Titre: Title:	Évaluation de l'impact de l'implantation d'un outil numérique en milieu hospitalier : approche pré- et post-implantation
Auteur: Author:	Nicolas Jouglet
Date:	2026
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Jouglet, N. (2026). Évaluation de l'impact de l'implantation d'un outil numérique en milieu hospitalier : approche pré- et post-implantation [Master's thesis, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/76790/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/76790/
Directeurs de recherche: Advisors:	Marcelin Joanis, & Louis-Martin Rousseau
Programme: Program:	Maîtrise recherche en mathématiques appliquées

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Évaluation de l'impact de l'implantation d'un outil numérique en milieu hospitalier : approche pré- et post-implantation

NICOLAS JOUGLET

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Mathématiques

Mai 2026

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Évaluation de l'impact de l'implantation d'un outil numérique en milieu hospitalier : approche pré- et post-implantation

présenté par **Nicolas JOUGLET**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Michel GENDREAU, président
Marcelin JOANIS, membre et directeur de recherche
Louis-Martin ROUSSEAU, membre et codirecteur de recherche
Nadia LAHRICHI, membre

DÉDICACE

À ma famille,
pour son soutien, sa patience et sa confiance tout au long de ce parcours.

À mes amis,
pour leur présence, leurs encouragements et les moments partagés qui ont rendu cette aventure plus belle.

À la vie

REMERCIEMENTS

Ce mémoire marque l'aboutissement de mon parcours de maîtrise et n'aurait pu voir le jour sans l'accompagnement de plusieurs personnes que je souhaite remercier ici.

Je tiens tout d'abord à exprimer ma sincère gratitude à mes directeurs de recherche, **Marcelin Joanis** et **Louis-Martin Rousseau**, pour leur disponibilité, la qualité de leur encadrement et leurs conseils précieux tout au long de ce travail.

Je souhaite également souligner que ce mémoire s'inscrit en partie dans la continuité des travaux réalisés par **Ilitéa Kina**, qui ont constitué une base utile pour cette recherche.

Je remercie aussi les établissements qui ont jalonné mon parcours, **Centrale Marseille** et **Polytechnique Montréal**, pour la formation et le cadre académique qui ont rendu ce travail possible.

Enfin, je tiens à remercier chaleureusement mes proches pour leur soutien, leur patience et leurs encouragements tout au long de cette aventure.

RÉSUMÉ

L'implantation d'outils numériques en milieu hospitalier est souvent présentée comme une voie prometteuse pour améliorer l'organisation des soins et soutenir la décision clinique. Pourtant, mesurer concrètement leur effet demeure difficile, car l'hôpital est un environnement complexe, traversé par de nombreuses variations de charge, de ressources et de pratiques. Dans ce contexte, attribuer une évolution des performances à un seul outil numérique exige une démarche méthodologique adaptée.

Ce mémoire propose un cadre d'évaluation fondé sur une distinction entre deux temporalités complémentaires : avant l'implantation et après l'implantation. Avant le déploiement, l'objectif n'est pas de démontrer un effet causal, mais d'estimer de façon structurée les effets plausibles de l'intervention afin d'éclairer la décision. Après le déploiement, l'enjeu devient d'analyser les changements observés et d'évaluer dans quelle mesure ils peuvent être reliés à l'outil implanté.

Le premier volet porte sur une analyse pré-implantation d'un algorithme de prédiction de la congestion des urgences à 24 heures. À partir de données historiques couvrant la période 2018-2021, un pipeline prédictif a été élaboré afin de produire un signal cohérent avec un usage opérationnel. La reformulation du problème à une échelle temporelle de 8 heures, l'optimisation des variables retardées et le réglage du modèle selon la métrique ROC-AUC ont permis d'obtenir un outil plus pertinent sur le plan décisionnel. Sur cette base, plusieurs scénarios d'utilisation ont été construits afin d'estimer les effets attendus de l'implantation sur les délais de prise en charge, puis de les traduire en gains organisationnels et économiques potentiels.

Le second volet porte sur une analyse post-implantation. En l'absence de randomisation, une approche quasi-expérimentale a été retenue, reposant principalement sur les séries temporelles interrompues. Cette analyse est complétée par un test de Mann-Whitney, des ruptures placebo et une approche bayésienne afin de mieux apprécier la robustesse des résultats. L'ensemble met en évidence un signal globalement favorable après l'intervention, tout en invitant à nuancer la portée causale des effets observés.

Au final, ce travail montre qu'il est essentiel de ne pas confondre estimation d'impact et inférence causale. Il propose un cadre cohérent pour articuler analyse prospective et évaluation rétrospective, en tenant compte des contraintes réelles de l'environnement hospitalier et des limites propres aux données disponibles.

ABSTRACT

The implementation of digital tools in hospital settings is often presented as a promising way to improve care organization and support clinical decision-making. Yet assessing their actual effect remains difficult, as hospitals are complex environments shaped by continuous variations in workload, available resources, and professional practices. In this context, attributing changes in performance to a single digital tool requires an appropriate methodological approach.

This thesis proposes an evaluation framework based on a distinction between two complementary time horizons: before implementation and after implementation. Prior to deployment, the objective is not to demonstrate a causal effect, but rather to provide a structured estimate of the plausible effects of the intervention in order to support decision-making. After deployment, the challenge becomes to analyze the observed changes and assess to what extent they can be linked to the implemented tool.

The first part focuses on a pre-implementation analysis of an algorithm designed to predict hospital congestion 24 hours in advance. Using historical data covering the period from 2018 to 2021, a predictive pipeline was developed in order to generate a signal consistent with operational use. Reformulating the problem at an 8-hour time scale, optimizing lagged variables, and tuning the model using the ROC-AUC metric resulted in a tool that is more relevant from a decision-making perspective. On this basis, several usage scenarios were developed to estimate the expected effects of implementation on care delays and to translate them into potential organizational and economic gains.

The second part examines a post-implementation analysis. In the absence of randomization, a quasi-experimental approach was adopted, primarily relying on interrupted time series analysis. This analysis is complemented by a Mann-Whitney test, placebo break tests, and a Bayesian approach in order to better assess the robustness of the results. Overall, the findings point to a generally favorable signal following the intervention, while also calling for caution regarding the causal interpretation of the observed effects.

This work therefore shows that pre-implementation and post-implementation approaches should not be confused, as they serve different purposes and rely on different assumptions and levels of evidence. Its main contribution is to propose a structured methodological framework that articulates these two evaluation horizons in order to better support decision-making and the analysis of digital interventions in hospital settings.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

TPR	True Positive Rate (sensibilité)
BEST	Bayesian Estimation Supersedes the t-Test
CHUM	Centre hospitalier universitaire de Montréal
DiD	Différences-en-différences
ECR	Essai contrôlé randomisé
EDWIN	Emergency Department Work Index
GLS	Generalized Least Squares
HAC	Heteroskedasticity and Autocorrelation Consistent
IRM	Imagerie par résonance magnétique
ITS	Interrupted Time Series
MCMC	Markov Chain Monte Carlo
NPV	Negative Predictive Value
OLS	Ordinary Least Squares
OMS	Organisation mondiale de la Santé
PCR	Polymerase Chain Reaction
QQ-plot	Quantile-Quantile Plot
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
SED	Simulation à événement discret

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

Au cours de l'histoire, la médecine et les hôpitaux ont évolué ensemble, en se transformant mutuellement. Les progrès médicaux et l'amélioration des traitements ont peu à peu conduit les hôpitaux à dépasser leur rôle initial de simples lieux d'accueil pour devenir des organisations complexes, mobilisant de nombreux professionnels autour de la prise en charge des patients. Cette évolution s'est encore accentuée à partir des années 2000 avec l'essor du numérique. Les hôpitaux sont alors devenus des environnements fortement informatisés, où les décisions cliniques et organisationnelles s'appuient de plus en plus sur des systèmes d'information, des logiciels d'aide à la décision et des outils de pilotage des ressources. L'évolution en santé ne relève donc plus uniquement du biomédical, mais aussi de transformations organisationnelles et numériques.

Dans ce contexte, l'enjeu ne réside plus seulement dans le développement de nouvelles connaissances ou de nouveaux outils, mais aussi dans la capacité des établissements à les intégrer efficacement à leurs pratiques. Or, dans un système hospitalier soumis à de fortes contraintes, marqué par une interdépendance importante entre services et par des conséquences potentiellement critiques sur la qualité et la sécurité des soins, l'introduction d'une innovation numérique soulève nécessairement des questions d'évaluation et de décision.

Dans cette complexité croissante, la décision d'implanter un logiciel ne relève plus uniquement de choix techniques, mais constitue une véritable décision organisationnelle engageant durablement le fonctionnement de l'établissement. Une telle implantation appelle donc une évaluation rigoureuse de l'impact du logiciel sur le fonctionnement du système hospitalier.

Afin de mener une telle évaluation, l'Organisation mondiale de la Santé (OMS) propose certaines pistes méthodologiques dans son guide *Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment*, publié en 2016 [1]. Dans celui-ci, deux temporalités sont considérées pour évaluer l'effet d'un outil numérique, avant et après son implantation. Cette distinction est importante car elle renvoie à deux questions différentes : dans

un premier temps, anticiper des effets plausibles avant le déploiement ; puis une fois l'outil numérique déployé, analyser les effets observés.

Les objectifs varient également avec la temporalité. Une estimation de l'effet avant implantation vise avant tout à formuler les attentes et à estimer si l'intervention peut plausiblement produire un impact. Dans ces conditions, l'évaluation constitue avant tout un outil d'aide à la décision, sans permettre de garantir la réalité effective des effets estimés. L'objectif ainsi que la méthodologie de l'étude post-implantation sont tout autres. Le logiciel étant déjà implanté, l'étude post-implantation cherche à mesurer un impact observé et à examiner dans quelle mesure il peut être attribué à l'intervention.

Dans cette perspective, ce mémoire propose un cadre méthodologique mis en œuvre à travers deux études de cas distinctes, chacune correspondant à l'une des deux temporalités de l'évaluation. Ce choix s'explique par le fait qu'un même projet ne permettait pas, dans le cadre de ce travail, d'illustrer de manière satisfaisante ces deux dimensions avec des données et des conditions d'analyse adaptées à chacune.

- La première étude de cas porte sur un algorithme de prédiction de la congestion hospitalière, conçu pour signaler en amont les situations de surcharge. Menée avant son déploiement, l'étude vise à estimer les effets plausibles de l'intervention à partir de scénarios contrefactuels implicites.

- La seconde étude de cas porte sur un outil de planification déjà déployé dans des départements d'oncologie. Menée à partir de données observées avant et après son implantation, elle vise à mesurer l'effet associé à l'intervention et à examiner dans quelle mesure celui-ci peut lui être attribué.

Ensemble, ces deux études mettent en évidence la complémentarité entre une logique d'estimation prospective et une logique d'évaluation rétrospective de l'impact. C'est à partir de cette double perspective que se définit l'objectif général du mémoire, soit proposer un cadre méthodologique permettant d'évaluer un outil numérique hospitalier selon la temporalité considérée. Plus

précisément, le travail vise, d'une part, à structurer l'estimation pré-implantation des effets plausibles d'une intervention et, d'autre part, à encadrer l'analyse post-implantation des effets observés à partir des données disponibles.

À ce titre, trois contributions principales se dégagent de ce travail. Premièrement, le mémoire établit un lien explicite entre l'évaluation pré-implantation et l'évaluation post-implantation d'une intervention, en montrant qu'elles répondent à des objectifs distincts, mobilisent des hypothèses différentes et reposent sur des niveaux de preuve spécifiques. Deuxièmement, il introduit une procédure de falsification fondée sur des ruptures placebo dans les séries temporelles, afin de tester la plausibilité d'un effet causal en vérifiant que des ruptures similaires ne sont pas détectées en l'absence réelle d'intervention. Troisièmement, il développe une analyse de robustesse du volet post-implantation reposant sur la construction de séries pré-intervention artificiellement améliorées, dans le but d'examiner à partir de quel niveau d'amélioration contrefactuelle les conclusions observées cessent d'être soutenues.

Le mémoire est structuré en six chapitres. À la suite de cette introduction, le chapitre 2 présente le cadre conceptuel et la revue de littérature consacrés à l'évaluation de l'impact des outils numériques en milieu hospitalier. Le chapitre 3 expose les objectifs de recherche, le cadre d'estimation et la méthodologie retenue pour articuler les volets pré- et post-implantation. Le chapitre 4 est consacré à l'analyse pré-implantation : il présente le protocole prédictif, explicite la chaîne d'action étudiée et propose une estimation des effets attendus à partir de plusieurs scénarios organisationnels. Le chapitre 5 porte sur l'analyse post-implantation, fondée sur une approche par séries temporelles interrompues complétée par plusieurs analyses complémentaires. Enfin, le chapitre 6 synthétise les principaux enseignements du mémoire, discute des limites générales, formule des recommandations méthodologiques et ouvre sur des perspectives de prolongement.

CHAPITRE 2 CADRE CONCEPTUEL ET REVUE DE LITÉRATURE

L'évaluation de l'impact d'un outil numérique en milieu hospitalier constitue un problème méthodologique exigeant. En effet, l'implantation d'un logiciel ne correspond pas à une intervention ponctuelle et isolée, mais à une transformation sociotechnique inscrite dans un système de soins complexe, où interagissent simultanément les organisations, les professionnels, les flux de patients et les ressources disponibles. Comme rappelé dans l'introduction, l'évaluation de l'effet ne répond pas aux mêmes objectifs selon la temporalité retenue, ce qui conduit à mobiliser des méthodes différentes. Afin de faciliter le suivi et la compréhension, le présent chapitre s'organise également autour de ces deux temporalités de l'évaluation.

Dans cette perspective, il convient d'examiner d'abord les approches mobilisables avant implantation, lorsque l'enjeu est d'estimer des effets plausibles sans disposer de données post-intervention, les approches mobilisables après l'implantation seront étudiées dans la seconde partie de ce chapitre.

2.1 Évaluer avant l'implantation : estimer des effets plausibles

Avant l'implantation d'un outil numérique, l'évaluation ne peut s'appuyer sur aucun effet observé, puisque l'intervention n'a pas encore eu lieu. La difficulté méthodologique consiste alors à produire une estimation suffisamment structurée et crédible des effets attendus pour éclairer la décision, sans pour autant prétendre démontrer que ces effets se réaliseront effectivement. En d'autres termes, il s'agit de construire un raisonnement permettant d'anticiper les conséquences plausibles du déploiement de l'outil.

Cette difficulté est renforcée par la nature même du milieu hospitalier. L'implantation d'un logiciel ne modifie pas seulement un support technique, mais peut affecter l'organisation du travail, la circulation de l'information, la coordination entre professionnels ou encore certaines décisions cliniques [2]. Dès lors, les effets attendus de l'intervention dépendent étroitement du contexte dans lequel elle s'insère, des usages envisagés et des mécanismes organisationnels par lesquels elle est supposée produire un changement.

2.1.1 Analyse pré-implantation : objet et portée

L'analyse pré-implantation a pour objet de situer l'intervention dans son contexte, de préciser le problème auquel l'outil est censé répondre et d'expliciter les mécanismes par lesquels il pourrait produire un changement. Elle ne vise donc pas uniquement à estimer des effets plausibles, mais aussi à clarifier les conditions de possibilité de ces effets avant le déploiement effectif de l'intervention.

Cette attention portée au contexte est cohérente avec la littérature sur l'implantation des technologies de santé. Le cadre **NASSS** (*Nonadoption, Abandonment, Scale-up, Spread and Sustainability*) souligne en effet que le succès d'un déploiement ne dépend pas uniquement des propriétés techniques de l'outil, mais aussi de la complexité de la situation clinique, des usages envisagés, des acteurs concernés, du contexte organisationnel et des conditions plus larges de mise en œuvre [3]. Des travaux empiriques vont dans le même sens en montrant qu'une phase de pré-implantation permet d'identifier en amont des barrières et des facilitateurs, puis d'adapter la stratégie de déploiement au contexte local [4, 5]. Dans cette perspective, l'analyse pré-implantation constitue un cadre d'analyse utile pour situer l'intervention dans son contexte d'application, expliciter les hypothèses retenues et comparer différents scénarios de déploiement. Elle sert ainsi avant tout à réduire l'incertitude décisionnelle en amont de l'implantation, en appréciant la valeur potentielle de l'intervention et les conditions dans lesquelles ses effets attendus pourraient se réaliser.

Dans ce cadre, la littérature ne propose pas une méthode unique, mais plusieurs démarches permettant de produire des ordres de grandeur plausibles des effets attendus. Parmi celles-ci figurent notamment les approches de simulation, les modélisations prédictives ex ante et le recours à des expériences externes comparables.

2.1.2 Simulation structurelle

Une première famille de méthodes consiste à modéliser explicitement le fonctionnement du système afin d'y simuler virtuellement l'intervention envisagée. Dans cette perspective, une approche fréquemment mobilisée dans la littérature est la simulation à événement discret (SED) [6]. La SED modélise le fonctionnement du service comme une succession d'événements spécifiques (par exemple : arrivée d'un patient, début d'une consultation, libération d'un lit). Elle permet ainsi d'intégrer la gestion des ressources limitées, la dynamique des flux et des files d'attente et les différentes règles de décision cliniques.

Une fois le modèle de référence (ou état initial) fonctionnel, c'est-à-dire un modèle reproduisant fidèlement le fonctionnement actuel du département, il devient possible de simuler virtuellement l'implantation d'un outil numérique à travers différents scénarios. Cette approche constitue un réel atout pour la décision : elle permet non seulement d'estimer un gain théorique, mais aussi d'identifier les effets de bord potentiels ou les modifications organisationnelles nécessaires pour que l'outil produise l'effet escompté.

2.1.3 Modélisation prédictive ex ante (empirique)

Alors que la simulation structurelle repose sur la représentation explicite des mécanismes internes d'un système, la modélisation prédictive consiste à s'appuyer directement sur les régularités observées dans les données historiques. Plutôt que de reconstruire le fonctionnement organisationnel à partir de ses composantes, cette démarche exploite les relations statistiques empiriquement identifiées afin d'anticiper l'évolution d'un indicateur dans un scénario hypothétique d'intervention.

Il ne s'agit donc pas, comme dans les méthodes quasi-expérimentales post-implantation, d'identifier un effet causal à partir d'une rupture observée ou d'une comparaison avec un groupe contrôle, mais d'estimer ce que deviendrait le système si certaines conditions étaient modifiées. Concrètement, un modèle prédictif est ajusté sur des données passées, puis mobilisé pour simuler

différents scénarios prospectifs. L'impact potentiel correspond alors à l'écart entre la trajectoire prédite en l'absence d'intervention et celle obtenue lorsque la modification envisagée est introduite. Cette logique se rapproche de ce que certains auteurs qualifient de prédiction contrefactuelle, par opposition à la prédiction strictement factuelle, qui vise uniquement à prolonger les tendances observées [7]. La validité de cette approche repose toutefois sur des hypothèses fortes, notamment la stabilité des relations empiriques entre variables et la capacité du modèle à conserver ses performances hors échantillon. En l'absence de données post-intervention, ces hypothèses ne peuvent être vérifiées ex ante et doivent donc être explicitement discutées.
D'un point de vue méthodologique, cette approche ne constitue pas un design d'identification causale au sens strict du cadre de Neyman-Rubin. Elle ne permet pas d'observer un contrefactuel réalisé, mais propose une estimation conditionnelle des effets attendus sous un ensemble d'hypothèses structurelles. Son intérêt réside principalement dans la réduction de l'incertitude décisionnelle : elle fournit un ordre de grandeur plausible de l'impact potentiel et permet ainsi d'éclairer la décision d'implantation lorsque l'expérimentation contrôlée n'est pas envisageable.

2.1.4 Benchmarking externe

Lorsque la simulation structurelle ou la modélisation prédictive ne sont pas envisageables, une autre possibilité consiste à s'appuyer sur des retours d'expérience externes. Dans la mesure où l'outil considéré a déjà été implanté dans d'autres établissements de santé, il peut être pertinent d'examiner les effets observés dans des contextes jugés comparables afin d'anticiper son impact potentiel dans le cadre étudié.

Cette démarche relève d'une forme de benchmarking externe : elle consiste à mobiliser des expériences antérieures comme source d'information pour éclairer une décision d'implantation locale. Son intérêt dépend toutefois fortement du degré de comparabilité entre les contextes considérés, qu'il s'agisse de l'organisation des services, des usages effectifs de l'outil ou encore des ressources disponibles. Il ne s'agit donc pas d'une identification causale locale, mais d'une extrapolation informée fondée sur des expériences comparables.

Ces différentes approches répondent à une même question : comment estimer, avant le déploiement, les effets plausibles d'un outil numérique sans disposer de données post-intervention. Une fois l'implantation réalisée, la question change toutefois de nature : il ne s'agit plus de construire une estimation prospective destinée à éclairer la décision, mais d'analyser un effet observé et d'examiner dans quelle mesure celui-ci peut être attribué à l'intervention.

2.2 Évaluer après l'implantation : interpréter un effet observé

Dans un environnement hospitalier complexe, cette attribution demeure délicate, car les indicateurs de performance peuvent être influencés par de nombreux facteurs concomitants [3, 4], tels que la saisonnalité, les changements de personnel, les réorganisations internes ou les épisodes épidémiques. La littérature s'est donc largement structurée autour de méthodes cherchant à interpréter cet effet observé dans un cadre causal ou quasi-causal.

C'est dans cette perspective que l'inférence causale devient pertinente. Son cadre de référence, l'essai contrôlé randomisé, repose toutefois sur des conditions souvent difficiles à réunir dans un système hospitalier réel [8]. Les contraintes éthiques, organisationnelles et opérationnelles limitent fréquemment la possibilité de comparer rigoureusement un groupe exposé et un groupe non exposé dans des conditions satisfaisantes. La littérature s'est ainsi tournée vers des approches alternatives permettant d'interpréter, de façon plus ou moins crédible selon les hypothèses retenues, un effet observé après implantation.

2.2.1 Inférence causale et estimation d'impact : cadre conceptuel

L'estimation de l'impact d'une intervention repose sur une logique d'inférence causale. Ce domaine des mathématiques vise à déterminer si des effets observés sont attribuables à une intervention plutôt qu'à des facteurs externes. Le modèle de Neyman-Rubin [9, 10] définit l'effet causal comme la différence entre le résultat qui serait observé en présence de l'intervention et celui qui serait observé en son absence, pour une même unité.

Pour une unité i, l'effet causal est défini par :

$$\tau_i = Y_i(1) - Y_i(0).$$

Où τ_i est l'effet causal individuel, $Y_i(1)$est le résultat potentiel de l'unité i si elle reçoit l'intervention et $Y_i(0)$est le résultat potentiel de la même unité i si elle ne reçoit pas l'intervention. En pratique, pour chaque unité un seul de ces deux résultats est observable, ce qui constitue le *problème fondamental de l'inférence causale*.

