

# Cahier de charge – Profil Big Data / ML

**Projet :** Kidjamo

**Date :** 14 août 2025

**Rédigé par :** Christian DJOUNDA NGAPGHO MOMO

**Version :** 1.0

## **1. Résumé exécutif**

Kidjamo est une solution de santé connectée qui combine bracelet IoT, application mobile/web et IA pour prévenir les crises drépanocytaires. Le projet livrera une architecture cloud sécurisée, des pipelines temps réel & batch, un modèle ML validé et des API de service pour l'alerte et le suivi.

## **2. Contexte du projet**

La drépanocytose provoque des crises vaso-occlusives liées à la déshydratation et aux infections. L'absence d'alertes précoces et de suivi continu entraîne des hospitalisations tardives et évitables. Kidjamo vise une surveillance proactive, inclusive et frugale.



## **3. Objectifs du projet :**

### **A. Objectifs spécifiques :**

- Mettre en place l'architecture cloud Big Data pour ingestion, stockage et traitement temps réel des données IoT et app.
- Développer des pipelines de données pour nettoyer, transformer et historiser les données.
- Construire et entraîner un modèle ML pour détecter les risques de crises et générer des alertes.
- Déployer les modèles en production et assurer le monitoring en continu.
- Garantir la sécurité et la conformité des données de santé (RGPD / HIPAA-like)

### **B. Objectifs fonctionnels :**

- Alertes patient/parents/médecin en cas de risque (push/SMS) si risque de crise détectée.
- Suivi quotidien (hydratation estimée, fièvre, Oxygénation, fréquence cardiaque).
- Console admin

- Portail soignant : tableaux de bord patients, historique, courbe et exports.

### **C. Objectifs techniques**

- Architecture data hybride (SQL/NoSQL + Data Lake)
- Pipelines streaming & batch des données IoT/mobile.
- Entraînement/serving ML avec MLOps.
- API sécurisées (chiffrement, IAM) et conformité.

### **D. Périmètre fonctionnel**

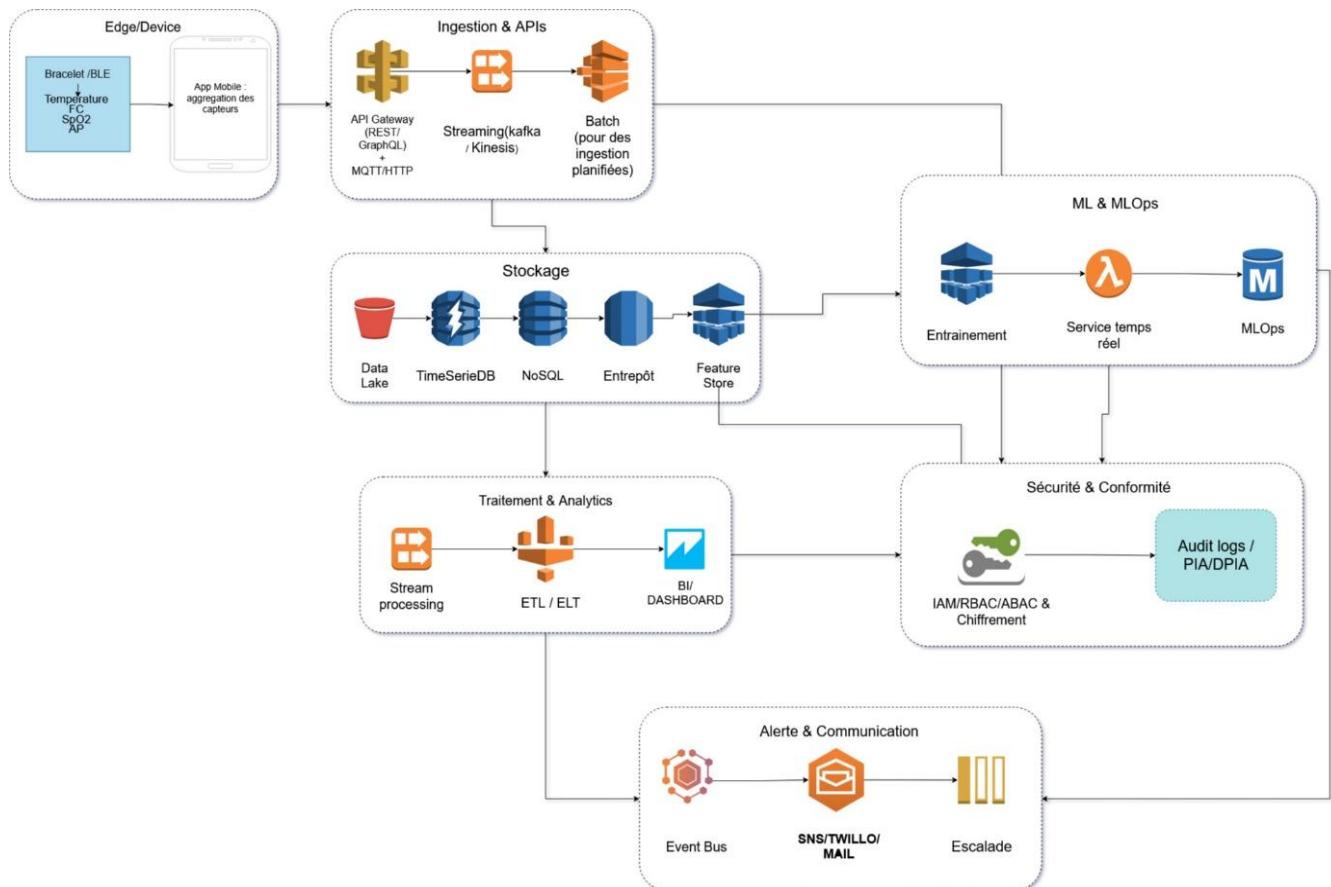
Fonctionnalité	Patient	Parent/Tuteur	Médecin/Soignant	Admin
<b>Suivi temps réel (Temp., SpO<sub>2</sub>, FC, activité)</b>	✓	✓ (lecture)	✓ (tableau patient)	—
<b>Journal santé (douleur, crises, médication)</b>	✓ (saisie)	✓ (co-saisie)	✓ (lecture)	—
<b>Alertes &amp; notifications (push/SMS)</b>	✓	✓	✓	✓ (règles)
<b>Téléconsultation/messagerie</b>	✓	✓	✓	—
<b>Tableaux de bord &amp; tendances</b>	✓ (perso)	✓ (enfant/perso)	✓ (cohorte)	✓
<b>Gestion comptes &amp; droits</b>	—	—	—	✓
<b>Export/Partage dossier</b>	✓	✓	✓	✓

