : 4), et donc peu d'activités différentes. De plus, on observe que les durées moyennes des activités (voir équation 3.2) et les nombres d'activités (voir équation 3.3) sont ordonnées entre les acteurs de façon inverse comme suit (MD-Ass = main droite de l'assistant, MG-Ass = main gauche de l'assistant, MD-Chir = main droite du chirurgien, MG-Chir = main gauche du chirurgien) :

$$\overline{durée_{MD-Ass}} = 285s > \overline{durée_{MG-Ass}} = 82.4s > \overline{durée_{MG-Chir}} = 34s > \overline{durée_{MD-Chir}} = 6s \tag{3.2}$$

$$\overline{N_{MD-Ass}} = 1.8 < \overline{N_{MG-Ass}} = 5.0 < \overline{N_{MG-Chir}} = 13.6 < \overline{N_{MD-chir}} = 58.5 \tag{3.3}$$

En annexe A, le tableau présente une description exhaustive de chaque label d'activité annotée dans le jeu de données *LapEx*. Finalement, nous proposons un modèle simplifié de la procédure au cours de l'étape de dissection du fundus, avec les triplets d'activité typiques pour chaque acteur :

- **Main droite du chirurgien :**
  - Pince électro-thermale ; disséquer ; ligament
  - Pince électro-thermale ; pousser ; estomac

  - Pince électro-thermale ; pousser ; ligament
- **Main gauche du chirurgien :**
  - Pince atraumatique ; tenir ; estomac
- **Main droite de l'assistant :**
  - Ecarteur de foie ; écarter ; foie
- **Main gauche de l'assistant :**
  - Pince atraumatique ; tenir ; ligament

3.4.1.2 | Validation de l'annotation

Dans le tableau 3.3 nous présentons les statistiques comparatives entre les deux versions d'une même vidéo annotées à plus d'un an d'intervalle. On lit d'abord la différence de nombre d'échantillons d'activités annotées entre les deux versions (Δ). Ainsi, pour le premier type d'activité, 60% d'échantillons en plus ont été annoté dans l'annotation sans le chirurgien comparé à celle avec le chirurgien. Le deuxième type d'activité a lui été annoté 67% échantillons de plus dans l'annotation avec le chirurgien que dans celle sans lui. Ces différences sont très importantes car les nombres d'échantillons annotés dans les deux cas sont faibles. Globalement, on observe que plus d'activités on été annotées lors de l'annotation sans le chirurgien que dans celle avec lui.

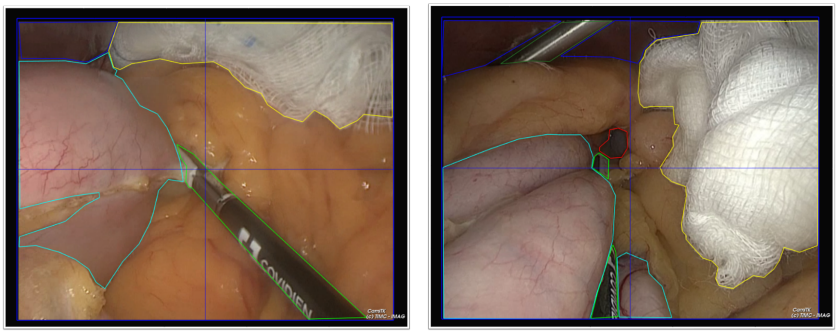
On lit ensuite la différence de durée moyenne des types d'activité entre les deux versions de l'annotation. Les valeurs positives signifient que la durée moyenne est plus longue pour la seconde annotation tandis que les durées négatives signifient que la durée moyenne est plus longue pour la première annotation. L'analyse de ce tableau montre que les activités de la main droite de l'assistant et de la main gauche du chirurgien ont durée plus longtemps en moyenne lors de la seconde répétition. La différence d'annotation est ici aussi non-négligeable.

Acteur	Verbe	Instrument	Cible	# d'activités annotées		
				avec chir.	sans chir.	Δ (en %)
MD chir.	verrouiller & séparer	pince électro-thermale	ligament	16	10	60.0
MD chir.	verrouiller & séparer	pince électro-thermale	meso-gastro postérieur	1	3	-66.7

Acteur	Verbe	Instrument	Cible	Δ de durée moyenne (en s)
MD Ass.	écarter	écarter de foie	liver	39.26
MG chir.	tenir	pince atraumatique	estomac	20.41
MD chir.	tirer	pince électro-thermale	compress	-1.95
MD chir.	pousser	pince électro-thermale	stomach	-0.04
MD chir.	verrouiller & séparer	pince électro-thermale	ligament	1.41
MD chir.	verrouiller & séparer	pince électro-thermale	meso-gastro posterieur	0.95

Tableau 3.3 – Statistiques comparatives sur les activités annotées pour l'évaluation de la variabilité inter-opérateur de l'annotation procédurale - MD chir. = Main droite du chirurgien - MG chir. = Main gauche du chirurgien - MD Ass. = Main droite de l'assistant.

3.4.2 | L'annotation de la pratique chirurgicale



(a) (b)

FIGURE 3.7 – illustrations (a) d'une bonne exposition et (b) d'une exposition insuffisante de la cible chirurgicale.

La figure 3.7 présente un exemple de bonne qualité d'exposition (voir Figure 3.7a), et de qualité d'exposition insuffisante (voir Figure 3.7b). A partir du travail de modélisation mené avec notre partenaire chirurgien et décrit en section 3.2, nous avons réuni un certain nombre de caractéristiques spécifiques aux bonnes gestions d'exposition, et aux gestions d'expositions insuffisantes. Dans le cas de la bonne exposition, on observe par exemple que la compresse est présente pour maintenir les tissus adipeux hors du champ d'action des outils ; que la pointe de l'outil est bien visible ; que la main gauche du chirurgien présente correctement l'estomac avec une tension correcte appliquée aux tissus. Bref, dans l'ensemble l'image est claire et présente une bonne organisation de l'espace chirurgical. A l'inverse, dans l'exemple de l'exposition insuffisante, on observe que la présence de la compresse ne permet pas de bien maintenir les tissus adipeux à l'écart de la zone de travail ; que les tissus adipeux apparaissent tant à droite qu'à gauche de l'image ; que la pince électro-thermale n'est pas bien visible et a peu d'espace pour maneuvrer ; que l'estomac est mal présenté par la main gauche du chirurgien. Bref, l'image est chaotique et plus difficile à comprendre.

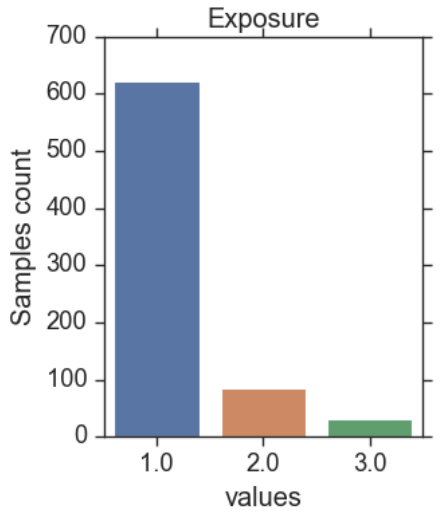
Au total, sur l'ensemble du jeu de données *LapEx*, 735 échantillons de qualité d'exposition ont été annotés. Dans la figure 3.8a on a la distribution de ces échantillons. On observe un net déséquilibre de la répartition des échantillons, avec la valeur 1 ou « bonne » exposition (620 échantillons) largement sur-représentée par rapport aux valeurs 2 ou exposition « suffisante » (85 échantillons) et 3 ou exposition « insuffisante » (30 échantillons) (voir tableau 3.4).

La figure 3.8b présente la distribution d'échantillons annotés par chirurgien. On retrouve ici encore le déséquilibre entre les 3 valeurs d'exposition annotées. On observe aussi que les échantillons ne sont pas distribués uniformément entre les deux chirurgiens : le chirurgien 0 a plus d'échantillons de valeur 1 tandis que le chirurgien 2 a plus d'échantillons de valeurs 2 et 3.

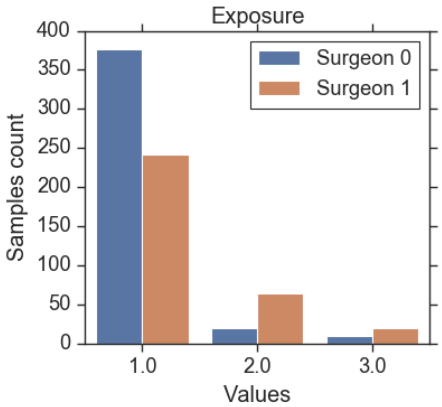
La figure 3.8c présente la distribution d'échantillons annotés par procédure. Ici également, les échantillons de sont pas distribués uniformément entre les différentes procédures. Le nombre d'échantillons varie entre 14 et 55 par chirurgie, avec une moyenne à 25 échantillons et un écart-type à 9.4 échantillons.

Cluster	# d'échantillons
Set de données complet	735
Exposition = 1	620
Exposition = 2	85
Exposition = 3	30
Chirurgien 0	407
Chirurgien 1	328

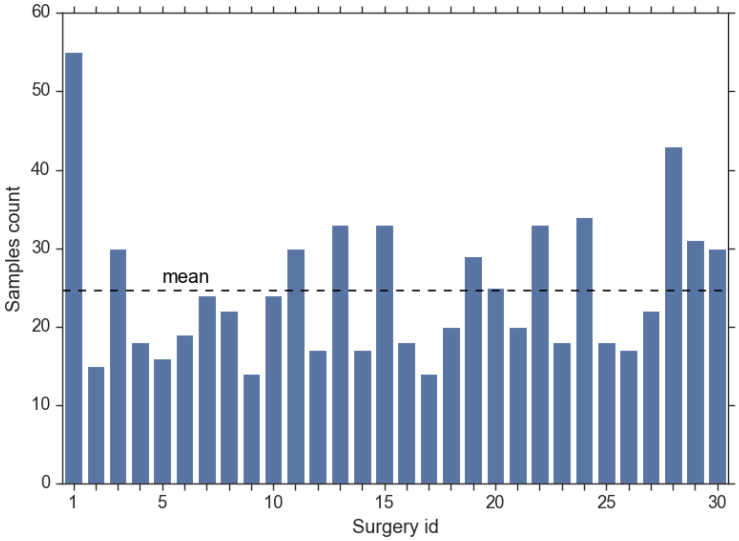
Tableau 3.4 – Nombre d'échantillons selon différents clusters dans le set de données *LapEx* - Les valeurs d'exposition sont 1=bonne, 2=suffisante et 3=insuffisante.



(a)



(b)



(c)

FIGURE 3.8 – Distribution (a) des échantillons annotés de l'exposition, (b) des échantillons annotés de l'exposition par chirurgien, et (c) des échantillons annotés de l'exposition par procédure pour le jeu de données *LapEx* - Avec les valeurs d'exposition 1=bonne, 2=suffisante et 3=insuffisante.

3.4.3 | La segmentation d'images

3.4.3.1 | Contenu de l'annotation

La tâche de segmentation a été menée par trois opérateurs sur l'interface de segmentation dédiée. L'interface de segmentation a été développée spécificiquement pour cette tâche dans l'environnement de développement CamiTK. Les trois opérateurs ont segmenté des nombres de vidéos et d'images différents :

— Opérateur 1 : 15 vidéos et 390 images segmentées,
— Opérateur 2 : 12 vidéos et 269 images segmentées,
— Opérateur 3 : 3 vidéos et 80 images segmentées.

La figure 3.9 représente les pourcentages de présence d'objets, et les pourcentages de surface d'image couverte par chaque objet sur l'ensemble des images segmentées du jeu de données *LapEx*. Ces ratios traduisent le taux de présence des objets labellisés, ainsi que la taille qu'ils occupent dans les images.

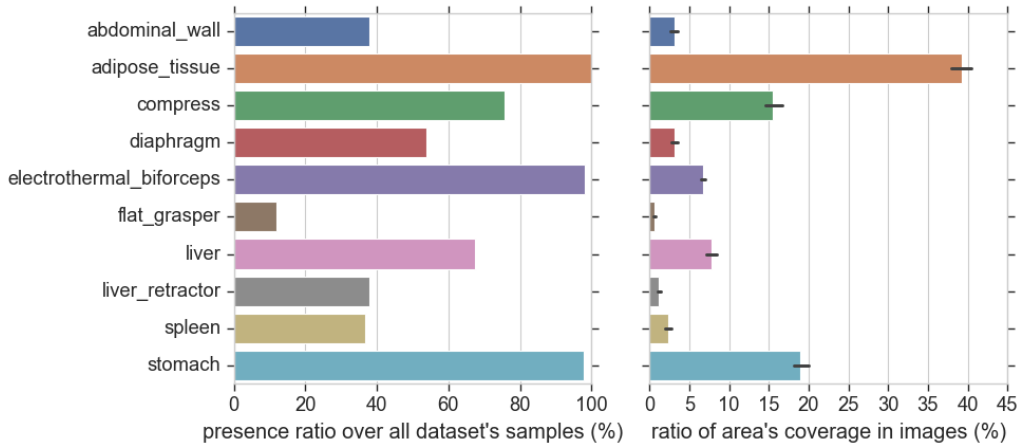


FIGURE 3.9 – Pourcentage de présence dans les images annotées et pourcentage de surface couverte par label et par image sur l'ensemble du jeu de données *LapEx*.

On observe que les tissus adipeux, la pince électro-thermale et l'estomac sont présents dans presque 100% des images, tandis qu'à l'inverse, la pince atraumatique n'apparaît que dans un peu plus de 10% des images. Lorsqu'on considère la place globalement occupée par les objets dans ces images, on constate que les tissus adipeux couvrent presque 40% de la surface des images en moyenne, viennent derrière l'estomac (18%) et la compresse (15%). 5 objets couvrent chacun moins de 5% des pixels : le diaphragme, la paroi abdominale, la rate, l'écarteur de foie et la pince atraumatique.

Le tableau 3.5 comptabilise le nombre de fois où différents types d'artefacts visuels ont été observés dans les images segmentés.

3.4.3.2 | Validation de l'annotation

Sur les 40 images initialement proposées pour l'évaluation de la variabilité inter-opérateurs, 35 ont finalement été segmentées par les 3 opérateurs. Les observations qui suivent concernent ces 35 images effectivement segmentées.

# d'observations de				# de chirurgies
Sang	Condensation	Flou de mouvement	Fumée	concernée
2	7	18	33	21

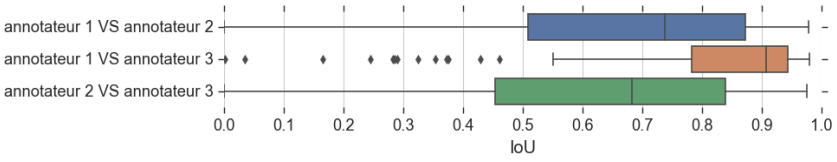
Tableau 3.5 – Nombre d'observations d'artefacts par catégorie et nombre de chirurgies concernées par les artefacts.

La figure 3.10a décrit la distribution des valeurs d'IoU sur l'ensemble des échantillons segmentés lorsqu'on compare les opérateurs deux à deux. Cette figure traduit la similarité et le recouvrement des segmentations entre les différents opérateurs. On observe que les opérateurs 1 et 3 ont en moyenne 91% des pixels en commun, les opérateurs 1 et 2 sont à $\overline{IoU} = 0.74$ en moyenne, et les opérateurs 2 et 3 sont à $\overline{IoU} = 0.68$ en moyenne. L'opérateur 1 a une IoU moyenne plus élevée avec les opérateurs 2 et 3 que n'en ont les opérateurs 2 et 3 entre eux. L'opérateur 1 présente donc une plus grande stabilité dans ses segmentations.

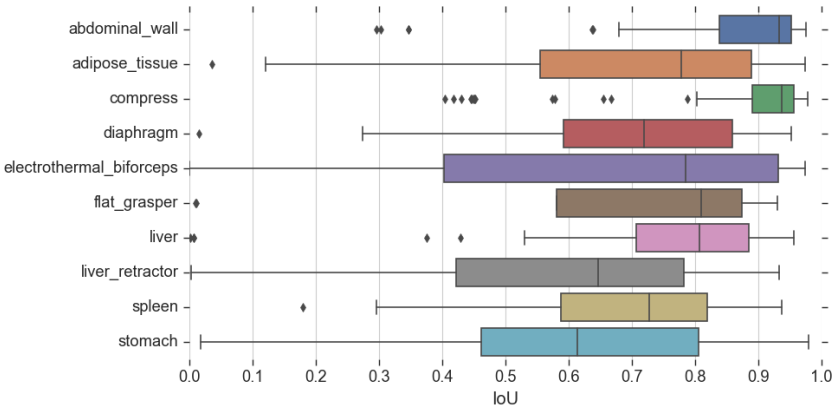
On remarque aussi que la distribution des valeurs de similarité est plus éparpillée entre les opérateurs 1 et 2, et entre les opérateurs 2 et 3, qu'entre les opérateurs 1 et 3. Ainsi, les opérateurs 1 et 3 montrent un plus haut niveau d'accord dans leurs segmentations.

De façon similaire, la figure 3.10b présente les distributions d'IoU sur l'ensemble des échantillons segmentées par objet labellisé. Pour chaque objet, on peut donc observer le niveau de similarité sur l'ensemble des images segmentées par les trois opérateurs. Ainsi, on voit que les deux objets segmentés avec le plus haut niveau de similarité sont la paroi abdominale ($\overline{IoU} = 0.935$), et la compresse ($\overline{IoU} = 0.94$). De plus, ces deux objets présentant des distributions resserrées, leur bonne segmentation est donc relativement systématique.

A l'inverse, les deux objets dont la segmentation présente le plus faible nivau d'accord



(a)



(b)

FIGURE 3.10 – Distributions de l'Intersection sur l'Union IoU pour (a) les différents opérateurs deux à deux, et (b) chaque label annoté par tous les opérateurs.

sont l'estomac ($\overline{IoU} = O.62$) et l'écarteur de foie ($\overline{IoU} = 0.64$), le diaphragme ($\overline{IoU} = 0.72$) et la rate ($\overline{IoU} = 0.73$) présentant aussi un faible niveau d'accord. Ces objets ont des distributions très éparpillées, ils ont donc un niveau de similarité entre opérateur très variable.

On observe aussi que les tissus adipeux et la pince électro-thermale ont un niveau de similarité satisfaisant ($\overline{IoU} = 0.78$), mais des valeurs très éparpillés. La qualité de ces segmentations est donc également variable.

Dans le tableau 3.6 sont comptés pour chaque objet et par couple d'opérateurs les cas où un échantillon d'objet a bien été reconnu mais où les segmentations proposées par les deux opérateurs ont eu un recouvrement nul. On observe :

— 7 cas de non-recouvrement entre les opérateurs 1 et 2, dont 5 pour la pince électro-thermale,
— 1 cas de non-recouvrement entre les opérateurs 1 et 3,
— 5 cas de non-recouvrement entre les opérateurs 2 et 3.

	Opérateur 1 VS Opérateur 2		Opérateur 1 VS Opérateur 3		Opérateur 2 VS Opérateur 3	
	Nombre	En %	Nombre	En %	Nombre	En %
Compresse	0	0.00	0	0.00	0	0.00
Diaphragme	0	0.00	0	0.00	1	0.11
Ecarteur de foie	1	0.10	0	0.00	0	0.00
Estomac	1	0.03	0	0.00	0	0.00
Pince atraumatique	0	0.00	1	0.20	0	0.00
Pince électrothermale	5	0.26	0	0.00	3	0.17
Tissus adipeux	0	0.00	0	0.00	1	0.03
Total	7	0.04	1	0.01	5	0.03

Tableau 3.6 – Nombre d'échantillons présentant une similarité régionale nulle ($IoU = 0$) pour chaque objet, et par couple d'opérateurs. Les pourcentages ont été calculés par rapport au nombre d'échantillons annotées en commun pour les deux opérateurs. Les objets ne rencontrant pas cette situation n'ont pas été présentés.

On constate d'une part que les problèmes de non-recouvrement concernent surtout l'opérateur 2 par rapport aux deux autres opérateurs. Et que l'objet majoritairement concerné par ce type de différence est la pince électro-thermale, avec 5 cas de non-recouvrement entre les opérateurs 1 et 2, et 3 cas de non-recouvrement entre les opérateurs 2 et 3. Il n'y a pas de non-recouvrement de la pince electro-thermale entre les opérateurs 1 et 3.

On notera que si la pince électro-thermale, et dans une moindre mesure le diaphragme, l'écarteur de foie, et la pince à traumatique n'ont présenté que quelques cas de non-recouvrement, ces échantillons présentant un non-recouvrement représente au moins 10% du nombre d'échantillons en commun entre les deux opérateurs. En revanche, ces cas deviennent négligeable quand on considère leur proportion vis à vis du nombre total d'échantillons annotés.

En résumé

— Dans notre jeu de données LapEx, nous avons annoté en moyenne 140 activités et 9 minutes par procédure.
— Nous avons annoté 735 échantillons de qualité d'exposition sur l'ensemble de notre jeu de données LapEx
— Ces échantillons présentent une distribution très déséquilibrée avec plus de 600 échantillons de « bonne » qualité et moins de 130 partagés entre les deux autres labels correspondant à une moins bonne qualité.
— 10 objets différents ont été segmentés dans notre jeu de données LapEx, pour un total de plus de 4500 échantillons annotés.
— Les trois objets les plus annotés et couvrant la plus grande partie des images sont les tissus adipeux, l'estomac, puis la compresse.

3.5 | Comparaison des jeux de données

De telles tâches d'annotation ont déjà été menées dans des conditions similaires par différents groupes de recherche de la communauté. Certains de ces groupes ont publié leur jeu de données annotées, généralement dans le cadre d'un challenge scientifique.

Je présenterai tout d'abord notre méthodologie de comparaison des jeux de données existants. Je présenterai ensuite ces jeux de données et j'étudierai les métadonnées existantes pour décrire la création de ces jeux de données. Enfin, nous comparerons quantativement le contenu annotés de ces jeux de données et du nôtre pour les trois type d'anntotations.

3.5.1 | Méthodologie de comparaison

Nous nous sommes intéressé ici aux jeux de données issus de vidéos laparoscopiques enregistrées dans des conditions d'entraînement sur banc d'essai, sur tissus vivants, ou dans des conditions de chirurgie réelle. Ces jeux de données devaient présenter au moins l'une des annotations de la procédure, de la pratique, ou du contenu visuel. Ces jeux de données devaient avoir été mis à disposition en libre accès entre 2014 et 2019. Il n'était pas nécessaire qu'une publication soit associée au jeu de données.

Catégorie	Id	Nom du paramètre
Organisation du Challenge	1	Nom et année du challenge
	2	Site internet du challenge
	3	Institution organisatrice et contact
	7	Avis éthique
	8	Conditions d'utilisation
Mission du Challenge	16	Champs d'application clinique
	18	Population cible
	19	Cible de l'algorithme
	20	Origine des donnnées
Conditions de l'Etude Clinique	22	Cohorte d'étude
	23	Contexte de l'étude clinique
	24	Centre
	25	Modalité d'imagerie
	26	Appareil d'acquisition
	27	Protocole d'acquisition
	28	Opérateur
Sets de Données du Challenge	30	Catégorie de méthode de génération de données d'entraînement
	31	# d'échantillons d'entraînement
	32	Caractéristiques des échantillons d'entraînement
	33	Protocole d'annotation des échantillons d'entraînement
	34	Annotateurs
	35	Méthode d'agrégation des échantillons annotés
	42	Méthode de pré-traitement des données
	43	Sources potentielles d'erreurs référencées

Tableau 3.7 – Liste des paramètres de métadonnées sélectionnés pour caractériser le processus de création d'un set de données annotées, à partir de l'étude de Maier-Hein et al. (Maier-Hein et al., 2018)

Afin de comparer ces jeux de données, nous avons considéré la qualité et la quantité des annotations. La qualité de l'annotation est centrale mais que très peu décrité ou évaluée comme l'explique Maier-Hein et al. (2018) qui s'intéressent à l'organisation de challenges scientifiques dans le domaine du traitement d'images médicales, ainsi qu'à la création des jeux de données associés. Un ensemble de métadonnées y sont décrites qui proposent un formalisme unificateur pour caractériser les différents aspects de tels challenges. Afin de bien situer l'apport de notre propre jeu de données à la communauté, j'ai tout d'abord considéré les métadonnées de l'étude de Maier-Hein et al. décrivant la création du jeu de données (voir tableau 3.7). J'ai comparé le nombre de métadonnées renseignées entre chacun de ces jeux de données et le nôtre. Ensuite, pour chaque type d'annotation, j'ai comparé quantitativement les données annotées entre les différents jeux de données à l'aide d'un ensemble de métriques (voir tableau 3.8).

Type	Métrique	Description
Procédure	# de labels	# d'évènements chirurgicaux de type différent annotés sur l'ensemble du set de données.
	# d'échantillons	# d'évènements chirurgicaux (phase, étape et activités) annotés sur l'ensemble du set de données.
	Statistiques sur le # d'échantillons par procédure	Moyenne et l'écart type des #s d'évènements chirurgicaux annotés par procédure.
	Durée annotée	Durée de vidéo annotée sur l'ensemble du set de données.
	Statistiques sur la durée annotée par procédure	Moyenne et l'écart type des durées annotées par procédure.
Qualité	# de labels	# d'aspects distincts de la qualité annoté sur l'ensemble du set de données.
	# d'échantillons	# d'échantillons d'évaluation annotée sur l'ensemble du set de données.
Visuel	# de labels	# de labels d'objets distincts annotés sur l'ensemble du set de données.
	# d'échantillons	# d'échantillons d'objets annotées sur l'ensemble du set de données.
	Statistiques sur le # d'échantillons par image	Moyenne et l'écart-type du # d'échantillons d'objets annotées par image.
	# d'images annotées	# d'images annotées sur l'ensemble du set de données.
	Statistiques sur le # d'images annotées par procédure	Moyenne et l'écart-type du # d'images annotées par procédure.
	Statistiques sur le # de pixels annotés par image	Moyenne et l'écart-type du # de pixels annotés par image.
	Statistiques sur la proportion de pixels annotés par image	Moyenne et l'écart-type de la proportion de pixels annotés sur la totalité des pixels de l'image.

Tableau 3.8 – Description des métriques utilisées pour comparer les annotations entre les différents sets de données, un échantillon correspond à une instance annotée d'évènement (phase, étape et/ou activité pour l'annotation procédurale) ou d'objet (pour l'annotation visuelle).

3.5.2 | Présentation des jeux de données

J'ai sélectionné 9 jeux de données mis à disposition de la communauté scientifique depuis 2014, en plus de notre jeu de données (voir tableau 3.9). 5 de ces jeux de données ont

Jeu de données			Procédure		Annotation	
Nom	Institution	Article associé	Type	#	Type	Granularité
JIGSAWS	John Hopkins Univ.	(Gao et al., 2014)	Tâches d'entraînement	103	Procédurale Niveau de pratique	Action OSATS
Endovis_2015_IST	KIT & UCL	/	Colorectale, robotisée	16	Visuelle	Suivi, segmentation
Cholec80	IRCAD	(Twinanda et al., 2016)	Cholecystectomie	80	Visuelle Procédurale	Etiquetage Phase
AtlasDione	Buffalo NY	(Sarikaya et al., 2017)	Tâches d'entraînement	99	Visuelle	Rectangle
Endovis_2018_RSS	Intuitive Surgical Inc.	/	Modèle porcin	15	Visuelle	Segmentation
CATARACTS	LATIM	(Al Hajj et al., 2019)	Phacoemulsification	25	Visuelle	Etiquetage
CaDIS	LATIM + Digital Surgery Ltd.	(Flouty et al., 2019)	Phacoemulsification	25	Visuelle	Segmentation
Endovis_2019_RMIS	DKFZ Heidelberg	/	Chir. digestives	16	Visuelle	Segmentation
Endovis_2019_SWSA	NCTD Dresden	/	Cholecystectomie	24	Visuelle Procédurale Niveau de pratique	Etiquetage Phase, action GOALS

Tableau 3.9 – Description des sets de données annotées

été mis à disposition dans le cadre des challenges **EndoVis** organisés depuis 2015 dans le cadre de la conférence MICCAI.

JIGSAWS a été publié en 2014 par la John Hopkins University (Baltimore, Etats-Unis) (Gao et al., 2014). 8 chirurgiens de niveaux variés ont effectué trois tâches d'entraînement robotisées sur banc d'essai plusieurs fois chacun. Pour chaque tâche, les activités effectuées pendant la procédure, ainsi qu'une évaluation du niveau de pratique basé sur le score structuré OSATS (voir section 2.3.1) ont été annotées.

Endovis 2015 Instrument Segmentation and Tracking (*Endovis_2015_IST*) a été mis a disposition conjointement par le Karlsruhe Institute of Technology (Karlsruhe, Allemagne) et le University College London (Londres, Royaume-Uni) en 2015. A ce jour, aucune publication n'est associée à ce jeu de données. Il a fait l'objet d'un challenge *EndoVis*. Ce jeu de données contient des extraits de chirurgie colorectale, et de tâches d'entraînement robotisées sur banc d'essai. Ce jeu de données présente soit des annotations de suivi d'instruments, soit la segmentation des instruments visibles.

Cholech80 a été publié par le laboratoire ICube de l'Université de Strasbourg en 2016 (Strasbourg, France) (Twinanda et al., 2019). 13 chirurgiens ont effectué 80 procédures de cholécystectomie. Ces chirurgies ont été annotées procéduralement et visuellement. L'annotation procédurale a consisté à reconnaître la phase en cours tout au long de la vidéo. L'annotation visuelle a consisté à étiqueter les outils présents dans les images des vidéos.

Atlas Dione a été publié par le Roswell Park Comprehensive Cancer Center (New-York, Etats-Unis) en 2017 (Sarikaya et al., 2017). 10 chirurgiens de niveaux variés ont effectué 6 tâches robotisées d'entraînement sur banc d'essai. Les instruments chirurgicaux ont été annotés visuellement en plaçant une « bounding box » sur chaque instrument visible.

Endovis 2018 Robotic Scene Segmentation (*Endovis_2018_RSS*) a été mis à disposition par l'entreprise Intuitive Surgical Inc. en 2018. A ce jour, aucune publication n'est associée à ce jeu de données. Il a fait l'objet d'un challenge *EndoVis*. Les vidéos ont été enregistrées lors de procédures robotisées sur modèle porcin. Les images ont été annotées en segmentant les outils robotisés dans les images, et en distinguant les différentes sous-parties des outils articulés.

CATARACTS a été publié par le laboratoire LATIM de l'Université Bretagne Occidentale en 2018 (Brest, France) (Al Hajj et al., 2019). Ce jeu de données a fait l'objet d'un challenge *EndoVis*. 3 chirurgiens de niveaux variés ont effectué 50 phacoémulsifications (procédure

de chirurgie opthalmique). La moitié de ces procédures (25 procédures) a été annotée en étiquetant les outils et structures anatomiques visibles dans les images des vidéos.

CaDIS, publié en 2019, est le fruit d'une collaboration entre le laboratoire LATIM auteur du jeu de données CATARACTS, et de l'entreprise Digital Surgery Ltd (Flouty et al., 2019). CaDIS est basé sur les mêmes données cliniques que CATARACTS mais propose un autre type d'annotation : ici, c'est une segmentation intégrale de certaines images qui a été effectuée.

Endovis 2019 Robust Medical Instrument Segmentation (*Endovis_2019_RMIS*) a été mis à disposition par le département Computer Assisted Medical Interventions (CAMI) du German Cancer Research Center en 2019 (DKFZ Heidelberg). A ce jour, aucune publication n'est associée à ce jeu de données. Il a fait l'objet d'un challenge *EndoVis*. Divers échantillons de chirurgie digestive ont été enregistrés à l'hôpital de Heidelberg. 250 images extraites de ces vidéos ont été annotées visuellement en effectuant une segmentation pixel-par-pixel.

Endovis 2019 Surgical Workflow and Skill Analysis (*Endovis_2019_SWSA*) a été mis à disposition par le National Center for Tumor Diseases en 2019 (Dresden, Allemagne). A ce jour, aucune publication n'est associée à ce jeu de données. Il a fait l'objet d'un challenge *EndoVis*. 30 opérations de cholécystectomie ont été effectuées par des chirurgiens experts. 24 d'entre elles ont fait l'objet d'une annotation procédurale au niveau des phases et des activités, et d'une annotation de la qualité basée sur le score GOALS (Vassiliou et al., 2005).

LapEx n'a pas encore été mis à disposition mais le sera courant 2020. Nous associerons cette mise à disposition à une publication décrivant le contenu du jeu de données, ainsi qu'une comparaison avec les jeux de données existants. Une description plus technique du contenu de ce jeu de données est donnée en annexe B.

3.5.3 | Analyse des métadonnées

Nous présentons ici une comparaison d'ordre qualitative sur le contexte dans lequel ces jeux de données ont été créés. Plus précisément, pour chaque jeu de données, nous évaluons le nombre de métadonnées fournies par ses créateurs sur la constitution du jeu de données (voir figure 3.11).

Dans le tableau 3.9 et la figure 3.11, les 10 jeux de données présentés (dont le nôtre) ont été ordonnés du plus ancien au plus récent. Sur l'ensemble des paramètres, on observe que, hormis pour *JIGSAWS*, plus le jeu de données est récent, plus le nombre de métadonnées renseignés est important. Notamment après la publication de (Maier-Hein et al., 2018) en

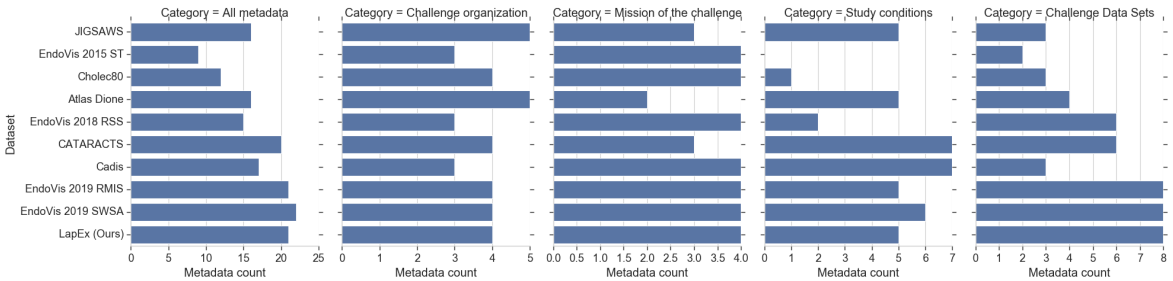


FIGURE 3.11 – Nombre de métadonnées renseignées par catégorie de métadonnées et au total pour les différents jeux de données

2018, le nombre de métadonnées renseignées respecte presque toutes les métadonnées de cette publication.

Les jeux de données pour lesquels le moins de métadonnées a été renseigné sont *EndoVis 2015 ST* et *Cholec80*. A l'inverse, les jeux de données pour lesquels le plus de métadonnées a été renseigné sont *CATARACTS*, *EndoVis 2019 RMIS*, *EndoVis 2019 SWSA* et *LapEx Dataset*.

Les deux catégories de métadonnées les moins renseignées sont les conditions de l'étude clinique, et les conditions d'annotation du jeu de données lui-même. Les métadonnées liées aux missions du challenge/jeu de données sont les mieux remplies dans l'ensemble, et les métadonnées sur l'organisation du challenge/jeu de données sont remplies de manière assez homogène sur l'ensemble des jeux de données.

3.5.4 | L'annotation procédurale

		LapEx_Dataset Chirurgie	Cholec80 Chirurgie	JIGSAWS Banc d'essai
# de labels		43	7	14
# d'échantillons		4198	545	1703
# d'échantillons	moyenne	139	6	16
par procédure	écart-type	53	0	5
Durée annotée (h)		4.5	51.2	2.5
Durée annotée	moyenne	9.0	38.4	1.5
par procédure (en min)	écart-type	3.5	17.1	0.7

Tableau 3.10 – Métriques comparatives sur les sets de données annotées procéduralement à partir de vidéos laparoscopiques

Dans le tableau 3.10 sont présentées différentes métriques comparatives sur les annotations procédurales des trois jeux de données *LapEx*, *Cholec80* et *JIGSAWS*. A noter que le jeu de données *Endovis 2019 SWSA* n'était pas disponible au moment de cette étude, nous ne présenterons donc pas de résultats pour ce jeu de données. Une première lecture rapide de ce tableau permet d'observer une grande diversité pour chaque métrique.

Le nombre de labels correspond au nombre de types distincts d'activités ou de phases qui ont été annotés dans chaque jeu de données. Tandis que *Cholec80* n'utilise que 7 types de phase différents, le double de labels différents est utilisé dans *JIGSAWS* pour annoter des activités. Dans notre jeu de données *LapEx*, ce sont plus de 40 labels qui ont été utilisés lors de l'annotation.

On observe la même tendance pour le nombre d'échantillons d'activités ou de phases annotées et pour la moyenne et l'écart-type du nombre d'échantillons par procédure. *Cholec80* présente beaucoup moins d'échantillons (545 au total et 6 en moyenne par procédure), que *JIGSAWS* qui en présente (1703 au total et 16 en moyenne par procédure) lui-même beaucoup moins que notre jeu de données *LapEx* (4198 au total et 139 en moyenne par procédure).

La durée totale annotée explique en partie cette tendance : nous avons annoté une durée presque deux fois plus longue dans *LapEx* que dans *JIGSAWS*, nous avons donc pu annoter

	# de labels	# de procédures	# d'échantillons
LapEx	2	30	1470
JIGSAWS	7	103	721

Tableau 3.11 – Statistiques sur l'annotation des compétences chirurgicales pour les sets de données *LapEx* et *JIGSAWS*.

un plus grand nombre d'échantillons d'activités. Dans *Cholec80* les phases annotées durent beaucoup plus longtemps que des activités. Cela explique en partie que nous ayons plus d'échantillons annotés, malgré une durée annotée clairement moins longue. La deuxième raison à cette tendance vient du nombre d'acteurs observés, et annotés pendant les procédures : tandis que dans *Cholec80* et *JIGSAWS* on se contente d'annoter un unique acteur, nous avons annoté trois, voire quatre acteurs en même temps. Notre densité d'annotation est donc beaucoup plus importante.

3.5.5 | L'annotation de la pratique chirurgicale

Le tableau 3.11 présente des statistiques générales pour l'annotation de la pratique dans les jeux de données *LapEx* et *JIGSAWS*. Les jeux de données *JIGSAWS* et *Endovis_2019_SWSA* ont ici évalué les compétences chirurgicales générales du chirurgien à travers les 7 composantes du score OSATS qui évalue le niveau d'expertise (une évaluation par vidéo), tandis que dans notre set de données *LapEx*, nous avons plus spécifiquement évalué la qualité de l'exposition ainsi que le profil de pratique du chirurgien. Encore une fois, le jeu de données *Endovis_2019_SWSA* n'étant pas disponible, nous ne présenterons pas de résultats pour ce jeu de données.

JIGSAWS présente beaucoup plus d'aspects différents des performances chirurgicales que *LapEx*, mais malgré un nombre de procédures plus de 3 fois plus important pour *JIGSAWS*, c'est *LapEx* qui présente le plus d'échantillons annotés. Plus précisément, on a dans *JIGSAWS* exactement une annotation de chacun des sept aspects pour chaque procédure, tandis que dans *LapEx*, comme expliqué en section 3.3.3.1, nous avons annoté la qualité d'exposition et le profil de pratique un nombre de fois variable par procédure. Ce qui explique que l'on arrive à deux fois plus d'échantillons pour *LapEx* (1470) que pour *JIGSAWS* (721).