Le cadre de référence pour l'inférence causale est la mise en place d'essais contrôlés randomisés (ECR) [11, 12]. Le principe est de constituer aléatoirement deux groupes, l'un recevant l'intervention (groupe expérimental), tandis que l'autre non (groupe contrôle). La randomisation élimine les biais et permet d'obtenir des groupes comparables en moyenne. Ainsi, la différence entre les deux groupes peut être interprétée comme un effet causal de l'intervention.

Face aux limites de l'approche des ECR, la littérature en interventions en santé s'est largement tournée vers le développement de méthodes alternatives dites quasi-expérimentales. Ces méthodes ne garantissent pas l'identification causale, mais cherchent à la rendre plausible sous un ensemble d'hypothèses explicites.

2.2.2 Séries temporelles interrompues

Lorsque des données sont disponibles avant et après l'intervention, l'approche la plus simple consiste à comparer les moyennes observées sur les deux périodes. Une telle comparaison repose toutefois sur une hypothèse particulièrement forte : celle selon laquelle, en l'absence d'intervention, le niveau de performance observé après implantation aurait été identique à celui observé avant implantation. Or, dans un contexte hospitalier, cette hypothèse est rarement crédible, car les indicateurs peuvent évoluer sous l'effet de tendances de fond, de fluctuations saisonnières ou d'autres changements concomitants.

Afin de dépasser cette limite, la littérature recommande fréquemment le recours aux séries temporelles interrompues (ITS). Contrairement à une simple comparaison avant-après, cette approche exploite la dynamique de la série dans le temps en estimant la trajectoire pré-intervention,

puis en examinant si l'implantation s'accompagne d'une rupture de niveau, d'un changement de tendance, ou d'une combinaison des deux. L'ITS permet ainsi de construire un contrefactuel plus crédible que la seule moyenne pré-intervention, même si son interprétation reste conditionnelle à plusieurs hypothèses, notamment l'absence de choc majeur concomitant à l'intervention et une modélisation adéquate de la structure temporelle de la série [13, 14].

2.2.3 Méthode des différences-in-différences (DiD)

L'impossibilité de constituer deux groupes distincts (test et contrôle) au sein d'un même département hospitalier rend la mise en place d'un essai contrôlé randomisé (ECR) impraticable. Toutefois, en adoptant une perspective plus globale, l'ensemble du département hospitalier peut être considéré comme le groupe expérimental. Il devient alors envisageable de le comparer à un département similaire appartenant à un autre hôpital, lequel ferait office de groupe contrôle. Ainsi la méthode DiD [15] mesure l'impact en comparant l'évolution d'un indicateur de performance avant et après l'intervention entre un groupe traité et un groupe contrôle. L'effet est estimé comme la différence entre ces deux variations. Cette méthode repose sur l'hypothèse centrale de tendances parallèles selon laquelle en l'absence d'intervention les deux groupes auraient suivi la même trajectoire. La difficulté de cette méthode réside principalement dans la recherche de ce groupe contrôle et dans le respect de l'hypothèse centrale (bien que non testable réellement).

2.2.4 Analyse contrefactuelle bayésienne post-intervention

Une alternative aux approches fondées sur une rupture observée consiste à recourir à une modélisation contrefactuelle prédictive, notamment via des approches bayésiennes de type *CausalImpact* [16]. L'objectif est d'estimer, à partir de la seule période pré-implantation, la trajectoire attendue de l'indicateur en l'absence d'intervention, en exploitant la capacité prédictive d'un modèle entraîné sur les données historiques. Les valeurs observées durant la période post-implantation sont ensuite comparées à ce scénario contrefactuel estimé. L'effet de l'intervention est mesuré comme l'écart entre les observations réelles et la prédiction conditionnelle du modèle, ce qui permet d'estimer un impact causal sous les hypothèses du cadre bayésien.

Dans l'ensemble de ces méthodes, l'inférence causale reste conditionnelle aux hypothèses du design et ne constitue pas une propriété intrinsèque des données. Les approches précédentes conservent ainsi, à des degrés divers, une ambition attributive : elles cherchent à rapprocher l'effet observé d'un contrefactuel crédible. D'autres méthodes, en revanche, poursuivent un objectif plus limité. Elles ne visent pas à identifier formellement un effet causal, mais à quantifier rigoureusement l'ampleur et l'incertitude d'une différence observée entre périodes.

2.2.5 Estimation bayésienne de différence moyenne (BEST)

La méthode BEST (Bayesian Estimation Supersedes the t-test), proposée par Kruschke (2013), repose sur une estimation bayésienne de la différence de moyennes entre deux groupes, généralement modélisée sous une vraisemblance de type Student-t, plus robuste aux valeurs extrêmes et aux distributions non normales.

Contrairement aux tests fréquentistes classiques, BEST fournit la distribution postérieure complète de la différence observée entre périodes, permettant d'obtenir une estimation ponctuelle, un intervalle crédible ainsi que la probabilité que la différence dépasse un seuil d'intérêt.

Cette approche ne constitue pas à elle seule un design d'identification causale, mais offre une quantification probabiliste rigoureuse d'une variation observée. Elle peut ainsi compléter une analyse post-implantation structurée en apportant une mesure explicite de l'incertitude associée à l'effet estimé.

2.3 Conclusion

Cette revue de littérature met en évidence que l'évaluation d'un outil numérique en milieu hospitalier ne peut être pensée indépendamment de la temporalité considérée. Avant implantation, l'enjeu consiste à produire une estimation prospective des effets plausibles de l'intervention afin d'éclairer la décision. Dans cette perspective, l'analyse pré-implantation permet de contextualiser l'outil, d'expliciter les mécanismes d'action attendus et de comparer plusieurs scénarios de déploiement. Les approches de simulation structurelle, de modélisation prédictive ex ante et de

benchmarking externe offrent ainsi des appuis méthodologiques complémentaires pour formuler des ordres de grandeur plausibles des effets attendus.

Après implantation, la question change de nature. Il ne s'agit plus d'anticiper des effets plausibles, mais d'interpréter un effet observé à partir des données disponibles. Dans un environnement hospitalier complexe, cette attribution demeure délicate, car les indicateurs de performance peuvent être influencés par de nombreux facteurs concomitants. Si les essais contrôlés randomisés constituent le cadre théorique de référence pour l'inférence causale, leur mise en œuvre est rarement compatible avec les réalités éthiques, organisationnelles et opérationnelles du milieu hospitalier. Dès lors, les méthodes quasi-expérimentales, telles que les séries temporelles interrompues ou les approches en différences-in-différences, apparaissent comme les principaux outils mobilisables pour analyser un effet observé après implantation, sous réserve d'hypothèses fortes et explicitement discutées. Les approches bayésiennes peuvent, quant à elles, compléter cette analyse en proposant soit un cadre contrefactuel prédictif, soit une quantification probabiliste de l'ampleur et de l'incertitude d'une différence observée.

Ainsi, la littérature conduit moins à retenir une méthode unique qu'à distinguer deux cadres d'évaluation complémentaires, correspondant à deux temporalités du cycle d'implantation. Le premier relève d'une logique prospective, orientée vers l'estimation et la contextualisation des effets plausibles avant implantation. Le second relève d'une logique rétrospective, centrée sur l'analyse et l'interprétation d'un effet observé après déploiement. Cette distinction structure le cadre méthodologique retenu dans ce mémoire et justifie, au chapitre suivant, l'articulation entre un volet pré-implantation et un volet post-implantation.

Tableau 2.1 Résumé des méthodes post-implantation

Méthode	Logique méthodologique	Hypothèse centrale	Menace principale	Type de conclusion
Essai contrôlé randomisé (ECR)	Design expérimental	Randomisation valide	Non-implémentabilité éthique/organisationnelle	Causale forte
Comparaison avant-après	Comparaison simple	Stabilité temporelle	Confusion temporelle	Descriptive
Séries temporelles interrompues (ITS)	Rupture temporelle structurée	Absence de choc simultané	Rupture non liée à l'intervention	Causale plausible
Différences-in-différences (DiD)	Comparaison inter-groupes	Tendances parallèles	Groupe contrôle non comparable	Causale conditionnelle
Analyse contrefactuelle bayésienne	Reconstruction contrefactuelle prédictive	Stabilité du modèle pré-intervention	Mauvaise spécification du modèle	Attribution probabiliste
BEST	Estimation bayésienne de différence	Modèle de vraisemblance valide	Non-stationnarité / dépendance	Différence probabiliste

Tableau 2.2 Résumé des méthodes pré-implantation

Méthode	Logique méthodologique	Hypothèse centrale	Menace principale	Type de conclusion
Simulation structurelle (SED)	Modélisation du système	Fidélité du modèle au réel	Simplification excessive	Scénarios plausibles
Benchmarking externe	Extrapolation comparative	Transférabilité contextuelle	Hétérogénéité institutionnelle	Extrapolation informée
Modélisation prédictive ex ante	Simulation prospective	Stabilité des relations statistiques	Rupture structurelle	Ordre de grandeur plausible

CHAPITRE 3 DESCRIPTION DE L'APPROCHE PROPOSÉE

La revue de littérature a conduit à distinguer deux cadres d'évaluation complémentaires, correspondant aux temporalités pré- et post-implantation. À partir de cette distinction, ce chapitre formalise le cadre méthodologique retenu dans ce mémoire.

Il présente d'abord les objectifs de recherche, puis explicite, pour chacun des deux volets, l'estimand visé, les hypothèses mobilisées et la portée des conclusions envisageables. Il précise enfin la stratégie empirique retenue pour mettre en œuvre ces deux cadres d'évaluation dans les chapitres suivants.

3.1 Objectifs de recherche

L'objectif de ce mémoire est de proposer un cadre méthodologique permettant d'évaluer l'implantation d'un outil numérique en milieu hospitalier selon la temporalité considérée. Il s'agit, d'une part, de structurer l'estimation pré-implantation des effets plausibles de l'intervention et, d'autre part, d'encadrer l'analyse post-implantation des effets observés à partir des données disponibles.

Deux objectifs spécifiques en découlent. Le premier consiste, en phase pré-implantation, à proposer une démarche permettant d'estimer l'effet potentiel d'un outil à partir des données historiques disponibles et de scénarios d'usage plausibles. Le second consiste, en phase post-implantation, à proposer une démarche permettant d'analyser l'évolution d'indicateurs observés après le déploiement de l'outil et d'en apprécier la plausibilité causale malgré les limites inhérentes aux données observationnelles en contexte hospitalier.

Au-delà de ces deux volets empiriques, ce mémoire cherche aussi à montrer que l'évaluation d'un outil numérique en milieu hospitalier ne peut être pensée de manière unique, mais doit être adaptée à la temporalité considérée. L'ambition n'est donc pas de défendre une méthode unique, mais de proposer un cadre cohérent reliant objectifs, hypothèses, données disponibles et type de conclusion que l'on peut raisonnablement tirer dans chaque situation.

3.2 Volet pré-implantation

3.2.1 Cadre d'estimation pré-implantation

Avant l'implantation, l'objectif est d'estimer un ordre de grandeur plausible de l'amélioration attendue si l'outil est utilisé dans les conditions envisagées.

L'effet étudié correspond à la différence entre la trajectoire de l'indicateur en l'absence d'implantation et la trajectoire simulée sous un scénario d'utilisation du logiciel. On peut l'écrire, de manière générale, comme :

$$\Delta_{\text{ex ante}} = \mathbb{E}[Y^{sim}(t)] - \mathbb{E}[Y^{base}(t)]$$

Où Y^{sim} correspond à la trajectoire simulée de l'indicateur conditionnellement au scénario d'intervention retenu, alors que $Y^{base}(t)$ désigne la trajectoire de référence en l'absence d'intervention.

Dans le cadre de ce mémoire, l'unité d'analyse correspond à une observation de la série Y, soit une période de 8 heures. Ce choix est cohérent avec le pipeline prédictif final, construit à cette même fréquence, ainsi qu'avec la logique opérationnelle des scénarios étudiés.

L'interprétation de cet effet repose sur plusieurs hypothèses. Premièrement, les relations statistiques observées dans les données historiques doivent rester suffisamment stables pour permettre une extrapolation vers les situations simulées. Deuxièmement, le modèle prédictif utilisé doit présenter une validité hors échantillon suffisante pour que les scénarios reposent sur un signal crédible. Enfin, il est supposé que les scénarios envisagés puissent effectivement être actionnés en réponse au signal fourni par l'algorithme.

Dans ce cadre, le niveau de preuve attendu reste limité. L'estimation produite doit être comprise comme une plausibilité structurelle conditionnelle aux hypothèses du modèle, et non comme la mesure d'un effet causal observé. Elle sert avant tout à éclairer une décision d'implantation en

fournissant des ordres de grandeur plausibles de l'amélioration attendue. Cette logique est proche de ce que Dickerman et Hernán décrivent comme une forme de *counterfactual prediction*, c'est-à-dire une prédiction portant sur ce qui se produirait dans un monde différent de celui observé, à partir d'un scénario hypothétique d'intervention [7].

3.2.2 Méthodologie pré-implantation

3.2.2.1 Contexte

Le cas pré-implantation étudié porte sur un logiciel de prédiction de la congestion des urgences à un horizon de 24 heures. L'outil est conçu comme un support à la décision permettant d'anticiper les épisodes de congestion et de favoriser une réponse organisationnelle précoce. Le projet s'inscrit dans la continuité de travaux préalables ayant permis de constituer une première base analytique ainsi qu'une première version du pipeline prédictif.

3.2.2.2 Données

L'analyse pré-implantation repose sur une base de données historiques regroupant 66 variables relevées à une fréquence de 15 minutes sur une période de plus de 3,5 ans allant de mi-2018 à 2021. Cette base rassemble les principales informations relatives au fonctionnement des urgences, notamment le triage, les consultations, les examens cliniques, les admissions à l'hôpital ainsi qu'un score représentatif de la congestion. Elle comprend également des variables temporelles permettant de situer chaque observation selon le mois, le jour de la semaine et l'heure. Ces données constituent le socle empirique à partir duquel est construit le modèle prédictif.

3.2.2.3 Approche pré-implantation retenue

En phase pré-implantation, l'effet du logiciel ne peut pas être observé directement puisqu'aucun déploiement réel n'a encore eu lieu. L'objectif n'est donc pas de mesurer un effet causal, mais d'estimer un effet potentiel plausible à partir des données disponibles. La méthodologie retenue repose sur deux composantes complémentaires : d'une part, une analyse du contexte d'implantation, visant à documenter la pertinence, la faisabilité et les conditions d'usage de l'intervention ; d'autre part, une estimation prospective de son effet potentiel à partir de données

historiques et de scénarios d'usage [17-19]. Dans cette perspective, l'analyse relève moins d'une évaluation d'effet observé que d'une démarche de prédiction contrefactuelle appliquée à un scénario d'usage hypothétique [7].

La première composante consiste à caractériser le contexte organisationnel dans lequel l'outil pourrait être implanté. Cette étape porte sur la problématique ciblée, les acteurs concernés, les conditions de mise en œuvre ainsi que les principaux freins et facilitateurs à l'utilisation du logiciel. Elle permet également de préciser les indicateurs qui serviront ensuite à l'évaluation et de définir une situation de référence à partir des données historiques.

La seconde composante repose sur une démarche de modélisation prospective. Elle consiste, dans un premier temps, à construire un modèle prédictif de la congestion à partir des données historiques, afin d'identifier les périodes susceptibles de correspondre à un épisode de surcharge. Dans un second temps, les prédictions issues de ce modèle servent de base à la définition et à la simulation de scénarios d'intervention organisationnelle. Le pipeline prédictif constitue ainsi un instrument intermédiaire de l'évaluation : il ne vise pas directement à mesurer l'impact du logiciel, mais à générer le signal sur lequel repose l'estimation prospective des effets des scénarios étudiés. Les détails techniques relatifs à la construction, à l'entraînement et à l'évaluation de ce pipeline sont présentés au chapitre suivant.

À partir de ce signal, plusieurs scénarios d'usage plausibles sont ensuite définis. Ces scénarios correspondent à des ajustements de ressources susceptibles d'être mobilisés lorsqu'un épisode de congestion est anticipé. Dans le cadre de ce mémoire, les scénarios retenus portent sur l'ajout d'un médecin urgentiste, l'ajout d'une civière, ainsi que l'ajout combiné des deux. L'effet de ces scénarios est d'abord exprimé sur l'indicateur de congestion, puis traduit en variations attendues de délais d'attente à l'aide des relations observées dans les données historiques. Cette étape permet de passer d'un indicateur synthétique peu interprétable à des ordres de grandeur plus concrets sur le fonctionnement du service.

Enfin, les variations estimées des délais sont utilisées pour construire une estimation économique simplifiée des scénarios étudiés. Celle-ci consiste à mettre en regard les coûts associés aux ajustements de ressources envisagés et les bénéfices potentiels liés à la réduction attendue des délais. La méthodologie pré-implantation s'organise ainsi en cinq étapes : analyse du contexte, construction d'un signal prédictif, définition des scénarios d'intervention, simulation de leurs effets sur la congestion, puis traduction de ces effets en impacts opérationnels et économiques.

Les résultats obtenus dans ce cadre doivent être interprétés avec prudence. Ils ne constituent pas une preuve causale, mais des ordres de grandeur plausibles, conditionnels aux hypothèses retenues et à la validité du modèle utilisé. Leur intérêt principal est d'éclairer la décision d'implantation en fournissant une estimation structurée de ce que l'outil pourrait produire dans des conditions réalistes d'utilisation. Là encore, la littérature sur la *counterfactual prediction* rappelle que ce type d'estimation repose sur des hypothèses fortes, non directement vérifiables dans les données observées [7].

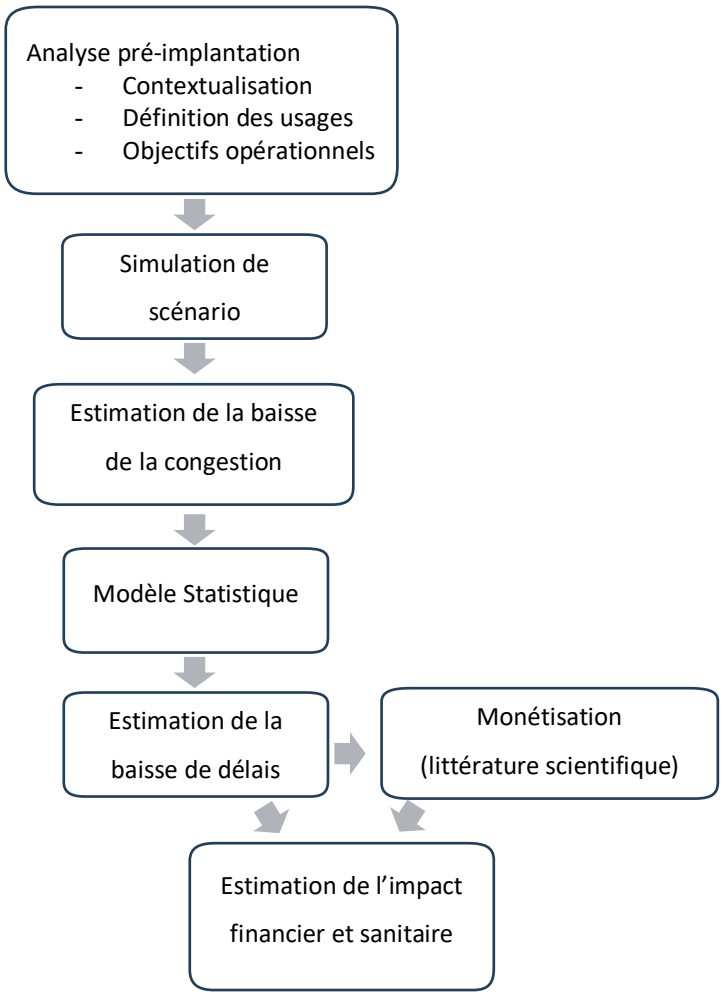

Figure 3.1 Schéma méthodologique de l'analyse pré-implantation

3.3 Volet post-implantation

3.3.1 Cadre d'estimation post-implantation

Après l'implantation, l'objectif n'est plus d'estimer un effet potentiel, mais d'évaluer si l'intervention s'accompagne d'un changement observable dans la dynamique de l'indicateur étudié. Plus précisément, l'effet d'intérêt correspond à l'écart entre la trajectoire observée après

implantation et la trajectoire qui aurait été attendue en l'absence d'intervention, compte tenu de la dynamique préexistante.

De manière générale, l'effet étudié peut être représenté comme :

$$\Delta^{\text{post}} = Y_{t\geq T_0}^{\text{obs}} - Y_{t\geq T_0}^{\text{cf}},$$

où $Y_{t\geq T_0}^{\text{obs}}$ désigne la trajectoire observée après la date d'implantation T_0, et $Y_{t\geq T_0}^{\text{cf}}$ la trajectoire contrefactuelle qui aurait été attendue en l'absence d'intervention à partir de la dynamique antérieure.

L'interprétation de cet effet repose sur plusieurs hypothèses. Il faut d'abord supposer que, sans intervention, la dynamique préexistante se serait prolongée de manière suffisamment régulière pour servir de référence contrefactuelle. Il faut ensuite considérer qu'aucun autre changement majeur, survenant au même moment, n'explique à lui seul la rupture observée. Enfin, la structure temporelle de la série doit être suffisamment bien prise en compte pour que les méthodes utilisées ne confondent pas rupture réelle et simple inertie de la série.

Le niveau de preuve attendu est ici plus élevé qu'en pré-implantation, mais reste limité par la nature observationnelle des données. L'objectif n'est pas de revendiquer une identification causale expérimentale stricte, mais d'évaluer la plausibilité d'une attribution de l'effet observé à l'intervention. Autrement dit, une rupture statistiquement et temporellement cohérente avec la date d'implantation constitue un indice compatible avec un effet de l'outil, sans suffire à elle seule à l'établir définitivement. Cette prudence est cohérente avec la littérature méthodologique sur les séries temporelles interrompues, qui insiste à la fois sur la force de ce devis lorsque la randomisation n'est pas possible, et sur la nécessité d'interpréter avec prudence les résultats lorsqu'ils reposent sur des données observationnelles et des hypothèses structurelles non testables [13, 14].

3.3.2 Méthodologie post-implantation

3.3.2.1 Contexte

Le second projet porte sur un outil numérique déjà implanté dans plusieurs établissements hospitaliers. Ce volet du mémoire s'appuie sur quatre rapports produits après le déploiement du logiciel dans différents systèmes hospitaliers. Ces rapports documentent l'évolution de plusieurs indicateurs observés avant et après l'implantation et constituent le matériau de départ de l'analyse post-implantation.

