Rapport de conception et d'analyse hospitalière

+ Étude d'impact et recommandations stratégiques

Projet Data — MVP Pitié-Salpêtrière — Promo 2026

Un seul document : conception, analyse hospitalière, étude d'impact et recommandations stratégiques.

---

1. Introduction et objectifs

1.1 Contexte

L'Hôpital Pitié-Salpêtrière, l'un des plus grands hôpitaux d'Europe, accueille chaque année plus de 100 000 patients aux urgences et gère plus de 1 800 lits d'hospitalisation. Les pics d'admission (hiver, épidémies, événements exceptionnels) rendent difficile la gestion du personnel et des équipements. La direction a demandé un **MVP (Minimum Viable Product)** de simulation et prévision des besoins hospitaliers pour anticiper ces pics et optimiser les ressources.

1.2 Objectifs du MVP

- **Générer un jeu de données fictif** inspiré de l'activité de la Pitié-Salpêtrière (tendances réalistes : saisonnalité, services, jour de la semaine).
- **Développer un prototype fonctionnel** : tableau de bord interactif, prévisions (admissions, occupation des lits), simulation de scénarios (épidémie, grève, canicule, afflux massif), recommandations automatiques.
- **Analyser les tendances d'admissions** et les périodes critiques, en s'appuyant sur la littérature (thèses Bouteloup 2020, Lequertier 2022).
- **Proposer une étude d'impact** et des recommandations stratégiques pour l'utilisation de l'outil en établissement.

Les données du projet sont **100 % synthétiques** ; aucune donnée réelle de patients n'est utilisée. Le prototype est une démonstration dont la validation opérationnelle nécessiterait des données réelles (PMSI, RPU) dans un cadre réglementaire (CEREES, CNIL).

---

2. Fonctionnalités du prototype et méthodologie de développement

2.1 Architecture et modules

Le prototype est organisé en modules complémentaires :

Module	Rôle
Données	

Module	Rôle
	Génération du jeu fictif (`src/data/generate.py`), chargement et agrégation (admissions par service, occupation quotidienne).
Analyse	Préparation des séries temporelles, indicateurs pour l'AED (notebooks).
Prédiction	Modèles de prévision des admissions et de l'occupation (Holt-Winters, Ridge, SARIMA, moyenne glissante, Boosting), modèle stock (occupation à partir des admissions).
Simulation	Scénarios paramétrés : épidémie grippe, grève, canicule, afflux massif ; durée configurable (14 à 90 jours).
Recommandations	Génération d'alertes et d'actions à partir des prévisions et des scénarios (seuils 85 % / 95 %, priorisation, actions concrètes).
Dashboard	Interface Streamlit : Flux & historique, Prévisions, Simulation de scénarios, Modèle Boosting, Recommandations.

2.2 Stack technique

- **Langage** : Python 3.
- **Visualisation et dashboard** : Streamlit, Plotly.
- **Modélisation** : pandas, numpy, statsmodels (Holt-Winters, SARIMA), scikit-learn (Ridge), XGBoost/Boosting pour le modèle d'apprentissage.
- **Données** : Fichiers CSV générés (admissions par date/service, occupation quotidienne), stockés dans `data/generated/` et `data/processed/`.

2.3 Conformité réglementaire et éthique (données de santé)

Le développement respecte les contraintes légales et éthiques liées aux données de santé (réf. consignes) :

- **Données fictives** : uniquement des données générées ; aucune donnée nominative réelle.
- **Finalité** : simulation et prévision des besoins (lits, personnel, matériel) pour la direction ; pas de suivi individuel des patients.
- **Minimisation** : seuls les champs nécessaires aux modèles et au dashboard sont générés (date, service, admissions, occupation_lits).
- **Périmètre d'usage** : prototype de démonstration ; une utilisation en production sur données réelles nécessiterait une analyse d'impact (AIP), un DPO et, le cas échéant, un hébergement des données de santé (HDS).

Le détail est documenté dans le fichier `CONFORMITE.md` à la racine du projet.

---

3. Analyse des tendances d'admissions

3.1 Périodes critiques (littérature et données fictives)

La littérature (Bouteloup 2020, urgences Pellegrin ; rapports DREES, Santé publique France) identifie des **pics saisonniers** : hiver (grippe, bronchiolite), début de semaine (lundi), et une sensibilité aux **jours fériés** et aux **vacances scolaires**. Les données fictives générées pour le MVP reproduisent ces tendances (indices mensuels : hiver +15 % à +18 %, été autour de 90–92 % ; jour de la semaine : week-end plus faible).