3.5.6 | La segmentation d'images

Ici je nomme échantillon une instance d'objet visible ayant été annotée dans une image. Et je nomme label un type d'objet visible annoté dans l'ensemble d'un jeu de données. Avant tout, on notera que nous comparons ici les trois types d'annotations visuelles que sont l'étiquetage, le suivi et la segmentation (voir section 3.3.1) avec des métriques communes (voir tableau 3.8). Les métriques sur les pixels (le # de pixels annotés par image et la proportion de pixels annotés par image) ne sont typiquement pas adaptées pour des données d'étiquetage ou de suivi et peu adaptées pour des données de « bouding box ». De fait, ces

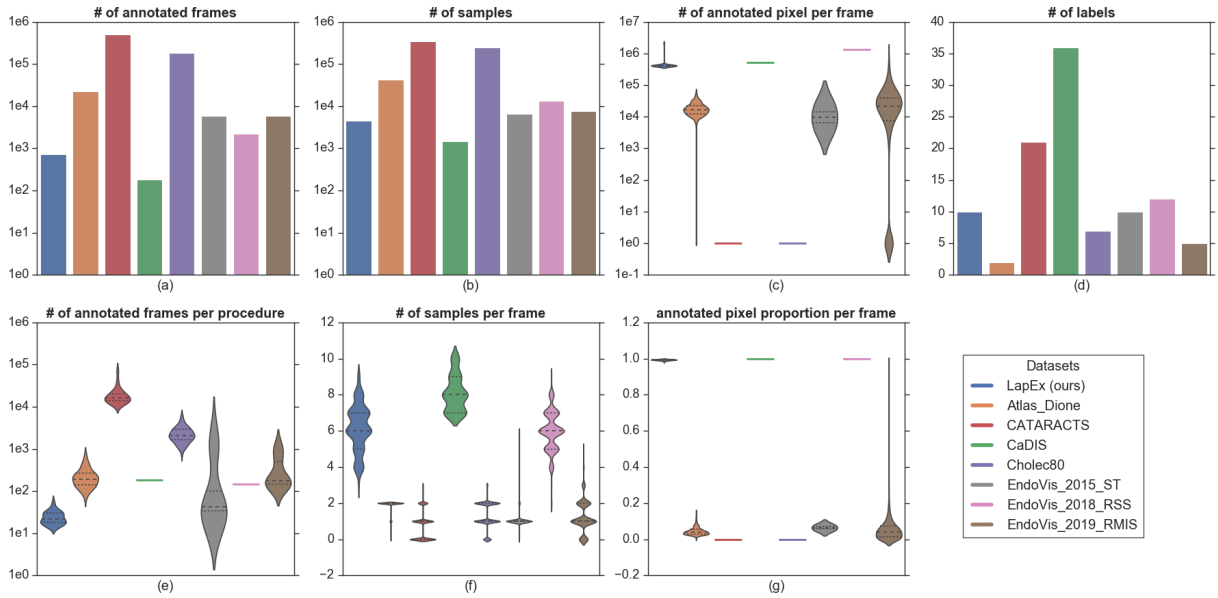


FIGURE 3.12 – Métriques comparatives sur les jeux de données annotées visuellement à partir de vidéos laparoscopiques.

métriques n'ont pas de sens pour les deux jeux de données *CATARACTS* et *Cholec80* (étiquetage) et doivent être interprétées avec précaution pour *Atlas Dione* (bounding box). De plus, on gardera en mémoire que le niveau de complexité de l'annotation a un fort impact sur les différentes métriques présentées, et biaise donc les valeurs observées.

Dans la figure 3.12, je compare le contenu des différents jeux de données présentant une annotation visuelle. Les graphiques (a) et (b) comparent respectivement le nombre d'images annotées et le nombre d'échantillons annotés. En terme d'images annotées, on a 89618 ± 175362 images annotées par jeu de données, avec deux jeux de données clairement plus gros : *CATARACTS* (494 868) et *Cholec80* (183 498) ; et deux jeux de données clairement plus petits : *LapEx* (735) et *CaDIS* (185). Une tendance similaire quoique moins marquée est observée pour le nombre d'échantillons annotés. On a donc 3 ordres de grandeur de différence entre les plus gros jeux de données et les plus petits en terme de nombre d'images annotées, et 2 ordres de grandeurs en terme de nombre d'échantillons annotés.

Le graphique (d) présente le nombre d'objets distincts labellisés dans chaque jeu de données. *CaDIS* distingue 36 objets, *CATARACTS* distingue 21 objets, tandis qu'*Atlas Dione* n'en a annoté que 2. Les autres comptent entre 5 et 12 labels distincts.

Le graphique (e) présente le nombre d'images annotées par procédure, on voit que pour *CATARACTS*, ce sont entre 1000 et 10 000 images qui sont annotées. A l'inverse notre jeu de données *LapEx* compte entre entre 20 et 40 images annotées par procédure. *EndoVis 2015 ST* a une distribution très étendue, qui exprime les multiples formes d'annotations présentes dans ce jeu de données. Deux jeux de données, *CaDIS* et *EndoVis 2018 RSS*, sont représentés pas une barre, car dans leurs cas, le nombre d'images annotées dans chaque procédure est fixe. En effet leurs vidéos ont toujours la même durée, leur fréquence d'annotation est également fixée.

Le graphique (f) présente le nombre d'échantillons par image et montre une tendance inverse au nombre total d'images annotées (graphique (a)) et d'échantillons annotés (gra-

hique (b)). Les jeux de données ayant le moins d'images et d'échantillons annotés au total sont ceux qui ont le plus d'échantillons annotés par image (*CaDIS*, *LapEx* et *EndoVis 2018 RSS*), tandis que ceux qui ont le plus d'images et d'échantillons annotés au total ont beaucoup moins d'échantillons annotés par image (*CATARACTS*, *Cholec80* et *Atlas Dione*).

Enfin le graphique (c) décrit le nombre de pixels annotés par image tandis que le graphique (g) décrit la proportion de pixels annotée/segmentée par image. Ces métriques ne sont pas applicable aux jeux de données *CATARACTS* et *Cholec80* pour lesquels les objets visibles ont été étiquetés mais non segmentés, ainsi que pour *Atlas Dione* pour lequel les instruments ont été encadrés. De fait, pour ces trois jeux de données, mais également pour *EndoVis 2015 ST* et *EndoVis 2019 RMIS*, la proportion de pixel est inférieure à 20% voire égale à 0. A l'inverse, les trois jeux de données *LapEx*, *CaDIS* et *EndoVis 2018 RSS* proposent des segmentations d'images exhaustives ou presque.

En résumé

— Afin de situer l'apport du jeu de données *LapEx* vis à vis des jeux de données existants, nous avons mené une comparaison des métadonnées et du contenu annoté des différents jeux de données.
— Depuis 2014, nous avons recensé 10 jeux de données annotées basés sur des vidéos ou des images de chirurgies (le nôtre compris).
— Ces jeux de données sont accompagnés de métadonnées les décrivant d'une façon incomplète, mais plus les jeux de données sont récents, plus les métadonnées proposées sont complètes.
— Sur ces 10 jeux de données, 9 présentent une annotation de type visuelle.

3.6 | Discussion sur l'annotation

A partir d'une cohorte de 30 patients, nous avons constitué un jeu de données cliniques constituée de 30 vidéos laparoscopiques issues de procédures de gastrectomie longitudinale. Nous avons établi une méthodologie d'annotation pour décrire les aspects procédural, de pratique et visuel de l'étape de dissection du fundus, une étape critique de cette technique chirurgicale. L'annotation procédurale est constituée des séquences d'activités effectuées par les mains du chirurgien et de son assistant. L'annotation de la pratique est constituée des évaluations de la qualité d'exposition lors de chaque activité de dissection, ainsi que du profil de pratique du chirurgien opérant. L'annotation visuelle est une segmentation complète des images correspondants à chaque évaluation de l'exposition. Nous avons également proposé une méthode de validation des annotations procédurale et visuelle.

Une fois ces différentes annotations complétées, nous avons pu comparer notre jeu de données annotées à ceux de la communauté scientifique, et mieux nous positionner par rapport à eux. La force de notre jeu de données réside dans ces trois formes d'annotation qui permettent de décrire différents aspects du contenu vidéo avec un haut niveau de détail. En effet, l'annotation procédurale des activités de 3 à 4 acteurs, de même que l'évaluation de la qualité d'exposition suivie tout au long de l'étape de dissection du fundus et la segmentation complète de certaines images de cette même étape, décrivent avec finesse le contenu de la vidéo. De plus, peu de travaux d'annotation ont proposé une validation quantitative de leur processus d'annotation.

3.6.1 | l'annotation des données procédurales

Hormis pour la durée annotée qui est clairement plus importante pour le jeu de données *Cholec80*, notre jeu de données *LapEx* présente une annotation plus complète, que ce soit en terme de types d'activités ou de nombre d'échantillons annotés. Le fait d'avoir annoté les activités de plusieurs acteurs en même temps rend notre connaissance de la procédure d'autant plus riche.

Lors de la tâche d'annotation procédurale, nous avons annoté les activités chirurgicales effectuées par les deux mains du chirurgien, ainsi que par la main de l'assistant tenant l'écarteur de foie. Dans 4 procédures (12, 20, 25 et 26), un quatrième acteur a également participé avec une seconde pince atraumatique.

On a observé une hiérarchisation inversée des activités selon leur nombre et leur durée, les activités des mains de l'assistant étant globalement plus longues et moins nombreuses que celles des mains du chirurgien. Cette hiérarchisation traduit le fait que le chirurgien effectue l'essentiel du travail, tandis que son assistant est là en soutien.

Cette hiérarchie des acteurs/mains selon leur durée d'action moyenne peut aussi être mise en relation avec la définition de l'exposition dans la scène chirurgicale en section 3.2.4.1. Précédemment, nous y avions expliqué que la gestion de l'exposition était nécessaire au bon avancement de la chirurgie. Or, cette gestion passe par les activités de tous les acteurs. La main droite du chirurgien, dont les activités sont au cœur du déroulement de la procédure, a pour but de remplir les objectifs chirurgicaux de l'opération. Les autres acteurs sont en soutien, et agissent quand l'état de l'exposition se détériore afin de l'améliorer. Ils effectuent

donc des actions plus longues dont l'objectif est de maintenir la bonne exposition, et ne changent d'activité que lorsque l'état de l'exposition le nécessite.

Si la durée moyenne des activités chaque acteur diffère, cela vient aussi du fait que chacun de ces acteurs gère l'exposition à une échelle différente, et agit pour corriger l'exposition à son niveau. Par exemple la main droite de l'assistant maîtrise la position du foie dans la cavité abdominale à l'aide de l'écarteur de foie. Cette activité dure longtemps (285 s en moyenne) et ne s'interrompt que lorsque le foie n'est plus écarté, et recouvre en grande partie l'estomac, ce qui suspend l'avancement de la procédure le temps de récupérer une exposition décente. Les activités de la main gauche du chirurgien ont pour but de présenter la grande courbure de l'estomac avec le bon niveau de tension appliqué aux tissus, afin de permettre le geste de dissection à la main droite du chirurgien. Ici, l'exposition est donc gérée plus finement, et ses variations plus fréquentes impliquent que les activités soient beaucoup plus courtes (39.7 s en moyenne).

Afin d'évaluer notre annotation procédurale, une des vidéos de dissection du fundus a été ré-annotée par le scientifique sans la présence du chirurgien avec un intervalle de temps de plus d'un an entre les deux annotations. On constate une claire différence entre les deux versions. Lors de la deuxième itération sans le chirurgien, plus d'activités ont été annoté pour la main droite du chirurgien, et les activités des mains gauches du chirurgien et droite de l'assistant ont duré plus longtemps en moyenne. Cela vient du fait que l'opérateur scientifique seul n'a pas su reconnaître correctement les début et fin de l'étape de dissection du fundus dans la vidéo, et a donc annoté plus que nécessaire. Cela soulève une difficulté spécifique à la gastrectomie longitudinale : dans cette technique de chirurgie digestive, les repères anatomiques restent très difficiles à reconnaître. Les différences de durée inférieures à 2s pour les activités de la main droite du chirurgien sont moins importantes mais restent néanmoins fortes, sachant que beaucoup d'activités de la main droite du chirurgien ont des durées de cet ordre de grandeur. La variabilité semble donc être importante.

Face à ces différences flagrantes entre les deux versions annotées, force est de constater la nécessité de mener cette annotation en présence du chirurgien pour limiter les erreurs, notamment celles liées à la reconnaissance du début et de la fin d'un évènement. Cette différence est flagrante quand on considère l'étape de dissection du fundus. Et le problème se pose également pour l'annotation des activités, même si ces erreurs s'expliquent plus par l'imprécision inhérente à l'annotations d'évènements tels que les activités qui sont courts et s'enchaînent rapidement.

Néanmoins, n'ayant pas pu mener de réelle étude de variabilité intra-opérateur en présence du chirurgien, le niveau de variabilité réelle du couple scientifique-chirurgien reste inconnu, même si on peut supposer que la variabilité sera réduite grâce à l'expertise du chirurgien. Dans le futur, il serait très intéressant de mener une étude de variabilité intra- et inter-opérateur plus poussée et plus fournie, afin de mieux quantifier cette variabilité. Une telle étude a déjà été menée par Huaulmé et al. (2019) dans le cas d'une série d'exercices robotisés basiques sur banc d'essai. Or dans nos conditions chirurgicales réelles, la variabilité et la complexité de l'environnement est beaucoup plus importante, et mériterait une étude dédiée.

Voici quelques pistes d'améliorations pour ces études de variabilité inter- et intra-opérateur :

— Ré-annoter en présence du chirurgien pour mieux imiter les conditions de l'annotation réelle.
— Répéter la ré-annotation plusieurs fois avec des intervalles de temps suffisants entre chaque annotation pour avoir une meilleure mesure de la variabilité intra-opérateur.
— Ré-annoter avec un ou plusieurs autres chirurgiens pour évaluer la variabilité inter-opérateur.
— Utiliser d'autres métriques d'évaluation telles que la distance basée sur un alignement temporel non-linéaire multidimensionnel utilisée par Huaulmé et al. (2017).

3.6.2 | L'annotation de la pratique chirurgicale

Le jeu de données *JIGSAWS* présente une annotation des performances chirurgicales recouvrant des aspects importants de la pratique chirurgicale robotisée car il s'agit d'une annotation basée sur les cinq composantes du score OSATS (Martin et al., 1997), du total de ces cinq composantes, et du niveau d'expérience du chirurgien. En comparaison, nos annotations de la qualité d'exposition et du profil de pratique semblent moins poussées.

Néanmoins, nous travaillons sur des enregistrements de chirurgies réelles menées uniquement par des chirurgiens experts, tandis que dans *JIGSAWS*, ce sont des enregistrements d'exercices d'entraînement sur banc d'essai menés par des chirurgiens d'un niveau de novice, intermédiaire ou expert. Il a été montré que le score OSATS n'était pas apte à distinguer entre des chirurgiens ayant un niveau d'expertise élevé (Hatala et al., 2015). De fait il n'était pas possible d'appliquer de tels scores à notre cas d'étude, c'est une des raisons qui nous a poussé à construire notre propre indicateur de qualité de pratique chirurgicale. Plus précisément, c'est sur la proposition de notre partenaire clinique que nous avons choisi et défini la qualité d'exposition comme indicateur de la pratique chirurgicale.

De plus, contrairement aux annotations proposés dans *JIGSAWS* qui ne sont évaluées qu'une seule fois par procédure, nous proposons une évaluation itérative tout au long de la procédure. On étudie donc l'évolution temporelle de cet indicateur de qualité de pratique, ce qui nous permet d'analyser plus finement la maîtrise que le chirurgien a de la procédure, avec ses forces et faiblesses. Il serait très intéressant d'étendre notre score d'évaluation au cours de la chirurgie en le complémentant avec un autre descripteur de la qualité de pratique, ou en affinant la notion d'exposition et en la scindant en sous-critères. En effet, dans la façon dont nous l'avons définie, cette gestion de l'exposition de la cible par le chirurgien nécessite une capacité d'anticipation du déroulement chirurgical, une capacité de reconnaissance de l'état de l'exposition et une capacité à rétablir l'exposition de la cible d'une façon optimale. Ces capacités pourraient faire l'objet d'études dédiées.

En section 3.3.3 nous avons défini trois niveaux pour l'annotation de la qualité d'exposition : « bonne qualité », « qualité suffisante » et « qualité insuffisante ». On notera l'absence du niveau « mauvais qualité » ou « qualité critique ». Ce niveau n'apparaît pas car les chirurgiens qu'on a évalués sont expérimentés et ont une longue carrière derrière eux. De plus, ils opèrent des patients dans des conditions réelles, et une « mauvaise qualité » indiquerait un danger réel pour la vie du patient. De tels cas sont rarissimes en routine clinique, et il est évidemment souhaitable de ne jamais les rencontrer. D'ailleurs, lorsqu'on observe la distribution des valeurs annotées de l'exposition pour *LapEx*, on observe un net déséquilibre

en faveur de la « bonne qualité ». C'est normal, ce sont des chirurgies réelles menées par des chirurgiens experts et qui se déroulent bien dans l'ensemble. Cette distribution est donc représentative de la réalité clinique, et le niveau « qualité insuffisante » signifie la possibilité d'une amélioration de l'exposition, et non un danger direct pour le patient.

La répartition des échantillons d'exposition sur les 30 chirurgies traduit un autre aspect de la réalité clinique : la variabilité des patients. Selon le patient, la taille du foie varie, tout comme la quantité de graisse autour de l'estomac. Le patient pourra aussi être atteint de cirrhose ou d'hypertension. Tous ces facteurs influent sur la complexité de la procédure et donc potentiellement sur la complexité de l'étape de dissection du fundus. L'étape en question pourra dès lors être plus ou moins longue, faire intervenir plus ou moins de gestes de dissection, et faire apparaître des états d'exposion de la cible chirurgicale de qualité varible.

Entre les deux chirurgiens, la différence de répartition entre les trois valeurs d'exposition vient de leurs spécialités différentes. En effet, si ces deux chirurgiens sont confirmés et ont des centaines de chirurgies à leur actif, ils n'ont pas la même spécialité. Le chirurgien 0 est spécialisé en cœlioscopie et pratique plusieurs centaines de chirurgies sous cœlioscopie chaque année. Le chirurgien 1 est spécialisé en laparotomie et pratique donc plus la chirurgie sous laparotomie (chirurgie ouverte) que sous cœlioscopie. On a donc deux profils de pratique bien distincts. Le fait d'observer une différence de répartition des valeurs de qualité d'exposition entre ces deux profils est une première preuve de la pertinence de notre score d'exposition pour distinguer différents profils de chirurgiens confirmés.

Une faiblesse de notre annotation de la qualité d'exposition est l'absence de validation. Nous n'avons pu mener une telle validation intra- ou inter-opérateur par manque de temps. Néanmoins il serait possible de mener de telles validations en ré annotant la qualité d'exposition dans une des chirurgies du jeu de données, soit avec les mêmes opérateurs (pour la validation intra-opérateur), soit avec d'autre opérateurs (pour la validation inter-opérateur). Une validation inter-opérateur serait particulièrement intéressante car elle nous permettrait d'évaluer à quel point la notion d'exposition et l'évaluation de sa qualité sont des notions reconnues et consensuelles chez les chirurgiens. De plus, de même que les entretiens que nous avons menés avec notre partenaire chirurgien lors de la phase de modélisation nous ont permis d'améliorer notre approche et d'étudier le point de vue du chirurgien, la rencontre avec d'autres chirurgiens nous permettrait d'enrichir d'autant notre démarche.

La notion de profil de pratique du chirurgien nous a permis d'annoter de façon simple lequel des deux chirurgiens opère. Dans une vision prospective, ce concept permettrait de caractériser la pratique de différents chirurgiens ou de différentes populations de chirurgiens, et potentiellement de les comparer (Forestier et al., 2018; Huaulmé et al., 2018). Cette « comparaison » de chirurgiens se rapproche cependant de la question de l'évaluation des chirurgiens en exercice évoquée en section 1.4 et présente les mêmes problématiques d'acceptabilité vis à vis des chirurgiens.

3.6.3 | La segmentation d'image

3.6.3.1 | Comparaison des jeux de données

En comparant l'annotation visuelle des différents jeux de données, on observe une grande diversité dans les types d'annotation. Cette diversité peut-être considérée par un compromis entre la fréquence et la densité d'annotation. Dans une vidéo, la fréquence d'annotation correspond à la quantité d'échantillons (un échantillon est un objet annoté dans une image) annotée par seconde, minute, ... La densité d'annotation correspond à la proportion de pixels annotés par image. Plus on augmente la fréquence et la densité d'annotation, plus l'annotation prend du temps. Il faut donc choisir si on préfère couvrir un nombre important d'images de la vidéo (fréquence), ou si notre objectif est plutot d'avoir une annotation détaillée sur moins d'images (densité). Ainsi, les jeux de données *CATARACTS* et *Cholec80* ont privilégié une haute fréquence d'annotation au détriment de leur densité d'annotation. A l'inverse, *LapEx* et *CaDIS* ont choisi une forte densité d'annotation, et ont donc une faible fréquence d'annotation. Les autres jeux de données se situent entre ces deux extrêmes, mais il semble que la tendance soit plutôt aux segmentations denses et à faible fréquence. Il serait intéressant d'étudier plus en détail les jeux de données au regard de ces notions de fréquence et de densité d'annotation. Il serait aussi envisageable de leur trouver une définition plus étendue permettant de comparer également les annotations procédurales et qualitatives.

Deux des huit jeux de données présentent des problèmes non-négligeables liés à leur choix d'annotation. Il m'a semblé important de les souligner pour éviter de les reproduire. Dans le cas d'*Atlas Dione*, nous avons choisi de traiter les rectangles d'annotation comme des segmentations, afin de pouvoir comparer ce jeu de données aux autres. Comparé à la tâche de segmentation, le rectangle introduit un nombre important de pixels qui sont des faux positifs ou des faux négatifs. Même si des valeurs indicatives ont été annotée pour décrire la présence d'occlusion ou de recouvrement, ce type d'annotation ne permet pas non plus de gérer parfaitement de tels problèmes. Par ailleurs, dans le cas d'*EndoVis 2018 RSS*, l'arrière-plan est annoté comme un objet unique alors qu'on y distingue différentes structures anatomiques, et que d'autres structures anatomiques sont correctement labellisées. Ce manque d'uniformité dans l'annotation est problématique.

3.6.3.2 | Focus sur *LapEx*

Dans notre jeu de données *LapEx*, on arrive à un total de 735 images segmentées (25 en moyenne) sur les 30 vidéos en associant une image segmentée à chaque échantillon d'évaluation de l'exposition. Si l'on avait voulu augmenter cette quantité, plusieurs stratégies auraient été possibles, on aurait pu :

— Augmenter la fréquence d'annotation sur l'ensemble de la vidéo.
— Définir des intervalles d'intérêt dans la chirurgie où mener la segmentation en priorité.
— S'appuyer sur l'annotation procédurale, et n'annoter que les segments de vidéo correspondant à une certaine activité.
— Annoter au préalable la présence d'un certain instrument ou organe, et n'annoter que des images dans lesquelles cet instrument ou organe était présent.

En ne segmentant qu'une image pour chaque échantillon de qualité d'exposition, nous ne considérons pas l'évolution temporelle du contenu visuel. Certes l'annotation de la pro-

cédure rend compte de l'évolution temporelle de la vidéo d'une certaine manière, mais cette annotation ne décrit pas les déplacements des objets visibles dans les images successives. Aussi, nous n'avons qu'une information visuelle statique pour chaque échantillon de qualité d'exposition. Les stratégies de segmentation proposées ci-dessus permettraient donc de répondre en partie à cette limite de notre annotation.

Afin de limiter le choix de l'opérateur, quant aux objets qu'il pouvait potentiellement segmenter dans les images, nous avons fixé la liste exhaustive des objets pouvant être segmentés. Cette liste a également permis de définir le niveau de granularité spatiale, c'est-à-dire à quel niveau de détail nous voulions que l'opérateur décrive le contenu de l'image. En effet il aurait été possible d'affiner le niveau de détail, en distinguant des sous-parties des instruments telles que le manche et les mâchoires, et pour chaque organe en distinguant différentes parties de ces organes. Par exemple, au lieu de segmenter l'estomac dans son ensemble, on aurait pu choisir de segmenter l'antre, la grande courbure, la petite courbure, le fundus, … Nous avons choisi de ne pas augmenter le niveau de détail car cela aurait grandement augmenté le temps d'annotation, et l'aurait aussi rendu plus complexe. D'autre part, la complexité des données ainsi annotées aurait été très élevée et peut-être trop élevée pour l'état de l'art des algorithmes actuels.

Les observations, objet par objet, des pourcentages de présence et de surface couverte par image (voir figure 3.9) méritent quelques explications. Tout d'abord, la présence quasi-systématique, et la place importante prise par les tissus adipeux, l'estomac, et la pince électro-thermale sont des résultats rassurants. En effet, il est cohérent d'observer une quantité importante de tissus adipeux dans l'abdomen de patients obèse, de même qu'il est cohérent d'observer systématiquement l'estomac et la pince électro-thermale, qui sont respectivement l'organe opéré, et l'outil principal de l'étape de dissection du fundus. Le fait que le foie apparaisse souvent mais couvre une faible surface montre que cet organe est bien écarté par l'écarteur de foie. Si on voit peu la pince atraumatique c'est également normal, car cet instrument a pour rôle de bien exposer la grande courbure de l'estomac et pour cela, il lui faut se placer loin en dehors du champ de vision de l'endoscope. De même que pour le foie, le diaphragme apparait souvent mais couvre une faible surface, mais dans ce cas c'est parce qu'il s'agit d'un repère anatomique important autour duquel il faut bien mener la dissection, et parce qu'il est en grande partie caché derrière l'estomac et la rate. La compresse elle, est une aide importante dans ces chirurgies pour « caler » les tissus adipeux et libérer l'espace opératoire. Elle couvre une surface importante car elle est placée au premier plan devant les tissus adipeux.

Nous avons omis de présenter un objet dans la liste présentée en figure 3.9 car il s'agit d'un artefact de segmentation. En section 3.3.4, nous avons défini des contraintes sur la tâche de segmentation, notamment le fait que tous les pixels de l'image doivent être labellisés. Or la segmentation étant manuelle, elle est par nature imparfaite, et certains pixels ne sont pas annotés. Nous avons donc groupés ces pixels sous le nom d'« espace interstitiel ». Nous ne le considérons cependant pas dans notre analyse car il ne correspond à aucun objet réellement annoté.

L'annotation des artefacts dans les images segmentées fait ressortir deux difficultés principales : la présence de fumée (dû à la dissection par effet thermique), et le flou dû aux

déplacements de la caméra. On constate que ces artefacts apparaissent dans plus de 2/3 des chirurgies, ce ne sont donc pas des phénomènes rares dans ce jeu de données. Il a donc fallu les gérer pendant l'annotation, ce qui a rendu la segmentation plus complexe. Un artefact qui n'a pas été noté mais qui pose problème est lié à l'éclairage de la scène : le haut de l'image qui est le plus éloigné de l'endoscope est donc le moins éclairé, et peut-être de fait difficile à segmenter correctement. La façon dont ces artefacts ont été annotés est très libre et certainement peu robuste. On aurait pu améliorer cette annotation en établissant une liste d'artefacts, dans laquelle chaque artefact serait clairement défini.

3.6.3.3 | Validation

Pour cette tâche de segmentation, les opérateurs ne sont pas des cliniciens mais des scientifiques, leur connaissances cliniques a priori sont donc inexistantes. C'est pourquoi la clarification du contexte clinique auprès des opérateurs était essentielle. Cela a permis de mettre à niveau les connaissances des opérateurs sur l'anatomie et la reconnaissance des structures anatomiques visibles dans la vidéo endoscopique lors de la gastrectomie longitudinale.

L'échantillon de 35 images utilisé pour l'étude de variabilité inter-opérateur a été proposé suite à cette introduction comme tutoriel. Ce sont donc les premières images qui ont été proposés aux opérateurs, sur lesquelles ils ont pris en main le logiciel de segmentation. Evidemment, cela introduit un biais dans l'étude de la variabilité inter-opérateur, car la prise en main du logiciel et l'adaptation au contexte clinique ont dû impacter la qualité des segmentations. Une étude de la variabilité intra-opérateur impliquant une ré-annotation de ces 35 images une fois le logiciel maîtrisé quelques mois plus tard permettrait de mieux évaluer ce biais.

Dans Zhou et al. (2017), le jeu de données contient des images de scènes de la vie quotidienne, avec des contenus variés et très différents de celui de nos images. En comparant deux versions segmentées à 6 mois d'intervalle d'un échantillon d'images, 17% des pixels des images étaient mal annotés en moyenne. Trois types d'erreurs sont considérées : des erreurs liées à la qualité de la segmentation et du contour, des erreurs liées aux labels donnés aux formes, et des erreurs liées au grand nombre d'objets visibles dans l'image. Zhou et al. (2017) utilisent comme métrique la précision (proportion de pixels correctement labellisés) sur l'ensemble de l'image et par label, ainsi que l'*IoU* (intersection sur l'union entre les pixels prédits et de la vérité terrain) sur l'ensemble de l'image et par label. Pour utiliser comme métrique d'évaluation la précision ou l'*IoU* telles qu'utilisées par Zhou et al. (2017), il nous aurait fallu une vérité terrain, ce que nous n'avons pas. Nous nous sommes donc contentés d'utiliser l'*IoU* comme métrique de comparaison pour nos trois opérateurs, sans faire d'hypothèse de rang entre eux trois.

La segmentation manuelle d'images est un processus présentant un certain nombre de difficulté, notamment lorsque les scènes annotées sont complexes. En comparant les segmentations des trois opérateurs sur un même échantillon de 35 images extraites des 735 échantillons, nous avons quantifié la variabilité inter-opérateur de cette tâche de segmentation. Cette comparaison nous a notamment permis de confirmer la stabilité de l'annotation menée par l'opérateur 1 vis-à-vis des deux autres. En effet, nous le considérions, a priori, comme le plus expérimenté des trois.

Pour évaluer les erreurs proposées par Zhou et al. (2017), il nous faudrait comparer nos segmentations à une vérité terrain, ce qui, dans notre cas, reviendrait à considérer les segmentations de l'opérateur 1 comme vérité terrain. Nos résultats ne nous semblent pas assez forts pour faire une telle hypothèse. Nous pourrions les renforcer en augmentant le nombre d'opérateurs inclus dans cette étude. Une autre possibilité serait de faire segmenter ce groupe d'image par un clinicien et de considérer son annotation comme notre vérité terrain. Néanmoins, le type d'erreur liée à la qualité de segmentation apparaîtrait très probablement, du fait de la complexité inhérente à ce type d'image, du manque d'habitude des opérateurs face à l'environnement chirurgicale, et des limites floues qui existent entre les différentes structures anatomiques. Le type d'erreur liée aux labels donnés aux formes devrait également apparaître, car il n'est pas toujours évident de savoir à quelle structure anatomique correspond telle forme ou telle autre. En revanche, comme nous n'avons jamais que quelques objets distincts visibles dans nos images, les erreurs liées au grand nombre d'objets visibles dans l'image ne devraient pas être observées.

Nous avons constaté que deux objets étaient segmentés de façon similaire par les trois opérateurs, ce sont la paroi abdominale et la compresse, tandis que les objets problématiques sont l'estomac et l'écarteur de foie. Lors de la segmentation, nous avions en effet noté la difficulté qu'il y avait à bien délimité l'estomac, notamment des tissus adipeux qui l'entourent, de même l'écarteur de foie se trouvaient presque systématiquement dans la partie supérieur de l'image segmentée qui est particulièrement sombre et donc difficile à segmenter avec précision. La rate et le diaphragme présentent également une qualité de segmentation très variable. De plus, la pince électro-thermale a été segmentée par les opérateurs sur des surfaces de l'image ne se recouvrant pas du tout dans 8 cas. Ces résultats nous donnent une idée du niveau de certitude associé à la segmentation de chaque objet, et nous laissent envisager de nouvelles stratégies de segmentation. On pourrait en effet imaginer pour ces objets « à risque » une redondance de la segmentation, une validation par un expert, voire de faire segmenter ces seuls objets par un expert. Ces informations pourraient également être utilisées comme carte de confiance en entrée de l'apprentissage d'un algorithme de segmentation, une telle information pourrait permettre de faire varier le niveau de précision avec lequel l'algorithme se fierait aux données de la vérité terrain.

3.6.4 | Les jeux de données en libre accès

L'analyse et la comparaison des différents jeux de données annotées existants dans notre domaine de recherche nous a permis d'observer l'évolution de ces jeux de données depuis 2014. On constate en premier lieu que les tâches d'annotation visuelle sont plus traitées que celles d'annotations procédurale et qualitative. On observe aussi que la complexité des tâches d'annotation proposées augmente pour les tâches d'annotation visuelle et procédurale. Une explication probable à ce phénomène viendrait de l'amélioration des algorithmes de reconnaissance qui remplissent des tâches de plus en plus complexes, et qui nécessitent, de fait, une complexification des jeux de données d'apprentissage.

Un fait marquant vis-à-vis de la création de ces jeux de données annotées est l'article de Maier-Hein et al. (2018). Notre analyse des métadonnées renseignées pour chaque jeu de données montre une nette augmentation de leur nombre entre les jeux de données mis en

ligne avant et après la publication de cet article. Cette augmentation est une amélioration importante pour les jeux de données annotées, dont le contenu est décrit de façon beaucoup détaillée. Selon moi, cette amélioration est due à la sensibilisation à l'importance de ces métadonnées (Wilkinson et al., 2016), ainsi qu'au formalisme proposé, qui est un outil très pertinent pour aider à cette formalisation.

Al Hajj et al. (2019) a été publié afin de décrire un jeu de données *CATARACTS* et les méthodes appliquées pour résoudre les tâches définies dans le cadre du challenge associé. *EndoVis 2019 RMIS* et *EndoVis 2019 SWSA* donneront lieu prochainement à des articles similaires. Dans ces études, les données proposées ont autant, voire plus d'importance que les méthodes qui leurs sont appliquées. La mise en place du formalisme de description des métadonnées traduit également l'importance croissante que prennent les données d'apprentissage dans notre domaine de recherche. A noter qu'*EndoVis 2019 SWSA* est le seul à proposer les trois formes d'annotation. Mais nous n'avons malheureusement pas pu récupérer de données et n'avons donc pas pu mener de comparaison pour ce jeu de données. Au vu des métadonnées disponibles, il semble que ce soit une annotation comparable à la nôtre avec néanmoins une annotation visuelle par étiquetage et une évaluation des compétneces chirurgicales à travers le score structuré OSATS (7 composantes)

Le jeu de données *CaDis* présente une évolution plus collaborative dans la création de jeux de données annotées de plus en plus complexe et qualitatifs. En effet, l'annotation visuelle proposée ici se base sur les données brutes ainsi que sur l'annotation proposées dans *CATARACTS*. Cette approche est un gain de temps important car il n'y a pas eu besoin de mettre en place d'étude clinique pour constituer le jeu de données brutes initiales d'une part, mais également car, en se basant sur les données d'annotation préexistantes (telles que la présence/absence d'instruments dans les images, ou l'annotation des phases), il a été possible d'établir une stratégie d'annotation plus pertinente et efficace. Bien entendu, une telle approche doit se faire en toute transparence, et avec l'accord des différentes entités ayant participé à la construction du jeu de données.

La validation du processus d'annotation n'est pas traitée suffisamment en profondeur. Dans l'étude de Maier-Hein et al. (2018), seulement trois des métadonnées traitent de cette validation : « Pratique d'annotation des cas d'entrainement de test », « Méthode d'agrégation des annotations des cas d'entrainement/ de test », « Sources d'erreurs potentielles ». Et seulement 3 des jeux de données étudiés (*CATARACTS*, *EndoVis 2019 RMIS* et *EndoVis 2019 SWSA*) décrivent ces métadonnées. De plus, les métadonnées proposées ne nous paraissent couvrir cette question de la validation que d'une façon partielle et omettent notamment de questionner la variabilité intra- et inter-opérateur. Selon nous, ce manque d'exhaustivité est problématique, car c'est la qualité des données même qui est en jeu. Et si on est incapable d'évaluer le niveau d'erreur ou la variabilité des données annotées fournies pour l'apprentissage d'algorithmes, comment être capable d'évaluer correctement les performances de ces algorithmes entraînés à remplir des tâches cliniques ? Nous espérons donc que la validation de l'annotation deviendra, à l'avenir, un sujet de préoccupation de plus en plus important (Wilkinson et al., 2016). Et c'est pourquoi nous avons proposé des approches de validation pour deux de nos trois annotations, notamment sur la question de la variabilité inter-opérateur.

En résumé

— La présence ou non d'un clinicien lors de la tâche d'annotation de la procédure modifie clairement le contenu annoté. La durée de l'étape de dissection du fundus semble notamment difficile à annoter pour un opérateur non-clinicien.
— La paroi abdominale et la compresse sont les deux objets avec le plus haut niveau d'accord entre les 3 opérateurs, tandis que l'estomac, l'écarteur de foie, la rate et le diaphragme présentent un plus haut niveau de désaccord.
— Mener une validation des tâches d'annotation permet de mieux comparer et évaluer leur qualité, mais de tels processus sont encore rares et incomplets.
— Une bonne validation nécessite de comparer un nombre suffisant d'opérateurs ou d'échantillons.
— Pour améliorer l'analyse des annotations, un enregistrement de la durée nécessaire pour chaque annotation serait un ajout intéressant.

3.7 | Conclusion

Dans cette étude nous avons décrit le processus complet de création d'un jeu de données chirurgicales per-opératoires annotées, depuis le protocole clinique permettant la mise en place d'une cohorte de patients, à la description détaillée du contenu annoté, en passant par la conception d'une méthodologie d'annotation. Ce travail est finalement essentiel car ces données qui ont été recuillies orientent fortement nos choix futurs d'études. Nous avons donc dû anticiper nos objectifs pour les autres études de cette thèse afin de créer des données en conséquence.

Nous avons surtout proposé une annotation très fournie et multiforme de ces vidéos laparoscopiques, nous nous avons en parti évalué la variabilité de ces annotations. De plus, nous avons mené une étude comparative du contenu des différents jeux de données annotées récents et en libre accès, ce qui nous a aussi permis de précisément décrire l'apport de notre travail à la communauté scientifique.

4

Étude 2 : Prédiction de descripteurs caractérisant la pratique du chirurgien par optimisation d'un algorithme d'apprentissage

Préambule

Dans cette étude, nous présentons une méthodologie de prédiction de la qualité d'exposition ainsi que du profil de pratique du chirurgien. Ces deux variables de sortie ont été prédites à partir des données d'annotation procédurale et de segmentation qui ont été mis au préalable sous la forme de variables dans une matrice d'entrée.

Nous introduirons en premier lieu cette étude et ses objectifs en section 4.1. Nous présenterons ensuite les variables d'entrée et de sortie en section 4.2. Puis nous décrirons notre méthodologie de prédiction, d'optimisation etd'étude d'impact des variables d'entrée en section 4.3. Nous présenterons nos résultats en section 4.4, nous discuterons nos résultats ainsi que nos choix méthodologiques en section 4.5 et nous conclurons enfin sur cette étude en section 4.6.

Sommaire

4.1 | Introduction

L'annotation du jeu de données LapEx présentée dans l'étude 1 nous a permis de mieux évaluer les informations contenues dans ces données. Les données annotées proposent aussi une information beaucoup plus facile à exploiter que les vidéos brutes. Nous avons notamment extrait de la vidéo des annotations sur l'aspect procédural, sur la qualité d'exposition, sur le profil de pratique, et sur le contenu visuel de la vidéo chirurgicale. Je vais désormais expliquer comment nous avons utilisé ces données dans un environnement d'apprentissage conçu pour prédire la qualité de gestion de l'exposition de la cible chirurgicale et le profil de pratique du chirurgien. Lors de cette conception, nous avons considéré les problématiques suivantes :

— Sous quelle forme doit-on présenter les données annotées pour pouvoir leur appliquer un algorithme prédictif ?
— Quelles précautions doit-on prendre pour que l'utilisation de ces données ne mènent pas à des résultats biaisés ?
— Quel traitement algorithmique permet de prédire correctement tant la qualité d'exposition que le profil de pratique ?

Nous nous sommes également intéressés à l'impact des données issues de l'annotation procédurale et de la segmentation sur les performances de la tâche de prédiction. Pour pouvoir mener une telle analyse, les contraintes suivantes étaient nécessaires :

— Les données utilisées doivent être interprétables.
— Le traitement algorithmique doit permettre d'analyser les liens entre les données en entrée et les données prédites en sortie.