Les séries disponibles étant hétérogènes, tant par leur longueur que par leur structure, l'objectif n'est pas ici de conduire une comparaison exhaustive de l'ensemble des cas. La démarche repose plutôt sur le choix d'une série de référence, retenue comme support principal de la démonstration méthodologique. Ce choix s'appuie sur plusieurs critères : qualité de mesure, absence de valeurs manquantes, stabilité relative de la série dans le temps, longueur suffisante de la période pré-intervention, ainsi que compatibilité avec les méthodes d'analyse retenues dans ce travail. L'objectif est ainsi moins de produire une synthèse comparative de tous les cas disponibles que de montrer, à partir d'un cas principal, comment la stratégie d'évaluation post-implantation peut être mise en œuvre et interprétée dans une situation réelle.

3.3.2.2 Limites des approches actuelles

Les rapports disponibles reposent principalement sur des comparaisons descriptives entre les périodes pré- et post-implantation. Ces approches permettent de constater une évolution apparente d'un indicateur, mais ne suffisent pas, à elles seules, à attribuer de manière crédible cette évolution à l'intervention. En particulier, elles ne tiennent ni compte de la dynamique préexistante de la série, ni de la structure temporelle des données, ni de l'incertitude associée aux écarts observés.

Cette limite est particulièrement importante lorsque les fenêtres temporelles retenues diffèrent d'un cas à l'autre, voire au sein d'un même rapport. Selon les périodes considérées, un même indicateur peut ainsi apparaître en amélioration ou en détérioration, ce qui fragilise l'interprétation des comparaisons avant/après. En outre, l'absence de quantification explicite de l'incertitude rend

difficile la distinction entre un changement structurant et une fluctuation compatible avec la variabilité naturelle des données.

L'enjeu ne porte donc pas principalement sur le choix des indicateurs, mais sur la méthode retenue pour interpréter rigoureusement leur évolution après implantation. Ces limites motivent le recours à une stratégie d'analyse explicitement structurée autour de l'attribution causale plausible, plutôt qu'à une simple comparaison descriptive entre périodes.

3.3.2.3 Approche post-implantation retenue

Les indicateurs analysés ayant déjà été définis dans les rapports disponibles, l'enjeu ne porte pas ici sur leur sélection, mais sur la méthode retenue pour analyser rigoureusement leur évolution après implantation.

L'objectif n'est pas uniquement de décrire une différence entre les périodes pré- et post-implantation, mais d'évaluer dans quelle mesure une modification observée peut être plausiblement attribuée à l'intervention [20]. Dans un cadre idéal, une telle évaluation reposerait sur un essai contrôlé randomisé. Toutefois, la nature organisationnelle du logiciel, déployé à l'échelle d'un service ou d'un établissement, rend ce type de dispositif peu réaliste. En l'absence de randomisation, l'évaluation post-implantation relève donc d'une logique quasi-expérimentale, fréquemment mobilisée lorsque les contraintes organisationnelles rendent un essai contrôlé difficile ou irréaliste.

La méthode principale retenue pour le volet post-implantation est l'analyse de séries temporelles interrompues (Interrupted Time Series, ITS), couramment utilisée pour évaluer des interventions mises en œuvre à un moment clairement identifié lorsqu'un essai randomisé n'est pas envisageable. Ce choix repose sur trois éléments. Premièrement, les données disponibles sont organisées sous forme de série temporelle avant-après, avec un point d'implantation connu. Deuxièmement, aucun groupe contrôle suffisamment comparable n'est disponible, ce qui limite la pertinence d'une approche en différences-in-différences. Troisièmement, l'objectif n'est pas de produire une preuve expérimentale, mais d'évaluer si les changements observés après implantation sont compatibles

avec l'intervention, sous des hypothèses explicites. Dans ce contexte, l'ITS offre le meilleur compromis entre adéquation aux données, lisibilité méthodologique et robustesse interprétative.

Cette approche consiste à étudier si la trajectoire d'un indicateur se modifie au moment de l'implantation du logiciel. Plus précisément, elle permet d'examiner le niveau initial de l'indicateur, la tendance observée avant l'intervention, une éventuelle rupture immédiate au moment du déploiement, ainsi qu'une possible modification de la tendance après implantation. L'idée est ainsi de comparer la trajectoire observée après intervention à celle qui aurait été attendue si la dynamique préexistante s'était poursuivie [13, 14, 21].

À partir de la série de référence décrite précédemment, plusieurs spécifications de modèles en série temporelle interrompue sont d'abord estimées afin d'examiner si la date d'implantation s'accompagne d'une rupture de niveau, d'une modification de tendance, ou des deux à la fois. L'analyse principale vise ainsi à quantifier l'ampleur de la rupture observée et à vérifier si celle-ci est cohérente avec la date d'intervention.

L'interprétation de cette analyse repose néanmoins sur plusieurs hypothèses [3, 14]. Elle suppose notamment une relative stabilité de la dynamique avant intervention, l'absence de choc majeur survenant exactement au moment de l'implantation, ainsi qu'une mesure cohérente de l'indicateur dans le temps. Dans ce cadre, l'objectif n'est pas d'établir une causalité expérimentale stricte, mais de construire une attribution causale plausible à partir de données observationnelles.

Afin de renforcer la robustesse de l'analyse principale, plusieurs méthodes complémentaires sont mobilisées. Un test de Mann-Whitney permet d'évaluer l'existence d'une différence globale entre les distributions observées avant et après implantation [22]. Des ruptures placebo sont ensuite introduites afin de vérifier que des ruptures d'ampleur comparable n'apparaissent pas à d'autres dates arbitraires [23]. Enfin, une approche bayésienne complète les analyses fréquentistes en apportant une lecture probabiliste de la différence observée entre les périodes pré- et post-implantation, ainsi qu'une quantification explicite de l'incertitude [24]. Leur mobilisation conjointe vise ainsi à mieux caractériser l'effet observé, sa robustesse et la portée de son

interprétation, sans pour autant revendiquer une preuve causale définitive.

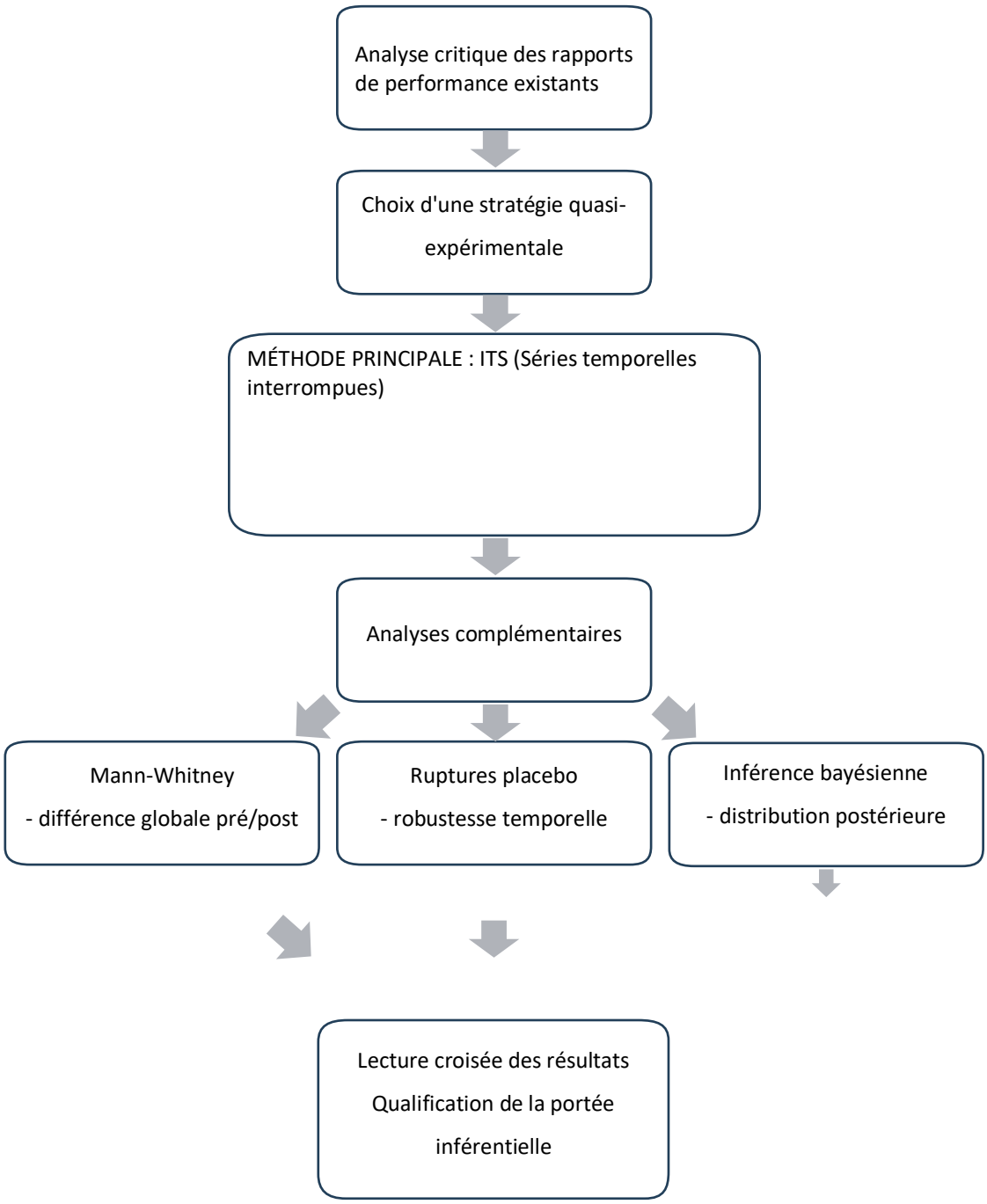

Figure 3.2 Schéma méthodologique de l'analyse post-implantation

3.4 Différence de portée entre les deux cadres

Les deux cadres d'estimation présentés ci-dessus n'autorisent pas le même type d'interprétation. En pré-implantation, l'analyse vise à produire un ordre de grandeur plausible des effets attendus sous certaines hypothèses de modélisation et d'usage. En post-implantation, elle cherche à déterminer si les évolutions observées sont compatibles avec un effet de l'intervention, sans pour autant revendiquer une identification causale expérimentale stricte.

Cette distinction est essentielle pour ajuster l'ambition de l'évaluation aux données effectivement disponibles. En amont du déploiement, l'enjeu est d'éclairer la décision à partir de scénarios plausibles ; après implantation, il s'agit d'interpréter un effet observé à partir d'un contrefactuel reconstruit de manière plus ou moins crédible. Le cadre retenu dans ce mémoire vise ainsi moins à défendre une méthode unique qu'à assurer une cohérence entre l'objectif de l'évaluation, les données mobilisées, les hypothèses retenues et la portée des conclusions envisageables.

Le tableau 3.1 synthétise ces différences de logique méthodologique, de données, de menaces de validité et de type de conclusion.

Tableau 3.1 Comparaison des méthodes pré- et post-implantation

	Pré-implantation (ex-ante)	**Post-implantation (ex-post)**
Objectif	Produire un **ordre de grandeur plausible** des gains *si* l'outil est utilisé et déclenche un ajustement organisationnel.	Estimer l'effet observé compatible avec l'implantation et en évaluer la robustesse, en l'absence de groupe contrôle.
Données disponibles	Données historiques **pré-implantation uniquement** (ex. 2018–2021), variables opérationnelles, indicateurs (congestion, délais).	Série temporelle d'indicateurs avec **période pré + post** et date d'implantation (T_0).

Tableau 3.1 Comparaison des méthodes pré- et post-implantation (suite et fin)

	Pré-implantation (ex-ante)	**Post-implantation (ex-post)**
Référence	**Trajectoire réellement observée**	Trajectoire contrefactuelle attendue sans intervention, principalement approchée par prolongation de la dynamique pré-intervention.
Trajectoire sous intervention	**Scénario(s) d'usage** : trajectoire simulée sous hypothèse que l'outil conduit à des ajustements (ressources, flux, décisions).	Trajectoire **réellement observée** après implantation.
Menaces de validité (principales)	Scénarios trop optimistes; changements structurels non anticipés; erreurs de calibration/transportabilité.	Choc externe simultané; saisonnalité non modélisée; autocorrélation mal traitée
Type de conclusion	**Plausibilité décisionnelle** : "ordre de grandeur attendu *si* …" (pas une preuve causale observée).	**Attribution causale plausible** : effet local compatible avec l'intervention, renforcé par robustesse temporelle (pas causalité expérimentale).
Sorties décisionnelles	Go/No-Go, priorisation, dimensionnement des ressources, estimation coûts/bénéfices attendus, conditions de succès (adoption).	Maintien/extension, ajustements de déploiement, estimation de l'effet réel, quantification de l'incertitude, limites et besoins de données/contrôles.

Les deux volets présentés dans ce chapitre sont ensuite mis en œuvre dans les chapitres suivants : le chapitre 4 est consacré à l'analyse pré-implantation, tandis que le chapitre 5 présente l'analyse post-implantation.

CHAPITRE 4 RÉSULTATS VOLET PRÉ-IMPLANTATION

Le présent chapitre met en œuvre le volet pré-implantation du cadre méthodologique proposé dans ce mémoire, à partir du cas d'un algorithme de prédiction de la surcharge des urgences au Centre hospitalier de l'Université de Montréal (CHUM). L'enjeu n'est pas d'observer un effet déjà réalisé, mais d'estimer, avant déploiement, les effets plausibles que pourrait produire un tel dispositif dans des conditions d'usage organisationnel réalistes.

Cette démarche suppose d'abord la construction d'un protocole prédictif permettant de générer un signal de congestion suffisamment robuste pour être mobilisé dans la suite de l'analyse. Un tel travail ne relève pas encore de l'estimation des effets proprement dite, mais constitue un préalable méthodologique indispensable à l'analyse pré-implantation.

Dans cette perspective, le chapitre procède d'abord à la construction du signal prédictif à partir des données historiques des urgences, en précisant le phénomène étudié, les variables disponibles et les choix méthodologiques retenus pour le pipeline. Il examine ensuite dans quelle mesure ce signal peut soutenir un usage décisionnel plausible au CHUM, avant d'en déduire une estimation structurée des effets attendus de plusieurs scénarios d'ajustement de ressources sur la congestion, les délais et les coûts associés.

4.1 Fonctionnement et congestion des urgences

4.1.1 Fonctionnement des urgences

Le service des urgences se distingue des autres services hospitaliers par un accès libre et non programmé, en réponse à des problèmes de santé aigus, graves ou imprévus. La prise en charge repose à la fois sur l'ordre d'arrivée des patients et sur le niveau de gravité évalué au triage. Ce niveau de gravité est déterminé à l'aide de l'échelle de triage présentée ci-dessous.

Tableau 4.1 Échelle de priorité aux urgences (1 à 5)

Niveaux	Délais de prise en charge médicale	Détails
1	Prise en charge immédiate	**Réanimation** : conditions mettant en jeu le pronostic vital ou l'intégrité d'un membre, nécessitant une intervention énergique et immédiate.
2	15 minutes	**Très urgent** : conditions menaçant la vie, l'intégrité d'un membre ou la fonction, exigeant une intervention médicale rapide.
3	30 minutes	**Urgent** : conditions souvent associées à un inconfort important ou à une incapacité à accomplir les activités quotidiennes.
4	60 minutes	**Moins urgent** : conditions variables selon l'âge et l'état du patient, présentant des risques de détérioration ou de complications.
5	120 minutes	**Non urgent :** conditions pouvant être aiguës, anciennes, ou s'inscrivant dans un problème chronique

Après leur arrivée, les patients sont ainsi orientés selon leur état : certains attendent une première évaluation médicale, tandis que les cas les plus graves peuvent être pris en charge immédiatement. Le médecin urgentiste assure ensuite l'évaluation initiale, prescrit les examens et traitements nécessaires, puis décide d'un retour à domicile, d'une observation, d'une orientation vers un spécialiste ou d'une admission à l'hôpital. Lorsque la situation l'exige, d'autres intervenants, notamment des médecins spécialisés, participent à la prise en charge.

Le chemin classique d'un patient aux urgences peut être résumé de la façon suivante :

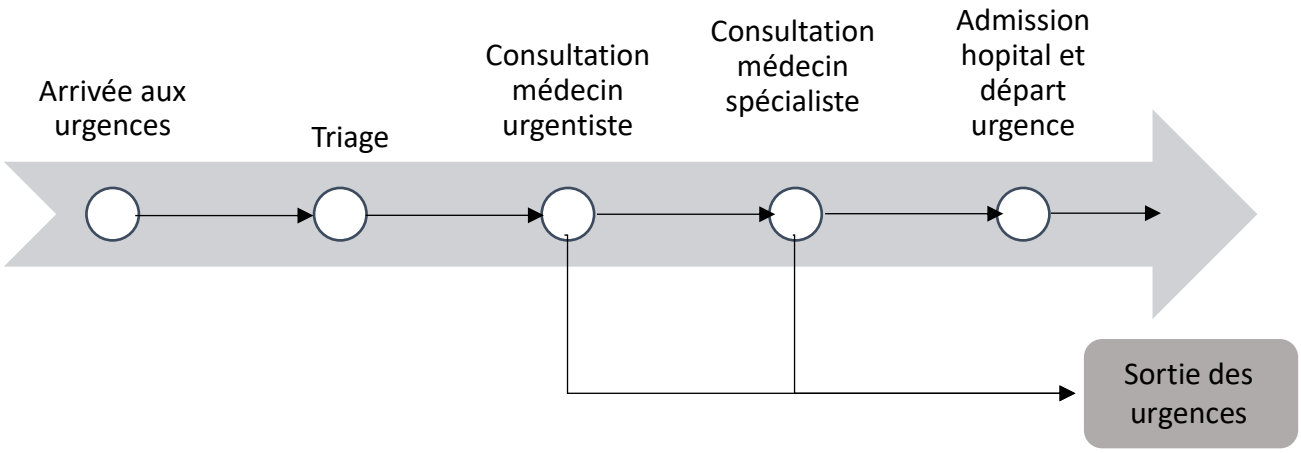

Figure 4.1 Étapes clés du parcours patient aux urgences

4.1.2 Engorgement des urgences

Dans un tel système, l'engorgement des urgences ne résulte pas d'un facteur unique, mais de l'accumulation de tensions à différentes étapes du parcours patient. Il peut survenir lorsque les flux d'arrivée excèdent temporairement la capacité de prise en charge, mais aussi lorsque certaines étapes en aval, notamment l'hospitalisation, ralentissent la sortie des patients du service. La congestion apparaît ainsi comme le produit d'un déséquilibre global entre les entrées, les ressources mobilisables et les possibilités de sortie.

Afin d'objectiver ce phénomène, Eitel propose en 2003 le score d'EDWIN (*Emergency Department Work Index*), conçu comme un indicateur synthétique du niveau de congestion des urgences. Ce score intègre plusieurs dimensions classiques de l'activité, notamment le nombre de patients par niveau de triage, le nombre de médecins disponibles, le nombre de civières et le nombre de patients hospitalisés encore présents aux urgences. Il est défini par :

$$EDWIN = \frac{\sum_{i=1}^{n} N_i \, T_i}{N_{\text{total}}(B - P)}$$

où N_i désigne le nombre de patients dans la catégorie de triage i, T_i le niveau de gravité associé, N_{total} le nombre total de médecins, B le nombre de civières disponibles et P le nombre de patients hospitalisés encore présents au service.

Dans le cadre de ce mémoire, le score d'Edwin n'est pas une mesure directement relevée dans la base de données, mais un indicateur reconstruit à partir de plusieurs variables décrivant l'activité des urgences. Son utilisation permet de traduire les données brutes du service — en particulier la répartition des patients selon le triage, les ressources médicales disponibles et l'occupation des civières — en une mesure synthétique de la congestion.

Dès lors, l'enjeu n'est pas seulement de mesurer cette surcharge en temps réel, mais aussi d'en reconstruire l'évolution dans le temps afin d'anticiper suffisamment tôt les épisodes critiques. C'est dans cette perspective que s'inscrit le pipeline prédictif étudié dans ce mémoire.

4.2 Données disponibles

Le présent chapitre s'appuie sur une base de données historique des urgences du CHUM, relevée toutes les 15 minutes entre juillet 2018 et décembre 2021. Cette base, intitulée `var_tot`, regroupe les principales informations relatives au fonctionnement du service, notamment le triage, les consultations, les examens cliniques, les admissions à l'hôpital ainsi que le score d'Edwin. Elle constitue le point de départ empirique du pipeline prédictif présenté dans la suite.

Cette base provient de travaux préalables ayant permis la collecte et le traitement initial des données du CHUM. Dans le présent mémoire, elle est utilisée comme support de départ pour la construction du pipeline prédictif retenu. Une description plus détaillée des variables et de leur organisation figure en annexe A.1.

4.3 Protocole prédictif de la surcharge des urgences

Le pipeline prédictif retenu dans ce mémoire vise à répondre à un double objectif : produire un signal suffisamment fiable pour anticiper les épisodes de congestion et construire un outil cohérent avec un usage opérationnel.

La démarche est présentée en deux temps. Un premier protocole est d'abord construit à partir des données observées à une fréquence de 15 minutes. Les résultats obtenus conduisent ensuite à une reformulation du problème à une échelle temporelle plus agrégée, à partir de laquelle est défini le modèle final retenu.

4.3.1 Construction des variables retardées

Dans un premier temps, le pipeline prédictif est construit à partir des données observées toutes les 15 minutes. Afin de tenir compte du fait que les variables explicatives n'influencent pas nécessairement immédiatement la congestion future, des versions retardées de chaque variable numérique sont générées. Pour chaque variable X_j, plusieurs lags candidats sont considérés dans une fenêtre de 24 heures précédant l'instant de prédiction, et le lag optimal L_j^* est retenu comme celui qui maximise la corrélation avec la cible future. Seule la version retardée correspondant à ce décalage est ensuite conservée pour l'apprentissage. Cette procédure permet de représenter, pour chaque variable, l'effet différé le plus informatif sur la cible, tout en évitant de conserver plusieurs versions laguées fortement corrélées d'une même variable. Afin d'éviter toute fuite d'information, la sélection des lags est effectuée uniquement sur le jeu d'entraînement.

Le calcul de ces lags est détaillé en annexe A.2.

4.3.2 Critères d'évaluation retenus

Afin d'obtenir une lecture complète des performances du modèle, plusieurs métriques issues de la matrice de confusion sont mobilisées. Parmi ces métriques, on retrouve :

- **La précision** : $\frac{TP}{TP+FP}$

La précision correspond au taux de bonne prédiction parmi les prédictions positives. Cette métrique est d'autant plus pertinente que le coût d'un faux positif est cher.