3.2 Stratégies hospitalières actuelles (benchmark)

Les établissements s'appuient souvent sur des **moyennes historiques** et des **règles empiriques** pour la planification des effectifs et des lits. Les modèles de prédiction du flux (GAM, ARIMA, régression avec calendrier) sont documentés dans la littérature (Bouteloup, Batal et al.) et montrent qu'une **anticipation du flux** permet de réduire les départs sans soins et les plaintes (Batal : −18,5 % et −30 % respectivement lorsque le planning est adapté à la prédiction). Notre MVP s'inscrit dans cette logique : fournir des prévisions et des seuils d'alerte pour ajuster les ressources.

---

4. Analyse statistique et dataviz

4.1 Modèle statistique (rappel consigne)

En data science, un **modèle statistique** est une représentation mathématique (généralement basée sur la théorie des probabilités) qui décrit la manière dont sont générées les données observées. Il repose sur des hypothèses concernant : (1) la distribution des variables, (2) les relations entre elles (corrélations, dépendances), (3) les paramètres (moyennes, variances, coefficients). L'objectif est d'**expliquer** les données, **prédire** des valeurs futures ou des probabilités, et **estimer** des grandeurs inconnues en quantifiant l'**incertitude** (intervalles de confiance).

4.2 Justification des visualisations implémentées

Le tableau de bord propose :

- **Courbes temporelles** (admissions, occupation) : lecture directe des tendances et des pics.
- **Répartition par service** (camembert) : part de chaque service dans le flux.
- **Heatmaps** (jour de la semaine × mois) : identification des combinaisons les plus chargées, en cohérence avec la littérature (lundi, hiver).
- **Prévisions avec intervalles de confiance (IC 95 %)** : fourchette plausible pour la décision ; la littérature (Bouteloup) recommande de privilégier la **borne haute** pour la planification afin d'éviter la sous-estimation du flux.
- **Backtest** (prévu vs observé) : évaluation de la qualité des modèles sur une période passée.

Ces choix sont justifiés par la nécessité de fournir aux décideurs des indicateurs **interprétables** et **opérationnels** (horizon court à moyen terme).

4.3 Modèles statistiques utilisés et applicabilité

Les modèles utilisés sont des **modèles de prédiction** (séries temporelles, régression, modèle stock) dont les hypothèses statistiques (distributions des résidus, relations linéaires ou additives, stationnarité) sont décrites dans la section 5 et dans le document *JUSTIFICATION-MODELES-PREDICTION.md*. Leur applicabilité est limitée au **contexte des données agrégées** (flux journalier, occupation) ; une extension à des données patient (PMSI, durée de séjour individuelle) requerrait d'autres modèles (réf. Lequertier 2022).

---

5. Modèles de prédiction

5.1 Familles envisagées et choix retenus

Famille	Choix dans le MVP	Justification
Séries temporelles	Holt-Winters (saisonnalité 7 j), SARIMA (optionnel)	Capture tendance et saisonnalité hebdo ; peu de variables exogènes.
Régression	Ridge (lags 1, 7, 14, moyenne j-7 à j-13, calendrier, température synthétique)	Aligné sur Bouteloup (GAM + lags 7–13) ; interprétable, régularisation L2.
Baseline	Moyenne glissante + tendance	Dernier recours si série courte ou échec des autres modèles.
Modèle stock	Occupation = f(admissions prédites, durée de séjour saisonnière)	Réf. Lequertier : durée de séjour variable (hiver +8 %, été −8 %) pour un réalisme accru.
Machine learning	Boosting (XGBoost/GBM) en onglet dédié	Apprentissage des patterns saisonniers (mois, saison_grippe) ; comparé au modèle principal via backtest.

L'ordre d'utilisation pour la prévision « meilleur modèle disponible » est : Holt-Winters → Ridge → SARIMA → Moyenne glissante. Le Boosting est proposé dans un onglet séparé avec métriques de validation (MAE, RMSE, % à ±10 %, biais).

5.2 Hypothèses et limites (synthèse)

- **Holt-Winters** : saisonnalité additive, tendance additive ; ne modélise pas explicitement les jours fériés ni les vacances.
- **Ridge** : relation approximativement linéaire ; température synthétique (courbe sinusoïdale), pas de données Météo France dans ce MVP.
- **SARIMA** : utilisé en fallback ; coût de calcul et risque de non-convergence.
- **Modèle stock** : agrégé ; pas de prédiction individuelle de la durée de séjour.
- **Données** : fictives ; les performances (ex. % à ±10 %) ne sont pas généralisables sans validation sur données réelles.