4.2 | Matériel

Dans l'étude précédente j'ai introduit notre jeu de données annotées LapEx. Pour mener à bien notre tâche de prédiction, il nous a fallu présenter ces données sous une forme adaptée au type d'algorithme que nous avons implémenté. Plus précisément, nous avons présenté les données sous la forme d'une matrice d'entrée X de taille $n_{ech} \times n_{var}$, et d'un vecteur de sortie y de taille $n_{ech} \times 1$, avec n_{ech} le nombre d'échantillons et n_{var} le nombre de variables. Dans la suite de cette étude, nous appelons *échantillon* une ligne de la matrice d'entrée ou du vecteur de sortien, correspondant à une instance de l'annotation de la qualité d'exposition. D'après notre étude du chapitre 3, on a donc $n_{ech} = 735$.

Nous avons appliqué un ensemble de transformations aux données annotées pour obtenir des variables numériques ou des labels, avec la contrainte que ces transformations devaient proposer des résultats sous une forme interprétable. En effet, comme notre but final est d'analyser nos résultats de prédiction et d'étudier les notions de qualité d'exposition et de profil de pratique, il nous fallait comprendre à quoi correspondaient les variables utilisées pour prédire ces notions.

4.2.1 | La forme des variables de qualité d'exposition et du profil de pratique

Le résultat de l'annotation de la qualité d'exposition montrait une répartition fortement déséquilibrée des échantillons entre les « bonnes » expositions (84%), les expositions « bonnes » (11%), et les expositions « insuffisantes » (4%). Ce fort déséquilibre est problématique pour la tâche de prédiction, aussi nous avons regroupé les expositions suffisantes et insuffisantes pour diminuer le déséquilibre. La variable décrivant la qualité d'exposition que l'on a cherchée à prédire est donc une variable binaire :

— 0 est la valeur correspondant aux expositions dites « perfectibles », elle couvre 16% des échantillons.
— 1 est la valeur correspondant aux « bonnes » expositions, elle couvre 84% des échantillons.

La variable décrivant le profil de pratique du chirurgien décrit soit le chirurgien 0 (55% des échantillons), soit le chirurgien 1 (45% des échantillons). C'est donc déjà une variable binaire et nous l'avons gardée comme telle.

Dans la suite de cette étude, nous appellerons *qualité* la variable binaire décrivant la qualité de gestion de l'exposition de la cible chirurgicale, et *chirurgien* la variable binaire décrivant le profil de pratique du chirurgien.

4.2.2 | Les descripteurs utilisés pour l'annotation procédurale

En ce qui concerne les données procédurales, nous considérons $t_{exp,i}$ l'instant auquel l'instance de *qualité* i a été annotée, et nous avons défini l'intervalle de temps T_i associé à cette même instance de *qualité* par :

$$T_i =]t_{exp,i-1}; t_{exp,i}] \tag{4.1}$$

Nous avons donc découpé la durée de la procédure en autant d'intervalles temporels qu'il y a d'instances de *qualité* annotées dans la procédure (voir figure 4.1).

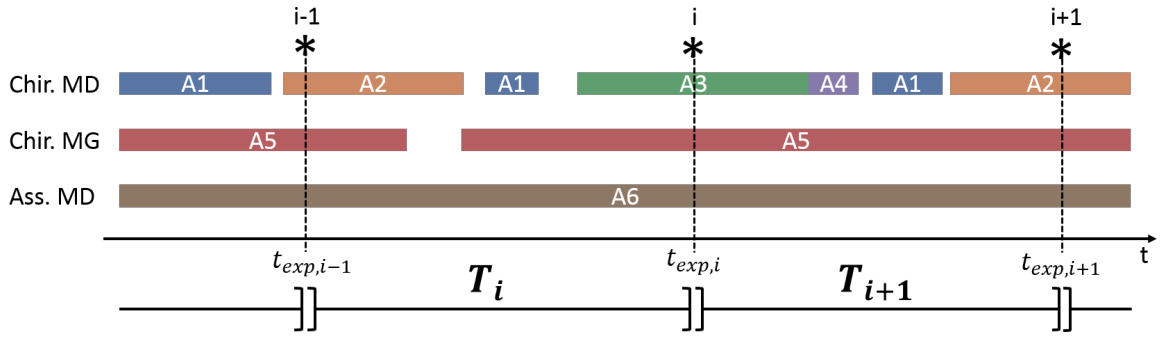


FIGURE 4.1 – Définition de l'intervalle de temps T_i associé à l'exposition i - avec (A1, ..., A6) les différentes activités, Chir. pour Chirurgien, Ass. pour Assistant, MD pour main droite et MG pour main gauche.

Afin de décrire le contenu procédural des étapes de dissection du fundus étudiées, nous avons proposé un ensemble de descripteurs pour chaque échantillon (voir tableau 4.1). Pour un échantillon donné, une activité intervient dans l'échantillon si elle commence, a lieu, ou finit pendant l'échantillon. Il n'y a pas de contraintes sur la durée de l'activité, qui peut être plus longue que la durée de l'exposition considérée. Les descripteurs définis dans ce qui suit sont calculés pour chaque échantillon.

Nom de la descripteur	Type	# de variables	# d'effecteurs
Moyenne/écart-type de la durée des activités	réel	2	4
Entropie d'ordre 0 des activités	réel	1	4
Entropie d'ordre 1 des activités	réel	1	4
Triplet de l'activité la plus courte/longue	label	6	4
Durée d'exposition	entier	1	/
# de schémas d'activités	entier	11	1
# de schémas de verbes	entier	10	1
# de schémas de cibles	entier	16	1

Tableau 4.1 – Description des descripteurs calculés à partir des annotations procédurales pour un total de 78 variables, les labels d'activités respectent le formalisme du triplet <verbe, instrument, cible> et sont basés sur le vocabulaire défini dans la section 3.2.3.

Tout d'abord, nous avons comme variable la durée associée à l'intervalle de l'échantillon, c'est à dire la durée de l'exposition. On calcule aussi la moyenne/écart-type des durées d'activité pour chaque effecteur (les deux mains du chirurgien et les deux mains de l'assistant). On calcule les entropies E_0 d'ordre 0 et E_1 d'ordre 1 des activités qui sont définies pour les séquences d'activités de chaque effecteur comme :

$$E_0 = -\sum_{i=1}^{A} n_{act,i} \times log_2(n_{act,i}) \tag{4.2}$$

$$E_1 = -\sum_{i=1}^{A}\sum_{j=1}^{A} n_{tr,i,j} \times log_2(n_{tr,i,j}) \tag{4.3}$$

Avec A le nombre de types d'activité différents, $n_{act,i}$ qui correspond au nombre d'apparitions de chaque type d'activité, et $n_{tr,i,i}$ qui correspond au nombre de transitions entre les différents types d'activité. Ainsi, l'entropie d'ordre 0 mesure la variabilité de la procédure en terme d'activités, et l'entropie d'ordre 1 quantifie la variabilité des transitions entre activités.

On a également isolé l'activité la plus courte et l'activité la plus longue pour chaque effecteur, et pour chacune de ces activités on décrit le triplet <verbe ;instrument ;cible> à l'aide de trois variables.

Huaulmé et al. (2017) définit la notion de plus long pattern/schéma séquentiel fréquent, qui lui permet d'obtenir les séquences d'évènements apparaissant plus d'un certain nombre de fois. Nous utilisons cette méthode pour sélectionner des schémas d'activités, de verbes, et de cibles récurrents dans nos procédures et ainsi caractériser les répétitions dans ces procédures. En annexe, on trouvera les plus longs schémas fréquents sélectionnés sur les activités (C.2), les verbes (C.3) et les cibles (C.1) sur l'ensemble du jeu de données. Au final, nous avons sélectionné 11 schémas d'activités, 10 schémas de verbes, et 16 schémas de cibles. Pour chaque schéma, nous avons défini une variable qui comptabilise son nombre d'apparitions sur l'échantillon considéré.

Finalement, en considérant toutes ces variables procédurales, on obtient une matrice X_{proc} de taille $n_{ech} \times n_{var,proc}$ avec $n_{var,proc} = 78$.

4.2.3 | Les descripteurs utilisés pour la segmentation

De même, nous avons réuni un ensemble de descripteurs caractérisant les objets segmentés dans chaque image. Les descripteurs sont décrits dans le tableau 4.2. Pour chaque objet, on considère l'image binaire de la forme couverte par sa segmentation. On calcule le périmètre et la surface couverte par cette forme en nombre de pixels. Les composantes x et y du barycentre de cette forme sont également calculées. On applique une analyse en composantes principales à l'image binaire pour déterminer les directions principales de cette forme, on obtient ainsi les coordonnées des deux vecteurs propres ainsi que le rapport des valeurs propres de ces directions principales. En appliquant la méthode des « local binary patterns » (Ojala et al., 2001), on quantifie la texture sur les pixels de cette forme. Enfin on calcule la couleur moyenne de la forme dans l'espace de couleur CIELAB.

Nom de la descripteur	Type	# de variables	# d'objets
Périmètre	entier	1	10
Surface	entier	1	10
Barycentre	entier	2	10
Directions principales	réel	5	10
Texture	entier	1	10
Couleur	entier	3	10

Tableau 4.2 – Description des descripteurs calculées pour chaque image segmentée pour un total de 130 variables.

Pour l'image segmentée associée à chaque échantillon on a calculé 13 variables pour

chacun des 10 objets segmentés dans les images. Nous avons donc une matrice X_{seg} de taille $n_{ech} \times n_{var,seg}$ avec $n_{var,seg} = 130$. On notera que tous les objets n'apparaissent pas nécessairement dans chaque image, aussi certaines valeurs de cette matrice ne sont pas renseignées et sont manquantes.

En résumé

— A partir de l'annotation des échantillons de qualité d'exposition, nous avons construit un vecteur binaire y_{qualit} de taille 735×1.
— En considérant le profil de pratique du chirurgien, nous avons construit un second vecteur binaire $y_{chirurgien}$ également de taille 735×1.
— Pour pouvoir mener un traitement algorithmique adapté, nous avons créé des variables à partir de les données d'annotation brutes. Ces variables ont été créées de façon à être interprétables, dans l'optique d'une analyse future.
— Les données issues de l'annotation procédurale ont été transformées pour construire une matrice X_{proc} de taille 735×78.
— Les données issues de la segmentation ont été transformées pour construire une matrice X_{seg} de taille 735×130.
— La concaténation de ces deux matrices nous a donné une matrice $X = X_{proc+seg}$ de taille 735×198.

4.3 | Méthodologie

A partir de cette matrice de variables caractérisant les aspects procédural et spatial des vidéos chirurgicales, nous avons pu envisager une analyse algorithmique et statistique des variables de *qualité* et du *chirurgien*. Afin de concevoir correctement notre approche algorithmique d'apprentissage machine ou « machine learning », nous devons correctement caractériser les données. Antonelli et al. (2019) présentent les problèmes principaux liés aux données qui doivent être considérés lors de la conception d'une approche d'analyse statistique. En nous inspirant de leurs observations, nous avons conçu notre méthodologie de prédiction de la qualité d'exposition en prenant en compte l'influence des problématiques propres à nos données.

4.3.1 | La gestion des valeurs manquantes

4.3.1.1 | Cas général

Lorsqu'un jeu de données est créé, il n'est pas rare d'y trouver des valeurs manquantes. De manière générale, les valeurs manquantes proviennent du processus d'acquisition des données et on distingue trois causes possibles (Wei et al., 2018) :

— Certaines valeurs peuvent se trouver en dehors des limites d'acquisition (missing not at random MNAR).
— Les valeurs manquantes proviennent d'une configuration particulière de l'expérience étudiée, et sont potentiellement expliquées par d'autres variables (missing at random MAR).
— Les valeurs manquantes peuvent également être dues à des erreurs ou des variations inattendues et stochastiques du dispositif d'acquisition (missing completely at random MCAR).

On peut illustrer ces trois causes à travers un exemple simple : un capteur mesurant la concentration en monoxyde de carbone dans l'air. Si la sensibilité du capteur est insuffisante, alors certaines concentrations faibles mais non nulles seraient mesurées comme valant 0, ce qui correspond à des valeurs MNAR. Autrement, il se pourrait qu'en présence d'un certain agent chimique, le capteur soit neutralisé et mesure des concentrations nulles. dans ce cas, en détectant la présence de cet agent perturbateur, on pourrait expliquer les valeurs manquantes MAR. Enfin, pendant une coupure de courant complètement indépendante de l'environnement étudié, le capteur ne peut plus effectuer de mesures pendant une durée donnée, et les valeurs manquantes MCAR sont alors impossibles à anticiper.

Pour résoudre ce problème de valeurs manquantes, l'approche la plus simple consiste à remplir toutes les valeurs manquantes avec une valeur choisie qui peut être la moyenne, la médiane, le minimum ou la moitié du minimum. Si on considère comme variable la concentration en monoxyde de carbone dans l'air, et qu'on remplace les valeurs manquantes par 0, qui correspond en réalité à l'absence de monoxyde de carbone, on se rend bien compte qu'on propose une information complètement fausse. On rencontrerait le même problème en remplissant les valeurs manquantes avec la moyenne ou la médiane. En revanche une valeur aberrante telle qu'une concentration négative pourrait être utilisée et représenterait

mieux ce qu'est une valeur manquante.

D'autres approches moins simplistes peuvent être envisagées, dans lesquelles on prend en compte le type de valeur manquante (MNAR, MAR et MCAR). Par exemple, pour traiter les valeurs MAR, manquantes pour des causes qui peuvent être explicitées et modélisées en considérant les autres variables existantes, les algorithmes de forêt aléatoire ou « random forest » et de K plus proches voisins ou « K-nearest neighbors » ont été utilisés avec succès. Les valeurs MNAR, manquantes car en deçà des seuils de détection, ont pû être modélisées en imputant leurs valeurs par régression quantile (Antonelli et al., 2019).

4.3.1.2 | Cas de l'étude LapEx

La matrice de données X_{seg} issue du calcul de descripteurs sur les données de segmentation présente des cas de valeurs manquantes (voir section 4.2.3). En effet cette matrice est construite avec 13 variables décrivant chacun des 10 objets pouvant potentiellement être segmentés, pour un total de 130 variables. Dans la réalité, les 10 objets n'apparaissent pas systématiquement dans les images segmentées. Si un objet n'apparait pas dans l'image associée à un échantillon, les 13 variables qui le décrivent sont manquantes et il nous a fallu remplir ces valeurs pour pouvoir ensuite traiter ces données.

Dans notre cas donc, une valeur est manquante quand l'objet qui lui est associé n'apparait pas dans l'image segmentée. Le remplacement de valeur que nous avons mis en place cherche, autant que possible, à traduire cette idée que l'objet n'est pas visible dans l'image. Ainsi, pour le périmètre et la surface, nous avons choisi la valeur 0 qui exprime bien le fait que l'objet ne couvre aucun pixel de l'image. Pour les coordonnées du barycentre, les coordonnées de vecteurs propres et le rapport des valeurs propres des directions principales et les composantes de la couleur nous avons choisi la valeur -1. Cette valeur aberrante ne correspond à aucun cas réel et peut donc être utilisée pour exprimer la notion d'absence de l'objet dans l'image. Le seul cas réellement problématique est celui de la texture qui est un réel couvrant potentiellement toute les valeurs possibles. Nous avons donc simplement choisi de remplacer cette valeur par la moyenne (Antonelli et al., 2019).

4.3.2 | L'association de variables de différents types

4.3.2.1 | Cas général

Ce problème est spécifique à notre approche. Les variables que nous traitons sont de types différents. Nous analysons en même temps des réels, des entiers, des durées et des labels (chaîne de caractères). L'association de ces différents types de données pose problème, notamment lorsqu'on doit les standardiser. Lors d'une standardisation on doit d'abord centrer les valeurs, puis leur soustraire leur variance. Ainsi chaque variable est centrée autour de 0, et on évite les problèmes d'échelle. Mais lorsqu'on traite des chaînes de caractères, cette standardisation n'est pas possible, et si on centre des durées autour de 0, certaines durées sont négatives, ce qui est aberrant. Nous devons donc procéder autrement pour gérer ces contraintes.

Pour résoudre ce problème, il est possible de convertir les différents types de données vers un type de données unique. La méthode de conversion est différente selon les types

de données, mais l'idée générale est de binariser les données. Pour les chaînes de caractères on effectue un encodage « one-hot » qui consiste à créer un vecteur binaire ayant autant de variables qu'il y a de valeurs possibles (de labels possibles), et d'indiquer par la valeur 1 le label observé, et par 0 tous les autres. Pour les nombres, réels ou entiers, on effectue un échantillonnage sur l'espace de définition de la variable, et on indique par la valeur 1 le niveau observé, et par 0 tous les autres.

On obtient ainsi un vecteur de valeurs binaires qui, du fait de la binarisation de l'information, compte beaucoup plus de variables que le vecteur initial.

4.3.2.2 | Cas de l'étude LapEx

Dans la matrice X_{proc}, les variables procédurales présentent deux cas : ce sont soit des réels, soit des labels (voir tableau 4.1). Dans la matrice X_{seg}, les variables spatiales sont soit des réels soit des entiers.

Pour toutes les variables réelles ou entières, nous avons appliqué une discrétisation à 10 niveaux, qu'elles soient procédurales ou spatiales, et nous avons appliqué un encodage one-hot aux variables procédurales se présentant sous la forme de labels.

4.3.3 | L'algorithme prédictif dans LapEx

Notre algorithme est défini comme un pipeline constitué de trois processus successifs : un prétraitement des données, suivi d'une réduction de dimension, et enfin une classification. Nous avons appliqué cette classification aux données $(X, y_{qualit}, y_{chirurgien})$.

4.3.3.1 | Le prétraitement adaptatif

Nous avons d'abord conçu un prétraitement adaptatif dépendant des types de données qui lui sont proposés en entrée (voir figure 4.2). Comme les descripteurs que nous traitons prennent des formes variées (labels, nombres réels et durées), nous avons choisi de traiter chacune de ces formes de données différemment. Ce choix a été fait afin de ne pas perdre d'informations ou de mal interpréter l'information contenue dans ces formes.

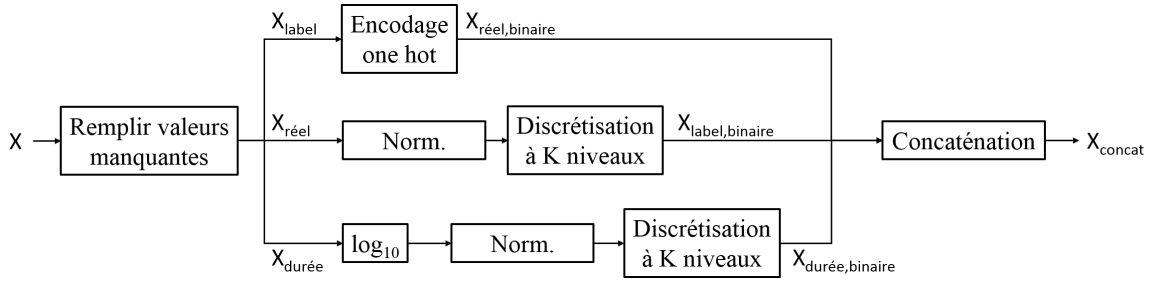


FIGURE 4.2 – Prétraitement adaptatif des données en entrée.

Les labels ont été binarisés dans un encodeur « one-hot ». Les nombres réels ont été traités en commençant par les centrer sur leur médiane, puis ils ont subi une mise à l'échelle en prenant pour référence l'écart interquartile, et enfin ces valeurs ont été discrétisées sur 10 niveaux en respectant une distribution homogène. Les valeurs de durées, exprimées en

millisecondes, sont des entiers qui ont subi le même traitement que les nombres réels. Toutefois, la distribution de ces valeurs couvrant plus de cinq ordres de grandeur, elles ont été converties au préalable sous une forme logarithmique en base 10.

Finalement, les trois types de descripteurs ont été transformés en séquences de valeurs binaires qu'on a concaténées.

4.3.3.2 | La réduction de dimension supervisée

La deuxième étape du pipeline est un algorithme de réduction de dimension utilisé dans le but de réduire le nombre de variables en entrée de l'étape de classification, tout en sélectionnant les plus pertinentes au regard de cette tâche de classification. Nous avons choisi une méthode appelée l'analyse linéaire discriminante (ALD). Cette méthode se rapproche de la célèbre analyse en composantes principales (ACP), et la transformation opérée par ces deux traitements prend la même forme :

$$X = T_A.P_A^T + R_A \tag{4.4}$$

Avec X la matrice $N \times M$ des données d'entrée, où N est le nombre d'échantillons, et M le nombre de variables. T_A est la matrice $N \times A$ de projection des échantillons en entrée dans le sous-espace des A composantes principales. P_A est la matrice $M \times A$ des vecteurs propres multipliés par leurs valeurs propres, les colonnes de P_A sont les A vecteurs propres non standardisés. R_A est la matrice $N \times M$ des résidus.

Il y a cependant deux différences fondamentales entre ces deux traitements. L'ACP est non-supervisée, et sélectionne les variables en fonction de la quantité de variance qu'elles représentent. A l'inverse, l'ALD est supervisée, et la sélection des variables se fait en fonction de leur pouvoir discriminant vis-à-vis de la variable à prédire en sortie.

On a défini le ratio de pouvoir discriminant *pd_ratio*, un hyperparamètre à optimiser. Cet hyperparamètre est défini par :

$$pd\_ratio \leq \Sigma e_i, \quad avec \quad pd_ratio \in [0;1] \tag{4.5}$$

Avec les e_i qui sont les valeurs propres sélectionnées par l'ALD. De plus, nous définissons une statistique a priori sur la distribution des valeurs de la sortie prédite pour contrer le fort déséquilibre entre les deux valeurs de la variable *qualité* : la valeur « bonne » couvre 84% des échantillons, quand la valeur « perfectible » en couvre 16% (voir section 4.2.1).

4.3.3.3 | La classification binaire

La dernière étape de classification de notre pipeline est une Machine à Vecteurs de Support ou « Support-Vector Machine » (SVM) avec pour noyau une « radial-basis function ».

Le principe du SVM est de chercher un hyperplan dans l'espace à N dimensions (N étant le nombre de variables) qui permette de classifier au mieux les échantillons $(\overrightarrow{x_i}, y_i) i \in [1; N]$. Le meilleur hyperplan est défini comme celui qui propose la plus grande distance entre les points des deux classes à prédire. Les vecteurs de supports sont les points les plus proches de l'hyperplan qui influencent fortement sa position, et à partir desquels est construit cet

hyperplan. Dans le cas où l'hyperplan peut être défini linéairement par son vecteur normal $\overrightarrow{w}$ et sa distance à l'origine b, cette définition est contrainte par la distance suivante :

$$\forall i \in [1;N], \begin{cases} \overrightarrow{w}.\overrightarrow{x_i} - b \geq 1, & \text{si } y_i = 1 \\ \overrightarrow{w}.\overrightarrow{x_i} - b \leq -1, & \text{si } y_i = -1 \end{cases} \tag{4.6}$$

Dans les cas où on ne trouve pas de solution linéaire satisfaisante à ce problème, des approches non-linaires existent. On définit alors la fonction de coût de Hinge :

$$h_i = max(0, 1 - y_i * (\overrightarrow{w}.\overrightarrow{x_i} - b)) \tag{4.7}$$

Avec cette fonction de coût on pénalise les points se trouvant du mauvais côté de l'hyperplan, et on obtient 0 si les points sont bien placés. On optimise notre hyperplan en minimisant le critère suivant :

$$\frac{1}{N}\sum_{i}^{N} h_i + C\|\overrightarrow{w}\|^2 \tag{4.8}$$

Ici le paramètre de coût C permet de trouver un compromis entre augmenter la marge, et assurer que les $\overrightarrow{x_i}$ se trouvent du bon côté de l'hyperplan.

On propose une approche non-linéaire en transformant le produit scalaire dans la fonction de coût de Hinge. On définit alors comme fonction noyau ou « kernel » la « radial basis function » gaussienne (rbf) :

$$k(\overrightarrow{x_i}, \overrightarrow{x_j}) = exp(-\gamma\|\overrightarrow{x_i} - \overrightarrow{x_j}\|^2), \quad avec \quad \gamma > 0 \tag{4.9}$$

Ainsi, nous avons choisi d'implémenter un SVM à noyau rbf pour remplir notre tâche de classification, et nous avons optimisé ses hyperparamètres C et γ.

4.3.4 | L'optimisation et la recherche des meilleurs paramètres

L'étape centrale pour concevoir un modèle est l'optimisation de l'algorithme et la sélection des valeurs d'hyperparamètres présentant les meilleures performances. Différentes stratégies existent pour mener à bien cette stratégie, et des choix doivent être faits. La stratégie que nous avons choisie est assez complexe, aussi nous allons l'introduire à travers 4 stratégies de difficulté croissante.

4.3.4.1 | Stratégie 1 : la validation croisée (VC) entraînement-validation

La stratégie basique de validation croisée (VC) nécessite une phase d'entraînement suivi d'une phase d'évaluation (voir figure 4.3). Pour ce faire, on définit la matrice X de données de taille $n_{ech} \times n_{var}$ et le vecteur y de taille $n_{ech} \times 1$ avec n_{ech} le nombre d'échantillons et n_{var} le nombre de variables.

On sépare les données en deux plis avec d'une part la matrice X_{train} de taille $n_{ech,train} \times n_{var}$ et le vecteur y_{train} de taille $n_{ech,train} \times 1$, et d'autre part la matrice X_{eval} de taille $n_{ech,eval} \times n_{var}$ et le vecteur y_{eval} de taille $n_{ech,eval} \times 1$, tels que $n_{ech} = n_{ech,train} + n_{ech,eval}$. En pratique on prend $n_{ech,train} \simeq 0.75 * n_{ech}$ et $n_{ech,eval} \simeq 0.25 * n_{ech}$.

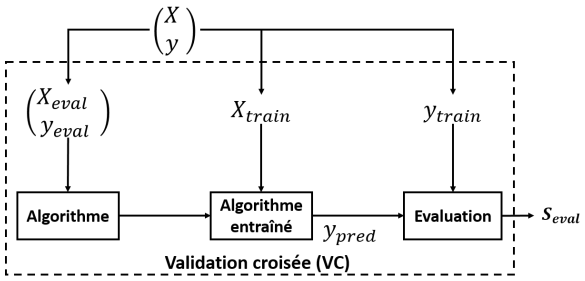


FIGURE 4.3 – La validation croisée Entrainement-validation

Pendant la phase d'entraînement, on présente chacun des échantillons du couple (X_{train}, y_{train}) à l'algorithme qui ajuste ses paramètres internes afin d'avoir la meilleure concordance avec ces $n_{ech,train}$ échantillons. Ensuite, pendant l'évaluation, on évalue la qualité de cet ajustement sur les $n_{ech,eval}$ échantillons. Ici, on présente la matrice X_{eval} n'ayant pas été utilisée pendant l'entraînement à l'algorithme, qui doit prédire les valeurs correspondantes du vecteur y_{pred} de taille $n_{ech,eval} \times 1$. La valeur de performance est obtenu en comparant le vecteur prédit y_{pred} à la vérité terrain y_{eval}grâce à un score S_{eval} que nous définierons en section 4.3.5.

Le principe essentiel de la VC est l'indépendance entre les données d'entraînement et les données d'évaluation pour éviter ce qu'on appelle le sur-apprentissage. En effet la qualité d'un modèle se mesure en évaluant sa capacité à prédire de nouvelles données, et non pas celles qu'il connait déjà. Autrement on ne mesure aucunement la capacité de généralisation du modèle étudié, et les résultats obtenus sont fortement biaisés. Nous reviendrons par la suite sur cette notion de biais lié aux données.

4.3.4.2 | Stratégie 2 : Leave-one-group-out LOGO

Dans la stratégie 1, nous avons séparé les données (X, y) en deux plis P_{train} avec (X_{train}, y_{train}) et P_{eval} avec (X_{eval}, y_{eval}). Si le pli P_{eval} compte suffisamment d'échantillons pour être représentatif de la variabilité du phénomène étudié, alors on peut considérer que la performance calculée est non-biaisée, mais c'est un cas idéal rarement rencontré. Dans la réalité, le nombre d'échantillons à disposition est insuffisant pour satisfaire cette hypothèse, et on doit trouver des approches alternatives :

On sépare désormais les données (X, y) en P plis avec pour données associées les matrices X_p de taille $n_{ech,p} \times n_{var}$ et les vecteurs y_p de taille $n_{ech,p} \times 1$ qu'on définit par :

$$n_{ech,p} = \frac{n_{ech}}{P} \tag{4.10}$$

Successivement, chacun des P plis sera le pli d'évaluation, tandis que les $P-1$ autres plis seront regroupés pour former le pli d'entraînement. On n'effectue donc plus une évaluation unique mais P évaluations, et la performance finale est désormais une statistique sur les P performances.

Ainsi, cette approche de VC plus complexe consiste à écarter successivement chaque groupe d'échantillons - chaque pli - qui servira pour l'étape de validation, d'où le nom de « Leave-one-group-out » ou LOGO (voir figure 4.4). En répétant la VC P fois, on présente les données sous différentes configurations à l'algorithme, ce qui augmente artificiellement la variabilité des données sur lesquelles l'algorithme s'entraîne et est évalué. La statistique

de performance obtenue au final est moins biaisée et plus représentative de la variabilité de données.

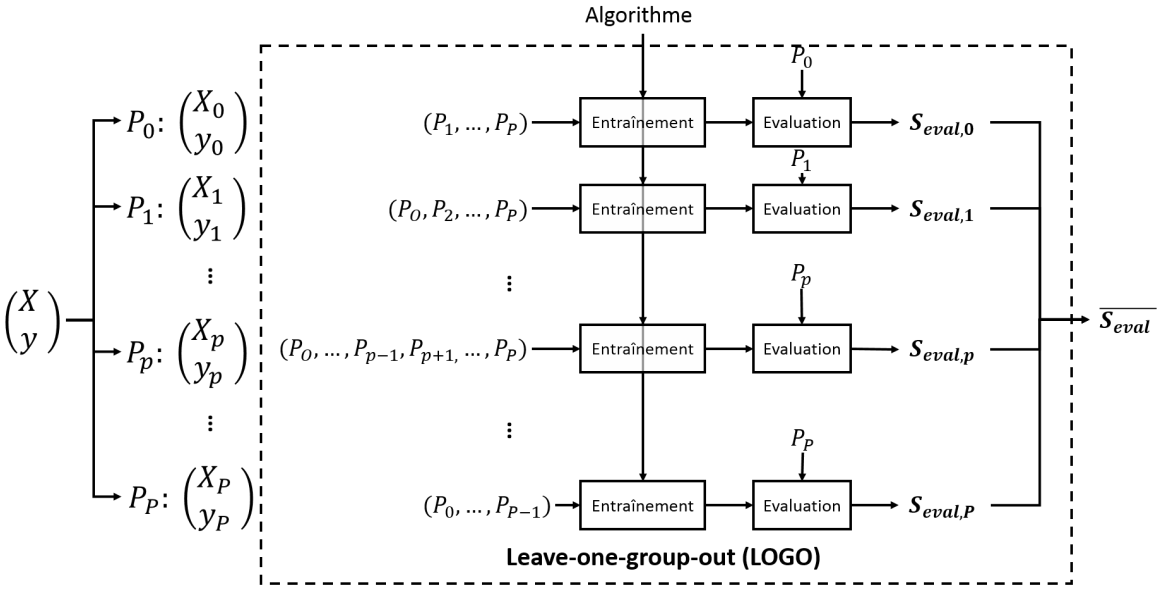


FIGURE 4.4 – La validation croisée Leave-one-group-out

4.3.4.3 | Stratégie 3 : la VC par recherche en grille ou « grid-search »

Dans la stratégie 2, nous avons évalué la statistique de performance d'un algorithme. Cette performance est-elle satisfaisante ? Une autre version de cet algorithme aurait-elle eu de meilleures performances ? On ne peut évaluer les performances d'un algorithme qu'en en testant plusieurs versions. Il faut donc évaluer plusieurs algorithmes pour finalement choisir le meilleur.

On considère donc M versions de notre algorithme et on cherche à trouver la version m_{best} ayant les meilleures performances. Ainsi en considérant une VC du type LOGO présentée dans la stratégie 2, on obtiendra une statistique de performance pour chaque version de notre algorithme et on pourra choisir la version avec les meilleures performances.

Cette méthode est appelée recherche en grille ou « grid search » car en général, on ne travaille pas sur des versions de l'algorithme, mais sur des hyperparamètres de l'algorithme qu'on teste sur un ensemble de valeurs (voir figure 4.5). Si l'algorithme présente des hyperparamètres, on mène donc une recherche en grille sur l'ensemble de combinaisons d'hyperparamètres possibles.

Cet stratégie est une méthode différente des deux précédentes, car son objectif n'est pas de gérer des problématiques de biais liés aux données mais de trouver le meilleur modèle pour remplir une tâche donnée. Cette méthode d'optimisation intègre cependant une méthode de VC (LOGO) pour gérer la répartition des données. Nous tenons à distinguer ces deux types de méthode étroitement imbriqués mais ayant des objectifs bien distincts.

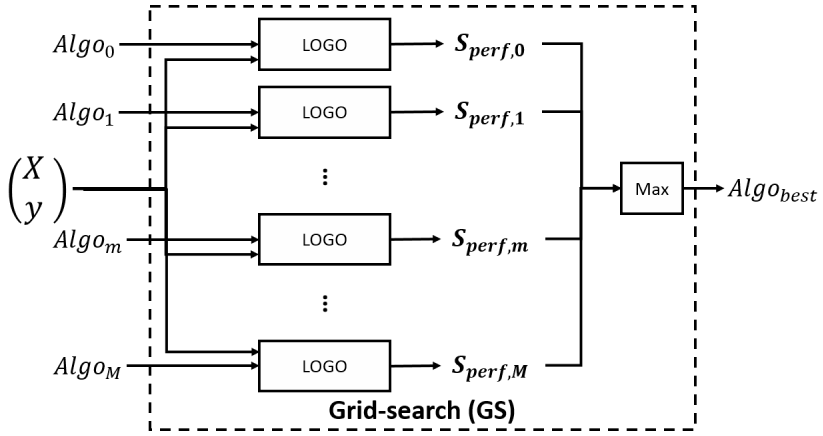


FIGURE 4.5 – L'optimisation par recherche en grille ou Grid-search

4.3.4.4 | Stratégie 4 : la VC imbriquée

La recherche en grille permet de sélectionner le meilleur modèle parmi une sélection de modèles. Par contre, cette méthode ne permet pas de correctement évaluer la performance du modèle « gagnant ». En effet, le modèle est bien sélectionné pour ses meilleures performances au cours de la grid-search VC, mais ces bonnes performances ont été obtenues sur un certain jeu de données. La situation est similaire à la stratégie 1, dans lequel on évaluait le modèle sur un pli de données différent du pli d'entraînement : l'optimisation des hyperparamètres menée dans la grid-search peut finalement être considérée comme une phase d'entraînement préalable, suite à laquelle on va mesurer objectivement les performances de la meilleure configuration d'hyperparamètres. C'est donc encore une fois un biais de données qui nécessite une extension de méthode de VC.

On considère donc nos donnée (X, y) sur lesquelles on effectue une VC avec deux plis : le pli d'optimisation avec ses données (X_{opt}, y_{opt}) et le pli de test avec ses données (X_{test}, y_{test}). Ici encore, les nombres d'échantillons valent : $n_{ech,opt} \simeq 0.75 * U$ et $n_{ech,test} \simeq 0.25 * U$. On mène une optimisation des hyperparamètres par grid-search sur le pli d'optimisation et à l'issue de cette optimisation on obtient un modèle gagnant. Puis on ré-entraine ce modèle sur l'ensemble du pli d'optimisation, et on teste ses performances sur les données du pli de test.

On peut faire la même remarque que dans la stratégie 2 sur la pertinence de la performance obtenue : pour avoir une performance bien représentative de la variabilité de les données, il faut répéter la VC suffisamment de fois. Ainsi on propose Q configurations des plis (P_{opt}, P_{test}), et on obtient une statistique de performance pour le meilleur modèle. Le modèle final est obtenu en effectuant, pour chaque hyperparamètre, une moyenne sur les Q meilleurs modèles pondérée par leur performance de test.

Cette méthode s'appelle la validation croisée imbriquée (Cawley and Talbot, 2010) parce qu'elle consiste en deux VCs, l'une imbriquée dans l'autre (voir figure 4.6) :

— *La VC extérieure* : on constitue Q configurations de VCs à deux plis, l'un avec classiquement 75% des échantillons est dédié à l'optimisation de l'algorithme, l'autre est dédié au test du modèle gagnant.

— *La VC intérieure* : cette VC s'effectue sur les données du pli d'optimisation de la VC

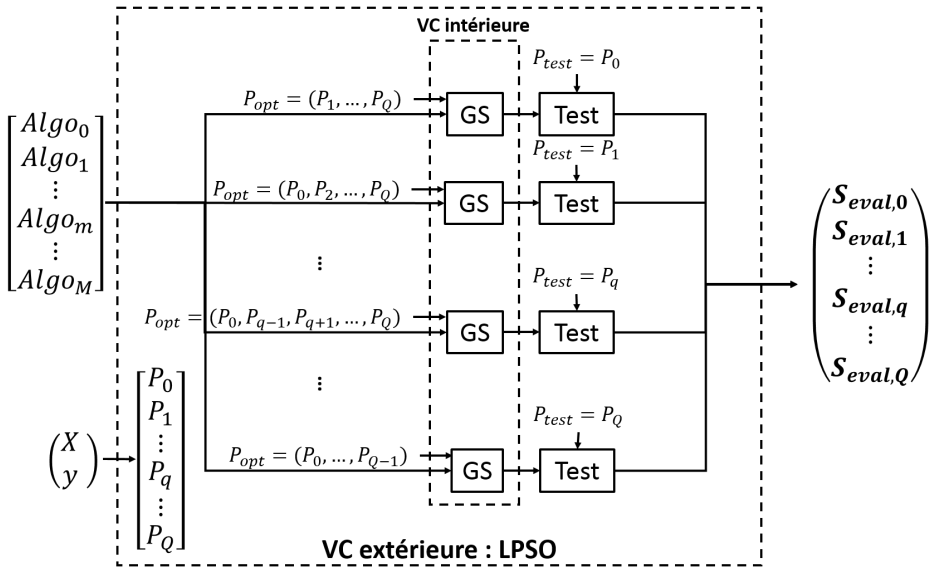


FIGURE 4.6 – La validation croisée imbriquée

extérieure. On mène une grid-search sur les hyperparamètres de l'algorithme évalué, au sein d'une VC de type LOGO (d'autres formes existent), et on sélectionne la combinaison d'hyperparamètres présentant les meilleures performances.

L'approche finale de VC imbriquée proposée ici fait penser à une poupée russe, avec des mécanismes de manipulations des données imbriquées les uns dans les autres. On notera que l'objectif principal de toute cette démarche est d'éviter des biais liés aux données, tels que le sur-apprentissage. On notera aussi que cette méthode est relativement coûteuse en temps de calcul car on répète de très nombreuses fois le processus d'apprentissage et de prédiction de l'algorithme. Plus précisément, en reprenant les notations des 4 stratégies précédents, l'algorithme est entrainé P*M*Q fois.

4.3.4.5 | Cas de LapEx

Nous avons développé une VC imbriquée afin de gérer les importants biais liés aux données lors de notre phase d'optimisation du modèle.

Dans la VC extérieure, nous avons séparé les données en deux groupes, un pli d'entraînement utilisé pour l'optimisation de notre algorithme, et un pli de test pour évaluer le modèle gagnant. Nous avons défini une stratégie de « leave-p-surgery-out » (LPSO) pour cette VC extérieure. Similaire à la LOGO, la LPSO crée des plis d'échantillons, sauf que la contrainte ne porte pas sur le nombre de plis par échantillon (voir équation 4.10). Dans notre LPSO il y a plusieurs contraintes :

— Les échantillons d'une chirurgie doivent tous être contenus par le même pli.
— On a p chirurgies par pli de test et $n_{chirurgie} - p$ chirurgies dans le pli d'optimisation. Nous avons fixé $p = 8$, ce qui nous fait 22 chirurgies dans le pli d'optimisation associé.
— On a fixé le nombre de configurations (P_{opt}, P_{test}) différentes à Q=20.

En respectant ces contraintes, nous avons respecté une hypothèse de dépendance entre les échantillons d'une même chirurgie. Et chaque chirurgie n'est apparue que dans un des

deux plis, ce qui nous a permis d'éviter un biais lié aux corrélations existantes entre les échantillons d'une même chirurgie.