- La valeur prédictive négative : $\frac{TN}{TN+FN}$

La valeur prédictive négative est le taux de bonnes prédictions parmi les prédictions négatives, c'est l'équivalent côté négatif de la prédiction.

- **La sensibilité** : $\frac{TP}{TP+FN}$

La sensibilité représente le taux d'observations positives détectées, elle répond à la question : parmi les observations positives, combien en ont été détectées ? Cette métrique est cruciale pour rater le moins possible d'observations positives.

- **La spécificité** : $\frac{TN}{TN+FP}$

La spécificité est l'équivalent de la sensibilité pour les observations négatives, elle est importante pour limiter les fausses alertes.

- **L'exactitude** : $\frac{VP+VN}{VP+VN+FP+FN}$

L'exactitude représente la proportion de prédictions correctes (vrais positifs et vrais négatifs) parmi l'ensemble des cas testés par le modèle.

À ces mesures s'ajoute le ROC-AUC, retenu comme critère principal pour le tuning des hyperparamètres, car il permet d'évaluer la capacité du modèle à ordonner correctement les périodes selon leur probabilité de congestion, sans dépendre d'un seuil de décision fixé a priori. Ce choix est cohérent avec un problème de classification binaire dans lequel l'objectif est avant tout de caractériser la séparation entre périodes congestionnées et non congestionnées.

4.3.3 Choix du seuil orienté décision

Une fois le meilleur modèle sélectionné au sens du ROC-AUC, le seuil de classification τ est déterminé a posteriori à partir des prédictions conservées lors de la validation croisée temporelle (*out-of-fold*). Cette étape est volontairement dissociée du tuning des hyperparamètres afin d'éviter

toute fuite d'information et de distinguer clairement deux problèmes de nature différente : d'une part, l'identification d'un modèle ayant une bonne capacité de discrimination ; d'autre part, le choix d'une règle de décision adaptée au contexte opérationnel.

Dans cette perspective, le choix du seuil peut être formulé comme un problème de décision. En effet, τ détermine directement l'arbitrage entre deux types d'erreurs : les faux positifs, lorsque le modèle anticipe une congestion qui ne se produit pas, et les faux négatifs, lorsqu'une congestion réelle n'est pas détectée. Dans le contexte étudié, ces deux erreurs n'ont pas les mêmes conséquences. Un faux positif peut conduire à mobiliser inutilement des ressources limitées, par exemple du personnel médical ou du matériel supplémentaire. À l'inverse, un faux négatif peut entraîner une sous-anticipation d'un épisode de surcharge, avec pour conséquence une dégradation des délais de prise en charge et des conditions de fonctionnement du service.

En associant un coût à chacun de ces deux types d'erreur, C_{FP}et C_{FN}, il est possible de formuler le choix du seuil comme la recherche d'un compromis minimisant un coût opérationnel attendu :

$$\tau^* = \arg\min_{\tau} \left[C_{FP}\, \mathbb{P}(\hat{Y}_\tau = 1, Y = 0) + C_{FN}\, \mathbb{P}(\hat{Y}_\tau = 0, Y = 1) \right].$$

En pratique, les valeurs exactes de C_{FP} et C_{FN} ne sont pas connues. Un critère empirique cohérent avec les contraintes du terrain a donc été retenu. Une spécificité minimale de 80 % est d'abord imposée afin de limiter le nombre de fausses alertes, puis, parmi les seuils satisfaisant cette contrainte, on choisit celui qui maximise la sensibilité. Cette règle reflète l'idée qu'un système générant trop d'alertes injustifiées risquerait d'être rapidement perçu comme peu crédible ou peu utile par les équipes, ce qui nuirait à son acceptabilité opérationnelle. À l'inverse, sous cette contrainte de spécificité, il demeure souhaitable de détecter le plus grand nombre possible d'épisodes réels de congestion.

Ainsi, le seuil n'est pas choisi pour maximiser mécaniquement une métrique agrégée, mais pour traduire un compromis décisionnel explicite entre fausses alertes et épisodes de congestion manqués. Cette approche inscrit le pipeline dans une logique cohérente avec l'objectif opérationnel

de détection précoce des surcharges, tout en assurant une construction méthodologique compatible avec cet usage.

4.3.4 Résultats du protocole à 15 minutes

Les choix méthodologiques présentés ci-dessus définissent un premier protocole prédictif appliqué à la base conservée à une fréquence de 15 minutes. Les résultats obtenus permettent d'évaluer dans quelle mesure cette première formulation du problème est compatible avec l'objectif opérationnel visé.

Tableau 4.2 Performances du modèle du Gradient Boosté

	Références		
Prédictions	Oui	Non	
Oui	9575	397	Précision = 86 %
Non	8727	16341	VPN = 65 %
	Sensibilité = 52 %	Spécificité = 98 %	Exactitude = 73 %

Les résultats montrent que le modèle privilégie fortement la réduction des faux positifs, comme en témoignent une précision de 86 % et une spécificité de 98 %. En revanche, sa sensibilité de 52 % indique qu'il ne détecte qu'un peu plus de la moitié des épisodes réels de congestion. Ainsi, bien que l'exactitude globale atteigne 73 %, le modèle apparaît trop conservateur pour un objectif opérationnel de détection, dans la mesure où il laisse échapper un nombre important de périodes réellement surchargées.

Ce résultat ne remet pas en cause l'intérêt du pipeline prédictif, mais suggère que la formulation temporelle initiale du problème n'est pas la plus adaptée à l'objectif poursuivi. Le protocole est donc révisé en modifiant l'échelle d'agrégation des données, afin d'obtenir une représentation plus compatible avec la dynamique de la congestion et avec l'horizon de prédiction retenu.

4.3.5 Reformulation temporelle du problème

Le pas de temps initial de 15 minutes apparaît en effet très fin au regard de l'horizon de prédiction fixé à 24 heures, et capte des micro-variations du score d'Edwin qui ne sont pas nécessairement les plus informatives pour l'anticipation de la congestion. Les données ont donc été agrégées à un pas de 8 heures afin d'obtenir une représentation plus stable du phénomène et plus cohérente avec la dynamique opérationnelle du service. Cette reformulation a également nécessité une reconstruction des variables retardées dans ce nouveau cadre temporel.

4.3.6 Résultats du modèle final à 8 heures

Une fois cette reformulation opérée, le protocole est réappliqué dans ce nouveau cadre temporel afin d'évaluer si cette représentation agrégée conduit à un comportement prédictif plus adapté à l'usage visé.

Tableau 4.3 Matrice de confusion du modèle sur période de 8 h

	Références		
Prédictions	Oui	Non	
Oui	540	82	Précision = 86,8 %
Non	92	381	VPN = 80,5 %
	Sensibilité = 85,4 %	Spécificité = 82,3 %	Exactitude = 84 %

Avec une sensibilité de 85,4 % et une spécificité de 82,3 %, le classificateur parvient à détecter la majorité des épisodes réels de congestion tout en maintenant un niveau de fausses alertes encore acceptable. La précision de 86,8 % et la valeur prédictive négative de 80,5 % confirment que les prédictions positives comme négatives sont globalement plus fiables. Ainsi, le modèle final à 8 heures présente un compromis plus pertinent pour un objectif opérationnel de détection des surcharges.

4.3.7 Comparaison des performances selon la fréquence

Les deux modèles ayant été construits selon une logique identique mais à des fréquences différentes, leur comparaison permet d'isoler l'effet de la reformulation temporelle sur les performances observées.

Tableau 4.4 Tableau comparatif des performances en fonction de la fréquence

Métrique	15 min	8 h
Prévalence	52 %	**57 %**
Exactitude	73 %	**84.1 %**
Precision / PPV	**96 %**	**86.8** %
Recall / Sensibilité	52 %	**85.4 %**
Spécificité	**98 %**	82.3 %
NPV	65 %	**80.5 %**

La comparaison des deux modèles met en évidence un changement net de compromis entre détection et fausses alertes. Le modèle à 15 minutes privilégie fortement la spécificité et la précision, ce qui traduit un comportement très conservateur : il génère peu de fausses alertes, mais manque une part importante des épisodes réels de congestion. À l'inverse, le modèle à 8 heures apparaît beaucoup plus équilibré. Il détecte une proportion bien plus importante des épisodes de surcharge, tout en maintenant un niveau de spécificité et de précision encore satisfaisant. L'amélioration la plus marquante concerne donc la sensibilité et, plus largement, la fiabilité globale des prédictions négatives, ce qui rend le modèle final plus pertinent dans une logique opérationnelle de détection précoce. En contrepartie, cette meilleure capacité de détection s'accompagne d'une hausse du nombre de fausses alertes, mais ce compromis semble plus adapté à l'objectif visé.

Au-delà de la reformulation temporelle, une dernière extension exploratoire consiste à tester l'apport d'informations externes au fonctionnement hospitalier. Cette analyse vise à déterminer si

l'enrichissement de la base par des variables exogènes modifie sensiblement les performances du modèle retenu.

4.3.8 Intégration exploratoire de variables exogènes

Enfin, une extension exploratoire du pipeline a consisté à enrichir la base analytique par des variables exogènes, en particulier des variables météorologiques et de pollution. Ces variables ont été intégrées dans le même cadre temporel que les variables hospitalières, de manière à permettre leur agrégation et, le cas échéant, la construction de versions retardées comparables aux autres prédicteurs. L'objectif de cette étape était d'évaluer si des sources d'information externes pouvaient améliorer la prédiction de la congestion au-delà des seules variables issues du fonctionnement hospitalier.

4.3.9 Effet de l'ajout de variables exogènes

L'apport empirique des variables exogènes ajoutées s'est révélé limité : les performances du modèle n'ont été que très faiblement modifiées. Ces variables n'ont donc pas été retenues dans la version finale.

Ce résultat suggère que l'information la plus utile pour la prédiction de la congestion était déjà majoritairement contenue dans les variables hospitalières et dans leur dynamique temporelle. Autrement dit, dans le cadre étudié, l'amélioration du modèle provient davantage du traitement temporel des données et du choix du seuil que de l'ajout de nouvelles sources exogènes.

4.3.10 Synthèse

Le protocole retenu aboutit ainsi à un modèle plus équilibré et plus cohérent avec un objectif opérationnel de détection anticipée de la congestion. La principale évolution concerne la sensibilité, qui augmente nettement entre les deux versions, tandis que la spécificité demeure à un niveau compatible avec l'usage visé. Ce résultat repose toutefois sur l'hypothèse que les variables mobilisées capturent une part suffisante de la dynamique du système pour permettre une prédiction

pertinente du niveau de congestion. Dans le cas étudié, le faible apport empirique des variables exogènes suggère que l'information la plus utile est déjà majoritairement contenue dans les variables hospitalières et dans leur structure temporelle.

Le modèle prédictif ainsi obtenu fournit un signal cohérent avec un objectif de détection anticipée. Il reste toutefois à déterminer dans quelles conditions ce signal peut être converti en action organisationnelle plausible. C'est l'objet de l'analyse pré-implantation présentée dans la section suivante.

4.4 Analyse pré-implantation

L'analyse pré-implantation vise à déterminer dans quelle mesure cet algorithme de prédiction de la surcharge des urgences peut soutenir un usage organisationnel plausible au CHUM. L'enjeu est d'examiner si le signal produit peut être converti en décision opérationnelle susceptible d'en atténuer les effets. Dans cette perspective, la présente section vise à expliciter le mécanisme d'action retenu dans le cas étudié et à en discuter la plausibilité.

4.4.1 Problème décisionnel et finalité de l'intervention

L'intérêt du protocole prédictif ne réside pas dans la seule détection d'un épisode de congestion, mais dans sa capacité à fournir suffisamment tôt une information exploitable pour permettre une action anticipée. Dans un contexte où la surcharge est associée à un allongement des délais de prise en charge et à une pression accrue sur les équipes, la question décisionnelle centrale devient la suivante : un signal de congestion produit 24 heures à l'avance peut-il soutenir une décision de gestion susceptible de réduire l'intensité de la surcharge ?

4.4.2 Contexte organisationnel et conditions d'usage du signal

Au CHUM, l'intérêt d'un signal prédictif de congestion dépend moins de sa seule performance statistique que de sa capacité à s'insérer dans un fonctionnement déjà fortement contraint. L'activité des urgences se caractérise par un volume élevé de patients, une variabilité marquée des

arrivées et une interdépendance forte entre les différentes étapes du parcours, de sorte qu'un épisode de surcharge résulte rarement d'un facteur isolé. Dans ce contexte, un signal prédictif n'a d'utilité que s'il permet d'anticiper une dégradation suffisamment tôt pour rendre possible un ajustement concret de l'organisation du service.

Cette possibilité repose sur l'existence d'un usage décisionnel plausible du signal. Les gestionnaires des urgences jouent ici un rôle central, dans la mesure où ce sont eux qui peuvent interpréter l'alerte produite par le modèle et la relier à un ajustement de ressources. L'enjeu n'est donc pas uniquement technique : il tient à la capacité du signal à être intégré dans une pratique effective de pilotage des urgences.

Cette chaîne d'usage demeure toutefois contrainte. La variabilité des flux, l'influence de facteurs externes difficilement anticipables, la charge de travail des équipes et la dépendance à des ressources qui ne relèvent pas toujours directement du service réduisent la marge de manœuvre réellement disponible. À l'inverse, la disponibilité de données opérationnelles riches et structurées, combinée à un contexte institutionnel favorable à l'innovation, rend crédible l'examen de scénarios d'usage fondés sur ce signal. L'objectif n'est donc pas de postuler qu'un effet se produira mécaniquement, mais d'identifier dans quelles conditions organisationnelles un tel effet peut raisonnablement être envisagé.

4.4.3 Mécanisme d'action retenu

L'analyse pré-implantation conduit à envisager l'intervention non comme l'ajout d'un simple algorithme, mais comme un dispositif d'aide à la décision reposant sur l'exploitation d'un signal prédictif. Le mécanisme d'action retenu est le suivant : un épisode de surcharge est anticipé par l'algorithme, cette anticipation peut conduire à un ajustement de ressources, puis cet ajustement est susceptible d'atténuer la congestion. L'effet potentiel du dispositif dépend ainsi de la possibilité de convertir le signal produit en action opérationnelle.

Dans ce contexte, trois actions organisationnelles sont envisagées :

- Scénario 1 : ajout d'un médecin, $N_{1\_scé} = N_{med} + 1$
- Scénario 2 : ajout d'une civière $B_{2\_scé} = B + 1$
- Scénario 3 : ajout d'un médecin et d'une civière $\begin{cases} B_{3\_scé} = B + 1 \\ N_{3\_scé} = N_{med} + 1 \end{cases}$

Où N_{med} correspond au nombre de médecin urgentiste présents et B au nombre de civières.

Ces scénarios constituent des configurations stylisées destinées à explorer l'effet potentiel d'ajustements organisationnels ciblés.

4.4.4 Indicateurs retenus et définition de la situation de référence

L'analyse pré-implantation s'appuie sur des indicateurs choisis en fonction des deux objectifs poursuivis dans ce mémoire : un objectif sanitaire, centré sur l'évolution de la congestion et des délais de prise en charge, et un objectif économique, centré sur l'estimation simplifiée des coûts susceptibles d'être évités. Dans cette perspective, le score d'Edwin est mobilisé comme indicateur synthétique de surcharge, puisqu'il correspond à la cible prédite par l'algorithme et constitue le point d'entrée de la simulation des scénarios. Cette mesure est ensuite complétée par des indicateurs de délais d'attente, plus directement interprétables du point de vue du fonctionnement du service, puis par une estimation des coûts associés aux variations observées. La situation de référence est construite à partir des données historiques observées avant toute intervention et sert de base à la simulation des effets plausibles des scénarios étudiés.

4.4.5 Portée de l'analyse pré-implantation dans ce mémoire

Ainsi, l'analyse pré-implantation a conduit à examiner la plausibilité du mécanisme d'action retenu et à introduire une situation de référence construite à partir des données historiques observées avant toute intervention. Cette référence sert de base à la simulation des effets plausibles des scénarios envisagés. Sa validité repose toutefois sur l'hypothèse qu'aucun choc externe majeur non modélisé n'affecte simultanément la congestion et les variables explicatives sur la période étudiée. Dans le cas présent, cette hypothèse est partiellement mise à l'épreuve par l'inclusion de la période de la

Covid-19, qui a pu modifier certaines relations observées dans les données. Sous cette réserve, l'analyse précédente permet de relier la performance prédictive du modèle à un usage organisationnel plausible dans le contexte du CHUM.

Une fois cette chaîne d'action explicitée et la situation de référence définie, l'enjeu n'est plus seulement de justifier l'intérêt potentiel du dispositif, mais d'en estimer les effets attendus. La section suivante porte donc sur la quantification de cet effet potentiel, à travers plusieurs scénarios d'ajustement de ressources et leurs conséquences plausibles sur la congestion, les délais de prise en charge et les coûts associés.

4.5 Calcul du score d'Edwin pour les scénarios envisagés

Afin d'étudier l'effet potentiel des scénarios présentés lors de l'analyse pré-implantation, il est nécessaire de revenir à l'indicateur utilisé par le protocole prédictif Edwin. L'examen de sa formule montre qu'aucun temps d'attente n'est pris en compte : le calcul repose uniquement sur des variables opérationnelles, telles que le nombre de médecins, le nombre de civières et le nombre de patients.

À partir de cette formulation, il devient possible de recalculer, pour chaque observation de la base de données, la valeur qu'aurait prise le score d'Edwin sous chacun des scénarios envisagés. Il s'agit toutefois d'une simulation contrefactuelle simplifiée : l'objectif n'est pas d'observer empiriquement les effets réels d'un ajout de ressources, mais d'estimer, de manière mécanique, comment le score aurait évolué si certaines composantes de la formule avaient été modifiées, toutes choses étant égales par ailleurs. Autrement dit, l'analyse suppose que la structure générale de la situation observée demeure inchangée et que seul l'ajustement considéré influe sur la valeur de l'indicateur.

Dans cette perspective, les trois scénarios étudiés sont simulés en modifiant directement les variables concernées dans la formule du score. Pour chaque observation, quatre valeurs sont ainsi considérées : le score observé et les trois scores contrefactuels associés aux scénarios d'ajustement de ressources.

$$\text{Le score observé : } \frac{\sum_{i=1}^{n} N_i * T_i}{N_{med} * (B-P)}$$

$$\text{Score du scénario 1 : } \frac{\sum_{i=1}^{n} N_i * T_i}{(N_{med}+1)(B-P)}$$

$$\text{Score du scénario 2 : } \frac{\sum_{i=1}^{n} N_i * T_i}{N_{med}(B+1-P)}$$

$$\text{Score du scénario 3 : } \frac{\sum_{i=1}^{n} N_i * T_i}{(N_{med}+1)(B+1-P)}$$

Compte tenu des ordres de grandeur observés, on s'attend à ce que l'ajout d'un médecin ait un impact plus important sur le score que l'ajout d'une civière. En effet, le nombre de médecins est de l'ordre de 6, tandis que le nombre de civières (via B−PB - PB−P) est de l'ordre de 60, ce qui implique qu'une variation unitaire représente une modification relativement plus importante dans le premier cas.

4.6 Lien score d'Edwin délais

4.6.1 Lien entre le score d'Edwin et les délais

Le score d'Edwin est pratique à calculer mais peu interprétable dans le sens où il ne mesure rien directement mais représente la congestion. Afin d'estimer un effet sanitaire, l'indicateur sélectionné durant l'analyse pré-implantation est : le délai d'attente. Différents délais d'attente peuvent être pris en compte en fonction du parcours du patient. Afin d'avoir une idée des délais disponibles, ils ont été représentés en suivant un parcours patient classique.

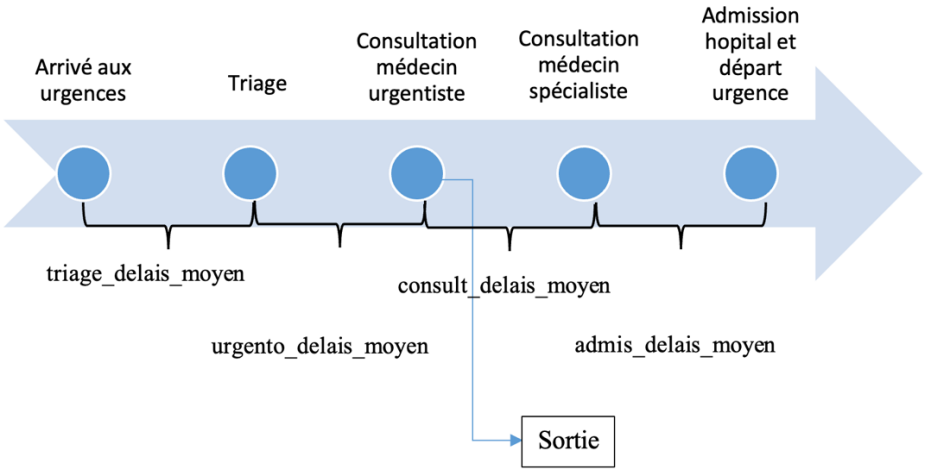

Figure 4.2 Délais d'un parcours patient

Ces délais sont calculés entre deux étapes du parcours patient. Afin d'avoir des délais depuis l'entrée du patient aux urgences, nous devons sommer les délais des différentes étapes. Ainsi nous construisons trois nouveaux délais, le délai de consultation d'un médecin urgentiste (*delai_urgento)*, le délai de consultation d'un médecin spécialisé (*delai_cons*) et enfin le délai total en cas d'admission (*delai_tot*). Afin de vérifier le lien entre ces délais et le score d'Edwin, il a fallu calculer l'AUC, la corrélation de Spearman ainsi que l'instabilité temporelle (auc_sd) entre tous les délais et le score.

	delai	n	auc	rho	auc_sd	na_rate
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	delai_cons	3837	0.850	0.627	0.198	0.0008
2	delai_tot	3837	0.832	0.692	0.160	0.0008
3	delai_urgento	3837	0.821	0.531	0.207	0.0008
4	urgento_delai_moyen	3837	0.819	0.528	0.207	0.0008
5	consult_delai_moyen	3837	0.770	0.511	0.176	0.0008
6	admis_delai_moyen	3837	0.730	0.541	0.158	0.0008
7	sang_delai_moyen	3837	0.689	0.332	0.201	0.0008
8	sangbunch_delai_moyen	3837	0.684	0.310	0.198	0.0008
9	con_med_delai_moyen	3837	0.682	0.332	0.140	0.0008
10	con_chir_delai_moyen	3837	0.677	0.315	0.213	0.0008
11	con_psy_delai_moyen	3837	0.663	0.288	0.180	0.0008
12	con_gyn_delai_moyen	3837	0.662	0.294	0.166	0.0008
13	admis_med_delai_moyen	3837	0.656	0.378	0.135	0.0008
14	triage_delai_moyen	3837	0.632	0.206	0.164	0.0008
15	rx_delai_moyen	3837	0.605	0.221	0.190	0.0008
16	admis_chir_delai_moyen	3837	0.598	0.218	0.161	0.0008

Figure 4.3 Liens entre le score d'Edwin et les variables de délai

Les délais les plus corrélés au score d'Edwin semblent être les délais totaux créés précédemment. Non seulement ces délais sont plus fortement corrélés au score d'Edwin mais ils sont aussi bien plus représentatifs des étapes des patients.