5.3 Évaluation et impact de l'utilisation des modèles

- **Critère opérationnel** : **% de jours à ±10 %** (écart relatif prédit vs observé), en cohérence avec Bouteloup (83,79 % sur Pellegrin avec modèle lag). Sur données synthétiques, ce pourcentage est souvent élevé (85–95 %) car les séries sont lisses ; sur données réelles, 70–85 % serait déjà un bon résultat.
- **Métriques complémentaires** : biais moyen, % de sous-estimation / surestimation. La **sous-estimation** est plus risquée pour la planification (Bouteloup) ; les recommandations du prototype privilégient la **borne haute** de l'IC 95 % lorsque disponible.
- **Impact attendu** : en permettant d'anticiper les pics (alertes 85 % / 95 %, recommandations), l'outil vise une meilleure répartition des ressources, une réduction des temps d'attente et une limitation des situations de saturation (réf. Batal et al. sur l'adaptation du planning).

Le détail des justifications méthodologiques, du protocole de validation et des références bibliographiques figure dans *JUSTIFICATION-MODELES-PREDICTION.md*.

---

6. Étude d'impact et recommandations stratégiques

(Partie intégrante de ce document.)

6.1 Efficacité de l'outil sur la gestion hospitalière

Le MVP fournit des **prévisions d'occupation** (taux, intervalles de confiance) et des **recommandations automatiques** (niveau normal, alerte, critique) avec des actions concrètes (renforts, report d'activité, vigilance sur les stocks). En situation de pic prévu, la direction peut ajuster les effectifs et les lits en amont. La littérature (Batal et al.) montre qu'adapter le planning à la prédiction du flux permet de réduire significativement les départs sans soins et les plaintes. L'efficacité réelle devra être mesurée en conditions réelles (données réelles, déploiement en établissement) avec des indicateurs tels que le taux d'occupation effectif, les délais de prise en charge et la satisfaction des équipes.

6.2 Comparaison avec les solutions existantes

Les solutions actuelles en milieu hospitalier vont des **tableaux Excel** et **moyennes historiques** à des **outils de pilotage** plus avancés (tableaux de bord d'activité, parfois modules de prévision). Les thèses intégrées (Bouteloup, Lequertier) et les rapports DREES/Santé publique France montrent que la **prédiction du flux** (urgences, admissions) avec des modèles statistiques ou de ML est un sujet actif. Notre MVP se différencie par : (1) la combinaison **prévisions + simulation de scénarios** (épidémie, grève, canicule, afflux massif) ; (2) des **recommandations** liées aux seuils d'alerte et à la borne haute de l'IC ; (3) une interface unique (Flux, Prévisions, Simulation, Boosting, Recommandations). Les axes d'amélioration possibles incluent l'intégration de données météo réelles, la prévision de la durée de séjour (réf. Lequertier) et une comparaison systématique des modèles (tableau MAE, RMSE, % ±10 %).

6.3 Recommandations stratégiques et évolutions futures possibles

- **Déploiement** : en cas de passage sur données réelles, mettre en place le cadre réglementaire (AIP, DPO, HDS si applicable) et des connecteurs sécurisés ; ne traiter que des données agrégées ou anonymisées dans le dashboard.
- **Évolutions fonctionnelles** : seuils d'alerte configurables, recommandations par service, export des recommandations (PDF/CSV), planification des lits et du personnel (ratios ETP, créneaux de déprogrammation).
- **Évolutions des modèles** : comparaison systématique Holt-Winters / Ridge / SARIMA / Boosting ; modèle ensembliste ; données météo réelles ; prédiction de la durée de séjour (LOS) pour affiner le modèle stock.

Ces pistes sont détaillées dans le document *PISTES-EVOLUTION.md* (dossier 05-reference).

---

7. Synthèse et perspectives

Le rapport a présenté les **fonctionnalités du prototype** (données, prédiction, simulation, recommandations, dashboard), la **méthodologie de développement** et la **conformité** (données de santé). L'**analyse des tendances** et l'**analyse statistique** ont été justifiées (dataviz, modèles statistiques, rappel de la définition d'un modèle statistique). Les **modèles de prédiction** (Holt-Winters, Ridge, SARIMA, modèle stock, Boosting) ont été décrits avec leurs hypothèses, limites et évaluation (métrique ±10 %, biais, IC 95 %). L'**étude d'impact** et les **recommandations**

stratégiques ont porté sur l'efficacité de l'outil, la comparaison avec l'existant et les évolutions futures.

Limites principales : données fictives ; pas de validation opérationnelle sur données réelles ; prévision agrégée (pas de prédiction individuelle de la durée de séjour). **Perspectives** : validation sur données réelles (PMSI, RPU) dans un cadre réglementaire ; renforcement des modèles (météo, durée de séjour, comparaison systématique) et des fonctionnalités (alertes configurables, recommandations par service, planification).

Projet Data Pitié-Salpêtrière — Promo 2026. Données fictives. Document unique : rapport de conception et d'analyse hospitalière + étude d'impact et recommandations stratégiques.