Dans la VC intérieure, nous avons appliqué la stratégie LOGO en séparant les données en P=10 plis. Ici, la façon dont nous séparons nos plis doit prendre en compte la corrélation entre les échantillons d'une même chirurgie, mais également le déséquilibre entre le nombre d'échantillons ayant une « bonne » exposition et le nombre de ceux ayant une exposition « perfectible ». Nous avons choisi de ne considérer que la deuxième contrainte liée au ratio de distribution D_y des valeurs binaires de la *qualité* sur le jeu de données. De fait, les 10 plis ont été constitués de façon à respecter la contrainte suivante :

$$d_{y,i} \in [D_y - \varepsilon; D_y + \varepsilon], \quad avec \quad \varepsilon \in [0; 1] \tag{4.11}$$

Où $d_{y,i}$ est le ratio de distribution des valeurs binaires de la *qualité* pour le pli i et ε est un réel qui définit à quel point la contrainte est forte. On a ainsi contraint une répartition des valeurs prédites de façon à ce que les deux valeurs binaires soient distribuées de façon relativement homogène entre les 10 plis.

Les valeurs de *qualité* de nos échantillons ne sont pas réparties de façon homogène entre toutes les chirurgies : certaines chirurgies présentent beaucoup d'échantillons tandis que d'autres en présentent beaucoup moins (voir figure 3.8). Ainsi dans la VC intérieure, nous avons dû séparer les échantillons des chirurgies entre différents plis pour respecter la contrainte définie par l'équation 4.11. Nous n'avons donc pas respecté l'hypothèse de dépendance entre les échantillons d'une même chirurgie définie précédemment pour la VC extérieure.

4.3.5 | Le choix du score d'évaluation

4.3.5.1 | Cas général

Lors d'une VC, on est amené à évaluer les performances de prédiction de l'algorithme en comparant les valeurs prédites par l'algorithme y_{pred} à la vérité terrain y. Des scores permettent de quantifier la qualité de cette prédiction. Les scores les plus simples sont l'exactitude ou « accuracy » Acc, la sensibilité Sn et la spécificité Sp définis par :

$$Acc = \frac{VP + VN}{VP + FP + VN + FN} \tag{4.12}$$

$$Sn = \frac{VP}{VP + FN} \tag{4.13}$$

$$Sp = \frac{VN}{VN + FP} \tag{4.14}$$

Avec VP = # de vrais positifs, VN = # de vrais négatifs, FP = # de faux positifs et FN = # de faux négatifs. Tandis que l'exactitude traduit la proportion d'échantillons correctement prédits, la sensibilité traduit la proportion de valeurs réellement vraies qui ont été bien prédites, et la spécificité à l'inverse traduit la proportion de valeurs réellement fausses qui ont été bien prédites.

Ainsi, en choisissant l'exactitude comme score d'évaluation, on effectue une évaluation qui ne prend pas en considération les proportions des deux valeurs ayant été bien prédites, tandis que la sensibilité et la spécificité proposent des évaluations asymétriques se focalisant uniquement sur l'une des deux valeurs.

4.3.5.2 | Cas de l'étude LapEx

Dans notre étude, l'étape d'optimisation menée dans la VC intérieure, mais également l'étape de test dans la VC extérieure ont été évaluées avec un score appelé précision optimisée PO (Hossin and M.N, 2015), et défini comme suit :

$$PO = Acc - \frac{|Sn - Sp|}{Sn + Sp} \tag{4.15}$$

Nous avons choisi ce score afin d'éviter un mauvais apprentissage dû au fort déséquilibre des valeurs de notre variable *qualité* à prédire. La précision optimisée, de par sa définition, accorde autant d'importance à la sensibilité et à la spécificité, qui mesurent respectivement le taux de bonne prédiction des valeurs positives et des valeurs négatives de la *qualité*. Si nous avions choisi l'exactitude comme score de d'optimisation, on aurait choisit notre modèle optimal comme celui prédisant parfaitement le label positif très dominant, et pas du tout le label négatif peu représenté.

4.3.6 | protocole expérimental d'optimisation

Notre objectif était d'obtenir le modèle présentant les meilleures performances pour remplir la tâche de prédiction de la variable *qualité*. Ce modèle est établi à partir du pipeline constitué du prétraitement, de l'ADL, et du SVM. Pour obtenir le meilleur modèle, nous avons optimisé les trois hyperparamètres : pd_ratio (hyperparamètre de la LDA), C et γ (hyperparamètres du SVM). Cette optimisation a été menée dans la VC imbriquée définie précédemment. La combinaison des trois hyperparamètres donnant les meilleurs résultats a été conservée pour constituer le modèle final.

Dans la grid-search, les hyperparamètres sont définis sur l'espace de définition suivant :

— $pd\_ratio \in [0, 1]$
— $C \in [10^{-1.5}; 10^{1.5}]$
— $\gamma \in [10^{-4.5}; 10^{-2.2}]$

Sur ces intervalles, chaque hyperparamètre peut prendre 10 valeurs qui sont définies de façon à avoir une distribution homogène dans une base-10 logarithmique. On a donc M=1000 combinaisons différentes qui sont évaluées dans la grid-search, et la combinaison d'hyperparamètres sélectionnée est celle qui présente la plus haute valeur de précision optimisée. Comme la grid-search est effectuée au sein de la VC intérieure qui comporte 10 configurations différentes, nous avons répété 10 fois la grid-search, et obtenu à chaque fois une combinaison d'hyperparamètres optimaux.

Pour gérer la variabilité entre les différentes chirurgies, nous avons constitué 20 configurations différentes de la VC extérieure sur lesquelles le protocole d'optimisation est mené en entier. Par la suite, nous appellerons *config_vc* chacune de ces 20 configurations de la

VC extérieure. Le modèle final est défini en moyennant chaque hyperparamètres des modèles sélectionnés dans les 20 *config_cv*s et en pondérant ces valeurs par les performances de leur modèle. Ces 20 *config_vc*s sont elles-mêmes répétées avec comme variables d'entrée X_{seg} pour les descripteurs spatiaux (S), X_{proc} pour les descripteurs procéduraux (P), et $X = X_{proc+seg}$ pour la concaténation des descripteurs spatiaux et procéduraux (S+P). Dans la suite, les trois configurations S, P et S+P seront appelées population de descripteurs en entrée (*PDE*). On a également défini la *sortie* comme étant la variable à prédire, et $sortie \in \{qualit; chirurgien\}$.

Ainsi, notre optimisation a pour but de définir un modèle optimal avec ses trois hyperparamètres pour chacun des trois *PDE* S, P et S+P pour la prédiction de la $sortie = qualit$.

Les performances finales de ces trois modèles sont obtenues en considérant uniquement la VC extérieure pour la prédiction de la variable de *qualité* d'une part, et de la variable du *chirurgien* d'autre part. Pour chacune des 20 configurations de la VC extérieure, chaque modèle a été entraîné sur le pli d'entraînement, puis testé sur le pli de test, qui nous donne les performances finales.

4.3.7 | Les clusters de descripteurs

Un fois le modèle optimisé, on peut l'entraîner et s'intéresser à ses paramètres internes, et plus précisément à ceux de l'ALD. Cette étape de réduction de dimension est caractérisée par ses vecteurs propres et ses valeurs propres. Les vecteurs propres relient notamment les variables en entrée de l'ALD à leur projection dans l'espace réduit de l'ALD. Et les composantes de ces vecteurs propres nous informent sur la façon dont les variables en entrée ont été sélectionnées. Les variables dont les composantes dans les différents vecteurs propres sont les plus grandes sont les plus discriminantes pour prédire la *sortie*.

En considérant les trois *PDE*s, les 20 $config_vc$s et les deux *sortie*s, on a 120 configurations possibles. Pour chacune de ces configurations, l'analyse des vecteurs propres de l'ALD nous permet d'extraire un cluster constitué des descripteurs les plus discriminants. Comme $config\_vc$ exprime la variabilité de notre jeu de données, nous pouvons moyenner la constitution de nos clusters sur ces 20 configurations. Nous obtenons ainsi 6 clusters « indépendants » $C(S+P, sortie)$, $C(S, sortie)$ et $C(P, sortie)$ pour $sortie \in qualit, chirurgien$. Nous avons également créé 6 clusters d'« intersection » à partir de ces 6 clusters indépendants :

Si on prédit la *qualité*, nous pouvons extraire les descripteurs les plus discriminants qui sont communs entre les *PDE*s :

— S+P et S : $C_{qualit}(S+P \cap S)$
— S+P et P : $C_{qualit}(S+P \cap P)$

De même, en considérant qu'on prédit le *chirurgien* :

— S+P et S : $C_{chirurgien}(S+P \cap S)$
— S+P et P : $C_{chirurgien}(S+P \cap P)$

Enfin, pour chaque *PDE*, on peut extraire les descripteurs les plus discriminants communs à la prédiction de la *qualité* et du *chirurgien* :

— S+P : $C_{S+P}(qualit \cap chirurgien)$
— S : $C_S(qualit \cap chirurgien)$
— P : $C_P(qualit \cap chirurgien)$

Nous nous sommes intéressés à ces différentes combinaisons de *PDE* et de *sortie* et nous avons étudié les clusters de descripteurs indépendants et d'intersection en comparant le nombre de descripteurs qu'ils contenaient.

En résumé

- Pour mener à bien la tâche de prédiction de la *qualité*, nous avons développé un algorithme constitué d'un prétraitement, d'une réduction de dimension, et d'une classification.
- Dans le prétraitement nous avons géré les différents types de variable présents dans la matrice de données d'entrée, et nous avons uniformisé leurs types de donnée en les binarisant.
- La réduction de dimension est une transformation appelée l'analyse linéaire discriminante ALD qui sélectionne, et transforme les variables en fonction de leur capacité à discriminer la variable binaire à prédire.
- La classification est une machine à vecteurs de supports SVM qui construit un hyperplan sur l'espace de définition des données en entrée afin de séparer au mieux les deux valeurs binaires.
- Un hyperparamètre *pd_ratio* a été défini pour l'ALD et deux pour le SVM : C et γ.
- Afin de trouver la meilleure configuration de ces 3 hyperparamètres, un environnement dédié appelé validation croisée imbriquée a été mis en place. La conception de cet environnement avait notamment pour objectif de manipuler correctement les données et d'éviter les biais aux données.
- Une fois l'algorithme optimisé et le modèle optimal sélectionné, nous nous sommes intéressés plus en détail à la façon dont les descripteurs en entrée étaient sélectionnés dans l'étape de réduction de dimension. Nous avons donc étudié différents groupes de descripteurs (spatiaux ou procéduraux) et de la variable prédite (*qualité* ou *chirurgien*).

4.4 | Résultats

4.4.1 | Optimisation de l'algorithme

	pd_ratio	C	γ
S+P	0.4	5	0.001
P	0.42	79	0.01
S	0.4	2.5	0.0018

Tableau 4.3 – Valeurs des hyper paramètres sélectionnés lors de la validation croisée imbriquée pour les trois population de descripteurs en entrée - S = spatial - P = procédural.

Nous avons obtenu les combinaisons d'hyperparamètres définissant le modèle optimal pour chaque *PDE*, ces valeurs sont données dans le tableau 4.3. Une fois que le modèle a été optimisé à prédire la *qualité* pour les trois *PDE*s, nous avons étudié les performances des trois modèles obtenus pour la prédiction de la *qualité* d'une part (voir figure 4.7a), et pour la prédiction du *chirurgien* d'autre part (voir figure 4.7b). Chaque colonne correspond à une des trois *PDE*s S, P et $S + P$. Nous ne donnons pas les résultats avec le score de précision optimisée utilisé dans l'optimisation, son interprétation n'est pas aisée. Nous lui avons préféré l'exactitude, la sensibilité et la spécificité pour étudier la capacité de nos modèles à prédire les deux valeurs de la *sortie*. Nous prédisons la *qualité* avec une exactitude de 0.68 et le *chirurgien* avec une exactitude de 0.72.

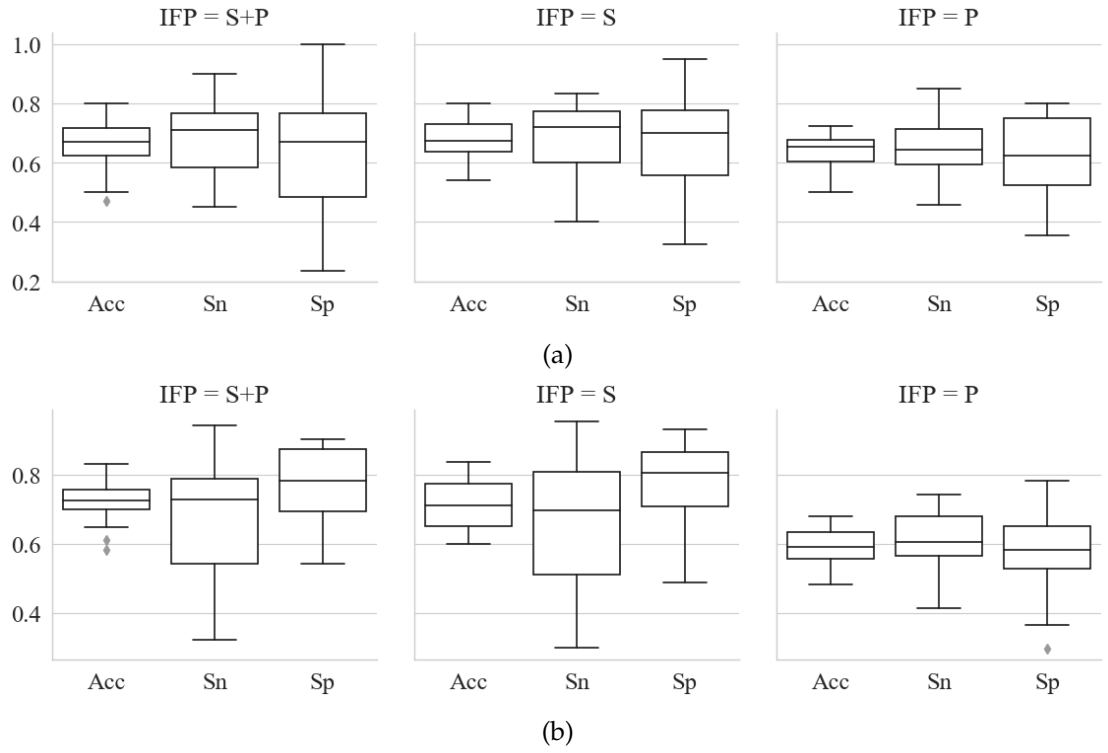


FIGURE 4.7 – Performances des meilleurs modèles pour les différentes *PDE*s et *sorties* - (a) Prédiction de $sortie = qualit$ - (b) Prédiction de $sortie = chirurgien$.

En nous intéressant aux performances de prédiction de la *qualité*, on observe que les descripteurs P procéduraux ont de moins bonnes performances que les descripteurs $S+P$ spatiaux et procéduraux ensemble. On constate également que les valeurs moyennes de spécificité et de sensibilité sont très proches, même si on observe des résultats légèrement meilleurs et plus stables pour la prédiction de la « bonne » *qualité* (sensibilité plus haute) que pour la prédiction de la *qualité* « perfectible » (spécificité plus basse). De façon similaire on observe des résultats légèrement meilleurs et plus stables pour la prédiction du *chirurgien* 0 (sensibilité plus haute) que pour la prédiction du *chirurgien* 1 (spécificité plus basse).

4.4.2 | Les clusters de descripteurs

Le tableau 4.4 présente des statistiques sur le nombre de descripteurs des clusters indépendants à gauche, et des clusters d'intersection à droite. Pour chaque cluster, la moyenne et l'écart type sont calculés sur les 20 $config_vc$s. Par exemple on peut lire :

— A gauche, la première ligne donne une moyenne de 29.9 descripteurs par cluster indépendant sur les 20 $config_vc$s pour le modèle entraîné sur les descripteurs $S+P$ à prédire la *qualité*.

— A droite, la première ligne donne une moyenne de 29.5 descripteurs communs entre les clusters des modèles entraînés à prédire la *qualité* à partir des descripteurs $S+P$ d'une part et des descripteurs S d'autre part.

Indépendant	**# de Descripteur**		**Intersection**	**# de Descripteur**	
clusters	**Moyenne**	**Ecart-type**	**clusters**	**Moyenne**	**Ecart-type**
$C(S+P, qualit)$	29.9	0.89	$C_{qualit}(S+P \cap S)$	29.5	0.92
$C(S, qualit)$	37.2	0.96	$C_{qualit}(S+P \cap P)$	0.4	0.49
$C(P, qualit)$	15.6	1.36	$C_{chirurgien}(S+P \cap S)$	29.4	0.79
$C(S+P, chirurgien)$	29.6	0.92	$C_{chirurgien}(S+P \cap P)$	0.25	0.43
$C(S, chirurgien)$	38.2	0.96	$C_{S+P}(qualit \cap chirurgien)$	19.2	1.47
$C(P, chirurgien)$	12.9	0.65	$C_S(qualit \cap chirurgien)$	26.2	1.88
			$C_P(qualit \cap chirurgien)$	11.6	0.74

Tableau 4.4 – Statistiques sur les clusters de descripteurs pour les différentes PDEs et $sortie$s, S+P=spatial+procédural, S=spatial, P=procédural.

Lorsqu'on observe les nombres de descripteurs des clusters indépendants, on voit que pour chaque *PDE*, le nombre de descripteurs sélectionnés est sensiblement le même pour les deux $sortie$s. De plus, les valeurs d'écart-type restent toujours sensiblement basses. Pour chaque *PDE*, on observe que le cluster d'intersection sur les $sortie$s $C_{PDE}(qualit \cup chirurgien)$ contient 70% des descripteurs des clusters indépendants correspondants. Enfin, les clusters d'intersection $C_{sortie}(S+P \cup S)$ sur les populations de descripteurs $S+P$ et S comptent presque le même nombre de descripteurs que leurs clusters indépendants correspondants $C(S+P, sortie)$, que la sortie soit la *qualité* ou le *chirurgien*. A l'inverse, les clusters d'intersection entre les *PDE* $S+P$ et P sont presque vides.

4.5 | Discussion

4.5.1 | Les descripteurs extraits

Les descripteurs que nous avons créés ont été choisis pour leur interprétabilité. C'est à dire le fait qu'ils soient facilement compréhensibles pour l'utilisateur et permettent une interprétation ayant un sens clinique. Nous verrons dans l'étude suivante que ce choix nous a permis de proposer des interprétations à portée clinique en nous basant sur l'analyse algorithmique de ces descripteurs.

Cette contrainte d'interprétabilité biaise notre choix de descripteurs : en effet, comme nous voulions obtenir certains résultats à portée clinique, nous avons créé nos descripteurs de façon à étudier ces résultats potentiels. De fait, nous sommes certainement passés à côté d'autres informations ou variables latentes qui pourraient être critiques. En faisant le choix de travailler directement sur les données brutes, nous raterions peut-être moins d'informations critiques, mais en supposant qu'on ait des performances de prédiction correctes avec une telle approche (ce qui n'est pas sûr), ces résultats seraient plus complexes à interpréter. Nous avons donc privilégié l'interprétabilité de nos analyses, en sachant que nous raterions d'autres informations.

Pour couvrir ces variables latentes ou d'autres aspects de la procédure chirurgicale, il serait intéressant de définir d'autres descripteurs. nous ne rentrerons pas dans le détail mais certaines études de notre état de l'art proposent de telles populations de descripteurs, tant pour l'aspect procédural que pour l'aspect visuel de la chirurgie. Nous pourrions nous en inspirer et évaluer les capacités prédictives de notre modèle à partir de descripteurs différents.

4.5.2 | Les performances de prédiction

L'évaluation de nos trois modèles (figure 4.7) montre qu'on peut prédire la *qualité* d'exposition dans la scène chirurgicale avec une exactitude de 0.68 et le profil de pratique du *chirurgien* avec une exactitude de 0.72. Ces performances doivent bien sûr être contextualisées en les comparant à l'état de l'art, mais également en étudiant les aspects de notre méthodologie qui ont le plus impacté ces performances. On discutera également des relations entre nos données observées grâce à notre modèle entraîné.

4.5.2.1 | comparaison avec l'état de l'art

Pour trouver un terrain de comparaison avec l'état de l'art, notre première exigence était de nous comparer avec des études ayant traité de la qualité ou de la pratique chirurgicale. Nous avons donc considéré les études traitant d'au moins un de ces aspects (voir section 2.2.1). Sur les 6 études obtenues, deux n'ont pas été considérées (El Ahmadieh et al., 2014; Ershad et al., 2016) car la comparaison avec nos travaux n'était pas envisageable. Les 4 études restantes sont décrites dans le tableau 4.5.

Trois études traitent de la qualité, et une du profil de pratique. Les études traitant de la qualité se sont basées sur les scores structurés OSATS et « Endoscopic Surgical Skill Quali-

Référence	Données	Méthode	Facette	Tâche chirurgicale	Performance	
					métrique	valeur
Matsuda et al. (2014)	Vidéo	Evaluation par 2 pairs	Qualité, ESSQ	Urologie, laparoscopique	Niveau d'accord	0.69
Malpani et al. (2015)	Vidéo	Crowdsourced	Qualité, OSATS	Entraînement robotisé	Exactitude	0.83
Khan et al. (2015)	Activités annotées + données d'accélérométrie	Lois topologiques et SVM	Qualité, OSATS	Entraînement chirurgie ouverte	Score F1	0.71
Huaulmé et al. (2018)	Activités annotées	Etude de schémas fréquent + agrégation	Profil de pratique	Entraînement à la suture en micro-robotique	Exactitude	0.94

Tableau 4.5 – Les études de prédiction de la qualité ou du profil de pratique chirurgicale et leurs performances.

fication » (ESSQ). Matsuda et al. (2014) ont mené une étude à grande échelle sur 1308 chirurgiens urologues de qualification chirurgicale, où chaque chirurgien était évalué par deux de ses pairs d'un groupe de 42 évaluateurs, pour passer la certification ESSQ. Outre la taille de l'échantillon de chirurgiens, et l'évaluation qui a été menée sur des vidéos de chirurgies réelles, cette étude a notamment montré que les deux évaluateurs ne s'accordaient sur leur évaluation que dans 69% des cas.

Ainsi, même pour des experts formés à mener des évaluations, il est toujours possible d'avoir des avis différents, et ce dans 31% des cas. Cette étude a été menée dans des conditions réelles et les évaluateurs ayant des sensibilités différentes, ils peuvent avoir une idée différente de ce qu'est la bonne façon de pratiquer, et donc préférer la pratique d'un chirurgien ou d'un autre. Néanmoins, si même pour des évaluateurs experts il est possible d'avoir un niveau d'accord de 0.69, alors il me semble acceptable que dans notre étude d'évaluation automatique de la qualité d'exposition sur des chirurgies réelles on ait une exactitude de 0.68.

J'insiste sur le fait que l'étude de Matsuda et al. (2014) a montré l'importante variabilité proposées par des chirurgiens experts quand ils évaluent leurs pairs. Ce résultat nous permet de relativiser l'ensemble des études qui, comme la nôtre, se base sur les données d'évaluation d'un ou de quelques chirurgiens pour mener leur tâche de prédiction automatique. Les résultats de prédiction obtenues dans ce type d'études ne reflètent en réalité que la capacité à reproduire l'évaluation d'un petit échantillon de chirurgien, mais ne sont en aucun cas une évaluation robuste et encore moins absolue de la qualité de pratique.

Malpani et al. (2015) ont, quant à eux, mené une étude d'évaluation par la foule sur des tâches d'entraînement robotisé. Dans cette étude, ils ont réussi, à partir des annotations proposées par les opérateurs anonymes, à reconstruire l'évaluation selon 7 axes du score OSATS avec une exactitude de 0.83 en moyenne. Autrement dit, la foule était capable de reproduire les évaluations détaillées du score OSATS proposées par un échantillon de 8 chirurgiens avec une exactitude de 0.83.

Cette étude montre avec quel niveau d'exactitude une méthode d'agrégation peut reconstruire un score structuré à partir d'annotations basiques proposées par une foule d'opérateurs néophytes. Pour ce faire, il a été nécessaire d'utiliser 49 annotations anonymes ainsi qu'un environnement dédié à la reconstruction du score pour reproduire la complexité de l'évaluation complète du score structuré menée par un chirurgien expert avec une exactitude de 0.83. L'objectif concrétisé ici est plus complexe que le nôtre car c'est un score à 7 composantes qui est obtenu, en revanche cet objectif a été rempli sur des tâches d'entraînement et non sur des extraits de chirurgies réelles comme nous.

Khan et al. (2015) proposent une approche automatique basée sur l'annotation procédurale d'activités et des données d'accélérométrie issues de tâche d'entrainement en chirurgie ouverte. A partir de ces données, une approche basée sur des lois topologiques et l'utilisation d'un SVM leur a permis de reconstruire l'évaluation selon 7 axes du score OSATS avec un score F1 valant 0.71 en moyenne. Ici la vérité terrain avait été établie par un expert sur l'ensemble des vidéos évalué. Aussi, cette approche automatique a permis de reproduire les évaluations du chirurgien avec un score F1 valant 0.71.

Cette étude est la plus proche de la nôtre en terme de méthodologie. Elle présente l'utilisation de données d'annotation procédurale, ainsi que de données d'accélérométrie décrivant les mouvements des instruments et pouvant se rapprocher de nos annotations spatiales. De plus, leur algorithme fait aussi intervenir un SVM à noyau rbf, même si la première étape de leur algorithme est plus poussée que la nôtre. Grace à leur approche, ils ont été capable de reproduire les 7 composantes du score OSATS avec un score F1 de 0 .71. Donc comme Malpani et al. (2015), l'objectif rempli est plus complexe que le nôtre, mais le cas d'étude est clairement simplifié.

Enfin, Huaulmé et al. (2018) ont également proposé une approche automatique basée sur l'annotation procédurale d'activités sur des tâche d'entraînement à la suture en chirurgie micro robotisée. A travers l'étude des longs schémas d'activités les plus fréquents et d'une méthode d'agrégations, ils ont été capable de distinguer les profils de pratique avec une exactitude de 0.94 sur une population constituée de deux chirurgiens novices et de deux ingénieurs roboticiens.

Cette étude est la seule à réellement s'attacher à distinguer les profils de pratique de différents opérateurs, et ce avec de très bonne performances comparés aux nôtre. En revanche, encore une fois leur cas d'étude est beaucoup plus simple car il s'agit d'exercices d'entraînement effectués qui plus est par deux chirurgiens juniors et deux non-chirurgiens.

Pour résumer, on constate d'une part la diversité des approches proposées et la difficulté qu'il y a donc à trouver des points de comparaison satisfaisants et pertinents. Néanmoins on peut dire que nos résultats sont comparables à ceux de (Huaulmé et al., 2018; Khan et al., 2015; Malpani et al., 2015). Et que, si nous proposons un résultat beaucoup plus simple que ces trois études, nous travaillons en revanche sur des données issues de chirurgies réelles et effectuées par des chirurgiens experts. En tout état de cause, l'objectif de prédiction que nous proposons pour la qualité d'exposition et le profil de pratique reste novateur en comparaison des nombreux travaux étudiant le niveau d'expertise.

4.5.2.2 | Choix et combinaisons de paramètres dans l'algorithme prédictif

Lors de la construction du pipeline et de la stratégie de VC, nous avons dû faire des choix pour pallier à certaines problématiques, et nous n'avons pas évalué la portée de tous ces choix. Nous discutons donc ici ces aspects méthodologiques qui sont résumés dans le tableau 4.6.

Etape	Problématique	Notre solution		A explorer
		Nom	Limitations	
Prétraitement	Valeurs manquantes	Valeur aberrante/ moyenne	Moyenne pour texture	RF/KNN
Prétraitement	Variables multi-types	Binarisation/ échantillonnage	Matrice creuse	?
Réduction de dimension	(Non-)supervisé	ALD	Echec quand $n_{var} \gg n_{ech}$	PCA/PLS
Classification	Performance de prédiction	SVM	Performances à améliorer	RF/NN
Optimisation	Sélection meilleurs hyperparamètres	Grid-search	Chronophage	Randomized-Search
Optimisation	Distribution *qualité* déséquilibrée	Précision optimisée	Score symétrique	Score F1
Validation croisée	Corrélation intra-chirurgie des échantillons	VC extérieure, LPSO	/	LOGO
Validation croisée	Distribution échantillons entre chirurgies	VC intérieure, LOGO, ε	Faible nombre d'échantillons	Plus de données

Tableau 4.6 – Présentation des principaux choix méthodologiques et de leurs alternatives - RF=Forêt aléatoire, KNN=K-nearest neighbors, PCA=Principal components analysis, PLS=Partial least squares, NN=Neural networks.

4.5.2.3 | Discussion sur l'algorithme prédictif

Lors du prétraitement des données, nous avons géré les valeurs manquantes en remplaçant, variable par variable, les valeurs manquantes par une valeur jugée aberrante, sauf pour la variable de texture pour laquelle nous avons pris la valeur moyenne. Nous avons fait ce choix méthodologique de façon arbitraire, et il serait intéressant d'étudier d'autres options plus évoluées telles que la forêt aléatoire ou le « K-nearest neigbors » proposés par Antonelli et al. (2019).

Lors du prétraitement, nous avons également cherché à homogénéiser les différents types de variable existants (réel, durée et label) en menant une binarisation par encodage « one-hot » pour les labels et par échantillonnage pour les réels et les durées. Nous avons justifié précédemment les raisons de ce choix (voir section 4.3.2), dont nous avions évalué l'impact dans nos travaux préparatoires, comparé à une absence de binarisation. Néanmoins il faut bien prendre en compte que la binarisation transforme la matrice de données X. D'une

part, cette dernière passe de 198 variables à presque 1200, ce qui fait qu'on passe d'une matrice ayant 7 fois plus d'échantillons que de variables à une matrice ayant deux fois plus de variables que d'échantillons. D'autre part, cette augmentation du nombre de variables distribue l'information entre toutes les variables et on se retrouve avec ce qu'on appelle une matrice creuse présentant un taux de 0 important comparé au taux de 1. On pourrait se demander s'il n'existe pas des traitements algorithmiques adaptés à cette forme de données.

Pour la réduction de dimension, nous avons choisi l'ALD au détriment de la PCA car l'ALD est supervisée et construit un espace de représentation axé sur les variables discriminantes au regard de la variable à prédire, tandis que la PCA mène cette construction indépendamment de cette variable à prédire. Au regard de l'étude sur les clusters de variables que nous avons menée (voir section 4.3.7) et qui se base justement sur les résultats de cette réduction de dimension, il nous parait difficile d'utiliser une méthode non-supervisée comme la PCA. En revanche, une autre méthode de régression supervisée à tester serait la régression des moindres carrés partiels ou « Partial least squares » (PLS) (Antonelli et al., 2019).

Dans notre étape de classification, nous avions choisi un SVM a noyau rbf non-linéaire qui donnait de bien meilleurs résultats que le SVM à noyau linéaire dans nos travaux préliminaires. En revanche, il serait intéressant de tester les performances de la méthode « Least absolute shrinkage and selection operator » (LASSO). Nous aurions également pu remplacer l'ensemble de notre pipeline par une forêt aléatoire.

4.5.2.4 | Discussion sur l'environnement d'optimisation

Pour tester ces différents choix, nous avons développé un environnement d'optimisation sous la forme d'une recherche en grille ou « grid-search » GS permettant de sélectionner les meilleurs hyperparamètres. Cependant cette approche implique un nombre d'itérations très élevé, et le temps d'optimisation se compte en dizaines d'heures. Aussi si nous voulons étendre le nombre de possibilités à évaluer, le temps d'optimisation sera allongé d'autant.

Pour éviter de nous retrouver dans cette situation, il existe une alternative valide à la recherche en grille : la recherche randomisée ou « randomized search ». Cette approche similaire à la recherche en grille, en diffère par son caractère non-exhaustif : au lieu d'explorer toutes les options possibles dans l'espace de définition des hyper paramètres, elle se contente d'explorer cet espace selon une distribution qu'on lui définit, ainsi qu'un nombre d'échantillons à tester. En fixant la distribution, on choisit d'orienter l'évaluation sur une partie ou une autre de l'espace de définition. En fixant le nombre d'échantillons, on détermine le nombre de fois qu'on entraine et teste l'algorithme et on maîtrise ainsi le temps de calcul. Cette approche permet d'obtenir des résultats d'optimisation similaires à ceux de la recherche en grille en maîtrisant le temps de calcul (Bergstra and Bengio, 2012).

La distribution des échantillons entre les valeurs « bonne » et « perfectible » de la *qualit* est fortement déséquilibrée, et ces échantillons sont également distribués de façon inhomogène entre les 30 chirurgies de notre jeu de données. Nous avons géré ce double déséquilibre en utilisant comme descripteur la précision optimisée PO. La précision optimisée nous a permis de définir la meilleure performance comme étant celle qui prédit correctement les deux valeurs de la *sortie* et non comme celle qui prédit correctement le plus grand nombre

d'échantillons. La stratégie de VC imbriquée que nous avons mise en place a également eu un impact, notamment en définissant le paramètre ε de façon à ce que les plis de la VC intérieure conservent des ratios de $qualit\ d_{y,i}$ similaires. Le fait que les valeurs de sensibilité et de spécificité observées dans la figure 4.7 soient aussi proches montre que nos choix ont été payants. On pourrait évaluer l'impact qu'aurait le score F1, une métrique d'évaluation similaire à la précision optimisée, mais donnant plus d'importance aux vrais positifs qu'aux vrais négatifs.

4.5.2.5 | Discussion sur la validation croisée

Pour évaluer les performances de notre modèle (VC extérieure) et pour optimiser ses hyperparamètres (VC intérieure) correctement, nous avons conçu une méthodologie de validation croisée imbriquée. La VC extérieure a pris la forme d'une « Leave-p-surgery-out » (LPSO) afin de prendre en considération la corrélation intra-chirurgie existante entre les échantillons d'une même chirurgie. Le paramètre p définissant le nombre de chirurgies dans chaque pli d'évaluation a été fixé à $p = 8$ afin de respecter le ratio 25%-75% d'échantillons communément accepté dans la littérature. Le paramètre Q définissant le nombre de configurations (P_{opt}, P_{test}) sur lesquelles on a choisi le modèle final a été fixé de façon arbitraire à $Q = 20$. On pourrait envisager de remplacer cette LPSO par une « Leave-one-group-out » (LOGO) qui ne respecterait pas cette contrainte de corrélation intra-chirurgie et qui disperserait donc les échantillons d'une même chirurgie dans différents plis.

Dans la VC intérieure, on a appliqué une LOGO avec $P = 10$, le nombre de plis. Ce choix ne respectant pas la contrainte de corrélation intra-chirurgie, permet cependant de gérer les problèmes liés à la distribution inhomogène des échantillons entre les chirurgies, ainsi qu'à la distribution déséquilibrée des valeurs « bonne » et « perfectible » de la variable de *qualité* sur les échantillons. Nous sommes ici dans un cas très problématique car, du fait de cette double VC, le nombre d'échantillons par pli est fortement réduit et les contraintes s'accumulent : en cherchant à gérer des biais liés à nos données, nous en avons créé d'autres. Pour résoudre ce problème, la solution serait d'avoir plus de données.

Un dernier problème plus technique mais relativement problématique vient du profil de matrice creuse obtenu suite à la binarisation de nos variables lors du prétraitement. Lors de l'étape d'optimisation de l'algorithme, on considère la matrice de données utilisée pour l'évaluation d'un algorithme entraîné au sein de la VC intérieure, et cette matrice compte 37 échantillons en moyenne (max=57, min=23). Ce faible compte d'échantillons est une conséquence logique de la VC imbriquée. Mais ce faible compte pourrait être problématique lorsqu'on sait que ce sont des échantillons d'une matrice qui se retrouve transformée en matrice creuse avant l'étape de réduction de dimension. Une matrice creuse comptant quelques dizaines d'échantillons et presque 1200 variables permet-elle un apprentissage valide ? Cette problématique n'est pas si prégnante lors de l'étape finale de test dans la VC extérieure, mais il se peut que la sélection du meilleur modèle dans la grid-search soit impactée. Si c'est le cas, il faudrait envisager, soit de ne pas binariser les variables initiales, soit de trouver une autre stratégie de VC qui manipule des plis avec plus d'échantillons.

4.5.3 | Analyse et interprétation clinique de l'impact des descripteurs dans le processus d'apprentissage

En étudiant nos résultats de prédiction, nous avons constaté que la prédiction de la *qualité* se repose plus sur les variables spatiales que sur les variables procédurales. Cela peut venir du fait que l'exposition est une notion fortement reliée au contexte spatial et moins au contexte procédural. Il semble en effet que la notion d'exposition dépende plus de la gestion visuelle et spatiale du chirurgien, et il est intéressant d'observer une telle différence de prédiction entre les deux catégories de descripteurs. Cela nous pousserait donc à croire que le contexte spatial et visuel de la chirurgie doit être pris en compte en priorité sur le contexte procédural lorsqu'on s'intéresse à la qualité d'exposition.

En revanche, cette constatation nous questionne car nous avions mené l'annotation de la qualité d'exposition de front avec l'annotation de la procédure. Et lors de ces annotations, nous devions donc être plus concentrés sur les évènements ayant lieu pendant la chirurgie que sur le contenu visuel. De plus, les résultats de l'étude de Huaulmé et al. (2018), et dans une moindre mesure ceux de Khan et al. (2015), ont montré qu'on pouvait avoir des bons résultats de prédictions à partir de données issues d'annotations procédurales. Cette primauté du contexte spatial sur le contexte procédural pour la prédiction de la *qualité* est donc surprenante et mériterait d'être validée.

En section 4.3.7, nous avons présenté une méthodologie d'extraction de clusters contenant les descripteurs les plus discriminants pour le modèle grâce à l'étape d'ALD. Nous avons pu constituer ces clusters grâce à la linéarité de l'étape de réduction de dimension ainsi que du prétraitement. Nous n'aurions pas pu obtenir de tels résultats aussi simplement en utilisant des méthodes non-linéaires telles que les réseaux de neurones ou les approches d'apprentissage profond.

Le fait que pour chaque population de descripeurs en entrée (PDE), on compte des nombres similaires de descripteurs pour les deux *sorties* et que les écart-types restent faibles nous laisse supposer que la composition de ces clusters de descripteurs reste stable pour les différentes $config_vc$s. De plus, le fait que les clusters d'intersection sur les deux *sorties* contiennent 70% de leurs clusters indépendants respectifs implique que ces 70% de descripteurs sont importants, tant pour la prédiction de la *qualité* que pour la prédiction du *chirurgien*. Les 30% de descripteurs restant sont spécifiques à chaque *sortie*, et caractérisent donc chaque *sortie* distincte. Ces descripteurs communs et distinctifs devraient être étudiés plus précisément pour comprendre ce qui relie et distingue la qualité d'exposition dans la scène chirurgicale et le profil de pratique du chirurgien.

Par ailleurs, le fait que la population de descripteurs $S + P$ ait sélectionné exclusivement des descripteurs spatiaux et aucun descripteur procédural est un résultat critique. En effet cela implique que l'apprentissage ne comptait que sur les descripteurs spatiaux pour mener à bien la tâche de prédiction, et aucunement sur les descripteurs procéduraux. Associé aux meilleurs résultats du modèle spatial comparé au modèle procédural, cela confirme bien que l'utilisation des descripteurs procéduraux n'apporte rien, voire perturbe l'apprentissage. Encore une fois, nos résultats impliquent que pour notre cas d'étude, l'aspect spatial prime sur l'aspect procédural.

Une analyse plus détaillée de ces descripteurs et de leur impact sur l'apprentissage et

la prédiction est nécessaire pour mieux comprendre ces résultats, les valider et proposer des conclusions à portée clinique. Plus précisément nous envisageons d'analyser plus en détails les descripteurs spatiaux qui donnent les meilleurs résultats, et de mieux étudier la sémantique liée à ces différents descripteurs pour interpréter leur comportement.