Les trois délais sélectionnés sont logiquement : *delai_urgento* (délai de l'entrée aux urgences jusqu'à la première consultation), *delai_cons* (délai de l'entrée aux urgences jusqu'à consultation d'un médecin spécialisé et enfin *délai_tot* qui correspond au délai entre l'arrivée et l'hospitalisation en cas d'une admission à l'hôpital.

4.6.2 Formalisation du lien Edwin → délais

Les scénarios étudiés permettent, dans un premier temps, d'obtenir une valeur simulée du score d'Edwin. Toutefois, l'indicateur d'intérêt dans cette analyse ne se limite pas au niveau de congestion mesuré par ce score : il porte avant tout sur les délais d'attente, qui constituent une mesure plus directement interprétable du point de vue du fonctionnement du service. Il est donc nécessaire d'établir un lien entre le score d'Edwin et les délais retenus.

Pour ce faire, une relation empirique est construite entre le niveau de congestion observé et les délais considérés. L'objectif est d'estimer, à partir des données historiques, le délai moyen associé à un certain niveau du score d'Edwin. Autrement dit, il s'agit de déterminer comment les délais évoluent lorsque la congestion mesurée par cet indicateur augmente ou diminue.

Plus précisément si l'on note E le score d'Edwin et D le délai d'intérêt, on considère la fonction :

$$m(e) = E[D \mid E = e]$$

où e désigne une valeur, ou plus généralement un intervalle de valeurs, du score d'Edwin. Cette fonction représente le délai moyen observé conditionnellement à un niveau donné de congestion. Elle permet ainsi de traduire une variation du score d'Edwin en une variation attendue du délai étudié.

Les figures suivantes présentent la distribution empirique des délais en fonction du score d'Edwin.

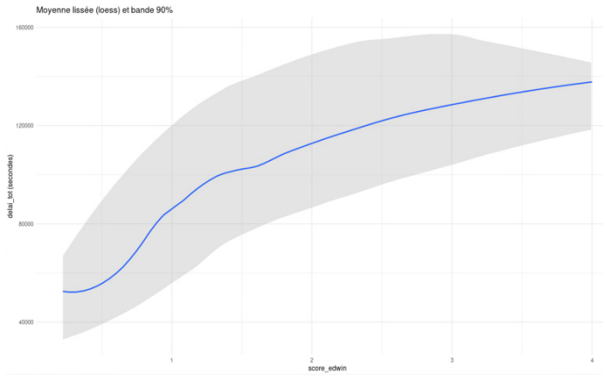

Figure 4.4 Moyenne lissée du délai delai_tot en fonction du score d'Edwin

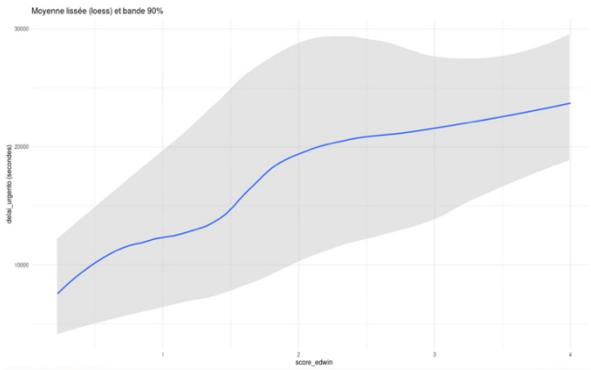

Figure 4.5 Moyenne lissée du délai delai_urgento en fonction du score d'Edwin

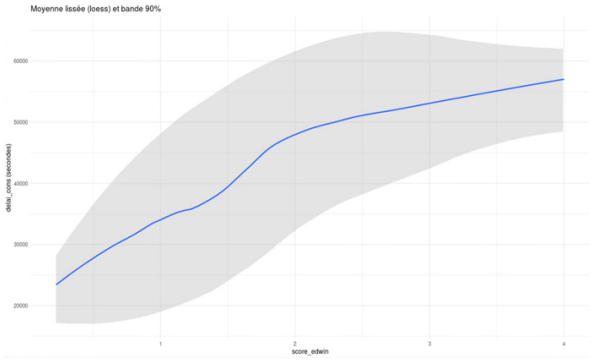

Figure 4.6 Moyenne lissée du délai delai_cons en fonction du score d'Edwin

Bien que la dispersion des observations demeure importante, les figures mettent en évidence une relation globalement positive entre le score d'Edwin et les délais d'attente : plus le niveau de congestion est élevé, plus les délais observés tendent à augmenter. Les courbes des moyennes empiriques présentent ainsi une tendance croissante, ce qui soutient l'hypothèse d'une association positive entre ces deux variables.

On observe par ailleurs une dispersion apparemment plus faible des délais pour les valeurs du score d'Edwin comprises entre 3 et 4. Ce phénomène doit toutefois être interprété avec prudence, car il peut autant refléter une caractéristique des données qu'un effectif plus limité dans cette plage de valeurs.

Ces résultats fournissent la base empirique permettant de relier, dans la suite de l'analyse, les scores d'Edwin observés ou simulés aux délais d'attente correspondants. Il devient ainsi possible d'estimer, pour chaque scénario envisagé, la variation attendue du délai moyen à partir de l'écart entre le score observé et le score simulé.

4.6.3 Hypothèse d'invariance structurelle

L'utilisation de cette relation empirique repose sur une hypothèse importante. On suppose que la relation historique entre le score d'Edwin et les délais d'attente reste approximativement valable lorsque le score d'Edwin est modifié par une intervention sur les ressources (par exemple l'ajout de civières ou de médecins).

Formellement, on suppose que la fonction :

$$E[D \mid E = e]$$

estimée à partir des données historiques constitue une approximation valable de la relation entre congestion et délais dans les scénarios simulés.

Cette hypothèse peut être qualifiée **d'hypothèse d'invariance structurelle**. Elle constitue une approximation permettant de transformer les variations de congestion simulées en ordres de grandeur plausibles de variation des délais.

Il est toutefois important de noter que le score d'Edwin intègre explicitement des composantes de capacité (notamment le nombre de médecins et de civières). Une modification directe de ces ressources pourrait donc modifier la relation entre Edwin et les délais. Les estimations présentées doivent ainsi être interprétées comme **des ordres de grandeur plausibles plutôt que comme des prédictions causales exactes**.

4.7 Estimation des effets plausibles

4.7.1 Effet sur les délais d'attente

Dans la continuité de l'analyse pré-implantation, l'évaluation des scénarios est restreinte aux périodes identifiées comme congestionnées par le modèle (622 périodes pour l'année 2021). Dans le cas du médecin, cette restriction correspond à la mobilisation ponctuelle d'un renfort lorsque le risque de surcharge est anticipé. Dans le cas de la civière, elle ne renvoie pas à une logique d'ajustement instantané, mais à l'examen contrefactuel de l'effet qu'une capacité additionnelle pourrait avoir durant ces mêmes périodes

En appliquant cette règle à l'année la plus récente de la base de données (2021), on obtient les résultats suivants :

Tableau 4.5 Estimation des effets sur les délais

		Edwin_1civ	**Edwin_1med**	**Edwin_civmed**
Délai consultation **médecin urgentiste**	Quantile 10 % (borne pessimiste)	-3.27	18.68	21.79
	Gain de délai moyen	0.63	46.13	48.86
	Quantile 90 % (borne optimiste)	9.35	72.33	75.54
Délai consultation **médecin spécialiste**	Quantile 10 % (borne pessimiste)	-9.59	24.38	28.78
	Gain de délai moyen	1.01	83.71	85.75
	Quantile 90 % (borne optimiste)	18.38	176.53	182.16
Délai total en cas d'**admission**	Quantile 10 % (borne pessimiste)	-53.30	57.66	67.54
	Gain de délai moyen	3.38	225.91	235.48
	Quantile 90 % (borne optimiste)	60.55	359.93	372.59

Les différences de délais ne sont calculées que sur les périodes considérées comme surchargées. De manière générale, on observe que tous les scénarios entraînent une baisse du délai moyen, ce qui constitue un signal encourageant. Toutefois, cette première lecture doit être nuancée : l'interprétation rigoureuse repose non pas sur la seule moyenne, mais sur la borne pessimiste définie par **le quantile 10 % de la distribution des gains simulés**.

Pour le scénario **Edwin_1civ** (ajout d'une civière), bien que la moyenne indique une légère amélioration, le quantile 10 % reste négatif. Cela signifie que, dans les scénarios les plus défavorables, l'intervention pourrait ne produire **aucun gain mesurable**, ce qui invite à interpréter l'amélioration moyenne avec prudence.

À l'inverse, les scénarios intégrant un **médecin supplémentaire** (Edwin_1med et Edwin_civmed) présentent systématiquement des effets beaucoup plus importants sur l'ensemble des délais. Les réductions moyennes sont substantielles, et surtout les quantiles 10 % sont tous **strictement positifs et élevés**, indiquant que l'amélioration persiste même dans les cas défavorables. Par exemple, en cas d'ajout d'un médecin seul, le délai total en cas d'admission diminuerait d'environ **210 minutes en moyenne**, et jusqu'à **356 minutes dans les situations les plus extrêmes** (borne optimiste, quantile à 90 %).

Rapporté au délai moyen observé lors des périodes surchargées (score Edwin > 1.54), le gain moyen représente une **réduction du délai d'attente d'environ 20 %**, ce qui constitue un effet opérationnel majeur pour un service d'urgence.

Maintenant que l'effet sanitaire de tels scénarios a été estimé grâce à une réduction du score d'Edwin et des délais d'attente, il est possible de se pencher sur l'effet économique d'une telle implantation.

4.7.2 Effet économique

Afin de déterminer l'effet économique du logiciel sur le département, il est nécessaire de déterminer les coûts et les bénéfices entraînés par les scénarios. Nous allons commencer par étudier les coûts inhérents aux scénarios présentés pour finir sur les bénéfices entraînés par de telles réductions de délais.

4.7.2.1 Coûts des scénarios

Dans l'évaluation des coûts, seuls les coûts directement liés à l'hôpital sont pris en compte.

- **Scénario 1 :** le coût supplémentaire pour l'ajout d'un médecin urgentiste est simplement calculable. Il suffit de multiplier le coût horaire par le nombre d'heures

$$C = C_{horaire} * n_{heure_{période}} * n_{période}$$

En estimant le coût horaire d'un médecin a 150 $ par heure, cela revient à un coût de $150 * 8 = 1200$ $ par médecin et par période de travail.

- **Scénario 2 :** le coût lié à l'ajout d'une civière est lui plus compliqué à estimer. En fonction des situations, s'il y a déjà assez de place ou s'il faut construire et moderniser de nouvelles salles, le coût lié à l'installation d'une nouvelle civière peut drastiquement changer. En plus du coût d'installation, faut-il ajouter un coût de fonctionnement annuel. Lorsqu'un patient se fait traiter, que ce soit une nouvelle ou une ancienne civière le coût est identique, l'ajout d'une civière n'entraîne pas un surcoût pour le traitement du patient, cependant cela peut ajouter un surcoût logistique.

En se rapportant aux prix proposés par MEDITEK qui est un fournisseur de matériel médical, le coût d'une civière varie entre 5 000 $ et 12 000 $, en fonction de leur équipement, du matelas, et de l'électronisation qu'elle comporte. Les civières utilisées aux urgences ne nécessitent pas un niveau d'équipement particulièrement élevé. Afin de retenir une estimation réaliste et cohérente avec les besoins opérationnels d'un service d'urgence, un coût unitaire de **7 000 $** par civière a été retenu. En estimant que celle-ci est amortie sur 10 ans, cela revient à un coût annuel de 700 $.

Le coût d'entretien annuel est d'environ 1,25 % soit 87,5 $ [25].

En sommant les coûts, on arrive à un coût annuel d'environ 800 $ ce qui est bien en dessous du coût de l'ajout d'un médecin.

- **Scénario 3 :** somme des deux coûts précédents

Le coût lié au scénario 3 est le coût d'un médecin par période multiplié par le nombre de périodes puis on ajoute 800 $ par an pour la civière.

Maintenant que nous avons les coûts entraînés par les scénarios, nous devons estimer l'effet d'une réduction des délais de traitement.

4.7.2.2 Gains économiques d'une réduction des délais

Les études menées mettent en lumière une hausse des coûts d'hospitalisation après un long temps d'attente aux urgences. Cette hausse varie en fonction des papiers et des méthodologies. Un premier article datant de 1994 estimait que la durée d'hospitalisation augmentait de 11,7 % [26] quand les patients attendant plus de 24 heures aux urgences avant leur admission à l'hôpital. Toutefois cette étude remonte à plus de trente ans, et compte tenu des changements organisationnels qu'a connus le système de santé depuis, ces résultats semblent difficilement transposables à notre contexte actuel.

Dans un article plus récent (2019) intitulé *Just a Minute: The Effect of Emergency Department Wait Time on the Cost of Care [27]*, les auteurs proposent une nouvelle étude permettant de lier délai de première consultation et durée d'hospitalisation.

En utilisant une approche par variables instrumentales, Woodworth et Holmes (2020) mettent en évidence un **effet causal marginal** du temps d'attente aux urgences sur les coûts hospitaliers. Plus précisément, ils montrent qu'une **augmentation de 10 minutes** du délai précédant la première évaluation par un médecin urgentiste est associée à une hausse moyenne des coûts de l'ordre de **6 % pour les patients les plus graves** (niveaux ESI 1–2) et de **3 % pour les patients modérément graves** (niveaux ESI 2–3), tandis qu'aucun effet statistiquement significatif n'est observé pour les patients moins sévères. Ces résultats portent sur des variations marginales du temps d'attente et concernent les patients dont le délai de prise en charge est influencé par les décisions de triage.

L'**Emergency Severity Index (ESI)** est une échelle de triage principalement utilisée aux États-Unis. Tout comme l'ETG, l'ESI classe les patients en 5 catégories, en fonction de la gravité et des ressources requises. La catégorie 1 correspond aux cas les plus graves tandis que la catégorie 5 représente les cas plus légers.

Sur la base de leurs estimations marginales, les auteurs discutent ensuite des **implications économiques potentielles** d'une réduction des temps d'attente, en proposant un **exercice illustratif** visant à fournir un ordre de grandeur des gains attendus à l'échelle d'un établissement.

Cet exercice repose sur une extrapolation linéaire des effets estimés par tranche de 10 minutes (un retard de 2*10 minutes entraîne 2 fois la hausse sur le prix d'un retard de 10 minutes) et sur une pondération par le coût moyen et la proportion de patients dans chaque catégorie de gravité. La réduction moyenne des coûts hospitaliers peut ainsi être exprimée comme une fonction pondérée des coûts moyens par niveau de triage et de la distribution des patients entre ces niveaux.

En utilisant les valeurs observées dans leur échantillon l'exercice suggère qu'une **réduction hypothétique de 30 minutes du temps d'attente moyen** pourrait être associée à une **diminution des coûts hospitaliers** de l'ordre de **7 % pour un hôpital américain moyen**, et potentiellement plus élevée pour les établissements accueillant une proportion plus importante de patients sévères. Il convient toutefois de souligner que cette estimation repose sur une extrapolation des effets marginaux et doit être interprétée comme un **ordre de grandeur**, et non comme une estimation causale directe.

4.8 Résultats

En appliquant les coûts présentés précédemment, nous obtenons le tableau suivant où le coût est en dollars canadiens et les délais en minutes.

Tableau 4.6 Estimation de l'effet sur les délais

	Edwin_1civ	Edwin_1med	Edwin_civmed
Coût	800 $	739 200 $	740 000 $
Gain délai de consultation d'un médecin urgentiste médian	0.63	46.13	48.86
Gain délai de consultation d'un médecin spécialiste	1.01	83.71	85.75
Gain délai total en cas d'admission médian	3.38	225.91	235.48

L'estimation des économies faites sur l'année 2021 est moins simpliste. D'après le tableau, la réduction médiane du délai de première consultation est supérieure à 30 minutes pour les deux scénarios incluant un médecin. Afin d'y associer une économie, il a fallu calculer le nombre de patients ayant bénéficié de cette réduction de délais ainsi que leur coût hospitalier moyen.

Dans leur étude Woodworth et Holmes prennent en compte le coût total de l'épisode hospitalier. N'ayant pas trouvé de données correspondant à cette mesure au Canada, j'ai dû le calculer. Ce coût total peut être approché par :

$$Cout\ total\ =\ ED\ (hopital)\ +\ honoraires\ ED\ +P(admission) * coût\ hopitalisation$$

Il comporte trois composantes, le coût hospitalier ($ED\ (hopital)$), les honoraires payés aux médecins et le coût d'hospitalisation. En effet le coût hospitalier d'une visite aux urgences (304 $ en 2018–2019) rapporté par l'ICIS exclut les honoraires médicaux. Dans une perspective "système de santé", on ajoute (i) une estimation des honoraires médicaux moyens par visite et (ii) le coût d'hospitalisation pondéré par la probabilité d'admission. Sous des hypothèses prudentes (honoraires ≈200 $, probabilité d'admission ≈12 %, coût moyen d'hospitalisation ≈10 000 $), le coût total moyen par visite s'établit autour de 1 700 $.

Sur l'année 2021, le délai de première consultation aurait été réduit d'au moins 30 minutes pour 17 732 si un médecin avait été ajouté aux périodes surchargées. Ainsi le coût total aurait été réduit de $17\,732 * 0.07 * 1700\,\$ = 2\,110\,108\,\$$.

4.9 Synthèse des résultats

Après reprise de la méthodologie d'entraînement et de test, et l'ajout de données exogènes l'algorithme obtient des résultats satisfaisants avec une exactitude de près de 85 % pour l'année 2021. Sur cette année, le modèle a détecté 540 (49 % des périodes) périodes en surcharge. Parmi elles 82 ne l'étaient pas réellement tandis que 92 périodes ont été prédites non surchargées alors qu'elles l'étaient.

À partir de ce modèle, trois scénarios ont été testés, le premier inclut l'ajout d'une civière, le deuxième l'ajout d'un médecin urgentiste et le dernier d'une civière et d'un médecin urgentiste.

Les scénarios incluant un médecin permettent une diminution du **délai total en cas d'admission** d'environ 20 % (200 min) pour au moins 50 % des patients tandis que le scénario ajoutant seulement une civière a un effet négligeable. Les scénarios contenant un médecin semblent bien plus impactants, mais sont aussi plus coûteux. Avec un coût horaire de 150 $ de l'heure, l'ajout d'un médecin urgentiste pour les périodes surchargées coûte environ 740 000 $ par an tandis que l'ajout d'une civière coûte seulement 800 $.

Malgré un coût élevé, les scénarios incluant un médecin permettent de réduire le délai de première consultation de plus de 30 min pour au moins 17 732 patients. En s'appuyant sur les travaux de Woodworth et Holmes, on peut estimer une baisse de 7 % du coût de ces patients ce qui représente une économie estimée $2\,110\,108$ $ sur l'année 2021. En y soustrayant les coûts (740 000 $), les économies potentielles pourraient être de près de 1 370 108 $ par an.

Pour conclure, l'ajout d'une **civière** ne permet pas de résoudre le problème de congestion des urgences mais offre une possibilité à bas coût de **réduire faiblement le délai d'attente**. Les scénarios comportant l'ajout d'un **médecin** permettent de faire de **bonnes économies** et **diminuent**

fortement les différents délais d'attente. Enfin, le scénario combinant l'ajout d'un médecin et d'une civière ne permet pas de démultiplier l'effet et est sensiblement proche du scénario d'ajout d'un médecin.

CHAPITRE 5 RÉSULTATS VOLET POST-IMPLANTATION

Le cas étudié dans ce chapitre concerne un logiciel de planification déjà implanté dans plusieurs départements d'oncologie en Amérique du Nord. Les données disponibles proviennent de quatre rapports produits après le déploiement du logiciel à la demande des établissements concernés. Conformément au cadre méthodologique présenté au chapitre 3, l'objectif n'est pas de conduire une comparaison exhaustive de l'ensemble des cas, mais d'appliquer la stratégie d'évaluation post-implantation à une série de référence choisie comme support principal d'analyse. Ce choix permet de montrer de manière détaillée comment les outils retenus peuvent être mobilisés, interprétés et articulés dans un cas réel.

5.1 Choix de la série

Afin d'illustrer les méthodes, une série temporelle issue de l'un des rapports est retenue comme support principal de l'analyse. Elle correspond à l'indicateur **nombre de minutes de traitement par période financière**. La série comprend **60 observations**, à raison d'une observation par période financière, soit **13 observations par an**. L'intervention a lieu à la fin de la période 46, ce qui laisse **46 observations pré-implantation** et **14 observations post-implantation**. La relative brièveté de cette période post-implantation s'explique par le délai limité entre la demande du rapport par le département et la date effective d'implantation.

Au vu de l'indicateur, une valeur plus élevée de celui-ci correspond à une meilleure performance du service. Les effets favorables de l'implantation se traduisent donc, dans l'interprétation des résultats, par une augmentation de la série.

Le choix de cette série repose sur plusieurs éléments. D'abord, les données présentent une qualité satisfaisante, sans aucune valeur manquante et une mesure relativement stable dans le temps. Ensuite, cet indicateur fournit une mesure directe de la performance opérationnelle du service, ce qui le rend particulièrement pertinent pour l'évaluation. Enfin, la longueur de la période pré-intervention permet d'appliquer de manière adéquate les méthodes d'analyse retenues dans ce travail.

D’autres séries temporelles étaient également disponibles, mais elles présentaient certaines limites, comme la présence de valeurs aberrantes, une période post-intervention trop courte ou une structure moins adaptée aux méthodes mobilisées ici. Le tableau 5.1 en présente quelques exemples à titre illustratif.