En résumé

— Notre choix de proposer des descripteurs interprétable est impactant pour la construction du modèle de prédiction.
— Nos performances de prédiction peuvent être comparés à ceux de quelques études traitant soit de la qualité d'exposition, soit du profil de pratique à travers des approches très variées. Si nous proposons une tâche de prédiction que ces études, notre cas d'étude clinique est lui issu de chirurgie réelle.
— Dans les étapes successives de notre pipeline algorithmique, de notre environnement d'optimisation, et dans notre processus de validation croisée, nous avons fait de nombreux choix dont il faudrait évaluer l'impact.
— L'analyse des clusters de descripteurs a été rendu possible par la linéarité des étapes de prétraitement et de réduction de dimension de notre algorithme.
— La qualité d'exposition et le profil de pratique sont prédits par 70% de descripteurs en commun, et n'ont que 30% de descripteurs qui les distinguent. Une étude approfondie de ses descripteurs communs et spécifiques permettrait de mieux comprendre ces deux notions.
— La primauté des descripteurs spatiaux comparés aux descripteurs procéduraux dans la tâche de prédiction est un résultat important mais questionnable.

4.6 | Conclusions

Dans cette étude, nous avons présenté un algorithme d'apprentissage classifiant une variable binaire à partir d'un vecteur de plusieurs centaines d'échantillons. Cet algorithme a trois étapes successives : une étape de préparation des données, une réduction de dimension, et une classification. La préparation des données gère surtout la combinaison de données ayant différents types en entrée. La réduction de dimension est une analyse discriminante linéaire avec un hyper-paramètre : le taux de variance. La classification est une machine à vecteurs de supports avec deux hyper-paramètres : C et γ.

Un environnement d'optimisation a été développé dans lequel les hyper-paramètres de l'algorithme sont optimisés : nous avons cherché la combinaison de valeurs sur l'espace de définition des hyper-paramètres donnant les meilleurs résultats en terme de prédiction de la variable en sortie. L'implémentation de cet environnement d'optimisation, une validation croisée imbriquée associée à une recherche de valeurs en grille, a nécessité de gérer des biais liés aux données :

— La corrélation des échantillons intra-chirurgie a été gérée en imposant qu'une chirurgie ne puisse être présente que dans un seul pli de la validation croisée extérieure (LPSO).
— Le déséquilibre de distribution des échantillons entre les deux valeurs de la variable a été géré en utilisant dans le processus d'optimisation la précision optimisée (PO) comme descripteur d'évaluation.
— La répartition inhomogène des échantillons de qualité d'exposition entre les différentes chirurgies a été gérée en mélangeant les échantillons des chirurgies dans la validation croisée intérieure (LOGO).
— La variabilité inter-chirurgie a été gérée avec une multiplication des configurations de données dans la validation croisée extérieure (LPSO).

L'optimisation de l'algorithme a été effectuée dans trois configurations de données : nous avons cherché à prédire la variable de qualité d'exposition à partir des variables spatiales, procédurales, et spatiales plus procédurales. On a ainsi obtenu une version différente de l'algorithme avec des valeurs d'hyper-paramètres différentes pour les trois configurations. Chacun de ces modèles a ensuite été évalué pour prédire la qualité d'exposition d'une part et le profil de pratique du chirurgien d'autre part. On a ainsi obtenu des résultats de performances pour six configurations différentes en faisant varier les variables en entrée : spatiales, procédurales, et spatiales plus procédurales ; et la variable binaire à prédire : la qualité d'exposition, et le profil de pratique du chirurgien.

L'étude d'influence de ces différentes configurations de données nous a permis d'observer notamment l'importance des variables spatiales par rapport aux variables procédurales pour la prédiction de la qualité d'exposition et du chirurgien manipulant. Cela nous incite à penser que le contenu visuel de la vidéo à plus d'importance que son contenu procédural pour ces tâches de prédiction. On a aussi observé que la tâche de prédiction du chirurgien manipulant était mieux réussie que la tâche de prédiction de la qualité d'exposition, nous en avons déduit que la notion de chirurgien est plus facile à prédire, et moins complexe, que la notion d'exposition de la scène chirurgicale.

Finalement, malgré des performances qui pourraient être améliorées, les résultats de prédiction de la qualité d'exposition et du chirurgien manipulant ont validé qu'il est possible de prédire à partir de données issues de vidéos laparoscopiques réelles des concepts tels que la qualité d'exposition dans la scène chirurgicale ou le profil de pratique du chirurgien.

5

Étude 3 : analyse et validation clinique des relations entre données d'entrée et de sortie de l'algorithme

Préambule

Dans cette étude, nous avons exploré les possibilités d'interprétations des relations entre nos données spatiales et nos variables de qualité d'exposition et de profil de pratique à partir de notre algorithme prédictif. Pour ce faire, nous avons introduit notre étude en section 5.1, puis nous avons créé plusieurs configurations de données en section 5.2, sur lesquelles nous avons mené une méthodologie aboutissant à la présentation d'interprétations algorithmiques et à l'évaluation de leur intérêt clinique en section 5.3. Les résultats de cette étude sont présentés en section 5.4 et discutés en section 5.5. Puis nous concluons sur cette étude en section 5.6.

Sommaire

5.1 | Introduction

Dans l'étude précédente, l'optimisation de l'algorithme nous a permis de caractériser le pouvoir prédictif de notre modèle. Nous avons également observé l'impact des différents types de données annotées sur la performance de prédiction. Pour ce faire, il a été nécessaire de construire des descripteurs à partir de nos données d'annotation. Nous allons pousser notre analyse plus loin en étudiant en détail les relations entre nos données d'entrée et de sortie.

Notre objectif général est de trouver une signification clinique à l'influence que nos descripteurs en entrée ont sur la pratique chirurgicale. Plus précisément, nos objectifs sont :

— D'analyser l'importance des variables en entrée dans la tâche de prédiction de la variable en sortie.
— De proposer, à partir de cette analyse, des interprétations ayant un sens clinique.
— De valider cliniquement la pertinence de nos interprétations.

5.2 | Matériel

Dans le chapitre précédent, nous avons constaté que notre modèle de prédiction basé sur les descripteurs spatiaux avait de meilleures performances que notre modèle basé sur les descripteurs procéduraux. Dans cette étude, nous nous focaliserons donc uniquement sur les variables spatiales, et notre matrice d'entrée sera donc $X = X_{seg}$ de taille $n_{ech} \times n_{var,seg}$, avec $n_{ech} = 735$ et $n_{var,seg} = 130$.

Dans le chapitre précédent, nous avons également défini les variables spatiales sous une forme particulière : les 10 objets visibles et segmentés au cours des vidéos ont été décrits à l'aide de 13 descripteurs les caractérisant spatialement (voir tableau 5.1). Voici la liste exhaustive de ces objets :

— Les entités anatomiques :
  La paroi abdominale, le diaphragme, l'estomac, le foie, la rate et les tissus adipeux.
— Les instruments chirurgicaux :
  La compresse, l'écarteur de foie, la pince atraumatique, et la pince électro-thermale.

Nom	#	Description
Barycentre	2	Coordonnées x et y du pixel central de la forme.
Couleur	3	Couleur des pixels de la forme dans l'espace de couleur CIE-Lab.
Périmètre	1	# de pixels du contour de la forme.
Surface	1	# de pixels de la forme.
Texture std	1	Ecart type de l'histogramme obtenue par application d'un algorithme de local binary pattern.
Valeur propre	1	ratio des 2 valeurs propres caractérisant la forme de l'objet.
Vecteurs propres	4	Coordonnées x et y des deux directions principales de la forme (calculé par une ACP).

Tableau 5.1 – Descripteurs spatiaux extraits pour chaque objet segmenté dans les images.

Chacune des 130 variables peut donc être analysée à travers sa représentation sémantique *descripteur-objet*.

Nous avons également défini un vecteur de sortie $y_{qualité}$ qui caractérise la qualité d'exposition de la cible chirurgicale, et un autre vecteur de sortie $y_{chirurgien}$ qui décrit le profil de pratique du chirurgien. Ces vecteurs ont tous deux une taille de $n_{ech} \times 1$.

Dans le protocole expérimental que nous décrirons par la suite, 6 groupes de données ont été définis à partir de nos données initiales $(X, y_{qualité}, y_{chirurgien})$ en fixant des contraintes basées sur les distributions de valeurs de $y_{qualité}$ et $y_{chirurgien}$ (voir tableau 5.2). Le groupe de données 1A correspond à l'analyse de la *qualité* sur le jeu de données complet, tandis que dans les groupes de données 1B et 1C nous avons cherché à caractériser la *qualité* sur les échantillons correspondant à chaque chirurgien distinct. Le groupe de données 2A correspond à l'analyse du *chirurgien* sur le jeu de données en entier, tandis que dans les groupes de données 2B et 2C, nous avons cherché à caractériser le *chirurgien* pour chaque niveau de *qualité*. Les « meilleures » et les « pires » chirurgies sont respectivement celles avec la plus haute/basse *qualité* moyenne. En d'autres termes, dans les groupes de données 2B et 2C, les 6 chirurgies ont été sélectionnées en se basant sur la *qualité* moyennée sur les échantillons de chaque chirurgie.

Groupe	Variable à prédire	Critère de sélection d'échantillons	Taille du groupe	
			# d'échantillons	# de chirurgies
1A	$y_{qualité}$	/	735	30
1B	$y_{qualité}$	$y_{chirurgien} == 0$	405	15
1C	$y_{qualité}$	$y_{chirurgien} == 1$	330	15
2A	$y_{chirurgien}$	/	735	30
2B	$y_{chirurgien}$	Echantillons des 3 chirurgies avec la plus haute *qualité* moyenne	156	15
2C	$y_{chirurgien}$	Echantillons des 3 chirurgies avec la plus basse *qualité* moyenne	155	15

Tableau 5.2 – Constitution des groupes de données à partir des données $(X, y_{qualité}, y_{chirurgien})$.

Nous pouvons aussi considérer que chaque groupe de données pose une question :

1A. « Qu'est-ce qui caractérise la qualité d'exposition de manière générale pour les deux chirurgiens ? »

1B. « Qu'est-ce qui caractérise la qualité d'exposition lorsque le chirurgien 0 pratique ? »

1C. « Qu'est-ce qui caractérise la qualité d'exposition lorsque le chirurgien 1 pratique ? »

2A. « Qu'est-ce qui caractérise le profil de pratique des chirurgiens quelle que soit la qualité d'exposition ? »

2B. « Qu'est-ce qui caractérise le profil de pratique des chirurgiens lorsque la qualité d'exposition est bonne ? »

2C. « Qu'est-ce qui caractérise le profil de pratique des chirurgiens lorsque le qualité d'exposition est perfectible ? »

En résumé

— Dans cette étude, nos variables d'entrée sont uniquement celles issues des annotations spatiales (au nombre de 130).
— Ces variables peuvent être représentées par le couple descripteur-objet.
— Nous avons considéré comme variables de sortie $y_{qualité}$ et $y_{chirurgien}$.
— A partir de ces variables d'entrée et de sortie, 6 configurations ou groupes de données différents ont été définis et considérés dans nos expérimentations.

5.3 | Méthodologie

Nous avons développé un environnement proposant une interprétation clinique et une validation clinique du traitement algorithmique opéré sur nos données. Nous n'avons appliqué notre traitement algorithmique que sur les variables spatiales. Aussi, comme nous n'avons plus ni labels ni durées dans nos variables, le prétraitement des données a pu être simplifié (voir section 4.3.3.1). Nous avons certes dû gérer les valeurs manquantes, mais nous nous sommes ensuite contentés d'une standardisation des variables en entrée. Les étapes de réduction de dimension (ALD) et de classification (SVM), elles, n'ont pas changé. L'objectif final de cette étude est de valider notre méthode en soumettant un questionnaire illustré par des extraits vidéos à des chirurgiens.

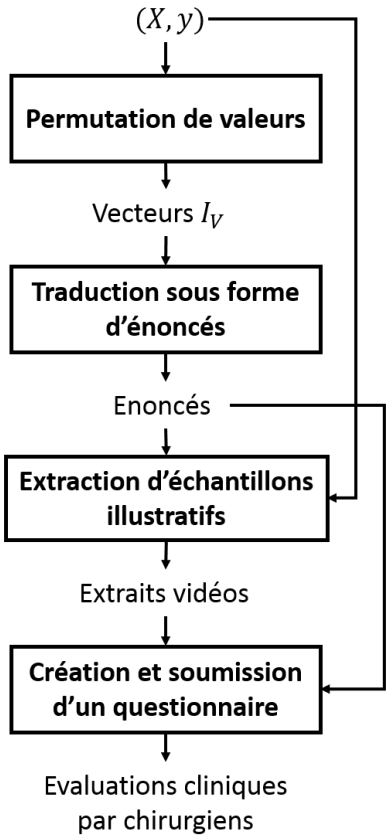


FIGURE 5.1 – Les étapes successives de la méthodologie d'analyse et de validation clinique.

Voici les 4 étapes de notre méthode (voir figure 5.1) :

1. section 5.3.1 : en considérant la tâche de prédiction des vecteurs de sortie y_{qualit} et $y_{chirurgien}$, nous avons mené une permutation de valeurs sur chaque variable v de nos données d'entrée X afin d'évaluer leur importance i_v pour cette tâche de prédiction. Nous avons ainsi obtenu un vecteur quantifiant l'importance des 130 variables que nous nommons vecteur I_V. En considérant chaque configuration de données introduite en section 5.2, ce sont 6 vecteurs I_V qui ont été calculés. Pour l'étape suivante de notre méthode, nous avons sélectionné les valeurs des vecteurs I_V ayant la plus grande « importance ».

2. section 5.3.2 : Nous avons traduit manuellement ces observations sous la forme d'*énoncé*s porteurs d'une signification clinique, en respectant certaines contraintes lors de la traduction.
3. section 5.3.3 : pour chaque énoncé, deux échantillons (1 par valeur de la sortie binaire) ont été sélectionnés semi automatiquement à partir d'une analyse des données d'entrée. Nous avons extrait un court segment vidéo correspondant à chaque échantillon, pour proposer une illustration au contenu de l'énoncé correspondant.
4. section 5.3.4 : Nous avons conçu un questionnaire pour valider la valeur clinique de notre étude, dans lequel chaque énoncé était illustré par ses deux extraits vidéos. Nous avons soumis ce questionnaire à un chirurgien expert et un interne sénior, qui ont tous deux évalué leur niveau d'accord pour chaque énoncé du questionnaire.

5.3.1 | Le vecteur d'importance des variables I_V

Il s'agissait d'observer l'impact des variables d'entrée sur les performances de la tâche de prédiction : pour chaque variable, nous avons effectué une permutation de valeurs et observé la modification de prédiction induite. La permutation de valeurs est une approche algorithmique qui consiste à mélanger les valeurs d'une variable et à observer l'influence de ce mélange sur la prédiction (Frank and Witten, 1998; Radivojac et al., 2004).

Pour éviter un biais de sur-apprentissage, nous avons implémenté la permutation de valeurs au sein d'un environnement de validation croisée entraînement-validation. Cet environnement se rapproche de celui développé dans notre étude précédente (voir section 4.3.4.5) car nous avons respecté l'hypothèse de dépendance entre les échantillons d'une même chirurgie. A noter que nous n'avons pas mené de validation croisée imbriquée ici, mais une simple séparation de nos données entre un pli d'entraînement et un pli de validation. Une validation croisée entraînement-validation suffit ici car il n'est pas question d'optimiser des hyper paramètres, mais simplement d'effectuer la permutation de valeurs.

Nous avons également défini le nombre Q comme le nombre de configurations entraînement-validation différentes de cette validation croisée dans lesquelles nous avons répété notre permutation de valeurs pour assurer une bonne représentativité de notre jeu de données. Ce nombre Q varie selon le groupe de données (voir tableau 5.3).

	1A	1B	1C	2A	2B	2C
Q (# de configurations de VC)	20	10	10	20	3	3
# total de chirurgies	30	15	15	30	6	6
# de chirurgies dans le pli d'entraînement	22	10	10	22	4	4
# de chirurgies dans le pli de test	8	5	5	8	2	2

Tableau 5.3 – Les différentes configurations de validation croisée VC entraînement-évaluation dans le processus de calcul de l'IV, en fonction du groupe de données.

Pour chaque configuration de validation croisée, nous avons d'abord entrainé notre modèle sur le pli d'entraînement. Sur le pli de test, nous avons ensuite évalué une performance de référence pour la prédiction à partir de nos données sans permutation. Sur le même pli

de test, nous avons successivement mené une permutation sur les valeurs de chaque variable. Pour les mêmes raisons liées au déséquilibre entre les deux valeurs de nos variables de sortie décrites en section 4.3.5, nous avons évalué les performances de prédiction avec la précision optimisée OP.

On considère l'amplitude de la modification de performance de prédiction qui résulte de la permutation de valeurs appliquée à la variable v. On définit l'importance i_v de la variable v comme cette amplitude moyennée sur les Q configurations de validation croisée. On défini le vecteur I_V qui est composé des valeurs d'importance i_v des variables d'entrée évaluées.

Protocole Expérimental. Nous avons mené la permutation de valeur au sein d'une validation croisée de type entraînement-test pour les 6 groupes de données (voir tableau 5.3), nous avons obtenu 6 vecteurs I_V de taille 130×1. On notera que comme le nombre d'échantillons varient selon les groupes de données, nous n'avons pas pu proposer autant de configurations de validation croisée pour les groupes 2B et 2C (Q=3 configurations) qui présentent beaucoup moins d'échantillons que les groupes 1B et 1C (Q=10 configurations) et encore moins que les groupes 1A et 2A (Q=20 configurations).

Nous avons mené deux expérimentations sur ces 6 vecteurs. Dans l'*expérimentation 1*, les variables étaient groupées par objet visible et nous avons ainsi obtenu les valeurs $\overline{i_v}$ moyennée sur les 13 variables de chaque objet. Dans l'*expérimentation 2*, nous avons uniquement conservé les variables de l'objet le plus important (celui avec la plus haute importance moyennée sur ses variables $\overline{i_v}$), et étudié l'importance i_v de ses 13 variables. Nous voulions étudier plus en détail l'impact de cet objet sur la tâche chirurgicale. Au final nous avons recueilli des observations sur 12 configurations expérimentales différentes. Nous avons donc cherché à répondre aux questions associées à chaque groupe de donnnées à travers l'objet le plus important (expérimentation 1), puis à travers les variables les plus importantes de cet objet (expérimentation 2).

5.3.2 | La traduction des observations des vecteurs I_V sous forme d'énoncés cliniques

Comme nous l'avons expliqué en section 5.2, chaque variable de nos données d'entrée est un *descripteur* spatial décrivant un *objet* visible dans une image. Et cet objet est soit une entité anatomique, soit un instrument chirurgical. Nous nous sommes donc focalisés sur cette description de la sémantique de nos variables pour obtenir des interprétations ayant une signification clinique, en considérant l'association *(descripteur-objet)* de chaque variable.

A partir des observations faites sur les vecteurs I_V, un scientifique maîtrisant très bien le contexte clinique de l'étude a traduit chaque observation sous la forme d'un *énoncé* ayant une signification clinique. En plus de ce besoin de sens clinique, chaque *énoncé* devait satisfaire les contraintes suivantes :

a Chaque observation ne devait donner lieu qu'à un unique énoncé.
b Un énoncé ne devait se référer qu'à un seul groupe de données.
c Un énoncé ne devait se référer qu'à un groupe de variables caractérisées par un seul *objet* et un ou plusieurs *descripteur*s de cet objet.
d Un énoncé devait se référer soit à la *qualité* soit au *chirurgien*.

5.3.3 | Illustration des énoncés à l'aide d'échantillons et de leurs vidéos associées

Nous avons sélectionné de façon semi-automatique des échantillons représentatifs pour chaque énoncé, avec une première phase de présélection automatique impliquant l'entraînement de notre modèle, puis une seconde phase de sélection manuelle effectuée par un opérateur spécialiste du domaine à l'issue de laquelle nous avons obtenu deux extraits vidéos illustrant l'énoncé associé.

Nous avons traité successivement chaque *énoncé E* caractérisé par son groupe de données, son objet et son groupe de descripteurs (voir paramètres 1,2 et 3 du tableau 5.4). Ces informations nous ont permis de définir les données (X_E, y_E) sur lesquelles a été menée la sélection d'échantillon : X_E est une matrice de taille $n_{ech,E} \times n_{var_E}$, et y_E est un vecteur de taille $n_{ech,E} \times 1$. $n_{ech,E}$ est le nombre d'échantillon du groupe de données associé à l'*énoncé E*. $n_{var,E}$ est le nombre de variables de l'*énoncé E*, défini par l'objet et les descripteurs associé à l'énoncé S et on a $n_{var,E} \in [1;13]$.

	Id	Paramètre	Description
Paramètres	1	Groupe de données	Les données (X, y) (voir section 5.2)
pour chaque	2	Objet	L'objet caractérisant les variables les plus importantes.
énoncé E	3	Descripteur	Le(s) descripteur(s) caractérisant les variables les plus importantes.
Paramètres	4	n_{sel}	# d'échantillons sélectionnés pour chaque valeur de sortie.
de la	5	t_{clip}	Durée du clip vidéo.
présélection	6	$t_{décalage}$	Décalage temporel du début du clip vidéo.

Tableau 5.4 – Paramètres de la présélection automatique d'échantillons illustrant les *énoncé*s.

Pour mener cette présélection automatique, nous avons entrainé le modèle sur les données (X_E, y_E). Dans ce modèle, le SVM classifie les échantillons à l'aide d'une fonction de coût retournant des valeurs positives pour la prédiction de la valeur 1, et des valeurs négatives pouzer la prédiction de la valeur 0 (voir section 4.3.3.3). Plus cette valeur est éloignée de 0, plus l'échantillon est décisif. Nous avons donc défini le nombre n_{sel} d'échantillons à sélectionner, on a extrait les n_{sel} échantillons ayant les plus hautes valeurs pour la valeur 1 de la variable de sortie y_E et les n_{sel} échantillons ayant les plus basses valeurs pour la valeur 0 de la variable de sortie y_E. Pour chaque *énoncé E*, la présélection automatique sélectionne n_{sel} échantillons pour la valeur 0 et pour la valeur 1 de la variable de sortie.

Une fois ces échantillons sélectionnés, nous les avons chacun par un clip vidéo extrait des vidéos chirurgicales de notre jeu de données. Ce clip est caractérisé par la date de l'échantillon associé, sa durée t_{clip} et son décalage temporel $t_{décalage}$. Le décalage temporel est défini comme l'écart temporel à la date de l'échantillon auquel commence le clip vidéo.

Enfin, un opératuer spécialiste mais non-clinicien a considéré chaque énoncé ainsi que les clips vidéos associés extraits automatiquement. Pour les deux valeurs de la variable de

sortie, il a choisi un échantillon parmi les n_{sel} afin d'illustrer au mieux l'idée clinique portée par l'énoncé E correspondant.

Protocole expérimental. Nous avons caractérisé le processus de sélection en fixant les paramètres $n_{sel} = 5$, $t_{clip} = 10s$ et $t_{décalage} = -7s$. Ainsi pour chaque énoncé, 5 échantillons ont d'abord été automatiquement présélectionnés pour les deux valeurs de sortie. Chaque échantillon a été illustré par un extrait vidéo de 10s commençant 7s avant la date de l'échantillon. L'opérateur a ensuite manuellement sélectionné un de ces 5 échantillons. Finalement, on obtient deux extraits vidéos pour chaque énoncé.

5.3.4 | Le questionnaire de validation clinique

Le questionnaire a été construit avec chaque énoncé et ses deux extraits vidéos associés. Pour évaluer ces énoncés, nous avons choisi d'utiliser une échelle de Likert à 5 niveaux : « absolument pas d'accord », « plutôt pas d'accord », « sans avis », « plutôt d'accord » et « complètement d'accord ». Ce questionnaire était d'un entretien libre avec le clinicien pour recueillir son avis sur la conception du questionnaire et la méthodologie proposée.

Protocole expérimental. Nous avons construit un questionnaire à partir des énoncés illustrés par leurs extraits vidéos. Les énoncés étaient présentés dans un ordre aléatoire. Nous avons soumis ce questionnaire à un chirurgien expert en cœlioscopie, et un interne sénior en cœlioscopie.

5.4 | Résultats

Tout d'abord, nous présentons dans le tableau 5.5 les performances de notre modèle pour la tâche de prédiction sur nos 6 groupes de données. Nous observons des performances similaires pour les 4 groupes *1A*, *1B*, *1C* et *2A*. Les groupes *2B* et *2C* présentent des performances presque parfaites, il semblerait que notre modèle se soit parfaitement ajusté à ces quelques échantillons.

Groupe de données	1A	1B	1C	2A	2B	2C
Exactitude	$0.64 \pm .05$	$0.66 \pm .15$	$0.73 \pm .05$	$0.60 \pm .08$	$0.99 \pm .02$	$0.97 \pm .03$
Sensibilité	$0.76 \pm .15$	$0.61 \pm .30$	$0.58 \pm .10$	$0.58 \pm .16$	$0.98 \pm .04$	$0.98 \pm .03$
Spécificité	$0.56 \pm .14$	$0.70 \pm .30$	$0.88 \pm .07$	$0.62 \pm .13$	1 ± 0	$0.95 \pm .04$

Tableau 5.5 – Performances de prédiction du modèle pour les 6 groupes de données.

5.4.1 | Observation des variables les plus importantes

Nous présentons les résultats de calcul des valeurs des vecteurs I_V après passage en logarithme base-10. Ces valeurs étant toute négatives, les éléments les plus importants sont ceux ayant les variations de performance les plus faibles, et donc les plus petites valeurs (i.e. les plus éloignées de 0).

D'une part, le tableau 5.6 présente les résultats de l'expérimentation 1 avec l'importance moyennée par objet visible, et le tableau 5.7 présente les observation obtenues à partir de ces résultats. D'autre part, le tableau 5.8 présente les résultats de l'expérimentation 2 avec les valeurs i_v des 13 variables de l'objet le plus important pour chaque groupe de données, et le tableau 5.9 présente les observations obtenues à partir de ces résultats.

Les deux tableaux 5.7 et 5.9 présentant les observations permettent de mieux comprendre et interpréter respectivement le contenu des tableaux 5.6 et 5.8. Nos observations sont spécifiquement focalisées sur l'analyse de l'impact des variables d'entrée sur la prédiction des variables de sortie *qualité* et *chirurgien*.

1A		1B		1C	
objet	$log_{10}(\overline{i_v})$	objet	$log_{10}(\overline{i_v})$	objet	$log_{10}(\overline{i_v})$
rate	-1.84	rate	-1.71	pince atraumatique	-1.43
pince électro-thermale	-1.80	tissus adipeux	-1.65	écarteur de foie	-1.41
pince atraumatique	-1.78	estomac	-1.64	paroi abdominale	-1.39
compresse	-1.77	pince atraumatique	-1.64	rate	-1.38
diaphragme	-1.75	pince électro-thermale	-1.62	tissus adipeux	-1.37
paroi abdominale	-1.73	écarteur de foie	-1.61	pince électro-thermale	-1.37
estomac	-1.71	compresse	-1.58	estomac	-1.36
tissus adipeux	-1.70	paroi abdominale	-1.57	foie	-1.36
écarteur de foie	-1.66	foie	-1.54	diaphragme	-1.34
foie	-1.48	diaphragme	-1.37	compresse	-1.34
2A		**2B**		**2C**	
objet	$log_{10}(\overline{i_v})$	objet	$log_{10}(\overline{i_v})$	objet	$log_{10}(\overline{i_v})$
estomac	-1.71	foie	-1.90	paroi abdominale	-2.73
compresse	-1.67	diaphragme	-1.86	écarteur de foie	-2.54
foie	-1.67	paroi abdominale	-1.82	pince atraumatique	-2.27
diaphragme	-1.65	rate	-1.82	diaphragme	-2.15
écarteur de foie	-1.61	estomac	-1.76	estomac	-2.10
tissus adipeux	-1.61	écarteur de foie	-1.75	rate	-2.09
paroi abdominale	-1.60	compresse	-1.59	tissus adipeux	-1.98
pince électro-thermale	-1.60	pince électro-thermale	-1.58	compresse	-1.89
rate	-1.58	tissus adipeux	-1.57	foie	-1.77
pince atraumatique	-1.55	pince atraumatique	-1.56	pince électro-thermale	-1.65

Tableau 5.6 – Importance des variables moyennée par objet visible.

Groupe de données	**Id**	**Observation**
1A	1	La rate est l'objet le plus important pour prédire la *qualité*.
1B	2	La rate est l'objet le plus important pour prédire la *qualité* lorsque le chirurgien 0 opère.
1C	3	La pince atraumatique est l'objet le plus important pour prédire la *qualité* lorsque le chirurgien 1 opère.
2A	4	L'estomac est l'objet le plus important pour prédire quel chirurgien opère.
2B	5	Le foie est l'objet le plus important pour prédire quel chirurgien opère lorsque l'exposition est bonne.
2C	6	La paroi abdominal est l'objet le plus important pour prédire quel chirurgien opère lorsque l'exposition pose problème.

Tableau 5.7 – Observations basées sur l'analyse des valeurs d'importance des variables moyennée $\overline{i_v}$ par objet visible (voir tableau ci-dessus).

rate 1A descripteur	$log_{10}(i_v)$	rate 1B descripteur	$log_{10}(i_v)$	pince atraumatique 1C descripteur	$log_{10}(i_v)$
texture	-2.32	surface	-1.91	color_a	-2.38
eigenvector_2_y	-2.13	eigenvector_1_y	-1.82	surface	-1.68
color_b	-2.11	color_b	-1.77	barycenter_y	-1.60
eigenvector_1_y	-1.97	barycenter_x	-1.74	color_l	-1.54
eigenvector_1_x	-1.93	perimeter	-1.73	barycenter_x	-1.45
perimeter	-1.87	color_l	-1.73	color_b	-1.41
barycenter_y	-1.83	eigenvalue	-1.72	texture	-1.38
eigenvector_2_x	-1.81	color_a	-1.71	eigenvector_2_x	-1.34
barycenter_x	-1.79	barycenter_y	-1.69	perimeter	-1.33
color_l	-1.73	texture	-1.67	eigenvector_2_y	-1.33
eigenvalue	-1.72	eigenvector_1_x	-1.66	eigenvalue	-1.31
color_a	-1.64	eigenvector_2_y	-1.65	eigenvector_1_y	-1.31
surface	-1.57	eigenvector_2_x	-1.55	eigenvector_1_x	-1.27

estomac 2A descripteur	log10(i_v)	foie 2B descripteur	log10(i_v)	paroi abdominale 2C descripteur	log10(i_v)
color_b	-2.07	eigenvector_1_y	-Inf	barycenter_x	-3.33
barycenter_x	-2.00	eigenvector_2_x	-Inf	barycenter_y	-3.33
color_l	-1.85	perimeter	-Inf	color_a	-3.33
color_a	-1.83	barycenter_y	-1.85	color_l	-3.33
surface	-1.73	barycenter_x	-1.78	eigenvalue	-3.33
eigenvector_2_x	-1.70	color_a	-1.78	eigenvector_1_x	-3.33
barycenter_y	-1.68	color_b	-1.78	eigenvector_1_y	-3.33
eigenvector_1_x	-1.66	color_l	-1.78	eigenvector_2_x	-3.33
eigenvector_1_y	-1.65	eigenvalue	-1.78	eigenvector_2_y	-3.33
texture	-1.64	eigenvector_1_x	-1.78	perimeter	-3.33
perimeter	-1.63	eigenvector_2_y	-1.78	surface	-3.33
eigenvalue	-1.56	surface	-1.78	texture	-3.33
eigenvector_2_y	-1.54	texture	-1.78	color_b	-1.72

Tableau 5.8 – Importance des variables de l'objet le plus important par groupe de données.

Groupe de données	Id	Observation
1A	7	La texture_std de la rate est le descripteur le plus important pour prédire la *qualité*.
1B	8	La surface de la rate est le descripteur le plus important pour prédire la *qualité* lorsque le chirurgien 0 opère.
1C	9	La color_a de la pince atraumatique est le descripteur le plus important pour prédire la *qualité* lorsque le chirurgien 1 opère.
2A	10	La color_b de l'estomac est le descripteur le plus important pour prédire quel chirurgien opère.
2B	11	Les eigenvector_1_x, eigenvector_2_y et périmètre du foie sont les descripteurs les plus importants pour prédire quel chirurgien opère lorsque l'exposition est bonne.
2C	12	Hormis la color_a, tous les descripteurs de la paroi abdominale sont importants pour prédire quel chirurgien opère lorsque l'exposition pose problème.

Tableau 5.9 – Observations basées sur l'analyse des valeurs d'importance i_v des variables de l'objet le plus important par groupe de données (voir tableau ci-dessus).

5.4.2 | Interprétation clinique et validation

En nous basant sur les observations des vecteurs I_V faites sur les 6 groupes de données et pour les deux expérimentations, nous avons constitué un questionnaire de 11 énoncés (voir tableau 5.10) avec leurs clips vidéos associés. Le questionnaire a été soumis à un expert et un interne sénior sous la forme d'un fichier power point avec les extraits vidéos incorporés (voir aperçus en annexe D). Les résultats sont présentés dans le tableau 5.11. On constate que l'expert était globalement d'accord avec 3 énoncés, pas d'accord avec 7 énoncés, et n'avait pas d'avis pour 1 énoncé. L'interne était globalement d'accord avec 5 énoncés, pas d'accord avec 5 énoncés, et n'avait pas d'avis pour 1 énoncé.

L'observation 12 du tableau 5.9 n'a pas donné lieu une traduction sous forme d'énoncé. Cette observation exprimait le fait que tous les descripteurs de la paroi abdominale ont la même importance, et se rapprochait trop de l'observation 6. Nous n'avons donc conservé et traduit que l'observation 6.

Obs. Id	Texte de l'énoncé
1	Afin d'optimiser l'exposition dans la scène chirurgicale, il est très important de bien gérer la rate.
2	Ici, pour bien gérer la qualité d'exposition, le chirurgien fait attention à sa gestion de la rate.
3	Ici, pour bien gérer la qualité d'exposition, le chirurgien fait attention à la manipulation de la pince a traumatique.
4	La différence de pratique entre les 2 chirurgiens de l'étude est surtout visible dans la façon de gérer l'estomac.
5	Dans les cas où la gestion de l'exposition est bonne, la différence de pratique entre les 2 chirurgiens de l'étude se voit surtout dans la gestion du foie.
6	Dans les cas où la gestion de l'exposition est plus compliquée, la différence de pratique entre les 2 chirurgiens de l'étude se voit surtout dans la gestion de la paroi abdominale.
7	Prêter attention à la texture de la rate aide beaucoup à évaluer la qualité d'exposition.
8	Ici, prêter attention à la surface de la rate dans l'image permet au chirurgien de bien gérer la qualité d'exposition.
9	Ici, prêter attention à la présence de couleur jaune ou bleu sur la pince a traumatique permet au chirurgien de bien gérer la qualité d'exposition.
10	Ici, prêter attention à la présence de couleur verte ou rouge sur l'estomac permet bien de distinguer la différence de pratique entre les 2 chirurgiens.
11	Dans les cas où la gestion de l'exposition est bonne, la différence de pratique entre les 2 chirurgiens s'observe surtout dans la forme et la place prise par le foie dans l'image.

Tableau 5.10 – Les énoncés cliniques basées sur les observations de nos résultats sur l'importance des variables, et qui ont été soumis à des cliniciens dans un questionnaire.

	Id de l'énoncé										
Clinicien	1	2	3	4	5	6	7	8	9	10	11
Chirurgien expert	++	++	−−	−−	−−	−−	−−	−−	+	o	−
Interne sénior	−−	++	++	++	−	−	−	++	o	−−	++

Tableau 5.11 – Résultats de la soumission du questionnaire à deux cliniciens - −−=« pas du tout d'accord », −=« plutôt pas d'accord », *o*=« sans avis », +=« plutôt d'accord », ++=« complètement d'accord »

5.5 | Discussion

Dans cette étude, nous avons investigué la notion de pratique chirurgicale à travers une analyse algorithmique sur différents sous-groupes de notre jeu de données, puis nous avons mené une interprétation clinique et une validation clinique des résultats de cette analyse. A la différence d'une grande majorité d'études de notre domaine de recherche qui se focalisent sur un aspect unique de la pratique chirurgicale : le niveau d'expertise ; nous nous sommes intéressés à deux autres indicateurs de la pratique chirurgicale : la qualité de gestion de l'exposition de la cible chirurgicale par le chirurgien, et le profil de pratique du chirurgien.

Notre objectif était de proposer une analyse quantitative de ces aspects de la pratique chirurgicale. Nous avons donc mené une permutation de valeurs sur nos variables d'entrée, afin d'évaluer leur importance i_v pour la tâche de prédiction (voir section 5.3.1). Nous avons proposé un cadre dédié à l'interprétation clinique des observations faites sur ces valeurs d'importance, et à la formalisation de ces interprétations sous forme d'énoncés (voir section 5.3.2). Nous avons illustré ces énoncés en sélectionnant semi automatiquement des clips vidéos issus des vidéos chirurgicales de notre jeu de données (voir section 5.3.3).

Enfin, nous avons proposé une nouvelle méthode de validation, dans laquelle nous avons confronté les résultats de notre analyse à l'avis de cliniciens. Cette confrontation a pris la forme d'un questionnaire que nous avons soumis à un chirurgien expert et un interne sénior, et dans lequel chaque proposition à évaluer prenait la forme d'un *énoncé* et de deux extraits vidéos associés (voir section 5.3.4).

5.5.1 | Les groupes de données

Les 6 groupes de données que nous avons définis proposaient différentes configurations de données d'entrée et de sortie. En créant ces différentes configurations, notre vision était d'obtenir des interprétations mettant en valeur les relations entre les façons d'opérer des chirurgiens (profil de pratique du chirurgien) et la qualité de gestion de l'exposition de la cible chirurgicale. Nous avons également proposé une présentation de ces groupes de données sous forme de questions.

Nous avons créé le groupe de données $1A$ pour étudier la *qualité* sur l'ensemble des données disponibles, tandis que les groupes de données $1B$ et $1C$ avaient pour but d'étudier les spécificités de chaque chirurgien dans sa gestion de la *qualité* d'exposition. Par symétrie, nous avons créé le groupe de données $2A$ pour étudier la pratique des deux chirurgiens sur toutes les données, et les groupes de données $2B$ et $2C$ pour étudier ce qui différencie la pratique des deux chirurgiens lorsque la *qualité* d'exposition est bonne ($2B$), et non-satisfaisante ($2C$). Autrement dit, nous avons questionné deux façons différentes de pratiquer le geste chirurgical afin de gérer au mieux l'exposition de la scène chirurgicale.

On notera que si les groupes de données $1A$ et $2A$ contiennent tous les échantillons de notre jeu de données, et ne diffèrent que par leur variable de sortie, les 4 autres groupes ne contiennent qu'une partie de ces échantillons. $1B$ et $1C$ contiennent tous deux les échantillons de 15 chirurgies, et $2B$ et $2C$ contiennent seulement les échantillons de 6 chirurgies.

La taille réduite de ces groupes de données impacte les performances de prédiction (voir tableau 5.5). La prédiction presque optimale pour $2B$ et $2C$ montre que notre modèle a par-

faitement prédit le *chirurgien*. Ces très bons résultats doivent être relativisés, en considérant notamment le peu d'échantillons à prédire, et les 3 configurations de validation croisée (voir tableau 5.3) sur lesquelles on a évalué les performances de prédiction : nous sommes face à des résultats très spécifiques.

5.5.2 | Le vecteur d'importance des variables I_V

En menant une évaluation de l'importance des variables à l'aide de la permutation de valeurs, nous pouvons facilement modifier voire complétement changer l'algorithme de prédiction utilisé. En effet, la permutation de valeurs est indépendante du modèle étudié puisqu'elle n'agit que sur les variables en entrée du modèle. En outre, cette approche ne présente pas de difficultés majeures dans sa conception. Nous pourrions donc facilement généraliser ce calcul de l'importance des variables à d'autres algorithmes et c'est une force de notre approche.

Pour chaque groupe de données, la pertinence du vecteur I_V calculé nous est donnée par le niveau de performance de la prédiction (voir tableau 5.5), et par le nombre de configurations de validation croisée sur lesquelles l'IV a été calculée (voir tableau 5.3). $1A$, $1B$, $1C$ et $2A$ ont des performances de prédiction relativement similaires, entre 0.6 et 0.73. Comme nous avons appliqué 20 configurations de validation croisée différentes sur $1A$ et $1B$ et 10 sur $1B$ et $1C$, nous avons bien pris en compte la variabilité des données et nous pouvons considérer que les vecteur I_V sur ces groupes de données presentent des résultats stables, notamment en comparaison des groupes de données $2B$ et $2C$.

En effet, les performances de prédiction presque parfaites observées sur $2B$ et $2C$, associées à seulement 3 configurations de validation croisée nous indiquent que les vecteurs I_V pour ces deux groupes sont forts mais instables, et très spécifiques à ces petits groupes de données. De fait, on constate dans le tableau 5.8 que les résultats de ces deux groupes de données ont une forme bien différente des autres groupes de données, avec notamment une distinction nette entre les variables importantes (eigenvector_1_x, eigenvector_2_x et périmètre pour $2B$) et les variables non-importantes (color_b pour $2C$). La pertinence de ces résultats nous paraît donc incertaine, mais nous profiterons de la validation clinique proposée par le questionnaire pour en rediscuter.

Afin d'effectuer correctement la permutation de valeurs, nous avons effectué une validation croisée, mais cette dernière a introduit une variabilité dans nos valeurs d'importance i_v. Nous n'avons présenté ici que l'importance moyenne $\overline{i_v}$, mais il serait également possible d'étudier la variance et la stabilité des valeurs d'importance i_v.

Par ailleurs, nous avons d'abord choisi de présenter l'importance moyennée par objet $\overline{i_v}$ avant de nous intéresser à l'importance des variables de l'objet le plus important. Nous avons présenté nos résultats de cette manière car nous n'avons pas observé de tendance plus forte qui aurait fait ressortir en priorité quelques variables précises. De plus, nous pouvions nous permettre de présenter nos résultats sous cette forme grâce à la sémantique descripteur-objet sous laquelle sont décrites nos variables. Cette représentation, spécifique aux données que nous avons annotées, nous a permis de proposer des observations interprétables à partir des vecteurs I_V portant sur ces descripteurs et objets. Cette représentation

est donc un point fort de notre méthodologie. Nous présenterons cependant des limites à l'interprétabilité de ces résultats par la suite.

Pour effectuer nos observations, nous avons considéré les résultats sous la forme des expérimentations 1 ($\overline{i_v}$ par objet) et 2 (i_v des 13 variables de l'objet le plus important), et nos observations portaient sur le(s) élément(s) ayant la plus haute valeur d'importance pour chaque groupe de données. On a ainsi obtenu 12 observations. Le choix de ne sélectionner que les variables de l'objet le plus important dans l'expérimentation 2 est discutable. Pour rappel, la valeur i_v correspond à la différence de performance observée en moyenne sur les Q configurations de validation croisée lorsqu'on applique une permutation de valeur. Dans le tableau 5.6, on observe par exemple pour le groupe de données 1A :

$$\begin{aligned} log_{10}(\overline{i_v}) &= -1.84, \quad \text{pour la rate} \\ log_{10}(\overline{i_v}) &= -1.80, \quad \text{pour la pince électrothermale} \end{aligned} \tag{5.1}$$

Ce qui signifie qu'entre ces deux objets on a une différence de performance moyenne inférieure à 0.002. Il est donc questionnable de n'avoir considérer que la rate comme objet le plus important, plutôt qu'une combinaison d'objets ayant une importance supérieure à un certain seuil par exemple. Cependant, en considérant plusieurs objet, nous aurions dû étendre notre analyse de l'expérimentation 2, et la forme des résultats dans cette expérimentation. Aussi, pour prendre mieux en compte les très fines différences de performance entre nos objets les plus importants sont très fines, il nous faudrait repenser la forme de nos résultats. Bien sûr, il aurait été plus simple d'interpréter des résultats montrant une tendance claire en faveur de certains objets ou descripteurs, mais ce n'a malheureusement pas été le cas.

Nous aurions également pu axer nos observations sur d'autres aspects de ces résultats, tels que le(s) élément(s) ayant la plus petite valeur ou en considérant la hiérarchie des objets, ou la répartition des organes et instruments dans les valeurs d'importance. A l'avenir, il serait intéressant de mieux formaliser et contraindre la forme prise par ces observations pour mieux maîtriser et automatiser cette étape de la méthode.

5.5.3 | La traduction des observations sur le vecteur I_V sous forme d'énoncés cliniques

Il n'a pas été sans difficultés d'interpréter les résultats sur l'importance des variables sous une forme clinique et en respectant la sémantique descripteur-objet. Ce processus a nécessité une analyse en profondeur du processus algorithmique ainsi qu'une bonne connaissance du contexte clinique.

Le fait que cette traduction ait été faite manuellement a permis de bien faire apparaître la dimension clinique dans les énoncés, mais elle a aussi introduit un biais dû à la perception personnelle de l'opérateur. Pour limiter la portée de ce biais, nous avons introduit certaines contraintes à respecter lors de la rédaction des énoncés. Nous aurions également pu limiter ce biais en impliquant plusieurs personnes dans ce processus, plutôt qu'une seule. Dans ce cas, la méthodologie de traduction aurait pu impliquer que chaque traducteur mène ses traductions indépendamment puis que, dans le cadre d'une discussion, chaque énoncé soit

discuter afin d'arriver à un accord, qu'il s'agisse d'un consensus ou d'un compromis. Une telle approche permettrait de confronter différentes versions des énoncés, ainsi que des différences d'interprétation des résultats pour trouver une solution prenant en considération plusieurs points de vue.

Pendant ce processus, nous avons rencontré des difficultés ou des subtilités liées à la sémantique du descripteur-objet. Par exemple, nous avons observé des configurations de descripteurs nécessitant des approches différentes pour être traduites sous forme d'énoncés : le barycentre d'un objet (descripteurs barycenter_x et barycenter_y) traduit la position moyenne de l'objet dans l'image, et, de fait, ses deux descripteurs ont des significations fortement entremêlées, même si ces deux coordonnées pourraient donner lieu à des significations cliniques séparées. Les 5 descripteurs eigenvalue, eigenvector_1_x, eigenvector_1_y, eigenvector_2_x, et eigenvector_2_y portent des informations sur la forme de l'objet, mais il semble compliqué de traduire en termes cliniques le sens spécifique de chacun de ces descripteurs. Nous n'avons pu traduire le descripteur de texture dans le contexte clinique, mais lors de la soumission du questionnaire à l'un des cliniciens, ce dernier nous a donné une piste. Selon lui, la texture est liée au ressenti tactile, et pas du tout au visuel. De même les énoncés 9 et 10 discutant des composantes de la couleur d'une forme ont été mal compris par les chirurgiens, qui cherchaient réellement ces couleurs dans les vidéos. Ainsi certains termes utilisés dans le contexte informatique prennent une signification tout autre dans le contexte clinique, et il faut y être attentif.

A l'inverse, aucune différence n'était faite du point de vue informatique entre instruments et organes, qui étaient indifféremment considérés commes des *objets*, alors que la différence est évidente dnas le contexte clinique. Il serait intéressant d'affiner la représentation objet-descripteur pour prendre en compte cette subtilité, ou au moins de proposer une différenciation au moment de la traduction.

Nous pourrions aussi envisager d'automatiser complètement notre méthodologie pour éliminer ce biais perceptif. Pour cela, il faudrait rendre automatique la génération des observations à partir des vecteurs I_V, ainsi que la traduction de ces observations sous forme d'énoncés. Comme nous l'avons fait remarquer précédemment, la génération des observations n'a pas été formalisée et si le processus de traduction est suffisamment contraint pour être mené manuellement, il ne l'est pas assez selon nous pour être mené automatiquement, au vu des subtilités liées à nos différents descripteurs. Un travail de formalisation de notre méthodologie est donc nécessaire pour tendre à l'automatisation complète, puis se poseront de surcroît des questions sur la capacité d'une approche algorithmique à traduire dans un langage intelligible des résultats eux-mêmes sélectionnés avec pertinence. Si nous voulons tendre vers une réelle application clinique, cette automatisation serait un apport important, non seulement dans une logique temps-réel, mais aussi pour assurer la stabilité du processus.

5.5.4 | Illustration des énoncés à l'aide d'échantillons et de leurs vidéo associées

Pour compléter les énoncés, nous avons sélectionné des clips vidéos issus des vidéos chirurgicales de notre jeu de données. Chaque clip était associé à un échantillon et servait de support visuel. Nous avons tenté d'extraire ces illustrations de façon aussi objective que possible en mettant en place une sélection automatique basée sur notre modèle. Ceci étant dit, si les performances de prédiction de notre modèle ne sont pas parfaites, la sélection des échantillons ne pourra pas l'être non plus. C'est pourquoi nous avons complété cette présélection automatique avec une sélection manuelle. De fait, le choix du paramètre n_{sel} fixant le nombre d'échantillon sélectionnés lors de la présélection automatique a été un compromis entre avoir une sélection complètement automatique d'une part, avec $n_{sel} = 1$, et laisser à l'opérateur une liberté complète dans le choix de l'échantillon d'autre part, avec $n_{sel} = n_{ech}$ le nombre d'échantillons de X.

Nous faisons ici une hypothèse forte, à savoir que le niveau de pertinence clinique des échantillons sélectionnés sur notre modèle entraîné est lié au niveau de prédiction de notre modèle. Il serait donc important de valider cette hypothèse sur laquelle repose nos résultats.

Dans le jeu de données complet, la durée séparant deux annotations de la *qualité* d'exposition variait sur l'intervalle $[2.3s; 361s]$ avec une statistique de durée de $19.1 \pm 21.9s$. En considérant la forte variabilité de ces durées, nous avons estimé que $t_{clip} = 10s$ donnerait une information suffisante sur les évènements caractéristiques de l'annotation de la *qualité*. Nous avons également décidé de fixer $t_{décalage} = -7s$, afin de donner plus de contenu sur les évènements précédant l'annotation de la *qualité*, plutôt que sur ceux lui succédant.

Quand nous lui avons soumis le questionnaire, l'un des cliniciens nous a fait remarquer qu'il trouvait les extraits vidéos trop courts pour pouvoir reconnaitre le chirurgien en train d'opérer. Il voyait cela comme un avantage lui permettant de se focaliser sur le geste pratiqué. Cette remarque nous pose question, car la variable de sortie *chirurgien*, ainsi que les énoncés portant sur cette dernière s'attachent justement à distinguer les profils de pratique des deux chirurgiens. A l'inverse, l'autre clinicien aurait préféré avoir des extraits vidéos plus longs qui lui aurait donné l'opportunité de distinguer les différents praticiens. Selon si on s'intéresse à la gestion de l'exposition ou à la caractérisation de la pratique du chirurgien, Il faudrait donc choisir respectivement une durée d'extrait plus courte ou plus longue pour accompagner l'énoncé.

Pour chaque énoncé, la présélection automatique nécessitait au préalable un entraînement de notre modèle sur le groupe de données et les variables associées à cet énoncé. Or, ce nombre de variables prend valeur dans $[1; 13]$, ce qui fait un nombre de variables très réduit comparé au $n_{var} = 130$ du jeu de données complet. Par exemple, l'énoncé 1 concerne les 13 variables de la rate, tandis que l'énoncé 11 se base uniquement sur 3 variables : eigenvector_1_x, eigenvector_2_y et perimeter pour l'estomac. Si ces variables ont été sélectionnées en raison de leur importance pour la tâche de prédiction, notre modèle a lui été initialement conçu et optimisé pour gérer une matrice de taille 735×130 et non une matrice de taille 730×13 (énoncé 1), voire une matrice de taille 156×3 (énoncé 11). Une étude d'impact de ces différents changements sur les performances du modèle, ainsi que sur le processus de sélection automatique des échantillons serait une perspective intéressante.

Contrairement au processus de calcul du vecteur I_V basé sur la permutation de valeurs, indépendante du modèle prédictif, l'approche que nous proposons ici pour sélectionner les échantillons est spécifique au SVM, car elle nécessite d'accéder à sa fonction de coût. Si nous changions notre algorithme prédictif, il nous faudrait trouver une méthode spécifique à ce nouvel algorithme. Par exemple, la forêt aléatoire, que nous avons discuté précédemment (voir section 4.5) résoud par sa nature même la question de l'importance des variables, et à travers une analyse fine des votes sur tous les arbres et dans chaque arbre, on pourrait envisager une méthode de sélection des échantillons.

L'étape de sélection manuelle des échantillons a été menée en visionnant les extraits vidéos associés, et en comparant leur contenu à l'énoncé associé. Cette approche est très exigeante pour l'opérateur, et les choix opérés sont complexes. En effet les échantillons présélectionnés par l'algorithme doivent être triés par l'opérateur, qui se trouve confronté à des choix de l'algorithme parfois difficilement compréhensibles. De plus, le choix final de l'opérateur laisse, encore une fois, place à sa subjectivité et se repose sur son interprétation personnelle. C'est donc un compromis entre laisser d'un côté l'opportunité à l'algorithme de sélectionner des échantillons d'une façon objective mais parfois incompréhensible du fait des performances de notre modèle, et de l'autre d'avoir une sélection subjective menée par l'opérateur, mais plus compréhensible.

5.5.5 | Le questionnaire de validation clinique

Nous avons soumis notre questionnaire à deux cliniciens et recueilli leurs opinions. La distribution des niveaux d'accord des deux cliniciens montre une première chose : c'est que les questions suscitent leur intérêt et qu'ils ne répondent pas « sans avis » à toutes les questions. Auquel cas, nous en aurions déduit que notre approche manquait de pertinence et n'intéressait pas les cliniciens. Suite à leur évaluation, les deux chirurgiens nous ont d'ailleurs confirmé qu'ils trouvaient cette démarche très intéressante. C'est donc un élément de validation important de notre méthode.

L'expert est d'accord avec 3 énoncés, tandis que l'interne est d'accord avec 5 énoncés. Cela nous montre que les cliniciens sont d'accord avec certaines des conclusions fournies par notre démarche algorithmique. On notera néanmoins qu'ils ne valident tous deux que l'énoncé 2 (voir tableau 5.10) qui caractérise une situation bien précise de pratique dans laquelle le chirurgien doit être attentif à la façon dont il gère la rate. Le fait de faire attention à la rate n'est pertinent pour le chirurgien que lorsqu'il se trouve à proximité de cet organe. Dans d'autres situations cet organe est ignoré. Donc si cet énoncé avait été accompagné d'autres vidéos extraits de situations l'illustrant moins bien, il nous semble que leurs avis auraient été différents. Nous aimerions donc savoir si l'avis que le chirurgien a sur l'énoncé proposé est impacté par le fait d'accompagner l'énoncé par des extraits vidéos ou non. Nous aimerions également savoir si l'avis que le chirurgien a sur l'énoncé proposé est impacté par le choix des extraits vidéos proposés.

Ces deux études d'impact permettraient aussi de valider une hypothèse que nous proposons ici : les analyses de la pratique clinique proposées avec notre approche peuvent en même temps être vraies dans des situations chirurgicales très spécifiques et fausses dans des situations chirurgicales plus générales. En effet, si on arrive à montrer qu'un énoncé

peut être validé ou invalidé par le chirurgien selon les situations de pratiques illustrées à travers les extraits vidéos l'accompagnant, alors cela signifie bien que nous ne traitons plus de cas généraux mais bien de situations particulières.

Il me semble que ce serait une avancée marquante dans l'analyse de la pratique chirurgicale, car nous serions désormais en mesure d'aider le chirurgien sur les détails de sa pratique. Passé un certain point de sa formation, le jeune chirurgien a bien acquis les principes généraux de la pratique chirurgicale, et ce qui lui pose encore problème, ce sont des situations spécifiques, des points de pratique bien précis. Or, sur ces points de pratiques, il n'existe aucune théorie et son seul soutien vient de ses propres expérimentations ou des chirurgiens plus expérimentés qui peuvent lui montrer l'exemple. Si donc nous arrivons à valider cette hypothèse, cela nous permettrait de proposer des supports de formation, non plus aux jeunes internes en début de formation, mais plutôt aux jeunes chirurgien en perfectionnement et à la recherche d'une autonomie complète.

Dans cette étude, la validation clinique n'a inclus que deux chirurgiens, nous aurions pu augmenter la robustesse de cette validation en augmentant le nombre de chirurgiens participants, et en variant leurs profils. De plus, le chirurgien expert qui a évalué notre approche est également un des deux chirurgiens qui ont opéré les patients de ntore cohorte. Il a donc analysé des échantillons vidéos provenant en partie de chirurgies qu'il avait lui-même effectué. Dans une étude future, il serait important d'éviter ce genre de biais et de bien distinguer les chirurgiens pratiquant de ceux qui évaluent.

En ce qui concerne nos résultats de validation, les deux chirurgiens n'étaient pas d'accord avec 7 énoncés (expert) et 5 énoncés (interne) mais surtout s'accordaient dans leur désaccord pour les 3 énoncés 5, 6 et 7. Ces désaccords traduisent potentiellement un manque de pertinence de notre approche, qui peut être dû :

1. à un niveau de performance insuffisant de notre modèle prédictif,
2. à une mauvaise traduction des résultats sur les vecteurs I_V,
3. à un mauvais choix d'extraits vidéos qui n'illustrerait pas de façon pertinente leur énoncé correspondant,

Dans le cas où l'explication 1 serait vraie, nous renvoyons le lecteur à la discussion du chapitre précédent (voir section 4.5) dans laquelle nous avons proposé différents axes possibles d'amélioration de notre modèle et de ses performances. On notera que ce niveau de performance a un impact sur le calcul des vecteurs I_V, mais également sur la présélection automatique des échantillons liés aux extraits vidéos. Dans le cas où l'explication 2 serait vraie, il s'agirait d'une part d'arriver à mieux comprendre pourquoi l'algorithme a produit de tels résultats, et d'autre part de réfléchir sur la façon dont nous avons traduit ces résultats algorithmiques sous une forme clinique ; ce qui nous ramène à notre discussion sur la traduction des observations sur les vecteurs I_V (voir section 5.5.3). Enfin pour le point 3, nous renvoyons le lecteur à notre discussion sur la sélection des échantillons (voir section 5.5.4), et notamment au compromis entre sélections semi-automatique et complètement automatique.

Les 3 énoncés 5, 6 et 7 pour lesquels les deux chirurgiens ont montré leur désaccord méritent d'être plus discuter. Il nous semble que l'énoncé 5 montre une limite de notre méthode

actuelle, dans le sens où, lors de la chirurgie, le foie est manipulé et déplacé par l'assistant et non par le chirurgien. Si le chirurgien guide et conseille son assistant, il n'a qu'un impact indirect sur la gestion du foie. Or dans l'ensemble de cette étude, nous avons considéré que tout ce qui se passait dans les images et donc dans les données, décrivait la pratique du chirurgien seul, mais pas de son assistant. Il faudra donc considérer cette subtilité à l'avenir. Nous ne voyons aucune explication plausible à l'énoncé 6 et pensons, comme les chirurgiens, qu'il s'agit réellement d'un mauvais résultat de notre approche. Enfin, l'énoncé 7 qui concerne la texture de la rate est peut-être plus faussé par notre mauvaise traduction du terme « texture », et si nous avions réussi à mieux traduire ce terme, à mieux comprendre ce qu'il exprimait, l'énoncé aurait pu être plus pertinent.

Le fait que l'expert et l'interne ne s'accordent que sur 4 de leur 11 évaluations est un signe de la faible robustesse de nos résultats. Nous aurions pu nous attendre à obtenir des résultats plus positifs, mais il semble que notre approche ne soit pas encore suffisamment aboutie et nécessite d'être améliorée afin de nous permettre d'obtenir des analyses cliniques réellement pertinentes. Ce manque de d'accord entre l'expert et l'interne est néanmoins intéressant car il traduit une forme de désaccord entre leurs visions respectives de la pratique du geste chirurgical. Cela est d'autant plus intéressant que l'interne sénior inclus dans l'étude à été formé par le chirurgien expert impliqué dans cette même étude. Même entre deux chirurgiens pratiquant ensemble en routine clinique et ayant les mêmes bases techniques, il existe des différences, notamment dans leur vision de la pratique chirurgicale, et de la qualité l'exposition plus précisément.

Deux énoncés ont chacun donné lieu à un « sans avis » de la part d'un chirurgien. Ce sont les énoncés 9 et 10 discutant de l'influence des composantes de couleur sur la qualité d'exposition et sur le profil de pratique. Face à ces considérations sur la coloration de la pince atraumatique et de l'estomac, les chirurgiens étaient désemparés, et il nous faudrait donc réussir à traduire de façon plus claire pour le clinicien ce que traduisent ces remarques sur la couleur des objets. D'autre part, nous avons noté que la réponse « sans avis » était problématique car elle voulait dire soit que le chirurgien avait autant d'arguments pour que d'arguments contre et choisissait une réponse 50/50, soit que le chirurgien ne comprenait pas la question ou ne la trouvait pas intéressante. A l'avenir, il serait pertinent de distinguer ces deux réponses pour que nos conclusions soient plus exactes.

Pour conclure, nous avons proposé dans cette étude une méthodologie basées sur des données issues de procédures chirurgicales réelles et menées sous cœlioscopie, permettant d'extraire des interprétations cliniques sous la forme de courts textes accompagnés d'extraits vidéos illustratifs. Ces interprétations ont été validées sous la forme d'un questionnaire soumis à deux chirurgiens qui ont donné leur avis sur la pertinence des interprétations proposées. A notre connaissance, aucune étude n'a encore proposé d'interprétations cliniques sous cette forme, ni n'a proposé de les faire valider par des chirurgiens sous la forme de questionnaire. Dans l'ensemble, nous avons donc proposé une approche novatrice et prometteuse pour l'étude et la compréhension de la pratique chirurgicale.

En résumé

— La qualité des résultats de prédiction a un impact sur les valeurs d'importance calculées, et donc sur la pertinence de nos énoncés cliniques
— La validation croisée a un impact important sur le calcul des valeurs d'importance et sur la sélection des échantillons illustratifs.
— Notre approche a le potentiel d'analyser des points spécifiques plus que des aspects généraux de la pratique chirurgicale.
— Malgré leurs bases de connaissances cliniques communes, les deux chirurgiens qui ont évalué notre questionnaire nous ont donné des résultats relativement différents. Cela montre la variabilité de la pratique chirurgicale, y compris dans le même service d'un même centre hospitalier.
— Cette méthodologie est novatrice et pleine de potentielles dans une optique d'analyse de la pratique chirurgicale.

5.6 | Conclusions

Dans cette étude, nous avons mené une permutation de valeurs sur nos variables d'entrée afin de quantifier l'importance de ces variables. En entraînant notre algorithme pour différentes configurations de données d'entraînement, nous avons pu caractériser l'importance de ces variables dans différentes configurations. Les observations de ces résultats ont été manuellement traduites sous forme d'énoncés cliniques en respectant certaines contraintes et en se basant sur la représentation sémantique des variables spatiales sous la forme objet-descripteur. Des extraits vidéos caractéristiques de chacun des énoncés ont été sélectionnés semi automatiquement dans le jeu de données, et viennent illustrer ces énoncés.

Nous avons ainsi construit un questionnaire illustré à l'adresse de chirurgiens experts. Ce questionnaire nous a permis d'évaluer la pertinence de notre approche consistant à extraire des interprétations cliniques complexes du fonctionnement d'un algorithme. Enfin, on notera que ce choix d'illustrer les échantillons à l'aide d'extraits vidéos a été validé par les deux cliniciens.

A l'aide d'un tel processus entièrement algorithmique, nous pensons qu'il sera possible dans un futur proche de créer automatiquement une telle association de textes descriptifs et de vidéos. Cette façon de présenter les analyses algorithmiques laisse envisager des applications cliniques dans le cadre de la formation, du retour sur expérience et, dans un futur plus lointain, de soutien à la pratique per-opératoire dans des situations chirurgicales spécifiques.

6

Conclusions & Perspectives

Préambule

Dans ce chapitre, je présenterai tout d'abord les contributions de mon travail de thèse en section 6.1, puis pour nos 3 études, je discuterai les limites ainsi que les perspectives que j'envisage sur l'annotation du jeu de données en section 6.2, sur la prédiction de la pratique chirurgicale en section 6.3 et sur l'analyse et la validation clinique de nos données en section 6.4. Enfin, je présenterai ma vision à court, moyen et long terme de l'évolution du domaine d'étude de la pratique chirurgicales en section 6.5.

Sommaire

6.1 | Résumé des contributions

Ce travail de thèse a présenté quatre apports distincts et cohérents :

1. un état de l'art centré autour de la notion de pratique chirurgicale et notamment des conclusions cliniques liées à cette notion,
2. La création d'un jeu de données annotées à partir de vidéos laparoscopiques et sa comparaison aux jeux de données en libre accès du domaine,
3. La prédiction de la qualité d'exposition et du profil de pratique du chirurgien,
4. l'interprétation clinique les relations entre ces données annotées et la pratique du geste chirurgical.

Dans notre état de l'art, nous avons étudié un corpus d'études portant sur différentes facettes de la pratique chirurgicale à partir de données extraites de l'environnement chirurgical. Nous avons abordé certains traits communs aux études de ce corpus, puis nous avons mené un travail de synthèse des différentes approches proposées en centrant notre discours sur la pratique chirurgicale. nous nous sommes plus particulièrement attachés à extraire et à décrire d'une façon cohérente les connaissances cliniques proposées dans les résultats d'analyses de ce corpus d'étude.

Nous avons ensuite mené une annotation détaillée au niveau de la procédure et des images de vidéos laparoscopiques, en y associant une annotation de la qualité d'exposition. Nous avons comparé notre jeu de données à ceux du domaine déjà existants, ce qui nous a permis d'évaluer les forces et faiblesses de notre approche. Plus précisément, la triple forme de notre annotation, ainsi que le niveau de détail de ces annotations font la richesse de notre jeu de données, et en comparaison de ces jeux de données, nous avons également proposé des approches de validation avancées de l'annotation de la procédure et de la segmentation.

A partir de ce jeu de données annotées, nous avons défini un ensemble de descripteurs interprétables, puis nous avons développé un algorithme et un environnement d'optimisation afin de prédire la qualité de gestion de l'exposition de la cible chirurgicale ainsi que le profil de pratique du chirurgien. Si nos performances de prédiction méritent d'être améliorées, nous sommes néanmoins les premiers à prédire des aspects de la pratique chirurgicale si peu étudiés et aussi complexes. Nous avons également constaté, sur notre cas d'étude, la primauté des descripteurs spatiaux sur les descripteurs procéduraux dans la tâche de prédiction.

Enfin, nous avons défini une méthodologie interprétant les analyses algorithmiques de nos données sous la forme d'énoncés cliniques accompagnés de clips vidéos et nous avons validé la pertinence de ces interprétations via un questionnaire d'évaluation soumis à deux cliniciens. De telles méthodologies d'interprétation et de validation cliniques n'ont jamais été proposées auparavant.

Finalement, nos résultats se présentent comme un travail préparatoire de faisabilité pour l'étude de nouveaux aspects de la pratique chirurgicale. Il serait possible de proposer de nouvelles plus-values cliniques notamment dans le cadre de la formation des chirurgiens ainsi que pour l'amélioration des connaissances cliniques portant sur la pratique chirurgicale. Notre travail est une preuve de concept qui ouvre la voie et permet d'envisager de nombreuses possibilités dans l'analyse et la compréhension de la pratique chirurgicale.

6.2 | Étude 1 : création du jeu de données issues de vidéos de chirurgies laparoscopiques

Une annotation de données peut toujours être plus étendue : que ce soit en menant la même tâche d'annotation sur de nouvelles données vierges, ou bien en améliorant et en détaillant l'annotation existante. Cette deuxième option me semble particulièrement intéressante, car elle pousse à densifier la quantité d'annotation et donc à proposer une information plus fournie sur les données initiales, mais aussi à imaginer de nouvelles formes d'annotation. Dans notre discussion autour de cette étude, nous tenterons tout d'abord d'imaginer de nouvelles formes d'annotation, puis nous proposerons une stratégie d'annotation basée sur nos annotations déjà existantes. Nous finirons en discutant de la validation du processus d'annotation.

6.2.1 | Vers de nouvelles formes d'annotation

6.2.1.1 | Approfondir la modélisation de la pratique chirurgicale

Dans notre état de l'art (voir section 2.2.1), nous avons vu que le niveau d'expertise chirurgicale était étudié de façon extensive, mais que la facon de modéliser cette notion variait très peu : elle était systématiquement définie sur deux ou trois niveaux (novice, intermédiaire et expert). Nous proposons ici quelques questions qui permettraient d'étendre et d'approfondir notre compréhension de cette notion :

1. Est-il possible d'établir avec précision le nombre de chirurgies caractéristique de chaque niveau d'expertise ?
2. En quoi la spécialité chirurgicale influence-t-elle le niveau d'expertise ?
3. Quels aspects de la pratique chirurgicales influencent le niveau d'expertise ?
4. Est-il possible de distinguer plus de niveaux d'expertise différents et ainsi d'affiner le niveau de granularité auquel est actuellement défini l'expertise ?

La question du nombre de chirurgies permettant de caractériser un niveau d'expertise pourrait faire l'objet d'un état de l'art, afin de décrire le spectre des propositions faites, et peut-être, justement de montrer que ce nombre de chirurgies peut être corrélé à la spécialité (Hedrick et al., 2009) ou à d'autres facteurs et ne peut donc être défini d'une façon absolue. On voit apparaître ici des éléments de réponse pour la deuxième question : on a certainement un lien à étudier entre le niveau d'expertise, le nombre de chirurgies la spécialité chirurgicale, voire la technique chirurgicale elle-même (Hedrick et al., 2009). Plus spécifiquement, il serait intéressant de connaître les différences entre des chirurgiens d'un niveau d'expertise équivalent mais issus de spécialités différentes.

L'étude de Nugent et al. (2012) que nous avions déjà décrite dans notre état de l'art (voir section 2.3.3) propose un autre élément de réponse à la question 3 : les capacités visuo-spatiales ont un impact sur le niveau de pratique des internes en chirurgie, ce qui peut influencer leur capacités d'apprentissage et donc leur niveau d'expertise. Les relations entre la courbe d'apprentissage et le niveau d'expertise seraient également intéressantes à étudier.

Pour répondre à la question 4, il serait intéressant de mener des travaux afin d'affiner ces niveaux d'expertise, à travers les différents stades de la formation du chirurgien par exemple (Vedula et al., 2016). A noter que l'étude des trois premières questions apportera très certainement des éléments de réponse à cette dernière question du fait des nouvelles connaissances qui apparaîtront et étaieront la modélisation du niveau d'expertise.

Dans notre étude, la modélisation de la pratique chirurgicale menée en compagnie de notre partenaire clinique nous a permis de proposer une étude des notions d'exposition de la scène chirurgicale, et du profil de pratique du chirurgien. Ce sont deux facettes de la pratique chirurgicale mais nous croyons qu'il est possible d'envisager encore d'autres façons de décrire ce phénomène complexe. Par exemple, l'idée de caractériser la pratique d'un individu ou d'un groupe d'individus, ou d'étudier la courbe d'apprentissage ont déjà été envisagées, et nous paraissent pleines de potentiel. Par ailleurs, une réflexion plus poussée sur le concept d'exposition, sur ses fondements et ses composantes pourrait permettre, à l'image des scores structurés, de proposer une décomposition de cette notion et d'y distinguer différents sous-aspects et sous-problématiques (voir la définition de la notion d'exposition en section 3.2.4.1).

A travers cette discussion, nous avons fait apparaître, en plus du niveau d'expertise, d'autre aspects de la pratique chirurgicale que sont la spécialité chirurgicale et la courbe d'apprentissage, la qualité d'exposition ainsi que le profil de pratique du chirurgien. Certaines facettes de la pratique chirurgicale dont nous discutons sont certes très proches et se recoupent, telles l'expertise et la pratique chirurgicale : un expert pratiquera à coup sûr mieux qu'un jeune chirurgien ; ou le profil de pratique et le niveau d'expertise : un chirurgien expert et un chirurgien junior d'un même centre hospitalier auront peut-être plus de similarités dans leurs pratiques que deux experts de deux centres différents. De façon générale, il me semble qu'il serait très intéressant de mener des études tranverses sur les relations et les différences entre ces différents aspects. Par la même occasion, on acquerrait une compréhension plus complète de la pratique chirurgicale dans son ensemble.

6.2.1.2 | le potentiel d'annotation du flux optique

Dans une vidéo, le flux optique correspond au mouvement relatif des objets visibles dans la scène vis à vis du point de vue de la caméra. Cette information traduit donc l'évolution temporelle de l'environnement spatial filmé par la caméra. Au vu de l'annotation que nous avons proposé de la procédure d'une part, et du contenu de certaines images d'autre part, ce flux optique caractérise un aspect essentiel et non-couvert par notre annotation : la temporalité de la vidéo. De fait, l'étude du flux optique permettrait par exemple, entre deux images segmentées, de caractériser les mouvements généraux de l'image, voire de caractériser spécifiquement les déplacements des différents objets observés. Cela permettrait de pallier une faiblesse de notre étude : en ne segmentant qu'une image par échantillon de qualité d'exposition, nous donnons certes une information détaillée à l'instant de l'annotation, mais c'est une information très partielle pour couvrir le phénomène complexe d'exposition.

Plus spécifiquement dans notre étude, nous avons étudié la qualité de gestion de l'exposition de la cible chirurgicale. Or, cette qualité dépend énormément de la vision que le chirurgien a de la scène et donc de la gestion du champ de vision de l'endoscope. Dans cette

optique, nous avions déjà considéré l'idée d'annoter les activités de la main de l'assistant tenant l'endoscope. Cependant, sous sa forme actuelle, le triplet <verbe ;instrument ;cible> n'est pas adapté pour proposer une telle annotation : le verbe serait toujours « filmer » et la cible serait « la scène chirurgicale ». Pour affiner cette description, il faudrait notamment détailler la scène chirurgicale et les différents objets visibles par exemple. On pourrait aussi imaginer de distinguer la cible focale et la vision périphérique.

Cette réflexion soulève un problème récurrent dans nos travaux : la granularité avec laquelle nous décrivons l'environnement chirurgical et plus précisément l'anatomie du patient est trop grossière, que ce soit dans l'annotation procédurale, la segmentation ou l'annotation du flux optique. Ce manque de finesse rend notamment nos interprétations cliniques de l'étude 3 trop imprécises. D'autant plus que nous nous intéressons à une étape chirurgicale focalisée sur le fundus, une partie de l'estomac pour laquelle nous n'avons pas annotés de repères anatomiques spécifiques. Il est donc difficile, dans une certaine mesure, de caractériser avec précision cette étape, si le vocabulaire utilisé ne permet pas de décrire correctement des sous-parties de ce fundus.

Pour décrire la gestion du champ de vision de l'endoscope, nous avions tenté une annotation dans laquelle nous qualifiions l'amplitude des déplacements de l'endoscope et donc des changements de plan par les adjectifs « immobile », « faible », « moyen », « fort ». Nous avons également utilisé les adjectifs « gauche », « droite », « haut », « bas », « zoom » et « dé zoom » pour décrire les directions possibles. Malheureusement nous nous sommes rendus compte que cette nomenclature était trop imprécise et laissait trop de place à l'interprétation.

Nous proposons ici une étude afin de qualifier de façon robuste les transformations du champ de vision endoscopique, de définir une sémantique permettant d'envisager une annotation et, à terme, de proposer une interprétation clinique telle que celle que nous avons proposée dans notre étude 3 (voir chapitre 5). Plus précisément, il faudrait étudier la faisabilité d'une reconnaissance des différents types de mouvement de l'endoscope (vertical, horizontal et en profondeur), ainsi que de la qualification de leur intensité à partir d'un algorithme de flux optique. Cette étude ne serait pas sans difficulté, et il faudrait notamment prendre en compte l'absence d'arrière-plan stable dans la scène chirurgicale. Mais si des telles informations peuvent être recueillies à partir du flux optique, on pourrait alors envisager de définir une sémantique caractérisant la direction et l'intensité des mouvements de l'endoscope. Dès lors, il serait possible d'envisager une annotation automatique des mouvements de l'endoscope.

Pour aller plus loin, l'analyse du flux optique est aussi limité par le fait qu'on filme en 2-dimensions une scène en 3-dimensions, et cette transformation implique nécessairement une perte d'information. On pourrait imaginer des approches multi-modales avec une technologie telle que l'échographie 3D et/ou avec un modèle 3D des organes observés. De telles approches orientées fusion de données ne sont pas à l'ordre du jour, du fait de nombreux défis à relever liés par exemple à l'information qualitative véhiculée par l'échographe ou à la plasticité d'un organe comme l'estomac (de fait difficile à modéliser en 3D). Dans une vision à long terme, un accès à de telles informations permettrait d'envisager un suivi temps-réel des organes tout au long de la procédure.

6.2.2 | Amélioration de la stratégie d'annotation

La façon dont une annotation est menée peut grandement influencer la forme et la qualité des données annotées ainsi que le temps passé à annoter. L'annotation étant extrêmement chronophage, le temps d'annotation est une problématique critique justifiant une grande partie des choix de stratégies d'annotation mises en place. Par exemple, au-delà d'approches basiques, la segmentation du jeu de données *CaDIS* (Flouty et al., 2019) s'est basée sur l'annotation procédurale *CATARACTS* et a ainsi densifié l'annotation déjà disponible sur les vidéos initiales de l'étude de (Al Hajj et al., 2019). L'annotation du jeu de données *CaDIS* a été menée de façon intelligente car les éléments annotés au préalable dans *CATARACTS* ont servi à construire une annotation plus pertinente et efficace. De façon similaire, c'est le manque de temps et de ressources qui nous ont poussés à n'annoter que la procédure des étapes de dissection du fundus et à ne segmenter que les images correspondant aux échantillons de qualité d'exposition.

6.2.2.1 | Définition de la répartition des échantillons

Il est important de bien penser la façon dont sont répartis les échantillons sur les données initiales. Leur répartition temporelle doit notamment être bien conçue. Selon la façon dont l'information annotée est distribuée, elle ne caractérisera pas de la même façon les données initiales. Ainsi, l'annotation proposée dans *CaDIS*, en plus de proposer un gain de temps, a été définie afin que les images segmentées rendent compte de la diversité des différentes phases chirurgicales et présentent toujours au moins un instrument visible. De même, nous avons axé notre segmentation d'image sur les échantillons de qualité d'exposition car c'est précisément le cœur de notre étude.

Dans cette dynamique, on peut envisager d'étendre notre annotation actuelle en définissant nos objectifs en amont et en prenant en compte les informations déjà annotées. Nos tâches d'annotation pourraient tout d'abord être étendues en définissant des règles de répartition des échantillons en fonction de l'annotation procédurale, et/ou en fonction des échantillons d'exposition annotés telles que :

— Une distribution homogène des échantillons selon les différents types d'activité.
— Une distribution homogène des échantillons selon les échantillons d'exposition.
— Une annotation d'échantillons lors des transitions entre activités.

En définissant des règles sur les distributions, on homogénéise la répartition des échantillons entre les différents types d'activité ou les échantillons d'exposition annotés et on pousse l'annotation à rendre compte de façon plus complète de la diversité procédurale de la chirurgie. En s'intéressant aux transitions entre activités, il s'agit plutôt de caractériser des instants de transition potentiellement critiques dans le déroulement de la chirurgie et donc intéressantes pour le clinicien. Par exemple, il serait particulièrement intéressant d'étudier des transitions telles que les repositionnements de la main gauche du chirurgien qui tient et expose l'estomac, ou bien les repositionnements de la compresse. Ces transitions d'activités correspondent à des changements d'états de la qualité d'exposition, où le chirurgien est insatisfait et cherche à améliorer l'exposition. Ce sont donc des instants critiques pour bien comprendre ce qui déclenche le besoin qu'a le chirurgien d'améliorer l'exposition de sa cible d'une part, et la façon dont il cherche à améliorer cette exposition d'autre part.

6.2.2.2 | Décomposition de la tâche d'annotation

Les validations que nous avons menées sur l'annotation de la procédure et sur la segmentation ont fait apparaître la forte variabilité des annotations, et donc des imprécisions dans la description du phénomène chirurgical. Cette variabilité est due en partie à la complexité de nos tâches d'annotation. En décomposant une tâche d'annotation complexe en tâches successives plus simples, on pourrait diminuer le risque d'erreur, où tout au moins le répartir et mieux le caractériser entre les différentes sous-tâches. Par exemple, la segmentation d'une image pourrait être découpée de la façon suivante :

1. Localisation : l'opérateur détecte chaque objet visible dans l'image et positionne un marqueur de présence à son endroit dans l'image.
2. Segmentation : l'opérateur, informé de la position des objets dans l'image, segmente la forme associée à chaque objet.
3. Validation : l'opérateur vérifie et assure la qualité des formes segmentées.