Tableau 5.1 Séries temporelles rencontrées

Direction / établissement	**Indicateur disponible**	**Longueur de la période post**	**Remarque méthodologique**
CHUM	Nombre d’heures travaillées par infirmières	1 an	Valeurs aberrantes
CHUM	Nombre de rendez-vous	~ 1 an	Valeurs aberrantes
CHUM	Nombre moyen de rendez-vous patients	~1 an	Indicateur de volume d’activité
Ottawa Hospital	Temps moyen de planification d’un rendez-vous	~1–2 mois	Série post-intervention courte
Ottawa Hospital	Temps d’attente administratif moyen pour l’ensemble des commis (taux mensuel)	~1–2 mois	Indicateur difficilement transposable à d’autres départements
Jewish General Hospital	Minutes totales de traitement délivrées	~6–12 mois	Série agrégée à l’échelle du service

La série retenue n’est donc pas présentée comme représentative de l’ensemble des cas, mais comme un **support méthodologique** permettant de montrer de façon détaillée la démarche d’analyse.

5.2 Résultats de l'analyse principale : séries temporelles interrompues

L'analyse principale repose sur une approche en séries temporelles interrompues (ITS), appliquée à la série retenue afin d'évaluer si l'implantation du logiciel s'accompagne d'une modification de sa dynamique. Trois spécifications emboîtées ont été considérées. Le premier modèle est un **modèle simple**, qui teste uniquement l'existence d'un saut de niveau au moment de l'intervention. Le deuxième est un **modèle à tendance commune**, qui ajoute une tendance temporelle linéaire identique avant et après implantation. Enfin, le troisième est un **modèle segmenté complet**, qui autorise à la fois un changement immédiat de niveau et une modification de la pente après l'intervention. Cette progression permet de distinguer trois situations : absence de tendance préalable, présence d'une tendance de fond inchangée, ou rupture plus profonde de la dynamique de la série. Les équations détaillées de ces trois modèles sont présentées en annexe B.1.

Les trois spécifications ont d'abord été estimées dans un cadre **OLS.** La validité des tests statistiques OLS (t-test, intervalle de confiance) repose sur l'hypothèse que les erreurs ε_t soient indépendantes et identiquement distribuées. Afin de vérifier cette hypothèse, un test de Durbin-Watson ainsi qu'un test de Ljung-Box ont été effectués sur les résidus. Les diagnostics (Durbin-Watson ≈ 0.6 et Ljung-Box $p < 0.05$) indiquent une forte autocorrélation, l'hypothèse d'indépendance et de distribution identique n'étant pas respectée, les inférences usuelles associées à l'OLS ne peuvent pas être interprétées directement.

Afin de tenir compte de cette autocorrélation, les erreurs standards ont été remplacées par les erreurs HAC de type Newey–West. Le détail du calcul de ces erreurs est en annexe B.2.

Tableau 5.2 Résultats du modèle OLS simple avec correction HAC

Modèle	**β_post**	**SE(β)**	**t**	**p-val**	**IC-95%**
OLS simple	649	130	4.98	6×10^{-6}	[394 ; 905]

Le modèle simple estime un saut de niveau immédiat de +649 minutes par période. Ce résultat est statistiquement significatif ($p = 6 \times 10^{-6}$) et suggère, pris isolément, une amélioration

immédiate associée à l'implantation. Il doit toutefois être interprété avec prudence, dans la mesure où cette spécification ne tient pas compte d'une éventuelle dynamique temporelle plus complexe.

Tableau 5.3 Résultats du modèle OLS à tendance commune avec correction HAC

Modèle	**β_post**	**SE(β)**	**t**	**p-val**	**IC-95%**
OLS tendance commune	294	243	1.21	0.227	[−183 ; 771]

Lorsqu'une tendance temporelle commune est introduite, le coefficient de saut β_{post} diminue pour atteindre **+294 minutes** et perd sa significativité statistique (p = 0,227). Comme l'intervalle de confiance [-183 ; 771], incluant zéro, cette spécification suggère que l'effet pourrait être confondu avec la dynamique naturelle de la série.

Tableau 5.4 Résultats du modèle OLS segmenté avec correction HAC

Modèle	**β_post**	**SE(β)**	**p-val**	**IC-95%**	**ϕ (min/période)**	**SE(ϕ)**	**p-val**	**IC-95%**
OLS segmenté	694	208	8.5×10^{-4}	[287 ; 1102]	−70	7	4×10^{-22}	[−85 ; −56]

Enfin le modèle le plus complet réévalue le saut de niveau à **+694 minutes** ($p = 8{,}5 \times 10^{-4}$) tout en révélant une modification structurelle de la pente ϕ de **-70 minutes** par période ($p = 4 \times 10^{-22}$).

Afin de formaliser la comparaison entre les trois spécifications, des **tests de Wald** ont ensuite été utilisés pour évaluer la significativité individuelle ou conjointe des coefficients associés à l'intervention. Dans ce cadre, le test permet de vérifier si l'ajout d'un paramètre de rupture de niveau, puis d'un paramètre de changement de pente, améliore significativement la modélisation de la série. Les résultats du test de Wald conduisent ici à retenir **le modèle segmenté complet** comme spécification principale. Les résultats détaillés des tests sont reportés en annexe B.3.

Le cadre OLS HAC corrige l'autocorrélation des données en modifiant l'erreur standard associée. Cependant, elle ne modélise pas clairement cette autocorrélation. Afin d'examiner la sensibilité des résultats à un traitement plus explicite de l'autocorrélation, une estimation GLS du modèle segmenté a également été réalisée. L'objectif n'est pas ici d'opposer deux méthodes concurrentes, mais d'évaluer dans quelle mesure l'interprétation des résultats dépend de la façon dont la structure temporelle est traitée.

Tableau 5.5 Résultats du modèle GLS segmenté

Paramètre	**Estimation (min)**	**Erreur standard**	**t/z**	**p-valeur**	**IC 95 %**
β_post (saut de niveau)	+287	290	0.99	0.326	[−281 ; 855]
ϕ = Δpente (POST − PRÉ)	−66	42	−1.59	0.119	[−147 ; 16]

L'estimation GLS conduit à une lecture plus prudente : les coefficients estimés restent globalement cohérents en signe avec ceux obtenus en OLS-HAC, mais leur significativité devient plus faible (p-value de 0,326 et 0,119) lorsque l'autocorrélation est traitée plus explicitement.

Dans l'ensemble, les résultats suggèrent qu'après implantation, la série étudiée a connu une amélioration immédiate de son niveau, mais que cette amélioration ne s'inscrit pas de façon nette dans la durée. En effet, le modèle segmenté estimé en OLS avec erreurs HAC, retenu comme spécification principale, met en évidence un saut de niveau positif et significatif de +694 minutes au moment de l'intervention, ce qui va dans le sens d'un effet initial favorable de l'implantation. Toutefois, ce même modèle estime également une modification négative de la pente post-intervention ($\phi = -70$ minutes par période), ce qui indique que le gain initial tend ensuite à s'éroder au fil du temps. Par ailleurs, le fait que le modèle à tendance commune ne retrouve pas d'effet significatif, et que l'estimation GLS conduise à des coefficients de même signe mais non significatifs, invite à une interprétation prudente. Ainsi, l'analyse principale est compatible avec l'hypothèse d'un effet bénéfique immédiat de l'implantation sur l'indicateur, mais elle ne permet

pas de conclure avec la même assurance à la persistance de cet effet lorsque la structure temporelle de la série est prise en compte plus explicitement.

5.3 Résultats des analyses complémentaires

5.3.1 Test de Mann-Whitney

L'ensemble des détails des calculs est présent en annexe B.6.

En complément de l'ITS, une analyse non paramétrique a été menée afin d'évaluer si les observations post-implantation occupent globalement une position plus élevée que les observations pré-implantation dans la distribution de la série. Le test de Mann-Whitney met en évidence une **dominance globale de la période post** sur la période pré. Les rangs cumulés sont plus élevés dans le groupe post que ce que sa taille plus réduite laisserait attendre, ce qui va dans le sens d'un déplacement global de la distribution après implantation. Dans le cas étudié, la somme des rangs est de **1144** pour la période pré et de **686** pour la période post, ce qui traduit déjà un avantage relatif du groupe post dans le classement global de la série.

Cette lecture est confirmée par la statistique de dominance de **Vargha–Delaney**, qui est ici **proche de A_{obs} = 0,9**. Autrement dit, une observation tirée de la période post-implantation a une probabilité nettement supérieure à 0,5 de dominer une observation tirée de la période pré. Pris isolément, ce résultat suggère donc une amélioration globale après intervention.

5.3.2 Ruptures placebo

Toutefois, cette conclusion doit être nuancée par la procédure de **ruptures placebo**. En raison de l'autocorrélation de la série, la significativité du test n'a pas été évaluée à partir d'une approximation asymptotique classique, mais à partir d'une p-value empirique construite sur l'ensemble des coupures contiguës admissibles. Cette procédure (détaillée en annexe B.7) vise à déterminer si la séparation observée à la date d'implantation est réellement spécifique à cette date, ou si des ruptures d'ampleur comparable peuvent être générées ailleurs dans la série du seul fait de

sa dynamique interne. Dans le cas présent, les ruptures placebo montrent que la séparation pré/post n'est **pas statistiquement spécifique** à la date d'implantation. Cela suggère que la tendance ou la saisonnalité de la série peut produire, à d'autres dates, des coupures d'ampleur similaire.

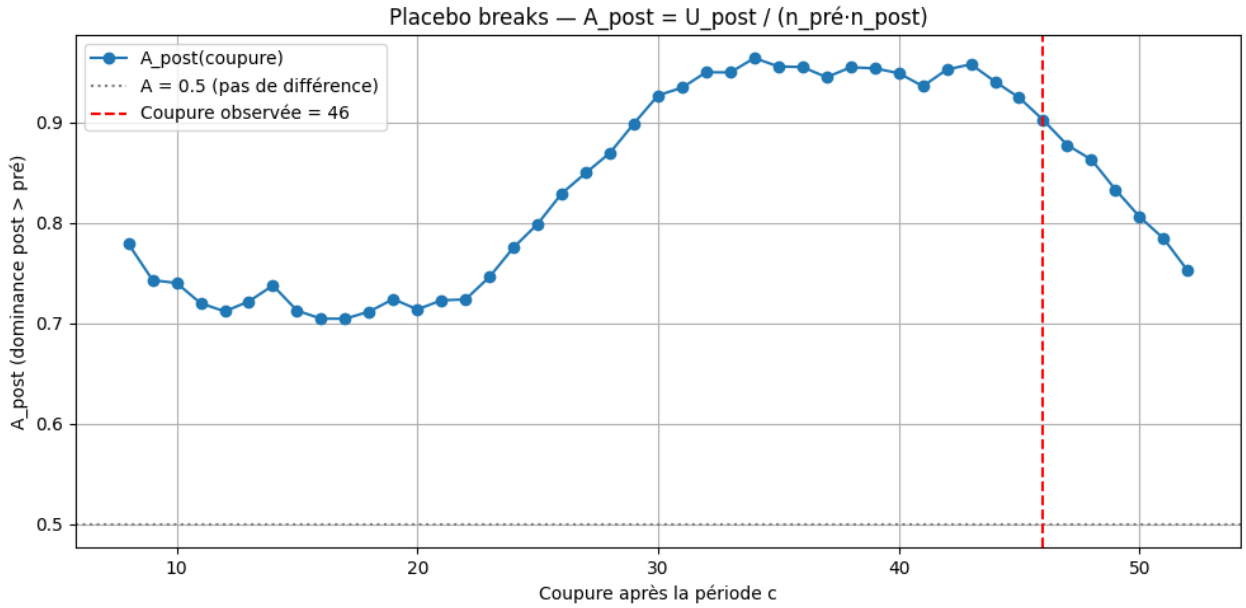

Figure 5.1 Ruptures placebo : statistique A selon la date de coupure

Ainsi, le test de Mann-Whitney confirme bien l'existence d'une **différence globale** entre les périodes pré- et post-implantation, mais la validation par ruptures placebo invite à la prudence quant à l'interprétation causale de cette différence. L'amélioration observée apparaît réelle sur le plan descriptif, sans pouvoir être considérée comme spécifiquement attribuable à la date d'intervention sur la seule base de cette méthode.

5.3.3 Inférence bayésienne

En complément des approches fréquentistes, une estimation bayésienne de type **BEST** (*Bayesian Estimation Supersedes the t-test*) a été mobilisée afin de quantifier l'écart entre les périodes pré- et post-implantation dans un cadre probabiliste. L'idée n'est plus seulement de tester l'existence d'une différence, mais d'estimer directement la distribution a posteriori de cette différence, ainsi

que la probabilité que la période post domine réellement la période pré. Dans ce cadre, les observations de chaque période sont modélisées par une loi de Student, ce qui permet de mieux prendre en compte la présence possible de valeurs atypiques qu'un modèle gaussien classique.

Plus précisément, le modèle repose sur une vraisemblance Student-t par groupe, avec des lois a priori faiblement informatives sur les moyennes, les paramètres d'échelle et les degrés de liberté. L'inférence est ensuite effectuée par **échantillonnage MCMC de type Hamiltonian Monte Carlo / NUTS**, ce qui permet d'obtenir un échantillon de la loi a posteriori des paramètres. Deux quantités d'intérêt sont retenues : d'une part la différence moyenne $\Delta = \mu_{post} - \mu_{pre}$, et d'autre part la **probabilité de dominance prédictive**, estimée à partir de pseudo-observations simulées sous le postérieur. Les détails complets de la spécification, des lois a priori, des équations du modèle et de la procédure d'inférence sont reportés en **annexe B.9**.

Les diagnostics de convergence sont satisfaisants, avec des valeurs de **R-hat égales à 1.0** et des **ESS élevés**, ce qui indique une exploration correcte de la distribution a posteriori (annexe B.10). L'adéquation du modèle est en outre vérifiée à l'aide de **posterior predictive checks**, notamment par l'examen des QQ-plots des résidus Student-t (annexe B.11), qui montrent un ajustement globalement compatible avec les données observées. Ces éléments sont détaillés en annexe, en particulier dans le tableau de convergence MCMC et les figures de diagnostic.

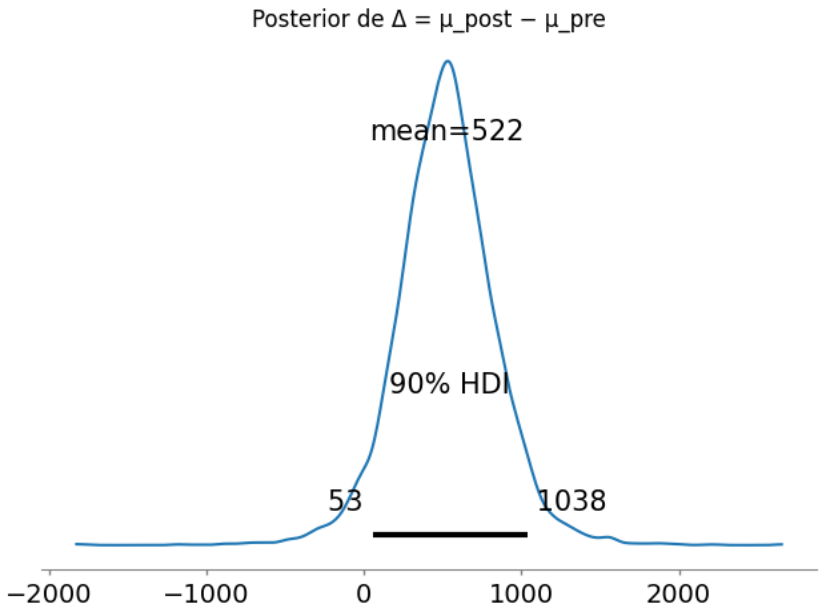

Figure 5.2 Distribution postérieure de la différence moyenne entre les périodes pré et post-implantation

Les résultats obtenus vont dans le sens d'une **supériorité probable de la période post-implantation**. La distribution a posteriori de Δ est centrée sur une valeur positive, avec une moyenne postérieure d'environ **522 minutes** et un **intervalle de crédibilité à 90 %** compris entre **53** et **1038** minutes. Autrement dit, sous le modèle retenu, la probabilité que la différence moyenne soit positive est proche de 1. La lecture bayésienne confirme ainsi, sur le plan descriptif, que la période post est associée à un niveau moyen plus élevé de l'indicateur que la période pré.

La probabilité de dominance prédictive va dans le même sens et suggère une séparation marquée entre les distributions pré- et post-implantation. Cette mesure complète utilement la différence moyenne, car elle renseigne non plus seulement sur un écart de localisation, mais sur la probabilité qu'une observation post tirée au hasard dépasse une observation pré. Dans le cadre présent, l'approche bayésienne renforce donc l'idée d'une amélioration post-implantation observable, tout en fournissant une quantification explicite de l'incertitude entourant cette conclusion. Il convient toutefois de rappeler que le modèle BEST **ne modélise pas explicitement la structure temporelle**

de la série et doit donc être interprété comme une comparaison **descriptive et probabilisée** des distributions marginales pré et post, plutôt que comme un modèle dynamique ou causal.

5.3.4 Analyse par série boostée

En complément des analyses précédentes, une analyse de robustesse a été menée à partir d'une **série pré-intervention artificiellement améliorée**, dite *série boostée*. Cette démarche ne constitue pas une méthode d'identification causale à part entière ; elle vise plutôt à évaluer la **sensibilité** de la conclusion obtenue en post-implantation à la définition du contrefactuel pré-intervention. Plus précisément, l'objectif est d'examiner à partir de quel niveau d'amélioration hypothétique de la période pré l'avantage observé de la période post cesse d'être soutenu.

Dans le cas étudié, une valeur plus élevée de l'indicateur correspond à de meilleures performances. La série modifiée est donc la **série pré-implantation**, tandis que la série post est conservée comme référence. On construit une famille de séries pré améliorées, indexées par un paramètre de boost β, appliqué de manière **multiplicative** observation par observation. Ce choix permet de raisonner en termes relatifs, ce qui est plus cohérent avec un indicateur d'activité que ne le serait un ajustement additif constant. Il préserve également la structure relative de la série, même s'il induit mécaniquement une augmentation proportionnelle de la dispersion. La procédure doit donc être interprétée comme la construction de **scénarios stylisés de sensibilité**, et non comme la simulation d'un mécanisme opérationnel réaliste. Les détails de cette transformation sont présentés en **annexe B.12**.

Pour chaque valeur de β appartenant à une grille de niveaux de boost, le modèle **BEST** est réestimé en conservant la même spécification — loi de Student par groupe, mêmes lois a priori, mêmes réglages d'échantillonnage MCMC et mêmes diagnostics de convergence — de manière à isoler l'effet du boost comme unique source de variation. La quantité d'intérêt principale devient alors la **probabilité de dominance du post sur la série pré boostée**, calculée à partir des distributions prédictives a posteriori. Le **point de bascule** $\beta^{\backslash *}$ est défini comme la plus petite valeur du boost pour laquelle cette probabilité atteint le seuil d'indifférence de 0,5. Il peut ainsi être interprété

comme l'amélioration minimale qu'il faudrait imposer artificiellement à la période pré pour neutraliser l'avantage observé du post. Les détails complets du recalcul bayésien pour chaque niveau de boost sont reportés en **annexe B.13**.

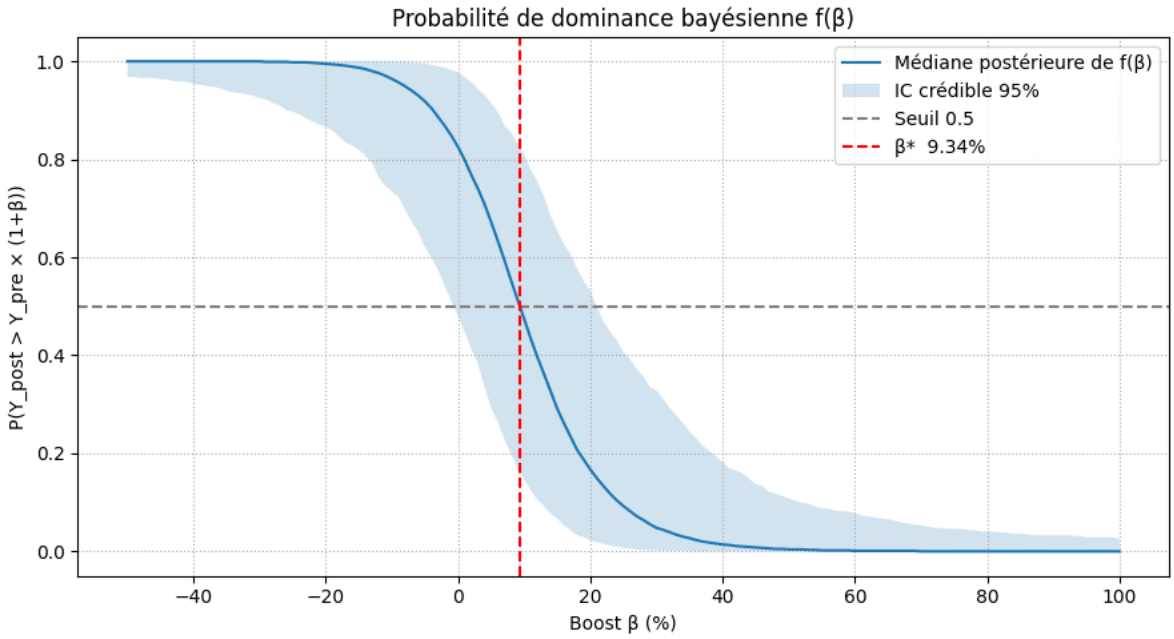

Figure 5.3 Probabilité de dominance bayésienne en fonction du boost

Les résultats sont synthétisés par la courbe de dominance bayésienne en fonction du boost. La médiane postérieure de $f(\beta)$ décroît de manière monotone lorsque la série pré est artificiellement améliorée, tandis que la zone ombrée représente l'intervalle crédible associé. Dans le cas présent, le point de bascule est atteint pour $\beta^{\backslash *} \approx 9{,}34\%$, ce qui signifie qu'une amélioration artificielle relativement modérée de la période pré suffirait à annuler l'avantage observé de la période post au sens de la dominance. Ce résultat ne remet pas en cause l'existence d'un écart favorable à la période post dans les données observées, mais il montre que cette conclusion est **modérément robuste** à une amélioration contrefactuelle du pré-intervention.

Ainsi, l'analyse par série boostée complète utilement l'approche BEST en apportant non plus seulement une quantification probabiliste de l'écart observé, mais une mesure explicite de sa **sensibilité**. Elle permet de passer d'une lecture statique de la différence pré/post à une lecture plus

dynamique, fondée sur la question suivante : jusqu'à quel point faudrait-il améliorer artificiellement la période pré pour que l'avantage post ne soit plus soutenu ? Dans cette perspective, elle renforce la prudence de l'interprétation : la supériorité post apparaît cohérente avec les autres résultats, mais demeure dépendante du cadre descriptif retenu et de la courte longueur de la période post-implantation.

5.4 Synthèse des méthodes

Dans l'ensemble, les résultats obtenus convergent vers l'existence d'une amélioration de l'indicateur après l'implantation du logiciel, une hausse de cet indicateur correspondant ici à une amélioration de la performance du service. L'analyse principale en séries temporelles interrompues suggère qu'un gain immédiat a pu se produire au moment de l'intervention. En particulier, le modèle segmenté estimé en OLS avec erreurs HAC, retenu comme spécification principale, met en évidence un saut de niveau positif, ce qui va dans le sens d'un effet initial favorable. Toutefois, cette amélioration ne paraît pas s'inscrire de manière nette dans la durée, puisque la pente post-intervention estimée est négative, suggérant une érosion progressive du gain initial. De plus, l'absence de significativité dans le modèle à tendance commune ainsi que dans l'estimation GLS invite à adopter une lecture prudente de l'ampleur et de la robustesse de cet effet.

Les analyses complémentaires confirment, sur un plan descriptif, que la période post-implantation est globalement associée à des valeurs plus élevées de l'indicateur que la période pré-implantation. Le test de Mann-Whitney, la statistique de dominance de Vargha–Delaney ainsi que l'approche bayésienne BEST vont tous dans le sens d'une supériorité probable de la période post. Ces résultats soutiennent donc l'existence d'un déplacement favorable de la distribution après intervention. Néanmoins, les ruptures placebo montrent que des séparations d'ampleur comparable peuvent également apparaître à d'autres dates de la série, ce qui limite la portée causale de cette différence. De la même manière, l'analyse par série boostée indique que l'avantage observé du post demeure sensible à la définition du contrefactuel pré-intervention, puisqu'une amélioration artificielle relativement modérée de la période pré suffit à neutraliser cette dominance.

Ainsi, pris dans leur ensemble, les résultats sont compatibles avec l'hypothèse d'une amélioration post-implantation, particulièrement visible à court terme et sur le plan descriptif. En revanche, ils ne permettent pas d'établir avec le même degré de certitude que cette amélioration soit entièrement spécifique à la date d'intervention ni qu'elle se maintienne de manière robuste dans le temps. L'interprétation qui se dégage est donc celle d'un signal globalement favorable à l'implantation, mais dont la portée causale et la stabilité temporelle doivent être nuancées au regard de l'autocorrélation de la série, de la courte période post-intervention et de la sensibilité des résultats aux choix de modélisation.

CHAPITRE 6 CONCLUSION

Ce mémoire présente une démarche d'évaluation des outils numériques en milieu hospitalier, fondée sur une distinction explicite entre deux temporalités : l'estimation pré-implantation, orientée vers l'aide à la décision, et l'évaluation post-implantation, centrée sur l'analyse des effets observés. Cette distinction structure l'ensemble du travail et conduit à mobiliser des méthodes différentes selon la question posée, les données disponibles et le niveau de preuve visé.

6.1 Apport du cadre d'évaluation proposé

L'apport principal de ce mémoire réside dans l'élaboration d'une démarche d'évaluation distinguant clairement deux situations souvent traitées de manière indistincte dans l'analyse des outils numériques en santé. Avant l'implantation, l'objectif n'est pas de démontrer un effet causal observé, mais de produire des ordres de grandeur plausibles de l'impact attendu afin d'éclairer une décision. Après l'implantation, l'enjeu devient différent : il s'agit d'examiner dans quelle mesure les changements observés dans les données peuvent être reliés de manière crédible à l'intervention, malgré les limites propres aux données observationnelles en milieu hospitalier.

Ce cadre permet ainsi d'éviter deux écueils fréquents. Le premier serait d'attribuer une portée causale excessive à une analyse pré-implantation fondée sur des scénarios et des hypothèses structurelles. Le second serait de considérer qu'une simple comparaison entre un avant et un après suffit à démontrer l'effet d'un logiciel une fois implanté. En distinguant explicitement ces deux temporalités, ce mémoire contribue à une évaluation plus solide des outils numériques en santé.

6.2 Enseignement de l'analyse pré-implantation

L'analyse pré-implantation a permis d'estimer les bénéfices plausibles d'un outil prédictif de congestion appliqué au service d'urgence du CHUM. Les simulations réalisées suggèrent qu'une utilisation opérationnelle de l'algorithme, combinée à un ajustement ponctuel des ressources lors des périodes de congestion anticipées, pourrait produire des gains organisationnels significatifs.

Les résultats indiquent notamment une réduction potentielle des délais de prise en charge et un bénéfice économique associé à une meilleure fluidité du parcours patient.

Ces résultats doivent toutefois être interprétés comme des ordres de grandeur conditionnels. Leur validité dépend de la capacité du modèle à capturer de manière suffisamment stable les dynamiques de congestion du service, alors même que les données d'entraînement couvrent une période marquée par d'importantes perturbations du système hospitalier. Par ailleurs, la simulation repose sur des hypothèses organisationnelles simplificatrices, notamment quant à la disponibilité des ressources et à l'absence d'effets comportementaux dans la gestion du service. Enfin, l'estimation de l'impact économique repose sur une chaîne d'hypothèses reliant congestion, délais et coûts hospitaliers, ce qui conduit à interpréter les gains estimés comme des indicateurs de potentiel d'efficience plutôt que comme des économies budgétaires directement observables.

Ainsi, l'évaluation ex ante ne constitue pas une démonstration d'efficacité, mais un outil d'aide à la décision permettant d'estimer, sous hypothèses explicites, les bénéfices plausibles d'une intervention avant son déploiement effectif.

6.3 Enseignement de l'analyse post-implantation

L'analyse post-implantation a permis d'examiner dans quelle mesure l'introduction du logiciel s'accompagne d'une évolution observable des indicateurs étudiés au sein du service d'oncologie. Contrairement au volet pré-implantation, qui repose sur une logique de simulation contrefactuelle, cette évaluation s'appuie directement sur les données observées avant et après l'intervention.

L'analyse principale par séries temporelles interrompues met en évidence un signal compatible avec une amélioration après implantation. Une estimation par régression OLS avec correction HAC suggère l'existence d'un saut initial et d'un changement de tendance, mais cette conclusion apparaît moins robuste lorsque la structure d'autocorrélation de la série est prise en compte de manière plus explicite via GLS. Cette sensibilité souligne la difficulté d'isoler un effet d'intervention dans un système caractérisé par une forte dépendance temporelle.

Les méthodes complémentaires permettent de mieux situer la portée de ce signal. Les ruptures placebo montrent que des séparations d'ampleur comparable peuvent apparaître à d'autres dates, ce qui limite la possibilité d'attribuer formellement la rupture observée à la seule implantation. La méthode BEST met en évidence une probabilité élevée d'amélioration associée à la période post-implantation, mais cette lecture demeure descriptive et conditionnelle au modèle retenu. Enfin, l'analyse par série pré-boostée montre qu'une amélioration artificielle modérée de la période pré suffit à neutraliser la dominance observée en post-intervention, ce qui suggère que l'écart observé reste d'ampleur limitée au regard de l'incertitude statistique.

L'analyse post-implantation ne permet donc pas d'établir une attribution causale formelle entre le logiciel et les évolutions observées. Elle met néanmoins en évidence un signal cohérent avec l'existence d'un effet favorable plausible associé à l'intervention.

6.4 Limites générales du travail

Les limites du présent mémoire tiennent d'abord à la nature des données mobilisées. Dans le volet pré-implantation, les données utilisées pour l'entraînement couvrent une période ancienne et en partie marquée par la pandémie, ce qui peut affecter la stabilité des relations apprises et la capacité du modèle à généraliser à d'autres contextes temporels. Dans le volet post-implantation, l'analyse repose sur une seule série retenue pour sa qualité relative, avec une fenêtre post-intervention courte, ce qui réduit la robustesse temporelle des conclusions et limite leur généralisation.

Une deuxième limite concerne les indicateurs et les mécanismes de traduction mobilisés. Dans le cas pré-implantation, le score d'Edwin fournit une représentation utile mais partielle de la congestion, et le lien entre surcharge, délais et coûts reste nécessairement simplifié. Dans le cas post-implantation, l'absence de groupe contrôle empêche d'exclure complètement l'effet de facteurs concomitants non observés. Dans les deux situations, les résultats doivent donc être interprétés comme conditionnels à la qualité du support empirique et au caractère raisonnable des hypothèses adoptées, plutôt que comme des estimations définitives de l'effet du logiciel.

6.5 Recommandations méthodologiques

Les analyses menées dans ce mémoire permettent de formuler plusieurs recommandations quant au choix des méthodes d'évaluation selon le stade de déploiement d'un outil numérique en santé.

Au terme de ce mémoire, trois recommandations méthodologiques principales peuvent être formulées pour l'évaluation des logiciels en santé. Premièrement, une estimation pré-implantation est pertinente lorsque l'objectif est d'éclairer une décision d'implantation plutôt que de démontrer un effet réel. Ce type d'approche doit être privilégié lorsque le logiciel n'est pas encore déployé, mais que des données historiques suffisamment riches permettent de modéliser le fonctionnement du service et de simuler des scénarios plausibles. Dans ce cadre, les résultats doivent être interprétés comme des ordres de grandeur plausibles et non comme des effets garantis.

Deuxièmement, une évaluation post-implantation fondée sur des séries temporelles interrompues est à privilégier lorsque l'on dispose d'un historique longitudinal clair avant et après l'intervention. Cette approche exige toutefois plusieurs conditions : une période pré-intervention suffisamment longue pour estimer la dynamique initiale, une période post-intervention assez étendue pour identifier une rupture stable, ainsi qu'une attention particulière portée à l'autocorrélation et aux changements concomitants. En l'absence de ces conditions, l'inférence causale demeure fragile, même si un signal d'amélioration peut être observé.

Troisièmement, lorsque les données sont limitées ou que le cadre causal est imparfait, il est recommandé de croiser plusieurs méthodes complémentaires plutôt que de s'appuyer sur une seule comparaison avant/après. Des approches comme les tests placebo, les comparaisons non paramétriques ou les méthodes bayésiennes peuvent utilement compléter une analyse principale, en apportant soit un test de robustesse, soit une mesure probabiliste plus directement interprétable. Une évaluation rigoureuse ne repose donc pas sur une méthode unique, mais sur un choix cohérent entre l'objectif poursuivi, les données disponibles et le niveau de preuve visé.

6.6 Ouverture

L’évaluation menée dans ce travail repose principalement sur des indicateurs organisationnels quantitatifs. Or, l’impact d’un outil numérique en milieu hospitalier ne se limite pas aux métriques observables telles que les délais ou les niveaux de congestion. Il comporte également une dimension humaine et organisationnelle plus difficilement quantifiable. Dans le cas post-implantation étudié, les retours du personnel suggèrent que l’outil aurait joué un rôle déterminant dans la gestion des périodes critiques. Cette perception ne constitue pas une preuve statistique d’efficacité, mais elle représente un signal qualitatif important, susceptible de révéler des effets non capturés par les indicateurs mesurés, tels que la réduction du stress organisationnel, l’amélioration de la coordination ou le sentiment de maîtrise face à l’incertitude.

Une piste d’approfondissement consisterait ainsi à développer une évaluation mixte combinant analyse quantitative et approche qualitative, par exemple à l’aide d’entretiens, de questionnaires ou d’indicateurs de bien-être organisationnel. Une telle démarche permettrait d’appréhender plus finement l’impact global de l’implantation, en articulant performance mesurée et expérience vécue par les acteurs du terrain.

RÉFÉRENCES

[1] World Health Organization, "Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment," 2016.

[2] T. Greenhalgh *et al.*, "Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies," (in eng), *J Med Internet Res,* vol. 19, no. 11, p. e367, Nov 1 2017, doi: 10.2196/jmir.8775.

[3] K. Baicker and T. Svoronos, "Testing the Validity of the Single Interrupted Time Series Design," *SSRN Electronic Journal,* 01/01 2019, doi: 10.2139/ssrn.3424248.

[4] D. R. Ayton *et al.*, "Barriers and enablers to the implementation of the 6-PACK falls prevention program: A pre-implementation study in hospitals participating in a cluster randomised controlled trial," (in eng), *PLoS One,* vol. 12, no. 2, p. e0171932, 2017, doi: 10.1371/journal.pone.0171932.

[5] T. A. Eaton, M. Kowalkowski, R. Burns, H. Tapp, K. O'Hare, and S. P. Taylor, "Pre-implementation planning for a sepsis intervention in a large learning health system: a qualitative study," (in eng), *BMC Health Serv Res,* vol. 24, no. 1, p. 996, Aug 28 2024, doi: 10.1186/s12913-024-11344-x.

[6] J. Jun, S. Jacobson, and J. Swisher, "Application of Discrete-Event Simulation in Health Care Clinics: A Survey," *Journal of the Operational Research Society,* vol. 50, pp. 109-123, 02/01 1999, doi: 10.1057/palgrave.jors.2600669.

[7] B. A. Dickerman and M. A. Hernán, "Counterfactual prediction is not only for causal inference," (in eng), *Eur J Epidemiol,* vol. 35, no. 7, pp. 615-617, Jul 2020, doi: 10.1007/s10654-020-00659-8.

[8] P. Craig, P. Dieppe, S. Macintyre, S. Michie, I. Nazareth, and M. Petticrew, "Developing and evaluating complex interventions: the new Medical Research Council guidance," (in eng), *Bmj,* vol. 337, p. a1655, Sep 29 2008, doi: 10.1136/bmj.a1655.

[9] J. Splawa-Neyman, D. Dabrowska, and T. Speed, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Statist. Sci.,* vol. 5, 11/01 1990, doi: 10.1214/ss/1177012031.

[10] D. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology,* vol. 66, pp. 688-701, 10/01 1974, doi: 10.1037/h0037350.

[11] A. B. Hill, "The clinical trial," (in eng), *Br Med Bull,* vol. 7, no. 4, pp. 278-82, 1951, doi: 10.1093/oxfordjournals.bmb.a073919.

[12] G. Imbens and D. Rubin, *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*. 2015, pp. 1-625.

[13] A. K. Wagner, S. B. Soumerai, F. Zhang, and D. Ross-Degnan, "Segmented regression analysis of interrupted time series studies in medication use research," (in eng), *J Clin Pharm Ther,* vol. 27, no. 4, pp. 299-309, Aug 2002, doi: 10.1046/j.1365-2710.2002.00430.x.

[14] J. L. Bernal, S. Cummins, and A. Gasparrini, "Interrupted time series regression for the evaluation of public health interventions: a tutorial," (in eng), *Int J Epidemiol,* vol. 46, no. 1, pp. 348-355, Feb 1 2017, doi: 10.1093/ije/dyw098.

[15] C. Wing, K. Simon, and R. A. Bello-Gomez, "Designing Difference in Difference Studies: Best Practices for Public Health Policy Research," *Annual Review of Public Health,* vol. 39, no. Volume 39, 2018, pp. 453-469, 2018, doi: https://doi.org/10.1146/annurev-publhealth-040617-013507.

[16] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. Scott, "Inferring causal impact using Bayesian structural time-series models," *The Annals of Applied Statistics,* vol. 9, pp. 247-274, 03/01 2015, doi: 10.1214/14-AOAS788.

[17] E. Murray *et al.*, "Evaluating Digital Health Interventions: Key Questions and Approaches," (in eng), *Am J Prev Med,* vol. 51, no. 5, pp. 843-851, Nov 2016, doi: 10.1016/j.amepre.2016.06.008.

[18] K. Skivington *et al.*, "A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance," (in eng), *Bmj,* vol. 374, p. n2061, Sep 30 2021, doi: 10.1136/bmj.n2061.

[19] I. J. MJ, H. Koffijberg, E. Fenwick, and M. Krahn, "Emerging Use of Early Health Technology Assessment in Medical Product Development: A Scoping Review of the Literature," (in eng), *Pharmacoeconomics,* vol. 35, no. 7, pp. 727-740, Jul 2017, doi: 10.1007/s40273-017-0509-1.

[20] C. J. Miller, S. N. Smith, and M. Pugatch, "Experimental and quasi-experimental designs in implementation research," (in eng), *Psychiatry Res,* vol. 283, p. 112452, Jan 2020, doi: 10.1016/j.psychres.2019.06.027.

[21] R. B. Penfold and F. Zhang, "Use of interrupted time series analysis in evaluating health care quality improvements," (in eng), *Acad Pediatr,* vol. 13, no. 6 Suppl, pp. S38-44, Nov-Dec 2013, doi: 10.1016/j.acap.2013.08.002.

[22] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics,* vol. 18, no. 1, pp. 50-60, 11, 1947. [Online]. Available: https://doi.org/10.1214/aoms/1177730491.

[23] A. Linden, P. R. Yarnold, and B. K. Nallamothu, "Using machine learning to model dose-response relationships," (in eng), *J Eval Clin Pract,* vol. 22, no. 6, pp. 856-863, Dec 2016, doi: 10.1111/jep.12573.

[24] J. K. Kruschke, "Bayesian estimation supersedes the t test," (in eng), *J Exp Psychol Gen,* vol. 142, no. 2, pp. 573-603, May 2013, doi: 10.1037/a0029146.

[25] G. Bektemur, N. Muzoglu, M. A. Arici, and M. K. Karaaslan, "Cost analysis of medical device spare parts," (in eng), *Pak J Med Sci,* vol. 34, no. 2, pp. 472-477, Mar-Apr 2018, doi: 10.12669/pjms.342.14245.

[26] P. Krochmal and T. A. Riley, "Increased health care costs associated with ED overcrowding," (in eng), *Am J Emerg Med,* vol. 12, no. 3, pp. 265-6, May 1994, doi: 10.1016/0735-6757(94)90135-x.

[27] L. Woodworth and J. F. Holmes, "JUST A MINUTE: THE EFFECT OF EMERGENCY DEPARTMENT WAIT TIME ON THE COST OF CARE," *Economic Inquiry,* vol. 58, no. 2, pp. 698-716, 2020, doi: https://doi.org/10.1111/ecin.12849.

ANNEXE A VOLET PRE-IMPLANTATION

A.1 DESCRIPTION DE LA TABLE VAR_TOT

1.Triage :

- Triage_nb
- Triage _delai_moyen

2 Consultation médecin urgentiste :

- Urgento_delai
- Urgento_nb

3 Consultation médecin spécialiste :

- Consult_nb :
  - Con_med_nb
  - Con_chir_nb
  - Con_gyn_nb
  - Con_mgen_nb
  - Con_psy_nb
- Consult_delai_moyen :
  - Con_med_ delai_moyen
  - Con_chir_ delai_moyen
  - Con_gyn_ delai_moyen
  - Con_mgen_ delai_moyen
  - Con_psy_ delai_moyen

4 Tests et examens cliniques :

- Radios :
  - Rx_nb
  - Rx_delai_moyen
- Échographie
  - Echo_nb
  - Echo_delai_moyen
- Scanner
  - Scan_nb
  - Scan_delai_moyen
- Irm
  - Irm_nb
  - Irm_delai_moyen
- Test pcr
  - Pcr_nb
  - Pcr_delai_moyen
- Sang
  - Sang_nb
  - Sang_delai_moyen
- Sangbunch
  - Sangbunch_nb
  - Sangbunch_delai_moyen
- Urine
  - Urine_nb
  - Urine_delai_moyen

5 Admission à l'hôpital

- Admis_nb :
  - Admis _med_nb
  - Admis _chir_nb
  - Admis _gyn_nb
  - Admis _mgen_nb
  - Admis _psy_nb
- Admis_delai_moyen
  - admis _med_ delai_moyen
  - admis _chir_ delai_moyen
  - admis _gyn_ delai_moyen
  - admis _mgen_ delai_moyen
  - admis _psy_ delai_moyen

6 Score d'edwin :

- score_edwin
- score_edwin_log
- edwin_12

Tableau A.1 Description des variables de var_tot

Variable	**Description**	**Type**
Triage_nb	Nombre total de patients en cours de triage	Numérique
Triage_delai_moyen	Délai moyen de triage des patients	Numérique
Urgento_delai	Délai moyen avant la consultation par un médecin urgentiste	Numérique
Urgento_nb	Nombre de patients en attente d'une consultation par un médecin urgentiste	Numérique
Con_med_nb	Nombre de patients en attente d'une consultation avec un médecin spécialiste en médecine interne	Numérique
Con_chir_nb	Nombre de patients en attente d'une consultation avec un chirurgien	Numérique
Con_gyn_nb	Nombre de patients en attente d'une consultation avec un gynécologue	Numérique
Con_mgen_nb	Nombre de patients en attente d'une consultation avec un médecin généraliste	Numérique

Tableau A.1 Description des variables de var_tot (suite)

Variable	**Description**	**Type**
Con_med_nb	Nombre de patients en attente d'une consultation avec un médecin spécialiste en médecine interne	Numérique
Con_chir_nb	Nombre de patients en attente d'une consultation avec un chirurgien	Numérique
Con_gyn_nb	Nombre de patients en attente d'une consultation avec un gynécologue	Numérique
Con_mgen_nb	Nombre de patients en attente d'une consultation avec un médecin généraliste	Numérique
Con_psy_nb	Nombre de patients en attente d'une consultation avec un psychiatre	Numérique
Con_med_delai_moyen	Délai moyen avant la consultation en médecine interne	Numérique
Con_chir_delai_moyen	Délai moyen avant la consultation en chirurgie	Numérique
Con_gyn_delai_moyen	Délai moyen avant la consultation en gynécologie	Numérique
Con_mgen_delai_moyen	Délai moyen avant la consultation en médecine générale	Numérique

Tableau A.1 Description des variables de var_tot (suite)

Variable	**Description**	**Type**
Con_psy_delai_moyen	Délai moyen avant la consultation en psychiatrie	Numérique
Rx_nb	Nombre de patient en attente d'une radiographie	Numérique
Rx_delai_moyen	Délai moyen avant la réalisation d'une radiographie	Numérique
Echo_nb	Nombre de patient en attente d'une échographie	Numérique
Echo_delai_moyen	Délai moyen avant la réalisation d'une échographie	Numérique
Scan_nb	Nombre de patient en attente d'un scanner	Numérique
Scan_delai_moyen	Délai moyen avant la réalisation d'un scanner	Numérique
Irm_nb	Nombre de patient en attente d'un IRM	Numérique
Irm_delai_moyen	Délai moyen avant la réalisation d'une IRM	Numérique
Pcr_nb	Nombre de patient en attente d'un test PCR	Numérique
Pcr_delai_moyen	Délai moyen avant le résultat d'un test PCR	Numérique
Sang_nb	Nombre de patient en attente d'une analyse sanguine	Numérique

Tableau A.1 Description des variables de var_tot (suite)