Présentée ainsi, la difficulté de la tâche complète a été répartie, et certaines tâches nous paraissent plus critiques. Ainsi la tâche de localisation permet d'assurer une certaine qualité de la segmentation qui suit, et il nous parait pertinent de placer un expert non-clinicien comme opérateur sur cette tâche. Pour le jeu de données *CATARACTS*, Al Hajj et al. (2019) ont proposé une telle étape d'annotation préalable menée de front par deux cliniciens. La tâche en question concernait la détection de la présence d'instruments dans les images, et le résultat de cette détection pour un instrument était 0 si les deux cliniciens n'avait pas détecté l'instrument, 1 s'ils l'avaient tous deux détecté, et 0.5 si seulement un des cliniciens l'avait détecté. Une telle approche pourrait nous inspirer en introduisant une forme de redondance de l'annotation.

La tâche 3 de validation doit nécessairement être bien menée pour s'assurer de la validité des segmentations et c'est pourquoi nous proposons de mettre un ou plusieurs cliniciens sur cette tâche, plutôt que sur la tâche 1 de localisation. D'autant plus que cette tâche 3 pourrait être dimensionnée de façon à ne pas représenter une charge de travail trop importante. Finalement, la tâche 2 de segmentation est la plus problématique, car c'est le travail le plus chronophage et le plus source d'erreurs. En effet, même si les objets visibles ont été localisés au-préalable, les limites entre objets, et notamment entre certaines structures anatomiques, sont floues et donc difficiles à segmenter. La présence potentielle d'artefacts au sein des images complexifie également ce travail d'annotation.

6.2.2.3 | L'apport de l'annotation par la foule

Dans l'exemple précédent, la segmentation est la plus chronophage, mais aussi la plus complexe car source d'erreurs. Nous proposons ici de l'adapter pour une annotation par la foule. Plus précisément, la tâche élémentaire proposée à l'opérateur anonyme serait la segmentation d'un objet unique dans une image, pour lequel on donnerait la position moyenne et le nom de l'objet à segmenter, et on récupérerait une image binaire de la forme segmentée de l'objet. Dans la logique de l'annotation par la foule, chaque tâche élémentaire serait menée par différents opérateurs (nombre à définir), et le résultat de la segmentation ne serait pas une image binaire mais une image en niveau de gris traduisant la segmentation

moyenne des différents opérateurs, ou plutôt la carte de chaleur définissant la probabilité de présence de l'objet dans l'image.

En amont de cette tâche noyau, il faudrait proposer une introduction aux opérateurs anonymes pour les préparer à la tâche attendue, ainsi qu'un tutoriel permettant d'évaluer la qualité de leur segmentation. Pour s'assurer de cette qualité, il serait intéressant de proposer des images témoins à l'opérateur à intervalle regulier pendant l'annotation. Par image témoin, j'entends des images pour lesquelles nous avons une vérité terrain nous permettant par la suite de mener une étude de variabilité.

Par ailleurs, notre étude de variabilité inter-opérateur sur la segmentation nous a montré que certains objets sont plus difficiles à segmenter que d'autres. En prenant ce fait en compte, il faudrait d'abord évaluer la pertinence de l'annotation par la foule sur des objets simples avant de s'intéresser aux objets plus complexes à segmenter.

La présentation des segmentations, non plus sous forme d'images binaires, mais de cartes de chaleur nous semble être très intéressante. D'une part, de telles données probabilistes se prêteraient très bien à une utilisation dans des algorithmes d'apprentissage profond. D'autre part, l'étude de telles cartes de chaleur permettrait de faire apparaître les niveaux d'accord entre les opérateurs sur la segmentation de différents objets. Nous pourrions ainsi étudier plus en profondeur la tâche de segmentation en fonction des objets, ou de l'environnement visuel dans lequel les objets sont segmentés. A travers une telle carte de chaleur, nous serions aussi mieux à même de quantifier la variabilité des segmentations et les erreurs potentielles faites par les opérateurs.

6.2.3 | Validation de l'annotation

A l'avenir je pense que si la validation des travaux d'annotation se systématisait, cela apporterait une meilleure connaissance des données annotées ainsi que du processus d'annotation lui-même. Nous pourrions mieux connaître les points forts et les points faibles d'une tâche d'annotation, tels que les objets particulièrement faciles/difficiles à annoter ou les types d'erreurs caractéristiques d'une tâche d'annotation. Nous aurions alors une meilleure idée de la qualité des données à disposition et de leur variabilité, ce qui nous permettrait d'orienter notre utilisation de ces données. Dans le cadre de la création d'une annotation plus détaillée, la stratégie d'annotation choisie se ferait en insistant par exemple sur certains aspects mals ou peu annotés, ou en évitant justement certains types d'annotation trop complexes. On pourrait également envisager de prendre en compte la qualité des annotations dans une approche de prédiction, où un coefficient de confiance pour les différents objets annotés pourrait être pris en compte dans l'apprentissage.

Forts de cette considération sur l'importance de la validation, nous sommes forcés de constaté que notre jeu de données présente une forte incohérence : nous n'avons mené aucune étude de validation inter- ni intra-opérateur pour l'annotation de la qualité d'exposition. Si nous n'avons pas pu mener cette validation, c'est en partie par manque de disponibilité de nos partenaires cliniques, ce qu'il nous faudrait donc prendre en compte dans la conception de nos études d'annotation futures impliquant des cliniciens, afin d'envisager des solutions alternatives. Cette absence de validation est un problème majeur, dès lors que notre objectif est de proposer des analyses cliniques portant sur la qualité d'exposition, et

que nous cherchons à prédire et à analyser cette notion. C'est la pertinence même de nos résultats qui est en jeu.

En ce qui concerne l'annotation procédurale et la segmentation, nous avons mené des études de variabilité inter-opérateur qui montrent une variabilité à ne pas négliger. Concernant l'annotation procédurale, notre étude a fait apparaître une claire variabilité, mais nous n'avons comparé que deux annotations d'une même procédure. Il nous faudrait plus de données pour conclure, ainsi que des métriques de comparaison plus adaptées que de simples statistiques descriptives. Notre étude de variabilité inter-opérateur sur la segmentation est déjà plus conséquente car elle porte sur 35 images entièrement segmentées par trois opérateurs, et a été menée avec la métrique d' « intersection over union ».

En reprenant notre proposition de stratégie d'annotation pour une tâche de segmentation par la foule, une étude de variabilité inter-opérateur serait extrêmement pertinente, voire même essentielle. Par exemple, faire annoter des images témoins à intervalles réguliers par des opérateurs anonymes permettrait d'obtenir de nombreux échantillons à partir desquels la variabilité inter-opérateur pourrait être estimée. Outre l'analyse de variabilité elle-même, nous envisageons plusieurs retombées à une telle étude :

— Tout d'abord, on pourrait mieux caractériser la segmentation de différents objets.
— On pourrait définir un niveau de confiance pour les différents opérateurs, à partir duquel on pourrait par exemple pondérer l'importance accordée aux segmentations de chaque opérateur dans les cartes de chaleurs finales, ou bien de leur proposer de tâches de segmentations niveau de difficulté cohérent avec leur niveau de confiance.
— On considère l'étape 3 de notre tâche d'annotation décomposée (voir ci-dessus section 6.2.2.2) dans laquelle nous proposons de valider les segmentations de l'étape 2. Dans le cadre d'une annotation par la foule, mener cette tâche de façon exhaustive serait extrêmement chronophage. Pour la réduire, on pourrait présélectionner les échantillons en fonction de la probabilité d'y rencontrer des erreurs de segmentation. A partir du niveau de confiance des différents opérateurs, nous pourrions axes cette étape de validation sur les échantillons les plus à risques.

6.3 | Étude 2 : Prédiction de descripteurs caractérisant la pratique du chirurgien par optimisation d'un algorithme d'apprentissage

Dans le cadre de notre étude sur la prédiction de la qualité d'exposition, nous avons particulièrement travaillé sur les problématiques liées à la gestion de notre jeu de données qui présente des contraintes fortes. Tout d'abord, nos échantillons sont distribués de façon inhomogène sur nos 30 procédures, on a aussi un fort déséquilibre entre nos valeurs « bonne » et « perfectible » de qualité d'exposition. Nous avons géré ces deux biais grâce à l'environnement de validation croisée, et avec succès, notamment quand on observe que notre modèle prédit aussi bien les deux valeurs malgré ce déséquilibre. Par ailleurs, nos variables d'entrée présentent de nombreuses valeurs manquantes et mélangent différents types de valeurs. Ces contraintes ont été gérées dans l'étape de prétraitement de notre algorithme.

En section 4.5, nous avons insisté sur l'intérêt qu'il y aurait à étendre les options de traitement des différentes étapes de notre algorithme, et à les comparer. De nombreuses options pourraient être testées que ce soit pour les étapes de prétraitement, de réduction de dimension ou de prédiction, et nous pensons que cela permettrait de trouver une combinaison donnant de meilleures performances que notre modèle, ou a minima de connaître le niveau de prédiction pour différentes options de traitement.

En explorant ces différentes options, il serait également possible améliorer les performances obtenues sur les prédictions issues des descripteurs procéduraux. Notre résultat de primauté des descripteurs spatiaux sur les descripteurs procéduraux est surprenant quand on sait que les données procédurales sont utilisées avec succès pour prédire le niveaux d'expertise dans de nombreuses études, et qu'aucune étude, à notre connaissance, ne propose de tels résultats à partir de descripteurs extraits d'une image unique. Il serait donc intéressant de trouver une configuration algorithmique proposant des performances de prédiction satisfaisantes à partir de données procédurales. Cela nous permettrait également d'étendre notre étude 3 à nos descripteurs procéduraux qui, jusqu'à maintenant, n'ont pas donné lieu à des analyses cliniques.

6.4 | Étude 3 : analyse et validation clinique des relations entre données d'entrée et de sortie de l'algorithme

Cette étude d'analyse statistique menant à des interprétations traduites en langage clinique et validées par des chirurgiens est le point d'orgue de ma thèse. Elle combine une analyse algorithmique relativement robuste à la présentation d'énoncés accompagnés d'extraits vidéos au sein d'un questionnaire soumis à des chirurgiens. A notre connaissance, aucune étude du domaine ne propose des résultats algorithmiques sous une telle forme, ni de validation clinique à travers un questionnaire soumis à des cliniciens.

Bien sûr notre étude présente de nombreux axes d'amélioration, mais cette preuve de concept nous semble très prometteuse et mérite d'être approfondie. Par exemple, il serait intéressant ici aussi de comparer différentes méthodes algorithmiques, car de meilleures performances de prédiction devraient logiquement être corrélées avec des analyses cliniques plus pertinentes. De plus, une approche algorithmique différente, en dehors de ses performances mêmes, peut proposer une approche analytique différente dont il serait intéressant d'étudier le potentiel en terme d'analyses cliniques. La validation clinique par questionnaire nous semble être un bon juge de la pertinence des résultats obtenus.

Par ailleurs, la formalisation des résultats à l'aide d'un énoncé et d'une vidéo a un potentiel très intéressant : c'est a priori un support pertinent pour valider nos résultats algorithmiques, comme nous l'ont confirmé les deux chirurgiens qui ont évalué le questionnaire. En interrogeant le chirurgien, on peut donc adapter ou réorienter notre méthode algorithmique en fonction de son opinion. On pourrait même envisager une méthode de recherche agile : de manière itérative, on répèterait des phases de développement suivies par la soumission de tels questionnaires auprès des chirurgiens, et on aurait ainsi un retour régulier du chirurgien sur l'orientation des recherches. Le questionnaire servirait alors d'interface avec les

cliniciens et permettrait de s'assurer que nos résultats sont pertinents à leurs yeux.

Dans une vision à plus long terme, et si nos résultats gagnent en pertinence et en robustesse, le suivi et l'analyse de critères comme la qualité d'exposition ou le profil de pratique pourraient permettre de caractériser la pratique d'un individu, voire d'étudier l'évolution de sa pratique. Par exemple dans le cadre de sa formation, on pourrait suivre en détail la transformation de la pratique de l'interne à travers un ensemble d'indicateurs. Ou bien en s'inspirant de l'étude de Vedula et al. (2016), on pourrait étudier la courbe d'apprentissage en proposant une série d'exercices d'entraînement à des internes de niveaux variés, et étudier les points forts et points faibles spécifiques à chaque année de formation. On se rapproche ici du profil de pratique tel que Huaulmé et al. (2017) l'ont envisagé : la caractérisation des profils de pratique a le potentiel d'expliciter les différences de pratique entre différents chirurgiens, ou entre différentes populations de chirurgiens issus d'hôpitaux différents.

Enfin on pourrait imaginer un outil d'assistance à la formation se basant sur un ensemble d'enregistrements chirurgicaux et faisant l'interface entre le chirurgien et ses internes. Le chirurgien pourrait focaliser l'attention de ses internes sur un aspect de la pratique à travers des extraits vidéos illustrant les notions enseignées, et ces extraits seraient sélectionnés grâce à une reconnaissance automatique de situations caractéristiques de ces mêmes notions (idée inspirée de l'application développée par Touch Surgery Labs [1]). Le support proposé aux internes pourrait alors prendre une forme plus évoluée de notre support énoncé plus vidéo : ce seraient des extraits de procédures chirurgicales sous-titrés. Ces sous-titres, à l'image de nos énoncés, prendrait la forme de conseils, de mise en garde ou de rappels, et s'adapteraient au contenu clinique de la vidéo.

6.5 | Perspectives

La présentation du contexte de ma thèse a fait apparaître un environnement clinique propice au développement de solutions pour la formation, l'évaluation et l'analyse de la pratique chirurgicale. Si les travaux de recherche du domaine proposent aujourd'hui des résultats prometteurs, les approches proposées ne permettent pas encore d'envisager des applications cliniques. Je vois trois limitations principales empêchant l'avènement de telles applications :

— Le temps-réel : pour être utilisable en routine clinique, les systèmes proposés doivent être capable de proposer un résultat exact et fiable au moment où le chirurgien l'attend, ou ne l'attend pas justement (s'il s'agit de système de détection de déviation liée à des risques par exemple).

— La plus-value clinique : l'idée est de soutenir le chirurgien et de faciliter son travail, ainsi que d'améliorer la prise en charge des patients. Si notre domaine de recherche en a le potentiel, rares sont les études actuelles qui proposent une valeur ajoutée claire pour le clinicien.

— La validation : à plusieurs reprises dans notre état de l'art, nous avons constaté la faible importance que revêtait la validation et l'évaluation des outils proposés. Or cet

1. https ://www.touchsurgery.com/

aspect est essentiel dans l'optique d'une utilisation clinique sûre et stable.

La formation des internes est également une tâche essentielle mais chronophage pour le chirurgien, ainsi tous les outils permettant de donner plus d'autonomie à l'interne dans sa formation, ou de proposer un enseignement plus efficace sont des apports réellement utiles pour le chirurgien et son interne. On peut aussi imaginer d'assister le chirurgien en l'aidant dans des tâches plus complexes qui sont au cœur de son métier. Et on pense naturellement aux opérations chirurgicales, mais aussi à la recherche clinique. J'insiste sur le fait qu'en facilitant le travail du chirurgien, en lui permettant de mieux se former et de pratiquer la chirurgie dans de meilleures conditions, on améliore d'autant la prise en charge du patient et on réduit les risques pour le patient.

6.5.1 | Vision à court terme

A court terme, les problématiques de notre domaine de recherche portent sur l'extraction automatique des données chirurgicales. Cette extraction est actuellement menée par annotation manuelle, mais les performances très prometteuses des approches par apprentissage profond laissent envisager la résolution de tâches de plus en plus complexes. Par exemple, les résultats ne cessent de s'améliorer quand on considère des tâches telles que la reconnaissance procédurale des phases, étapes et activités ou le suivi et la segmentation d'objets et notamment d'instruments chirurgicaux à partir de vidéos laparoscopiques. Pour permettre une amélioration des performances de ces algorithmes, de tels jeux de données en accord avec les réglementations cliniques sont encore nécessaires bien qu'à l'avenir, ces algorithmes puissent de plus en plus se baser sur des données vidéos brutes non-annotées pour proposer directement le résultat attendu, et sans passer par une sémantique particulière. Mais il me semble que l'existence d'une telle sémantique, ou d'une forme d'interpétabilité devrait-être conservées si l'on veut être à même d'analyser et de comprendre la pratique chirurgicale et de faire le lien avec le clinique.

En évoluant, ces méthodes seront capables de traiter d'aspects de plus en plus complexes de la chirurgie. Cette amélioration laisse envisager de répondre de mieux en mieux à la contrainte de la plus-value clinique : en prenant mieux en compte la complexité de l'environnement chirurgical, ces travaux seront plus à même de répondre aux attentes des chirurgiens. Ces systèmes gagneront également en rapidité et des applications en peropératoire seront donc plus envisageables. A noter que certaines études sont limitées par leurs nécessité d'avoir accès à une information complète sur la procédure avant de fournir un résultat. Dans de tels cas, le temps-réel n'est pas envisageable, et nous noterons que notre approche n'est pas concernée par cette contrainte. Par ailleurs, il nous semble que face à la variabilité du patient et de la procédure chirurgicale, les questions de la fiabilité et de la robustesse de tels systèmes soient trop complexes pour être résolues rapidement.

6.5.2 | Vision à moyen terme

A moyen terme, il sera donc possible d'envisager des études porteuses d'une plus-value clinique de plus en plus importante. Ces études pourront accompagner le clinicien dans certaines tâches complexes liées à la formation, ou à la chirurgie elle-même. Dans mon état

de l'art 2.3, mon objectif était de montrer la cohérence des connaissances rassemblées dans différents travaux, mais aussi de montrer l'étendue des possibilités, dès lors qu'on applique des approches quantitatives sur des données peropératoires afin d'analyser la pratique chirurgicale. Les conclusions à portée clinique que nous avons extraits de ces différentes études ne proposent, à l'heure actuelle, que des confirmations sur des principes basiques de la pratique chirurgicale.

A travers notre étude 3 (voir chapitre 5), nous avons montré le potentiel d'une assistance au chirurgien dans le cadre de chirurgies réelles et de situations chirurgicales très spécifiques. En gagnant en maturité et en se reposant sur une annotation de données plus fournie et plus adaptée, on entrevoit l'aide que de telles approches pourraient donc apporter à la formation des chirurgiens. Cela est d'autant plus vrai qu'en consultant les livres de technique chirurgicale, on y lit des explications très générales sur les procédures per-opératoires, mais pas ou peu de descriptions détaillées sur les points critiques de ces chirurgies, sur les situations à éviter, ... Les internes en chirurgie sont donc confrontés à un manque réel d'informations sur la conduite à tenir dans des situations chirurgicales bien précises. Si ces analyses de situations chirurgicales très spécifiques se confirment et sont validées cliniquement, elles pourraient aussi être inclues dans des guides de bonnes pratiques chirurgicales ou des aides-mémoires pour le chirurgien.

D'autres usages possibles seraient de prendre en charge des tâches simples mais utiles dans le processus de formation des internes, telles qu'une aide à l'évaluation et au retour sur expérience en se basant sur des scores structurés (Wolf, 2013). Avec la structuration des jeux de données hospitalières, on peut imaginer des outils à l'usage des chirurgiens et de leurs internes leur permettant d'échanger sur des exemples de procédure. A travers l'analyse des trajectoires, du regard, des forces appliquées et de la procédure, il serait envisageable d'aider le chirurgien à orienter l'attention et l'apprentissage de ses internes sur certains aspects de la pratique [2].

En ce qui concerne l'assistance au chirurgien au cours de la chirurgie en elle-même, des outils simples dédis à des tâches précises et pertinentes sont déjà en cours de développement. Je pense notamment aux travaux de Twinanda et al. (2019) sur la prédiction du temps chirurgical restant qui répondent à une problématique logistique et organisationnelle très pratique de disponibilité du bloc opératoire. Cependant, dans l'optique de garantir la sécurité du patient, de tels outils devront avant tout être validés, et l'installation d'outils dans le bloc opératoire me semble être un niveau de complexité bien supérieur à l'installation d'outils dédiés à la formation.

Le passage du support de formation à la chirurgie réelle s'accompagne aussi d'un gros saut de complexité lié à la variabilité de l'environnement chirurgical, au patient et aux multiples acteurs impliqués. De plus on ne considère plus de jeunes chirurgiens en formation mais des chirurgiens expérimentés. Aussi, pour pouvoir les assister dans leur pratique, il faudra être capable de distinguer avec beaucoup plus de finesse les niveaux d'expertise, et de caractériser leurs différents profils de pratique (Forestier et al., 2018; Huaulmé et al., 2017). C'est ce niveau de subtilité qui doit être atteint si on veut soutenir le chirurgien expérimenté dans sa pratique quotidienne. Bref, c'est dans une vision à long terme qu'il me semble

2. https ://digitalsurgery.com/

raisonnable d'envisager des outils de suivi de la pratique du chirurgien en peropératoire.

6.5.3 | Vision à long terme

A long terme, l'aide à la formation des chirurgiens pourra prendre la forme d'une méta-analyse de la courbe d'apprentissage similaire à l'étude de Vedula et al. (2016). Il serait intéressant d'envisager des transformations de fonds de la formation liées à des études ayant analysé l'évolution des compétences de l'interne au cours de sa formation. Dans les années à venir, les questions de la formation continue des chirurgiens et de leur évaluation en routine clinique vont devenir de plus en plus actuelles, et avec le perfectionnement des outils pour la formation et l'évaluation des internes, on tendra de plus en plus vers des outils matures pour de telles applications. De telles transformations impliquent évidemment de relever les défis liés au milieu chirurgical, au contexte socio-professionnel, institutionnel et légal dont nous ne discuterons pas ici (voir section 1.4).

En ce qui concerne l'avènement d'outils dédiés à l'assistance du chirurgien en peropératoire, on peut résumer les conditions nécessaires à leur mise en place comme suit. Cela nécessitera l'extraction automatique des données issues de procédures chirurgicales dans des conditions de complexité réelles, en gérant notamment les artefacts associés à ces conditions réelles. Ces données devront permettre des analyses/évaluations suffisamment fines et spécifiques de la pratique du chirurgien expérimenté lors de procédures chirurgicales menées sur de vrais patients, ce qui implique de pouvoir également gérer la variabilité liée à ces patients. Ces analyses et évaluations pourront ensuite être proposées au chirurgien ou aux membres de l'équipe chirurgicale d'une façon adaptée, compréhensible, et sans surcharger l'environnement du bloc opératoire. Et tout ce processus devra être effectué en temps réel, en apportant un service médical réel au patient qu'il faudra avoir valider pour assurer son acceptabilité, sa robustesse et sa fiabilité. On constate donc que c'est bien dans une vision à long terme qu'on peut envisager de résoudre chacune de ces contraintes séparément, puis dans leur globalité.

A

Statistiques détaillées de l'annotation procédurale

Voici quelques observations du contenu de ces tableaux :

— tout d'abord, on notera que les activités ayant un verbe 'idle' ou une cible 'idle' sont des activités où rien ne se passe. Cet artefact d'annotation apparaissant de très nombreuses fois correspond en réalité aux transitions durant plus de 0.4s entre des activités réellement annotées.

— Sur l'ensemble des activités, chaque acteur utilise presque toujours le même instrument chirurgical. On n'observe que quelques inversions d'instruments entre les deux mains du chirurgien. Cette faible variabilité du nombre d'instruments utilisés s'explique par le fait que nous n'avons annoté qu'une seule étape de la procédure chirurgicale.

— Lorsqu'on s'intéresse aux activités des deux mains de l'assistant, et de la main gauche du chirurgien, on observe une forte corrélation entre l'instrument et le verbe. Cette forte prévisibilité du triplet d'activité traduit la spécificité du rôle et des activités de chaque acteur.

Instrument	Verb	Target	# of samples
Actor = Surgeon left-hand			
flat grasper	idle	idle	382
flat grasper	holding an object	stomach	332
flat grasper	pushing an object	stomach	26
electrothermal bipolar forceps	idle	idle	11
electrothermal bipolar forceps	sealing & dividing	lipome of his	7
flat grasper	holding an object	ligament	6
flat grasper	holding an object	lipome of his	6
electrothermal bipolar forceps	pushing an object	stomach	5
flat grasper	pushing an object	liver	3
flat grasper	pulling an object	compress	2
electrothermal bipolar forceps	pushing an object	compress	1
flat grasper	holding an object	adhesion	1
flat grasper	holding an object	compress	1
flat grasper	pushing an object	ligament	1

Tableau A.1 – Nombre d'apparition de chaque triplet d'activité par acteur sur l'ensemble du set de données *Lapex* (1/2).

Instrument	Verb	Target	# of samples
Actor = Surgeon right-hand			
electrothermal bipolar forceps	idle	idle	1571
electrothermal bipolar forceps	pushing an object	stomach	660
electrothermal bipolar forceps	sealing & dividing	ligament	450
electrothermal bipolar forceps	sealing & dividing	lipome of his	109
electrothermal bipolar forceps	pulling an object	compress	100
electrothermal bipolar forceps	pushing an object	ligament	99
electrothermal bipolar forceps	sealing & dividing	meso-gastro posterior	70
electrothermal bipolar forceps	sealing & dividing	adhesion	64
electrothermal bipolar forceps	pushing an object	compress	62
electrothermal bipolar forceps	coagulating a tissue	ligament	32
electrothermal bipolar forceps	holding an object	ligament	9
electrothermal bipolar forceps	pushing an object	lipome of his	4
flat grasper	idle	idle	4
electrothermal bipolar forceps	coagulating a tissue	meso-gastro posterior	3
electrothermal bipolar forceps	pushing an object	adhesion	3
electrothermal bipolar forceps	pushing an object	liver	3
electrothermal bipolar forceps	coagulating a tissue	adhesion	2
flat grasper	holding an object	stomach	2
electrothermal bipolar forceps	coagulating a tissue	Spleen	1
electrothermal bipolar forceps	holding an object	compress	1
electrothermal bipolar forceps	pulling an object	lipome of his	1
flat grasper	holding an object	ligament	1
flat grasper	pulling an object	compress	1
Actor = Assistant right-hand			
liver retractor	idle	idle	66
liver retractor	retracting an object from some place	liver	52
liver retractor	retracting an object from some place	stomach	1
Actor = Assistant left-hand			
flat grasper	idle	idle	23
flat grasper	holding an object	ligament	16
flat grasper	pushing an object	ligament	4

Tableau A.2 – Nombre d'apparition de chaque triplet d'activité par acteur sur l'ensemble du set de données *Lapex* (2/2).

B

Description technique des données LapEx

Cette description technique a été associée à la base de données mise à disposition au sein d'un fichier "README.txt".

```
============================
LapEx: Laparoscopic Exposure
============================

This folder contains:
- LapEx dataset:
    - 30 procedures of Sleeve Gastrectomy, and for each procedure:
        - "frames" directory
        - "seg" directory
        : activities.csv
        : exposures.csv file
    - metadata
        : instrument.csv
        : segmented_objet.csv
        : target.csv
        : verb.csv
    : README.txt

===================
Dataset Description
===================

For more detailed description please refer to:
- The publication:
???
- The challenge website:
???
- Any questions regarding the dataset can be sent to:
```

???

==========
"frame" directory:

The laparoscopic video covers the surgical step of the "Dissection of Fundus" and was recorded with a frequence f=25Hz. The content of the laparoscopic video is proposed as a sequence of images (JPG files) in RGB format. Each image file is a frame named by its timestamp in ms. Frames showing views outside of the abdominal cavity were replaced by black frames.

==========
"activities.csv" file:

This file contains the activities annotated for all surgical actors along the "Dissection of Fundus" step.

File shape:

```
       Actor ; Start timestamp ; End timestamp ; Verb ; Instrument ; Target
Act 1    ...
Act 2    ...
...
Act N    ...
```

Each annotated activity is a line of CSV file which is described by its actor, its start and end timestamps, and by its triplet <verb;instrument;target>. Timestamps are given in ms and are synchronized with the timestamps of the JPG files in directory "frames". Label "idle" is used to describe the non-active state of the actor between two effectively annotated activities. This label is used for verb and target.

==========
"exposures.csv" file:

This file contains the quality of exposure annotated at different points along the procedure. Each annotation was evaluated as an activity with verb "sealing & dividing" or "coagulating a tissue" was annotated.

File shape:

```
            Timestamp ; level of quality
Exposure 1    ...
Exposure 2    ...
...
Exposure M    ...
```

Timestamps are given in ms and are synchronized with the timestamps of the JPG files in directory "frames".

Level of quality takes value in:
- 1: good quality
- 2: acceptable quality
- 3: non-acceptable quality

==========
"seg" directory:

This directory contains the results of the pixel-wise segmentation performed on specific frames of the video. Each segmented frame corresponds to an annotated quality of exposure, so we have as much segmented images as annotated qualities of exposure. Each segmentation is a JPG file in level of gray named with its timestamps in ms followed by '_seg'. Each pixel is labelled with a given gray-value.

==========
"metadata" directory:

This directory contains metadata about the procedural annotation and the segmentation.

"instrument.csv", "verb.csv" and "target.csv" contain the exhaustive list of labels used in the triplet <verb;instrument;target> of the "activities.csv" files for each procedure.

"segmented_objet.csv" contains the exhaustive list of the segmented objects with their associated level of gray along the whole dataset.

==========
License and References
==========

???

==========
Acknowledgement
==========

???

C

Description des longs schémas d'activités procédurales les plus fréquents

#	cible 1	cible 2	cible 3	cible 4	cible 5
1	ligament	ligament	ligament	ligament	/
2	ligament	ligament	ligament	stomach	ligament
3	ligament	ligament	stomach	ligament	stomach
4	ligament	stomach	ligament	ligament	ligament
5	ligament	stomach	ligament	ligament	stomach
6	ligament	stomach	ligament	stomach	ligament
7	ligament	stomach	ligament	stomach	stomach
8	stomach	ligament	ligament	stomach	ligament
9	stomach	ligament	stomach	ligament	ligament
10	stomach	ligament	stomach	ligament	stomach
11	stomach	ligament	stomach	stomach	ligament
12	stomach	stomach	ligament	stomach	ligament
13	stomach	lipome of his	stomach	lipome of his	stomach
14	stomach	adhesion	stomach	adhesion	stomach
15	stomach	meso-gastro posterior	stomach	meso-gastro posterior	stomach
16	lipome of his	stomach	lipome of his	stomach	lipome of his

Tableau C.1 – Liste des longs schémas de cibles les plus fréquents sur le jeu de données Lapex

#	verbe	instrument	cible
1	sealing & dividing	electrothermal bipolar forceps	ligament
	sealing & dividing	electrothermal bipolar forceps	ligament
2	sealing & dividing	electrothermal bipolar forceps	ligament
	pushing an object	electrothermal bipolar forceps	stomach
	sealing & dividing	electrothermal bipolar forceps	ligament
3	sealing & dividing	electrothermal bipolar forceps	ligament
	pushing an object	electrothermal bipolar forceps	ligament
	sealing & dividing	electrothermal bipolar forceps	ligament
4	sealing & dividing	electrothermal bipolar forceps	ligament
	pushing an object	electrothermal bipolar forceps	stomach
	pulling an object	electrothermal bipolar forceps	compress
5	sealing & dividing	electrothermal bipolar forceps	ligament
	pushing an object	electrothermal bipolar forceps	stomach
	pushing an object	electrothermal bipolar forceps	stomach
6	sealing & dividing	electrothermal bipolar forceps	ligament
	pushing an object	electrothermal bipolar forceps	stomach
	pushing an object	electrothermal bipolar forceps	stomach
	sealing & dividing	electrothermal bipolar forceps	ligament
7	sealing & dividing	electrothermal bipolar forceps	ligament
	pushing an object	electrothermal bipolar forceps	stomach
	pulling an object	electrothermal bipolar forceps	compress
	pushing an object	electrothermal bipolar forceps	stomach
8	pushing an object	electrothermal bipolar forceps	stomach
	pushing an object	electrothermal bipolar forceps	stomach
	sealing & dividing	electrothermal bipolar forceps	ligament
9	sealing & dividing	electrothermal bipolar forceps	lipome of his
	pushing an object	electrothermal bipolar forceps	stomach
	pushing an object	electrothermal bipolar forceps	lipome of his
10	sealing & dividing	electrothermal bipolar forceps	adhesion
	pushing an object	electrothermal bipolar forceps	stomach
	pushing an object	electrothermal bipolar forceps	adhesion
11	sealing & dividing	electrothermal bipolar forceps	meso-gastro posterior
	pushing an object	electrothermal bipolar forceps	stomach
	pushing an object	electrothermal bipolar forceps	meso-gastro posterior

Tableau C.2 – Liste des longs schémas d'activités les plus fréquents sur le jeu de données Lapex, les labels d'activités respectent le formalisme du triplet <verbe, instrument, cible>

#	verbe 1	verbe 2	verbe 3	verbe 4
1	sealing & dividing	sealing & dividing	/	/
2	sealing & dividing	pushing an object	sealing & dividing	/
3	sealing & dividing	pushing an object	pushing an object	/
4	sealing & dividing	pushing an object	pushing an object	sealing & dividing
5	sealing & dividing	pushing an object	pushing an object	pushing an object
6	pushing an object	pushing an object	pushing an object	sealing & dividing
7	pushing an object	pushing an object	pushing an object	pushing an object
8	sealing & dividing	pushing an object	pulling an object	/
9	pushing an object	pulling an object	pushing an object	pushing an object
10	pushing an object	pulling an object	pushing an object	pulling an object

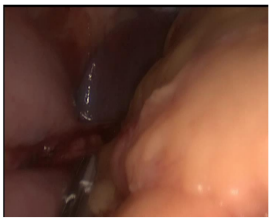
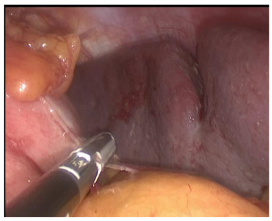
Tableau C.3 – Liste des longs schémas de verbes les plus fréquents sur le jeu de données Lapex

D

Questionnaire de validation clinique

Enoncé #1 / 11

Afin d'optimiser l'exposition dans la scène chirurgicale, il est très important de bien gérer la rate.

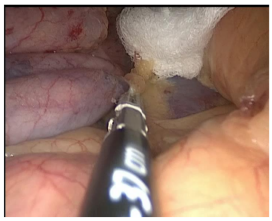



A compléter :
Vous êtes plutôt pas d'accord

Slide Suivante

Enoncé #2 / 11

Ici, pour bien gérer la qualité d'exposition, le chirurgien fait attention à sa gestion de la rate.

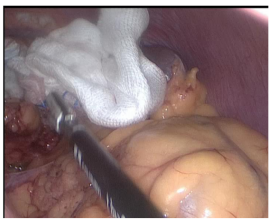
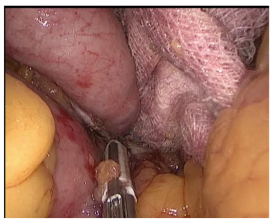



A compléter :
Vous êtes sélectionner un élément

Slide Suivante

Enoncé #3 / 11

Ici, pour bien gérer la qualité d'exposition, le chirurgien fait attention à la manipulation du flat grasper.

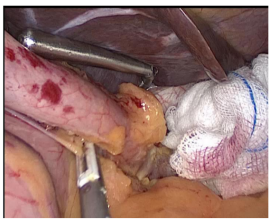
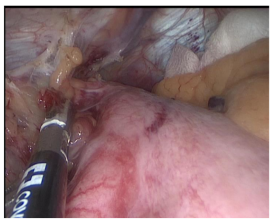



A compléter :
Vous êtes sélectionner un élément

Slide Suivante

Enoncé #4 / 11

La différence de pratique entre les 2 chirurgiens de l'étude est surtout visible dans la façon de gérer l'estomac.

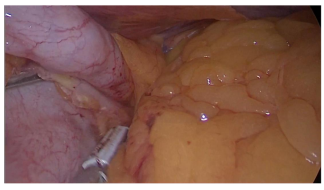
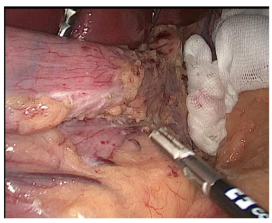



A compléter :
Vous êtes sélectionner un élément

Slide Suivante

Enoncé #5 / 11

Dans les cas où la gestion de l'exposition est bonne, la différence de pratique entre les 2 chirurgiens de l'étude se voit surtout dans la gestion du foie.

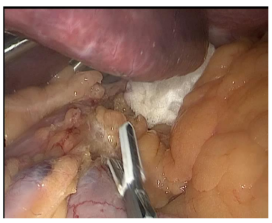
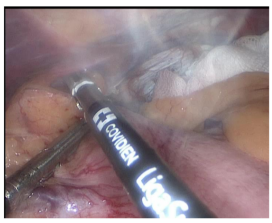



A compléter :
Vous êtes sélectionner un élément

Slide Suivante

Enoncé #6 / 11

Dans les cas où la gestion de l'exposition est plus compliquée, la différence de pratique entre les 2 chirurgiens de l'étude se voit surtout dans la gestion de la paroi abdominale.

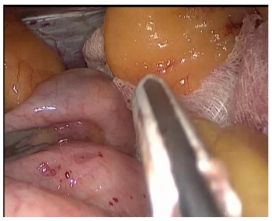
A compléter :
Vous êtes sélectionner un élément

Slide Suivante

Tableau D.1 – Questionnaire de validation clinique sans les extraits vidéos - diapositives 1 à 6

Enoncé #7 / 11

Prêter attention à la texture de la rate aide beaucoup à évaluer la qualité d'exposition.




A compléter :
Vous êtes sélectionner un élément

Slide Suivante

Enoncé #8 / 11

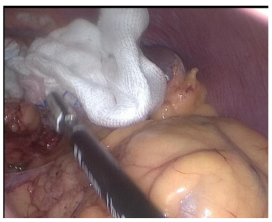
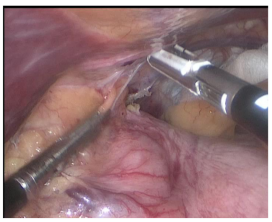
Ici, prêter attention à la surface de la rate dans l'image permet au chirurgien de bien gérer la qualité d'exposition.


A compléter :
Vous êtes sélectionner un élément

Slide Suivante

Enoncé #9 / 11

Ici, prêter attention à la présence de couleur jaune ou bleu sur le flat grasper permet au chirurgien de bien gérer la qualité d'exposition.

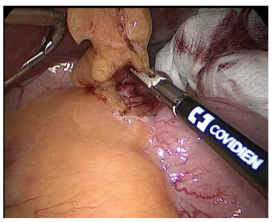



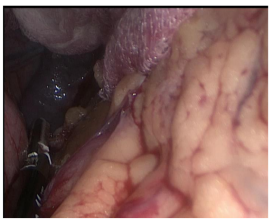
A compléter :
Vous êtes sélectionner un élément

Slide Suivante

Enoncé #10 / 11

Ici, prêter attention à la présence de couleur verte ou rouge sur l'estomac permet bien de distinguer la différence de pratique entre les 2 chirurgiens.



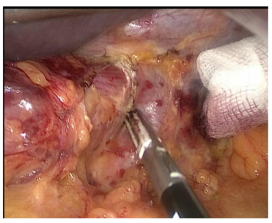


A compléter :
Vous êtes sélectionner un élément

Slide Suivante

Enoncé #11 / 11

Dans les cas où la gestion de l'exposition est bonne, la différence de pratique entre les 2 chirurgiens s'observe surtout dans la forme et la place prise par le foie dans l'image.

A compléter :
Vous êtes sélectionner un élément

Slide Suivante

Tableau D.2 – Questionnaire de validation clinique sans les extraits vidéos - diapositives 7 à 11

Bibliographie

[1] Monty A. Aghazadeh, Isuru S. Jayaratna, Andrew J. Hung, Michael M. Pan, Mihir M. Desai, Inderbir S. Gill, and Alvin C. Goh. External validation of Global Evaluative Assessment of Robotic Skills (GEARS). *Surgical Endoscopy*, 29(11) :3261–3266, November 2015. ISSN 1432-2218. doi : 10.1007/s00464-015-4070-8.

[2] Nava Aghdasi, Randall Bly, Lee W. White, Blake Hannaford, Kris Moe, and Thomas S. Lendvay. Crowd-sourced assessment of surgical skills in cricothyrotomy procedure. *Journal of Surgical Research*, 196(2) : 302–306, June 2015. ISSN 00224804. doi : 10.1016/j.jss.2015.03.018.

[3] Narges Ahmidi, Gregory D. Hager, Lisa Ishii, Gabor Fichtinger, Gary L. Gallia, and Masaru Ishii. Surgical Task and Skill Classification from Eye Tracking and Tool Motion in Minimally Invasive Surgery. In Tianzi Jiang, Nassir Navab, Josien P. W. Pluim, and Max A. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, Lecture Notes in Computer Science, pages 295–302. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15711-0.

[4] Narges Ahmidi, Masaru Ishii, Gabor Fichtinger, Gary L. Gallia, and Gregory D. Hager. An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data. *International Forum of Allergy & Rhinology*, 2(6) :507–515, 2012. ISSN 2042-6984. doi : 10.1002/alr.21053.

[5] Narges Ahmidi, Piyush Poddar, Jonathan D. Jones, S. Swaroop Vedula, Lisa Ishii, Gregory D. Hager, and Masaru Ishii. Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. *International Journal of Computer Assisted Radiology and Surgery*, 10(6) :981–991, June 2015. ISSN 1861-6429. doi : 10.1007/s11548-015-1194-1.

[6] Siti Nor Zawani Ahmmad, Eileen Su Lee Ming, Yeong Che Fai, Suneet Sood, and Anil Gandhi. Objective Measurement for Surgical Skill Evaluation. *Jurnal Teknologi*, 78(7-5), July 2016. ISSN 2180-3722. doi : 10.11113/jt.v78.9461.

[7] Hassan Al Hajj, Mathieu Lamard, Pierre-Henri Conze, Soumali Roychowdhury, Xiaowei Hu, Gabija Maršalkaitė, Odysseas Zisimopoulos, Muneer Ahmad Dedmari, Fenqiang Zhao, Jonas Prellberg, Manish Sahu, Adrian Galdran, Teresa Araújo, Duc My Vo, Chandan Panda, Navdeep Dahiya, Satoshi Kondo, Zhengbing Bian, Arash Vahdat, Jonas Bialopetravičius, Evangello Flouty, Chenhui Qiu, Sabrina Dill, Anirban Mukhopadhyay, Pedro Costa, Guilherme Aresta, Senthil Ramamurthy, Sang-Woong Lee, Aurélio Campilho, Stefan Zachow, Shunren Xia, Sailesh Conjeti, Danail Stoyanov, Jogundas Armaitis, Pheng-Ann Heng, William G. Macready, Béatrice Cochener, and Gwenolé Quellec. CATARACTS : Challenge on automatic tool annotation for cataRACT surgery. *Medical Image Analysis*, 52 :24–41, February 2019. ISSN 1361-8415. doi : 10.1016/j.media.2018.11.008.

[8] Joseph Antonelli, Brian L. Claggett, Mir Henglin, Andy Kim, Gavin Ovsak, Nicole Kim, Katherine Deng, Kevin Rao, Octavia Tyagi, Jeramie D. Watrous, Kim A. Lagerborg, Pavel V. Hushcha, Olga V. Demler, Samia Mora, Teemu J. Niiranen, Alexandre C. Pereira, Mohit Jain, and Susan Cheng. Statistical Workflow

for Feature Selection in Human Metabolomics Data. *Metabolites*, 9(7), July 2019. ISSN 2218-1989. doi : 10.3390/metabo9070143.

[9] David P. Azari, Lane L. Frasier, Sudha R. Pavuluri Quamme, Caprice C. Greenberg, Carla M. Pugh, Jacob A. Greenberg, and Robert G. Radwin. Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. *Annals of Surgery*, 269(3) :574–581, March 2019. ISSN 1528-1140. doi : 10.1097/SLA.0000000000002478.

[10] James Bergstra and Yoshua Bengio. Random Search for Hyper-Parameter Optimization. page 25, 2012.

[11] David Bouget, Max Allan, Danail Stoyanov, and Pierre Jannin. Vision-based and marker-less surgical tool detection and tracking : a review of the literature. *Medical Image Analysis*, 35 :633–654, January 2017. ISSN 1361-8415. doi : 10.1016/j.media.2016.09.003.

[12] The COlon cancer Laparoscopic or Open Resection Study Group. Laparoscopic surgery versus open surgery for colon cancer : short-term outcomes of a randomised trial. *The Lancet Oncology*, 6(7) :477–484, July 2005. ISSN 1470-2045. doi : 10.1016/S1470-2045(05)70221-7.

[13] Gavin C. Cawley and Nicola L. C. Talbot. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, 11(Jul) :2079–2107, 2010. ISSN ISSN 1533-7928.

[14] Carolyn Chen, Lee White, Timothy Kowalewski, Rajesh Aggarwal, Chris Lintott, Bryan Comstock, Katie Kuksenok, Cecilia Aragon, Daniel Holst, and Thomas Lendvay. Crowd-Sourced Assessment of Technical Skills : a novel method to evaluate surgical performance. *Journal of Surgical Research*, 187(1) :65–71, March 2014. ISSN 0022-4804, 1095-8673. doi : 10.1016/j.jss.2013.09.024.

[15] Richard I. Cook and David D. Woods. Operating at the Sharp End : The Complexity of Human Error. In Marilyn Sue Bogner, editor, *Human Error in Medicine*, pages 255–310. CRC Press, 1 edition, 1994. ISBN 978-0-203-75172-5. doi : 10.1201/9780203751725-13. URL `https://www.taylorfrancis.com/books/9781351440219/chapters/10.1201/9780203751725-13`.

[16] Shanley B. Deal, Thomas S. Lendvay, Mohamad I. Haque, Timothy Brand, Bryan Comstock, Justin Warren, and Adnan Alseidi. Crowd-sourced assessment of technical skills : an opportunity for improvement in the assessment of laparoscopic surgical skills. *The American Journal of Surgery*, 211(2) :398–404, February 2016. ISSN 0002-9610, 1879-1883. doi : 10.1016/j.amjsurg.2015.09.005.

[17] Shanley B. Deal, Dimitrios Stefanidis, Dana Telem, Robert D. Fanelli, Marian McDonald, Michael Ujiki, L. Michael Brunt, and Adnan A. Alseidi. Evaluation of crowd-sourced assessment of the critical view of safety in laparoscopic cholecystectomy. *Surgical Endoscopy*, 31(12) :5094–5100, December 2017. ISSN 1432-2218. doi : 10.1007/s00464-017-5574-1.

[18] Tarek Y. El Ahmadieh, James Harrop, H. Hunt Batjer, Daniel K. Resnick, and Bernard R. Bendok. Blinded peer assessment of surgical skill is feasible and can predict complication rates : a step toward measuring surgical quality. *Neurosurgery*, 74(6) :N12–14, June 2014. ISSN 1524-4040. doi : 10.1227/01.neu.0000450232.06740.ef.

[19] Pouya Entezami, Lauren E. Franzblau, and Kevin C. Chung. Mentorship in Surgical Training : A Systematic Review. *HAND*, 7(1) :30–36, March 2012. ISSN 1558-9447. doi : 10.1007/s11552-011-9379-8.

[20] M. Ershad, R. Rege, and Ann Majewicz Fey. Automatic and near real-time stylistic behavior assessment in robotic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 14(4) :635–643, April 2019. ISSN 1861-6429. doi : 10.1007/s11548-019-01920-6.

[21] Marzieh Ershad, Zachary Koesters, Robert Rege, and Ann Majewicz. Meaningful Assessment of Surgical Expertise : Semantic Labeling with Data and Crowds. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Lecture Notes in Computer Science, pages 508–515. Springer International Publishing, 2016. ISBN 978-3-319-46720-7.

[22] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2) :303–338, June 2010. ISSN 1573-1405. doi : 10.1007/s11263-009-0275-4.

[23] O. Faiz, J. Warusavitarne, A. Bottle, P. P. Tekkis, A. W. Darzi, and R. H. Kennedy. Laparoscopically Assisted vs. Open Elective Colonic and Rectal Resection : A Comparison of Outcomes in English National Health Service Trusts Between 1996 and 2006. *Diseases of the Colon & Rectum*, 52(10) :1695, October 2009. ISSN 0012-3706. doi : 10.1007/DCR.0b013e3181b55254.

[24] Mahtab J. Fard, Sattar Ameri, R. Darin Ellis, Ratna B. Chinnam, Abhilash K. Pandya, and Michael D. Klein. Automated robot-assisted surgical skill evaluation : Predictive analytics approach. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 14(1) :e1850, 2018. ISSN 1478-596X. doi : 10.1002/rcs.1850.

[25] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Evaluating surgical skills from kinematic data using convolutional neural networks. *arXiv :1806.02750 [cs]*, 11073 :214–221, 2018. arXiv : 1806.02750.

[26] Evangello Flouty, Abdolrahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes-Hurtado, Hinde Taleb, Santiago Barbarisi, Gwenole Quellec, and Danail Stoyanov. CaDIS : Cataract Dataset for Image Segmentation. *arXiv :1906.11586 [cs]*, July 2019. arXiv : 1906.11586.

[27] Germain Forestier, François Petitjean, Pavel Senin, Fabien Despinoy, Arnaud Huaulmé, Hassan Ismail Fawaz, Jonathan Weber, Lhassane Idoumghar, Pierre-Alain Muller, and Pierre Jannin. Surgical motion analysis using discriminative interpretable patterns. *Artificial Intelligence in Medicine*, 91 :3–11, September 2018. ISSN 0933-3657. doi : 10.1016/j.artmed.2018.08.002.

[28] Céline Fouard, Aurélien Deram, Yannick Keraval, and Emmanuel Promayon. CamiTK : A Modular Framework Integrating Visualization, Image Processing and Biomechanical Modeling. In Yohan Payan, editor, *Soft Tissue Biomechanical Modeling for Computer Assisted Surgery*, Studies in Mechanobiology, Tissue Engineering and Biomaterials, pages 323–354. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-29014-5.

[29] N. K. Francis, N. J. Curtis, J. A. Conti, J. D. Foster, H. J. Bonjer, G. B. Hanna, M. Abu-Hilal, F. Agresta, S. A. Antoniu, A. Arezzo, C. Balagúe, L. Boni, N. Bouvy, T. Carus, B. Edwin, M. Diana, G. Faria, D. Ignjatovic, N. de Manzini, F. M. Margallo, L. Martinek, N. Matveev, Y. Mintz, K. Nakajima, D. E. Popa, P. J. Schijven, P. Sedman, E. Yiannakopoulou, and on behalf of the EAES committees. EAES classification of intraoperative adverse events in laparoscopic surgery. *Surgical Endoscopy*, 32(9) :3822–3829, September 2018. ISSN 1432-2218. doi : 10.1007/s00464-018-6108-1.

[30] Eibe Frank and Ian H. Witten. Using a permutation test for attribute selection in decision trees. pages 152–160, 1998.

[31] Isabel Funke, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. Video-based surgical skill assessment using 3D convolutional neural networks. *arXiv :1903.02306 [cs]*, March 2019. arXiv : 1903.02306.

[32] Michel Gagner, Colleen Hutchinson, and Raul Rosenthal. Fifth International Consensus Conference : current status of sleeve gastrectomy. *Surgery for Obesity and Related Diseases*, 12(4) :750–756, May 2016. ISSN 1550-7289. doi : 10.1016/j.soard.2016.01.022.

[33] Sandeep Ganni, Sanne M. B. I. Botden, Magdalena Chmarra, Richard H. M. Goossens, and Jack J. Jakimowicz. A software-based tool for video motion tracking in the surgical skills assessment landscape. *Surgical Endoscopy*, 32(6) :2994–2999, June 2018. ISSN 1432-2218. doi : 10.1007/s00464-018-6023-5.

[34] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Bejar, David D Yuh, Chi Chiung Grace Chen, Rene Vidal, Sanjeev Khudanpur, and Gregory D Hager. JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) : A Surgical Activity Dataset for Human Motion Modeling. page 10, 2014.

[35] Khurshid R. Ghani, David C. Miller, Susan Linsell, Andrew Brachulis, Brian Lane, Richard Sarle, Deepansh Dalela, Mani Menon, Bryan Comstock, Thomas S. Lendvay, James Montie, James O. Peabody, and Michigan Urological Surgery Improvement Collaborative. Measuring to Improve : Peer and Crowd-sourced Assessments of Technical Skill with Robot-assisted Radical Prostatectomy. *European Urology*, 69 (4) :547–550, April 2016. ISSN 1873-7560. doi : 10.1016/j.eururo.2015.11.028.

[36] Goh Alvin C, Goldfarb David W., Sander James C., Miles Brian J., and Dunkin Brian J. Global Evaluative Assessment of Robotic Skills : Validation of a Clinical Assessment Tool to Measure Robotic Surgical Skills. *Journal of Urology*, 187(1) :247–252, January 2012. doi : 10.1016/j.juro.2011.09.032.

[37] Ernest D. Gomez, Rajesh Aggarwal, William McMahan, Karlin Bark, and Katherine J. Kuchenbecker. Objective assessment of robotic surgical skill using instrument contact vibrations. *Surgical Endoscopy*, 30(4) : 1419–1431, April 2016. ISSN 1432-2218. doi : 10.1007/s00464-015-4346-z.

[38] Richard J. Gray, Kanav Kahol, Gazi Islam, Marshall Smith, Alyssa Chapital, and John Ferrara. High-Fidelity, Low-Cost, Automated Method to Assess Laparoscopic Skills Objectively. *Journal of Surgical Education*, 69(3) :335–339, May 2012. ISSN 1931-7204. doi : 10.1016/j.jsurg.2011.10.014.

[39] A. Harvey, J. N. Vickers, R. Snelgrove, M. F. Scott, and S. Morrison. Expert surgeon's quiet eye and slowing down : expertise differences in performance and quiet eye duration during identification and dissection of the recurrent laryngeal nerve. *American journal of surgery*, 207(2) :187–193, February 2014. ISSN 0002-9610. doi : 10.1016/j.amjsurg.2013.07.033.

[40] Rose Hatala, David A. Cook, Ryan Brydges, and Richard Hawkins. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS) : a systematic review of validity evidence. *Advances in Health Sciences Education*, 20(5) :1149–1175, December 2015. ISSN 1573-1677. doi : 10.1007/s10459-015-9593-1.

[41] Traci Hedrick, Florence Turrentine, Hilary Sanfey, Bruce Schirmer, and Charles Friel. Implications of laparoscopy on surgery residency training. *The American Journal of Surgery*, 197(1) :73–75, January 2009. ISSN 0002-9610. doi : 10.1016/j.amjsurg.2008.08.013.

[42] Daniel Holst, Timothy M. Kowalewski, Lee W. White, Timothy C. Brand, Jonathan D. Harper, Mathew D. Sorenson, Sarah Kirsch, and Thomas S. Lendvay. Crowd-Sourced Assessment of Technical Skills : An Adjunct to Urology Resident Surgical Simulation Training. *Journal of Endourology*, 29(5) :604–609, October 2014. ISSN 0892-7790. doi : 10.1089/end.2014.0616.

[43] Daniel Holst, Timothy M. Kowalewski, Lee W. White, Timothy C. Brand, Jonathan D. Harper, Mathew D. Sorensen, Mireille Truong, Khara Simpson, Alyssa Tanaka, Roger Smith, and Thomas S. Lendvay. Crowd-Sourced Assessment of Technical Skills : Differentiating Animate Surgical Skill Through the Wisdom of Crowds. *Journal of Endourology*, 29(10) :1183–1188, October 2015. ISSN 1557-900X. doi : 10.1089/end.2015.0104.

[44] T. Horeman, S. P. Rodrigues, F. Willem Jansen, J. Dankelman, and J. J. van den Dobbelsteen. Force Parameters for Skills Assessment in Laparoscopy. *IEEE Transactions on Haptics*, 5(4) :312–322, 2012. ISSN 1939-1412. doi : 10.1109/TOH.2011.60.

[45] T. Horeman, J. Dankelman, F. W. Jansen, and J. J. van den Dobbelsteen. Assessment of Laparoscopic Skills Based on Force and Motion Parameters. *IEEE Transactions on Biomedical Engineering*, 61(3) :805–813, March 2014. ISSN 0018-9294. doi : 10.1109/TBME.2013.2290052.

[46] Tim Horeman, Sharon P. Rodrigues, Frank-Willem Jansen, Jenny Dankelman, and John J. van den Dobbelsteen. Force measurement platform for training and assessment of laparoscopic skills. *Surgical Endoscopy*, 24(12) :3102–3108, December 2010. ISSN 1432-2218. doi : 10.1007/s00464-010-1096-9.

[47] Mohammad Hossin and Sulaiman M.N. A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5 :01–11, March 2015. doi : 10.5121/ijdkp.2015.5201.

[48] Arnaud Huaulmé, Sandrine Voros, Laurent Riffaud, Germain Forestier, Alexandre Moreau-Gaudry, and Pierre Jannin. Distinguishing surgical behavior by sequential pattern discovery. *Journal of Biomedical Informatics*, 67 :34–41, March 2017. ISSN 1532-0464. doi : 10.1016/j.jbi.2017.02.001.

[49] Arnaud Huaulmé, Kanako Harada, Germain Forestier, Mamoru Mitsuishi, and Pierre Jannin. Sequential surgical signatures in micro-suturing task. *International Journal of Computer Assisted Radiology and Surgery*, 13(9) :1419–1428, September 2018. ISSN 1861-6429. doi : 10.1007/s11548-018-1775-x.

[50] Arnaud Huaulmé, Fabien Despinoy, Saul Alexis Heredia Perez, Kanako Harada, Mamoru Mitsuishi, and Pierre Jannin. Automatic annotation of surgical activities using virtual reality environments. *International Journal of Computer Assisted Radiology and Surgery*, June 2019. ISSN 1861-6429. doi : 10.1007/s11548-019-02008-x.

[51] Angelo Iossa, Mohamed Abdelgawad, Brad Michael Watkins, and Gianfranco Silecchia. Leaks after laparoscopic sleeve gastrectomy : overview of pathogenesis and risk factors. *Langenbeck's Archives of Surgery*, 401(6) :757–766, September 2016. ISSN 1435-2451. doi : 10.1007/s00423-016-1464-6.

[52] Xianta Jiang, Bin Zheng, Geoffrey Tien, and Margaret Atkins. Pupil response to precision in surgical task execution. *Studies in health technology and informatics*, 184 :210–4, February 2013. doi : 10.3233/978-1-61499-209-7-210.

[53] Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. *arXiv :1802.08774 [cs]*, February 2018. arXiv : 1802.08774.

[54] P Joice, G.B Hanna, and A Cuschieri. Errors enacted during endoscopic surgery - a human reliability analysis. *Applied Ergonomics*, 29(6) :409–414, December 1998. ISSN 00036870. doi : 10.1016/S0003-6870(98)00016-7.

[55] Darko Katić, Chantal Julliard, Anna-Laura Wekerle, Hannes Kenngott, Beat Peter Müller-Stich, Rüdiger Dillmann, Stefanie Speidel, Pierre Jannin, and Bernard Gibaud. LapOntoSPM : an ontology for laparoscopic surgeries and its application to surgical phase recognition. *International Journal of Computer Assisted Radiology and Surgery*, 10(9) :1427–1434, September 2015. ISSN 1861-6429. doi : 10.1007/s11548-015-1222-1.

[56] Aftab Khan, Sebastian Mellor, Eugen Berlin, Robin Thompson, Roisin McNaney, Patrick Olivier, and Thomas Plötz. Beyond Activity Recognition : Skill Assessment from Accelerometer Data. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1155–1166, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3574-4. doi : 10.1145/2750858.2807534. URL `http://doi.acm.org/10.1145/2750858.2807534`. event-place : Osaka, Japan.

[57] Georgina S. J. Kirby, Paul Guyver, Louise Strickland, Abtin Alvand, Guang-Zhong Yang, Caroline Hargrove, Benny P. L. Lo, and Jonathan L. Rees. Assessing Arthroscopic Skills Using Wireless Elbow-Worn Motion Sensors. *JBJS*, 97(13) :1119, July 2015. ISSN 0021-9355. doi : 10.2106/JBJS.N.01043.

[58] Timothy M. Kowalewski, Bryan Comstock, Robert Sweet, Cory Schaffhausen, Ashleigh Menhadji, Timothy Averch, Geoffrey Box, Timothy Brand, Michael Ferrandino, Jihad Kaouk, Bodo Knudsen, Jaime Landman, Benjamin Lee, Bradley F. Schwartz, Elspeth McDougall, and Thomas S. Lendvay. Crowd-Sourced Assessment of Technical Skills for Validation of Basic Laparoscopic Urologic Skills Tasks. *Journal of Urology*, 195(6) :1859–1865, June 2016. ISSN 0022-5347, 1527-3792. doi : 10.1016/j.juro.2016.01.005.

[59] Rajesh Kumar, Amod Jog, Anand Malpani, Balazs Vagvolgyi, David Yuh, Hiep Nguyen, Gregory Hager, and Chi Chiung Grace Chen. Assessing system operation skills in robotic surgery trainees. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 8(1) :118–124, 2012. ISSN 1478-596X. doi : 10.1002/rcs.449.

[60] Florent Lalys and Pierre Jannin. Surgical process modelling : a review. *International Journal of Computer Assisted Radiology and Surgery*, 9(3) :495–511, May 2014. ISSN 1861-6429. doi : 10.1007/s11548-013-0940-5.

[61] Z. Lin, M. Uemura, M. Zecca, S. Sessa, H. Ishii, M. Tomikawa, M. Hashizume, and A. Takanishi. Objective Skill Evaluation for Laparoscopic Training Based on Motion Analysis. *IEEE Transactions on Biomedical Engineering*, 60(4) :977–985, April 2013. ISSN 0018-9294. doi : 10.1109/TBME.2012.2230260.

[62] C. Loukas and E. Georgiou. Multivariate Autoregressive Modeling of Hand Kinematics for Laparoscopic Skills Assessment of Surgical Trainees. *IEEE Transactions on Biomedical Engineering*, 58(11) :3289–3297, November 2011. ISSN 0018-9294. doi : 10.1109/TBME.2011.2167324.

[63] Constantinos Loukas. Video content analysis of surgical procedures. *Surgical Endoscopy*, 32(2) :553–568, February 2018. ISSN 1432-2218. doi : 10.1007/s00464-017-5878-1.

[64] Constantinos Loukas and Evangelos Georgiou. Performance comparison of various feature detector-descriptors and temporal models for video-based assessment of laparoscopic skills. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 12(3) :387–398, 2016. ISSN 1478-596X. doi : 10.1002/rcs.1702.

[65] Constantinos Loukas, Constantinos Rouseas, and Evangelos Georgiou. The role of hand motion connectivity in the performance of laparoscopic procedures on a virtual reality simulator. *Medical & Biological Engineering & Computing*, 51(8) :911–922, August 2013. ISSN 1741-0444. doi : 10.1007/s11517-013-1063-4.

[66] Jay M. MacGregor and Robert Sticca. General Surgery Residents' Views on Work Hours Regulations. *Journal of Surgical Education*, 67(6) :376–380, November 2010. ISSN 1931-7204. doi : 10.1016/j.jsurg.2010.07.008.

[67] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, Carolin Feldmann, Alejandro F. Frangi, Peter M. Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A. Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H. Menze, Henning Müller, Peter F. Neher, Wiro Niessen, Nasir Rajpoot, Gregory C. Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel Aziz Taha, Fons van der Sommen, Ching-Wei Wang, Marc-André Weber, Guoyan Zheng, Pierre Jannin, and Annette Kopp-Schneider. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1) :1–13, December 2018. ISSN 2041-1723. doi : 10.1038/s41467-018-07619-7.

[68] Anand Malpani, S. Swaroop Vedula, Chi Chiung Grace Chen, and Gregory D. Hager. A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. *International Journal of Computer Assisted Radiology and Surgery*, 10(9) :1435–1447, September 2015. ISSN 1861-6429. doi : 10.1007/s11548-015-1238-6.

[69] Picard Marceau, Simon Biron, Roch-André Bourque, Martin Potvin, Frédéric-Simon Hould, and Serge Simard. Biliopancreatic Diversion with a New Type of Gastrectomy. *Obesity Surgery*, 3(1) :29–35, February 1993. ISSN 1708-0428. doi : 10.1381/096089293765559728.

[70] J. A. Martin, G. Regehr, R. Reznick, H. Macrae, J. Murnaghan, C. Hutchison, and M. Brown. Objective structured assessment of technical skill (OSATS) for surgical residents. *BJS*, 84(2) :273–278, 1997. ISSN 1365-2168. doi : 10.1046/j.1365-2168.1997.02502.x.

[71] Tadashi Matsuda, Hiroomi Kanayama, Yoshinari Ono, Akihiro Kawauchi, Hiroaki Mizoguchi, Ken Nakagawa, Masatsugu Iwamura, Masanobu Shigeta, Tomonori Habuchi, Toshiro Terachi, and the Referee Committee of the Endoscopic Surgical Skill Qualification System in Urological Laparoscopy. Reliability of Laparoscopic Skills Assessment on Video : 8-Year Results of the Endoscopic Surgical Skill Qualification System in Japan. *Journal of Endourology*, 28(11) :1374–1378, November 2014. ISSN 0892-7790, 1557-900X. doi : 10.1089/end.2014.0092.

[72] Ashleigh Menhadji, Corollos Abdelshehid, Kathyrn Osann, Reza Alipanah, Achim Lusch, Joseph Graversen, Jason Lee, Stephen Quach, Victor Huynh, Daniel Sidhom, Isabelle Gerbatsch, Jaime Landman, and Elspeth McDougall. Tracking and Assessment of Technical Skills Acquisition Among Urology Residents

for Open, Laparoscopic, and Robotic Skills Over 4 Years : Is There a Trend ? *Journal of Endourology*, 27(6) : 783–789, December 2012. ISSN 0892-7790. doi : 10.1089/end.2012.0633.

[73] Erica L. Mitchell, Dae Y. Lee, Nick Sevdalis, Aaron W. Partsafas, Gregory J. Landry, Timothy K. Liem, and Gregory L. Moneta. Evaluation of distributed practice schedules on retention of a newly acquired surgical skill : a randomized trial. *The American Journal of Surgery*, 201(1) :31–39, January 2011. ISSN 0002-9610. doi : 10.1016/j.amjsurg.2010.07.040.

[74] Emmeline Nugent. The evaluation of fundamental ability in acquiring minimally invasive surgical skill sets. *MD theses*, January 2012.

[75] Emmeline Nugent, Hazem Hseino, Emily Boyle, Brian Mehigan, Kieran Ryan, Oscar Traynor, and Paul Neary. Assessment of the role of aptitude in the acquisition of advanced laparoscopic surgical skill sets : results from a virtual reality-based laparoscopic colectomy training programme. *International Journal of Colorectal Disease*, 27(9) :1207–1214, September 2012. ISSN 1432-1262. doi : 10.1007/s00384-012-1458-y.

[76] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. A Generalized Local Binary Pattern Operator for Multiresolution Gray Scale and Rotation Invariant Texture Classification. In Sameer Singh, Nabeel Murshed, and Walter Kropatsch, editors, *Advances in Pattern Recognition — ICAPR 2001*, Lecture Notes in Computer Science, pages 399–408. Springer Berlin Heidelberg, 2001. ISBN 978-3-540-44732-0.

[77] WHO Consultation on Obesity, editor. *Obésité : prévention et prise en charge de l'épidémie mondiale ; rapport d'une consultation de l'OMS ; [Consultation OMS sur l'Obésité, Genève, 3 - 5 juin 1997]*. Number 894 in OMS, série de rapports techniques. Genève : Organisation mondiale de la santé edition, 2003. ISBN 978-92-4-220894-8. OCLC : 255470791.

[78] Vanessa N. Palter and Teodor P. Grantcharov. Simulation in surgical education. *CMAJ*, 182(11) :1191–1196, August 2010. ISSN 0820-3946, 1488-2329. doi : 10.1503/cmaj.091743.

[79] Igor Pernek and Alois Ferscha. A survey of context recognition in surgery. *Medical & Biological Engineering & Computing*, 55(10) :1719–1734, October 2017. ISSN 1741-0444. doi : 10.1007/s11517-017-1670-6.

[80] X. Pouliquen. Gestes de base en chirurgie laparoscopique de l'adulte. *EMC - Techniques chirurgicales - Appareil digestif*, 4(1) :1–25, January 2009. ISSN 02460424. doi : 10.1016/S0246-0424(09)44167-0.

[81] Mary K. Powers, Aaron Boonjindasup, Michael Pinsky, Philip Dorsey, Michael Maddox, Li-Ming Su, Matthew Gettman, Chandru P. Sundaram, Erik P. Castle, Jason Y. Lee, and Benjamin R. Lee. Crowdsourcing Assessment of Surgeon Dissection of Renal Artery and Vein During Robotic Partial Nephrectomy : A Novel Approach for Quantitative Assessment of Surgical Performance. *Journal of Endourology*, 30(4) :447–452, November 2015. ISSN 0892-7790. doi : 10.1089/end.2015.0665.

[82] Predrag Radivojac, Zoran Obradovic, A. Keith Dunker, and Slobodan Vucetic. Feature Selection Filters Based on the Permutation Test. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Machine Learning : ECML 2004*, Lecture Notes in Computer Science, pages 334–346, Berlin, Heidelberg, 2004. Springer. ISBN 978-3-540-30115-8. doi : 10.1007/978-3-540-30115-8_32.

[83] H. Rafii-Tari, C. J. Payne, J. Liu, C. Riga, C. Bicknell, and G. Yang. Towards automated surgical skill evaluation of endovascular catheterization tasks based on force and motion signatures. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1789–1794, May 2015. doi : 10.1109/ICRA.2015.7139430.

[84] James Reason. *Human Error*. Cambridge University Press, October 1990. ISBN 978-0-521-31419-0. Google-Books-ID : WJL8NZc8lZ8C.

[85] C. Richards, J. Rosen, B. Hannaford, C. Pellegrini, and M. Sinanan. Skills evaluation in minimally invasive surgery using force/torque signatures. *Surgical Endoscopy*, 14(9) :791–798, September 2000. ISSN 1432-2218. doi : 10.1007/s004640000230.

[86] Lee Richstone, Michael J. Schwartz, Casey Seideman, Jeffrey Cadeddu, Sandra Marshall, and Louis R. Kavoussi. Eye metrics as an objective assessment of surgical skill. *Annals of Surgery*, 252(1) :177–182, July 2010. ISSN 1528-1140. doi : 10.1097/SLA.0b013e3181e464fb.

[87] Raul J. Rosenthal. International Sleeve Gastrectomy Expert Panel Consensus Statement : best practice guidelines based on experience of >12,000 cases. *Surgery for Obesity and Related Diseases*, 8(1) :8–19, January 2012. ISSN 1550-7289. doi : 10.1016/j.soard.2011.10.019.

[88] D. Sarikaya, J. J. Corso, and K. A. Guru. Detection and Localization of Robotic Tools in Robot-Assisted Surgery Videos Using Deep Neural Networks for Region Proposal and Detection. *IEEE Transactions on Medical Imaging*, 36(7) :1542–1549, July 2017. ISSN 0278-0062. doi : 10.1109/TMI.2017.2665671.

[89] C Schaaf, A Lannelli, and J Gugenheim. État actuel de la chirurgie bariatrique en France. *e-memoires de l'Académie nationale de chirurgie*, (Vol. 15, fasc. 2) :104–107, 2015. ISSN 1634-0647. doi : 10.14607/emem.2015.2.104.

[90] Y. Sharma, T. Plötz, N. Hammerld, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, and I. Essa. Automated surgical OSATS prediction from videos. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 461–464, April 2014. doi : 10.1109/ISBI.2014.6867908.

[91] Yarden Sharon, Thomas Sean Lendvay, and Ilana Nisky. Instrument Orientation-Based Metrics for Surgical Skill Evaluation in Robot-Assisted and Open Needle Driving. *arXiv :1709.09452 [cs]*, September 2017. arXiv : 1709.09452.

[92] Ravikiran B. Singapogu, Lindsay O. Long, Dane E. Smith, Timothy C. Burg, Christopher C. Pagano, Varun V. Prabhu, and Karen J. L. Burg. Simulator-based assessment of haptic surgical skill : a comparative study. *Surgical Innovation*, 22(2) :183–188, April 2015. ISSN 1553-3514. doi : 10.1177/1553350614537119.

[93] S. T. A. Snaineh and B. Seales. Minimally invasive surgery skills assessment using multiple synchronized sensors. In *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 314–319, December 2015. doi : 10.1109/ISSPIT.2015.7394351.

[94] Renata Sánchez, Omaira Rodríguez, José Rosciano, Liumariel Vegas, Verónica Bond, Aram Rojas, and Alexis Sanchez-Ismayel. Robotic surgery training : construct validity of Global Evaluative Assessment of Robotic Skills (GEARS). *Journal of Robotic Surgery*, 10(3) :227–231, September 2016. ISSN 1863-2491. doi : 10.1007/s11701-016-0572-1.

[95] Ranil Sonnadara, Neil Rittenhouse, Ajmal Khan, Alex Mihailidis, Gregory Drozdzal, Oleg Safir, and Shuk On Leung. A novel multimodal platform for assessing surgical technical skills. *The American Journal of Surgery*, 203(1) :32–36, January 2012. ISSN 0002-9610, 1879-1883. doi : 10.1016/j.amjsurg.2011.08.008.

[96] Takahisa Suzuki, Hiroyuki Egi, Minoru Hattori, Masakazu Tokunaga, Hiroyuki Sawada, and Hideki Ohdan. An evaluation of the endoscopic surgical skills assessment using a video analysis software program. *Surgical Endoscopy*, 29(7) :1804–1808, July 2015. ISSN 1432-2218. doi : 10.1007/s00464-014-3863-5.

[97] Lingling Tao, Ehsan Elhamifar, Sanjeev Khudanpur, Gregory D. Hager, and René Vidal. Sparse Hidden Markov Models for Surgical Gesture Classification and Skill Evaluation. In Purang Abolmaesumi, Leo Joskowicz, Nassir Navab, and Pierre Jannin, editors, *Information Processing in Computer-Assisted Interventions*, Lecture Notes in Computer Science, pages 167–177. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-30618-1.

[98] Jérémie Thereaux, Nicolas Veyrie, Nicola Corigliano, Abdelhalim Aissat, Stéphane Servajean, and Jean-Luc Bouillot. Chirurgie bariatrique : techniques chirurgicales et leurs complications. *La Presse Médicale*, 39 (9) :945–952, September 2010. ISSN 0755-4982. doi : 10.1016/j.lpm.2010.01.015.

[99] Tony Tien, Philip H. Pucher, Mikael H. Sodergren, Kumuthan Sriskandarajah, Guang-Zhong Yang, and Ara Darzi. Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair. *Surgical Endoscopy*, 29(2) :405–413, February 2015. ISSN 1432-2218. doi : 10.1007/s00464-014-3683-7.

[100] A. P. Twinanda, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy. RSDNet : Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations. *IEEE Transactions on Medical Imaging*, 38(4) :1069–1078, April 2019. ISSN 0278-0062. doi : 10.1109/TMI.2018.2878055.

[101] Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. EndoNet : A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *arXiv :1602.03012 [cs]*, February 2016. arXiv : 1602.03012.

[102] Munenori Uemura, Morimasa Tomikawa, Ryuichi Kumashiro, Tiejun Miao, Ryota Souzaki, Satoshi Ieiri, Kenoki Ohuchida, Alan T. Lefor, and Makoto Hashizume. Analysis of hand motion differentiates expert and novice surgeons. *Journal of Surgical Research*, 188(1) :8–13, May 2014. ISSN 0022-4804, 1095-8673. doi : 10.1016/j.jss.2013.12.009.

[103] Munenori Uemura, Pierre Jannin, Makoto Yamashita, Morimasa Tomikawa, Tomohiko Akahoshi, Satoshi Obata, Ryota Souzaki, Satoshi Ieiri, and Makoto Hashizume. Procedural surgical skill assessment in laparoscopic training environments. *International Journal of Computer Assisted Radiology and Surgery*, 11(4) : 543–552, April 2016. ISSN 1861-6429. doi : 10.1007/s11548-015-1274-2.

[104] Melina C. Vassiliou, Liane S. Feldman, Christopher G. Andrew, Simon Bergman, Karen Leffondré, Donna Stanbridge, and Gerald M. Fried. A global assessment tool for evaluation of intraoperative laparoscopic skills. *American Journal of Surgery*, 190(1) :107–113, July 2005. ISSN 0002-9610. doi : 10.1016/j.amjsurg.2005.04.004.

[105] S. Swaroop Vedula, Anand Malpani, Narges Ahmidi, Sanjeev Khudanpur, Gregory Hager, and Chi Chiung Grace Chen. Task-Level vs. Segment-Level Quantitative Metrics for Surgical Skill Assessment. *Journal of Surgical Education*, 2016. ISSN 1931-7204. doi : 10.1016/j.jsurg.2015.11.009.

[106] P. Verhaeghe, A. Dhahri, Q. Qassemyar, and J.-M. Regimbeau. Technique de la gastrectomie longitudinale (« sleeve gastrectomy ») par laparoscopie. */data/traites/t01/40-53311/*, February 2011. ISSN 0246-0424.

[107] Simone L. Vernez, Victor Huynh, Kathryn Osann, Zhamshid Okhunov, Jaime Landman, and Ralph V. Clayman. C-SATS - Assessing Surgical Skills Among Urology Residency Applicants. *Journal of Endourology*, 31(S1) :S–95, September 2016. ISSN 0892-7790. doi : 10.1089/end.2016.0569.

[108] Ziheng Wang and Ann Majewicz Fey. SATR-DL - Improving Surgical Skill Assessment And Task Recognition In Robot-Assisted Surgery With Deep Neural Networks. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2018 :1793–1796, July 2018. ISSN 1557-170X. doi : 10.1109/EMBC.2018.8512575.

[109] Robert A. Watson. Use of a Machine Learning Algorithm to Classify Expertise : Analysis of Hand Motion Patterns During a Simulated Surgical Task. *Academic Medicine*, 89(8) :1163, August 2014. ISSN 1040-2446. doi : 10.1097/ACM.0000000000000316.

[110] Runmin Wei, Jingye Wang, Mingming Su, Erik Jia, Shaoqiu Chen, Tianlu Chen, and Yan Ni. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports*, 8(1) : 1–10, January 2018. ISSN 2045-2322. doi : 10.1038/s41598-017-19120-0.

[111] Lee W. White, Timothy M. Kowalewski, Rodney Lee Dockter, Bryan Comstock, Blake Hannaford, and Thomas S. Lendvay. Crowd-Sourced Assessment of Technical Skill : A Valid Method for Discriminating Basic Robotic Surgery Skills. *Journal of Endourology*, 29(11) :1295–1301, November 2015. ISSN 0892-7790, 1557-900X. doi : 10.1089/end.2015.0191.

[112] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco

Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1) :1–9, March 2016. ISSN 2052-4463. doi : 10.1038/sdata.2016.18.

[113] Remi Wolf. *Quantification of the 'quality' a surgical gesture from 'a priori' knowledge*. Theses, Université de Grenoble, June 2013. URL `https://tel.archives-ouvertes.fr/tel-00965163`.

[114] Q. Zhang and B. Li. Relative Hidden Markov Models for Video-Based Evaluation of Motion Skills in Surgical Training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6) :1206–1218, June 2015. ISSN 0162-8828. doi : 10.1109/TPAMI.2014.2361121.

[115] Qiang Zhang and Baoxin Li. Video-based Motion Expertise Analysis in Simulation-based Surgical Training Using Hierarchical Dirichlet Process Hidden Markov Model. In *Proceedings of the 2011 International ACM Workshop on Medical Multimedia Analysis and Retrieval*, MMAR '11, pages 19–24, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0991-2. doi : 10.1145/2072545.2072550. URL `http://doi.acm.org/10.1145/2072545.2072550`. event-place : Scottsdale, Arizona, USA.

[116] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing Through ADE20K Dataset. pages 633–641, 2017. URL `http://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_Scene_Parsing_Through_CVPR_2017_paper.html`.

[117] Aneeq Zia and Irfan Essa. Automated surgical skill assessment in RMIS training. *International Journal of Computer Assisted Radiology and Surgery*, 13(5) :731–739, May 2018. ISSN 1861-6429. doi : 10.1007/s11548-018-1735-5.