Variable	**Description**	**Type**
Sang_nb	Nombre de patient en attente d'une analyse sanguine	Numérique
Sang_delai_moyen	Délai moyen avant le résultat d'une analyse sanguine	Numérique
Sangbunch_nb	Nombre de patient en attente d'un bilan sanguin	Numérique
Sangbunch_delai_moyen	Délai moyen avant le résultat d'un bilan sanguin groupé	Numérique
Urine_nb	Nombre de patient en attente d'une analyse urinaire	Numérique
Urine_delai_moyen	Délai moyen avant le résultat d'une analyse urinaire	Numérique
Admis_med_nb	Nombre de patient en attente d'une admission en médecine	Numérique
Admis_chir_nb	Nombre de patient en attente d'une admission en chirurgie	Numérique
Admis_gyn_nb	Nombre de patient en attente d'une admission gynécologie	Numérique
Admis_mgen_nb	Nombre de patient en attente d'une admission médecine générale	Numérique
Admis_psy_nb	Nombre de patient en attente d'une admission en psychiatrie	Numérique

Tableau A.1 Description des variables de var_tot (suite et fin)

Variable	**Description**	**Type**
Admis_med_delai_moyen	Délai moyen avant admission en médecine	Numérique
Admis_chir_delai_moyen	Délai moyen avant admission en chirurgie	Numérique
Admis_mgen_delai_moyen	Délai moyen avant admission en médecine générale	Numérique
Admis_psy_delai_moyen	Délai moyen avant admission en psychiatrie	Numérique
score_edwin	Score global d'évaluation Edwin (continu)	Numérique
score_edwin_log	Version logarithmique du score Edwin (binaire ou log-transformée)	Numérique / Logique
edwin_12	Score Edwin calculé sur les 12 dernières heures	Numérique
edwin_24	Score Edwin calculé sur les 24 dernières heures	Numérique

A.2 DÉTAIL TECHNIQUE DE LA CONSTRUCTION DES VARIABLES RETARDÉES

Les données étant observées à une fréquence de 15 minutes, des variables retardées sont construites afin de tenir compte du fait que l'effet des variables explicatives sur la congestion future peut être différé dans le temps. Pour chaque variable numérique X_j, un ensemble de lags candidats est considéré sur une fenêtre de 24 heures précédant l'instant de prédiction, soit :

$$l \in \{0,1,\dots,96\}$$

où chaque pas correspond à un intervalle de 15 minutes. Pour chaque lag l, la variable retardée est définie par :

$$X_j^{(l)}(t) = X_j(t-l)$$

Le lag optimal L_j^*associé à la variable $X_j$est ensuite sélectionné comme celui maximisant la corrélation absolue avec la cible future $Y(t)$:

$$L_j^* = \arg \max_{l \in \{0,\dots,96\}} | \operatorname{corr}\,(X_j(t-l), Y(t)) |$$

Une seule version retardée est ensuite conservée pour chaque variable, soit :

$$\tilde{X}_j(t) = X_j(t - L_j^*)$$

La base finale utilisée pour l'apprentissage est donc constituée des variables transformées $\tilde{X}_j(t)$et de la cible $Y(t)$. Afin d'éviter toute fuite d'information, la sélection des lags L_j^*est réalisée uniquement sur le jeu d'entraînement.

```
[TEST] ROC-AUC = 0.9182 | PR-AUC(AP) = NA | seuil = 0.558391

-------------------------------------------
TEST (8h) — Matrice de confusion (Prédictions x Références)
-------------------------------------------
          Reference
Prediction  no yes
       no  381  92
       yes  82 540

-------------------------------------------
TEST (8h) — Métriques classiques
-------------------------------------------
Prévalence          :   57.7 %
Accuracy            :   84.1 %
Precision / PPV     :   86.8 %
Recall / Sensibilité :   85.4 %
Spécificité         :   82.3 %
NPV                 :   80.5 %
N (total)           : 1095
TP=540 | FP=82 | FN=92 | TN=381
```

Figure A.1 Performances du modèle prédictif

ANNEXE B VOLET POST-IMPLANTATION

B.1 MODÈLE D'ITS CONSIDÉRÉS

Modèle 1 : modèle simple

La première spécification, dite **modèle simple**, repose sur une unique variable indicatrice *post*, égale à 1 pour les observations post-implantation et à 0 sinon. Cette régression vise à mesurer directement la différence de niveau moyen avant et après l'intervention, sans contrôle additionnel.

$$y_t = \alpha + \beta \, \text{post}_t + \varepsilon_t \qquad (3.19),$$

- α: niveau de base avant intervention,
- β: changement de niveau ("effet immédiat") après l'intervention,
- ε_t: terme d'erreur.

Modèle 2 : modèle avec tendance commune

La deuxième spécification, **modèle avec tendance commune**, ajoute une variable de tendance linéaire en fonction du temps afin de tenir compte de l'évolution naturelle du service. L'ajout de cette tendance permet de distinguer le changement ponctuel de niveau attribuable à l'intervention d'une éventuelle tendance naturelle.

$$y_t = \alpha + \beta \, \text{post}_t + \delta \text{t} + \varepsilon_t \qquad (3.20),$$

- $\delta \, t$: tendance linéaire avant l'intervention,

Modèle 3 : modèle segmenté

Enfin, la troisième spécification, **modèle segmenté**, autorise non seulement un saut mais aussi une rupture de pente après la date d'intervention. Ce modèle permet de vérifier si l'effet observé est

uniquement un saut de niveau, ou s'il s'accompagne d'une modification structurelle de la dynamique de l'indicateur

$$y_t = \alpha + \beta \,\text{post}_t + \delta\, t + \phi\,(t - T_0)_+ + \varepsilon_t \qquad (3.21),$$

Ici, ϕ représente la modification structurelle de la tendance. Si ϕ est significatif, la dynamique temporelle de l'indicateur change après l'intervention.

B.2 ERREURS HAC

- Les **écarts-types HAC (Newey-West, lag = 4)**, robustes à l'hétéroscédasticité et à l'autocorrélation, mieux adaptés aux séries temporelles.

Forme sandwich (long-run variance) :

$$\text{Var}_{\text{HAC}}(\hat{\beta}) = (X'X)^{-1}\ \hat{S}\ (X'X)^{-1} \qquad (3.24)$$

où $\hat{S}$ agrège les **autocovariances résiduelles** jusqu'à un **lag** L avec des **poids** (kernel de Bartlett par défaut).
En notant x_t la ligne t de X(vecteur colonne) et $\hat{u}_t$ les résidus OLS :

$$\hat{S} = \hat{\Gamma}_0 + \sum_{\ell=1}^{L} w_\ell(\hat{\Gamma}_\ell + \hat{\Gamma}_\ell'), \hat{\Gamma}_\ell = \sum_{t=\ell+1}^{n} \hat{u}_t\ \hat{u}_{t-\ell}\ x_t x_{t-\ell}'.$$

- **Terme contemporain** :

$$\hat{\Gamma}_0 = \sum_{t=1}^{n} \hat{u}_t^2\ x_t x_t'.$$

- **Poids de Bartlett** :

$$w_\ell = 1 - \frac{\ell}{L+1}, \ell = 1, \dots, L.$$

L'**écart-type HAC** d'un coefficient est alors :

$$SE_{\text{HAC}}(\hat{\beta}_j) = \sqrt{[(X'X)^{-1}\,\hat{S}\,(X'X)^{-1}]_{jj}} \tag{3.25}$$

Ce choix vise à garantir la validité de l'inférence statistique dans des conditions de légère autocorrélation dans la série.

B.3 TESTS DE WALD

Cette structure emboîtée permet de tester, à l'aide de **tests de restriction (test de Wald robuste HAC)**, si les paramètres additionnels apportent une amélioration significative du modèle. Autrement dit, on cherche à savoir si l'ajout d'une tendance temporelle ou d'un changement de pente post-intervention améliore réellement la capacité explicative du modèle, ou si ces effets sont statistiquement négligeables.

Deux comparaisons principales ont été réalisées :

- Segmenté (niveau + pente) vs Tendance + saut

L'hypothèse nulle $H_0: \varphi = 0$(absence de changement de pente post-intervention) a été testée à l'aide d'un test de Wald robuste (HAC). Le résultat ($\chi^2 = 93{,}53, p \approx 4{,}0 \times 10^{-22}$) conduit à **rejeter très fortement** H_0, indiquant que la variation de pente post-intervention améliore significativement l'ajustement.

Par ailleurs, la tendance temporelle commune est significative dans le modèle "tendance + saut" ($p(t_{center}) = 0{,}0376$).
Ces résultats confirment que le modèle segmenté décrit plus fidèlement la dynamique temporelle observée.

- Segmenté (niveau + pente) vs Saut seul

Le modèle “saut seul” ne prend en compte ni la tendance globale ni la modification de pente. Or, la tendance est significative dans le modèle “tendance + saut” ($p = 0{,}0376$), et la variation de pente l’est également dans le modèle segmenté ($p(t_{post}) = 4{,}0 \times 10^{-22}$). Ces deux résultats invalident les restrictions des modèles plus simples.

B.4 PASSAGE AUX GLS

$y = X\theta + \varepsilon$ avec $Var(\varepsilon) = \Omega$

On multiplie **les deux côtés** du modèle par $\Omega^{-1/2}$:

$$\Omega^{-1/2} y = \Omega^{-1/2} X\theta + \Omega^{-1/2}\varepsilon$$

Avec $Var(\Omega^{-1/2}\varepsilon) = I$

Alors : $y^* = X^*\theta + u$ avec $u \sim iid(0, I)$

où :

- $y^* = \Omega^{-1/2}\, y$
- $X^* = \Omega^{-1/2} X$

En remplaçant :

- $\|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\theta}\|^2 = (\mathbf{y} * -\mathbf{X} * \boldsymbol{\theta})\top(\mathbf{y} * -\mathbf{X} * \boldsymbol{\theta})$
- $\|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\theta}\|^2 = \ (\Omega^{-1/2}(y - X\theta))^\top(\Omega^{-1/2}(y - X\theta))$

- En regroupant :

$$\|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\theta}\|^2 = (y - X\theta)^\top\Omega^{-1}(y - X\theta)$$

B.5 RÉSULTATS MODÈLES GLS

Les coefficients estimés par GLS pour les différentes spécifications sont présentés ci-dessous.

Tableau B.1 Résultats modèle simple (saut de niveau) GLS.

Paramètre	Estimation (min)	Erreur standard	t/z	p-valeur
β_post (saut de niveau)	+323	248	1.30	0.198

Dans le modèle ne tenant compte que du saut de niveau, le coefficient associé à la variable POST est estimé à $+323$minutes (SE = 248). Ce coefficient n'est pas statistiquement significatif ($p = 0{,}198$), et l'intervalle de confiance à 95 % $[-163; 808]$inclut zéro.

Ce résultat contraste avec l'estimation OLS simple, où le saut apparaissait fortement significatif, et suggère que l'effet observé dans le modèle naïf est en grande partie amplifié par l'autocorrélation non modélisée.

Tableau B.2 Résultats modèle avec tendance commune GLS

Paramètre	Estimation (min)	Erreur standard	t/z	p-valeur	IC 95 %
β_post (saut de niveau)	+120	275	0.44	0.664	[−419 ; 659]

Lorsque l'on introduit une tendance temporelle commune, l'estimation GLS conduit à un saut de niveau encore plus faible ($+120$ minutes, SE = 275), très largement non significatif ($p = 0{,}664$). L'intervalle de confiance $[-419; 659]$est large et centré autour de zéro.

B.6 DÉTAILS DU TEST DE MANN-WHITNEY

Le test de Mann-Whitney repose sur le classement conjoint des observations issues des périodes pré- et post-implantation. Les valeurs sont ordonnées de la plus faible à la plus élevée, puis un rang est attribué à chacune. En cas d'égalité, le rang associé correspond à la moyenne des rangs concernés. Le tableau complet des observations classées est présenté ci-dessous.

Tableau B.3 Classement des observations pour le test de Mann-Whitney

Groupe	Value	Rank
Pré	4021	1
Pré	4577	2
Pré	4752	3
Pré	4830	4
Pré	4903	5
Pré	4915	6
Pré	4934	7
Pré	5823	38
Pré	5824	39
Pré	5832	40
…	…	…
Post	5971	46
Post	5987	47
Post	6029	48
Post	6106	49
Pré	6125	50
Post	6148	51
Post	6157	52
Post	6222	53
Pré	6243	54
Pré	6243	55
Post	6331	56
Post	6400	57
Pré	6420	58
Post	6521	59
Post	6540	60

À partir de ce classement, les sommes de rangs sont calculées pour chaque groupe :

$$T_{\text{pré}} = 1144, T_{\text{post}} = 686.$$

Avec

$$n_{\text{pré}} = 46, n_{\text{post}} = 14,$$

on obtient les rangs moyens :

$$M_{\text{pré}} = \frac{T_{\text{pré}}}{n_{\text{pré}}} = 24{,}87, M_{\text{post}} = \frac{T_{\text{post}}}{n_{\text{post}}} = 49{,}0.$$

La statistique de Mann-Whitney associée au groupe post est alors donnée par :

$$U_{\text{post}} = T_{\text{post}} - \frac{n_{\text{post}}(n_{\text{post}} + 1)}{2} = 686 - \frac{14 \times 15}{2} = 581.$$

Cette statistique peut être normalisée sous la forme de la statistique de dominance de Vargha–Delaney :

$$A_{\text{obs}} = \frac{U_{\text{post}}}{n_{\text{pré}} n_{\text{post}}} = \frac{581}{46 \times 14} = 0{,}9022.$$

Ainsi, environ 90 % des paires (post, pré)sont ordonnées dans le sens post $>$ pré, ce qui traduit une séparation marquée entre les deux périodes au niveau de la coupure observée.

B.7 VALIDATION PAR RUPTURES PLACEBO

La mise en œuvre classique du test de Mann-Whitney repose sur une approximation asymptotique normale de la statistique U, qui suppose notamment l'indépendance des observations. Cette hypothèse n'est pas réaliste ici en raison de l'autocorrélation de la série, mise en évidence par une statistique de Durbin–Watson égale à :

$$DW = 0{,}62.$$

Dans ces conditions, une **p-value empirique fondée sur des ruptures placebo** est préférée à l'approximation asymptotique usuelle.

Le principe consiste à comparer la coupure observée entre les périodes pré- et post-implantation à un ensemble de coupures contrefactuelles obtenues en faisant varier le point de rupture le long de la série temporelle ordonnée. Pour chaque coupure admissible c, la série est divisée en deux groupes contigus $[1, \dots, c]$et $[c + 1, \dots, N]$, avec une taille minimale de 8 observations par groupe.

Pour chaque coupure c, on calcule une statistique de dominance :

$$A_c = \frac{U_{\text{post}}^{(c)}}{n_{\text{pré}} n_{\text{post}}} \in [0,1].$$

La p-value empirique bilatérale est ensuite définie par :

$$p_{\text{emp}} = \frac{1}{|\ C\ |} \sum_{c \in C} \mathbf{1}(|\ A_c - 0{,}5\ | \geq |\ A_{\text{obs}} - 0{,}5\ |),$$

où C désigne l'ensemble des coupures admissibles.

Dans le cas étudié, la coupure observée entre les périodes pré- et post-implantation conduit à :

$$A_{\text{obs}} = 0{,}9022, \qquad p_{\text{emp}} = 0{,}378.$$

Ainsi, bien que la séparation observée soit forte en valeur absolue, elle n'apparaît pas statistiquement spécifique à la date d'implantation. La figure 5.1 montre en effet que plusieurs coupures placebo produisent des valeurs de $A(c)$ au moins aussi extrêmes que celle observée à la date réelle d'intervention. La portée inférentielle associée à cette rupture demeure donc limitée.

B.8 DÉTAILS DU MODÈLE BAYESIEN BEST

Afin de comparer les périodes pré- et post-implantation dans un cadre probabiliste, une estimation bayésienne de type **BEST** (*Bayesian Estimation Supersedes the t-test*) est utilisée. Le modèle repose sur une représentation séparée des deux groupes, chacun étant modélisé par une loi de Student, plus robuste aux valeurs extrêmes qu'une loi normale.

On note $Y_{\text{pré}} = \{y_1, \dots, y_{n_{\text{pré}}}\}$ l'ensemble des observations pré-implantation et $Y_{\text{post}} = \{z_1, \dots, z_{n_{\text{post}}}\}$ l'ensemble des observations post-implantation. Le modèle s'écrit :

$$y_i \sim t_\nu(\mu_{\text{pré}}, \sigma_{\text{pré}}), i = 1, \dots, n_{\text{pré}},$$
$$z_j \sim t_\nu(\mu_{\text{post}}, \sigma_{\text{post}}), j = 1, \dots, n_{\text{post}}.$$

Ici :

- $\mu_{\text{pré}}$ et μ_{post} désignent les moyennes des deux groupes ;

- $\sigma_{\text{pré}}$et σ_{post}leurs paramètres d'échelle ;
- νles degrés de liberté de la loi de Student.

La quantité d'intérêt principale est la différence moyenne entre les deux périodes :

$$\Delta = \mu_{\text{post}} - \mu_{\text{pré}}.$$

Le modèle permet également de définir une probabilité de dominance entre les deux groupes, à partir de simulations prédictives postérieures :

$$P(Y_{\text{post}} > Y_{\text{pré}}).$$

Des lois a priori faiblement informatives sont spécifiées sur les paramètres de localisation, d'échelle et de forme. En pratique, ces lois a priori sont choisies de manière à laisser les données guider l'inférence tout en stabilisant l'estimation. La loi a posteriori s'écrit, à une constante de normalisation près :

$$p(\theta \mid Y) \propto p(Y \mid \theta)\, p(\theta),$$

où $\theta = (\mu_{\text{pré}}, \mu_{\text{post}}, \sigma_{\text{pré}}, \sigma_{\text{post}}, \nu)$.

Cette distribution n'admet pas de forme fermée simple. L'inférence est donc réalisée par **échantillonnage MCMC**, ce qui permet d'obtenir un ensemble de tirages de la loi a posteriori des paramètres et, par conséquent, de toute quantité dérivée d'intérêt comme Δ ou la probabilité de dominance

B.9 CONVERGENCE DE L'ÉCHANTILLONNAGE

Tableau B.4 Convergence MCMC

Paramètre	Mean	SD	HDI_3%	HDI_97%	ESS_bulk	ESS_tail	R_hat
mu_pre	5492.66	69.14	5366.91	5625.57	7771	5719	1.0
mu_post	6127.47	80.69	5978.07	6282.82	7734	5591	1.0
sigma_pre	430.55	60.29	317.32	547.61	6264	4583	1.0
sigma_post	279.00	66.34	168.04	404.87	7688	5286	1.0
nu_minus1	26.16	25.99	0.85	74.52	5775	5129	1.0
Delta	634.80	107.17	437.71	838.46	8041	5723	1.0

B.10 QQ-PLOTS DES RESIDUS STUDENT-T

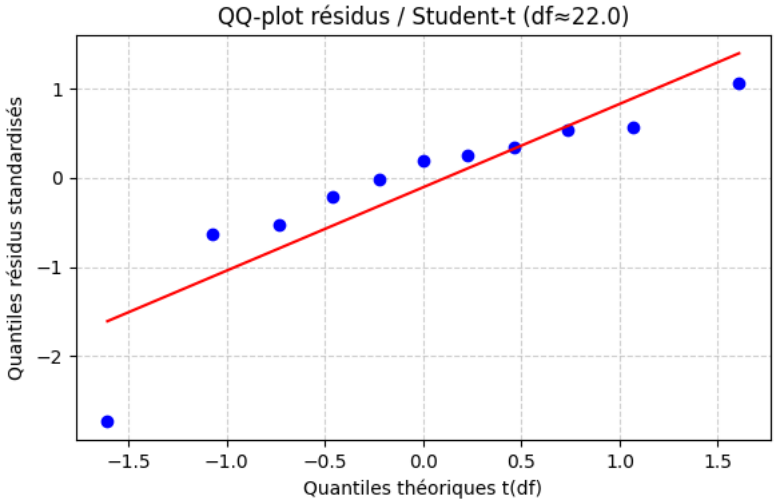

Figure B.1 QQ-plot résidus / Student-t

B.11 ANALYSE DE ROBUSTESSE PAR SÉRIE BOOSTÉE

Afin d'étudier la sensibilité des conclusions à la définition du contrefactuel pré-intervention, une famille de séries pré artificiellement modifiées est construite. Le principe consiste à améliorer progressivement la série pré-implantation à l'aide d'un paramètre de **boost** β, tout en conservant la série post inchangée.

Soit $Y_{\text{pré}} = \{y_1, \dots, y_{n_{\text{pré}}}\}$la série pré-implantation observée. Pour une valeur donnée de $\beta \geq 0$, on définit la série pré boostée par transformation multiplicative :

$$Y_{\text{pré}}^{(\beta)} = \{(1+\beta)y_1, \dots, (1+\beta)y_{n_{\text{pré}}}\}.$$

Cette transformation permet de raisonner en termes relatifs et de générer une famille de scénarios contrefactuels indexés par β. La série post, notée Y_{post}, est conservée inchangée.

B.12 MÉTHODE DE LA SÉRIE BOOSTÉE

Pour chaque valeur de β, le modèle bayésien BEST est réestimé en remplaçant $Y_{\text{pré}}$par $Y_{\text{pré}}^{(\beta)}$. On définit alors la fonction de dominance :

$$f(\beta) = P(Y_{\text{post}} > Y_{\text{pré}}^{(\beta)}),$$

où la probabilité est estimée à partir des distributions prédictives postérieures issues du modèle.

La fonction $f(\beta)$mesure donc, pour chaque niveau de boost, la probabilité qu'une observation post dépasse une observation pré artificiellement améliorée.

Le **point de bascule** β^*est défini comme la plus petite valeur de βpour laquelle cette probabilité atteint le seuil d'indifférence de 0,5 :

$$\beta^* = \inf\{\beta \geq 0 : f(\beta) \leq 0.5\}.$$

Cette quantité peut être interprétée comme le niveau minimal d'amélioration artificielle qu'il faut appliquer à la période pré pour annuler la dominance du post.

En pratique, $f(\beta)$est évaluée sur une grille discrète de valeurs de β, puis β^*est approché numériquement à partir de cette grille ou par interpolation locale entre deux valeurs adjacentes. La procédure consiste donc à :

1. choisir une grille $\{\beta_1, \dots, \beta_K\}$;

2. construire $Y_{\text{pré}}^{(\beta_k)}$ pour chaque k;
3. réestimer le modèle BEST pour chaque β_k;
4. calculer $f(\beta_k)$;
5. identifier la première valeur pour laquelle $f(\beta_k) \leq 0.5$.

Cette démarche fournit une mesure de sensibilité des conclusions bayésiennes à une modification du contrefactuel pré-intervention, sans modifier la structure du modèle de comparaison entre groupes.