Domaine	Tâches principale
<b>Collecte &amp; Ingestion</b>	IoT => Cloud (API), Streaming temps réel, Batch processing.
<b>Stockage &amp; Architecture Data</b>	SQL(TimescaleDb), NoSQL (DynamoDB), Data Lake Cloud
<b>Traitement &amp; Big Data</b>	ETL/ELT, nettoyage, Spark/Streaming Analytics/Kinesis
<b>Machine Learning &amp; IA</b>	Préparation, entraînement, déploiement modèle, API ML
<b>Visualisation &amp; Exploitation (Optionnel)</b>	TB, KPI santé, export PDF/Excel

## Hors périmètre (V2+) :

- Marketplace capteurs avancés...

### 4. Architecture technique (vue logique)



### 5. Estimation volumétrie & dimensionnement :

**Formule :** points/jour = (nb\_patients × fréquence\_Hz × 86 400 × nb\_variables).

Hypothèses : 6 variables ; 24 o/point compressé.

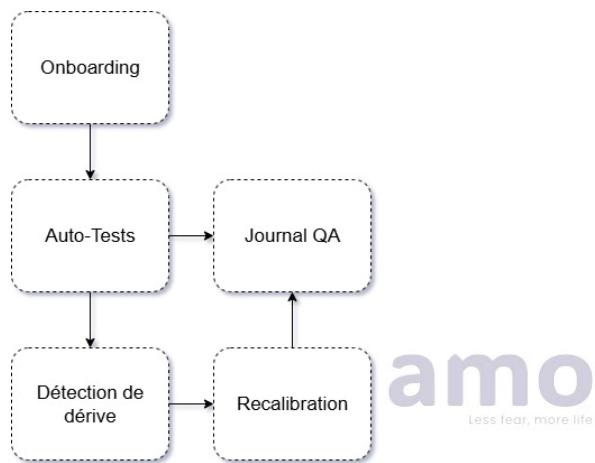
Scénario	Patients	Fréq. (Hz)	Points/jour	Stockage/jour	Stockage/mois
Pilote (MVP)	300	0,2	~31 104 000	~0,7 Go	~21 Go
Croissance	2 000	0,5	~518 400 000	~11,6 Go	~348 Go
Ambitieux	5 000	1,0	~2 592 000 000	~58 Go	~1,74 To

**Critères d'acceptation :** capacité d'ingestion > 2× pic ; latence P95 < 5 s (pilote) ; coût mensuel ≤ budget (FinOps).

## **6. Protocole de calibration capteurs & assurance qualité**

Onboarding	Auto-Test	Détection de dérive	Journal QA / capteur
Vérification matériel, calibration guidée (référence clinique si possible, sinon repos 5 min).	Quotidien (PPG/IMU/Temp.), hebdomadaire (offsets T°/humidité).	Cohérence inter-signaux, filtres robustes (Hampel).	score_signal, % valides, flatline, pertes BLE.

**KPIs** : % échantillons valides  $\geq 95\%$  /jour ; alerte recalibration si dérive  $\text{SpO}_2 > \pm 2\%$  ou  $T^\circ > 0,5^\circ\text{C}$  sur 24 h.

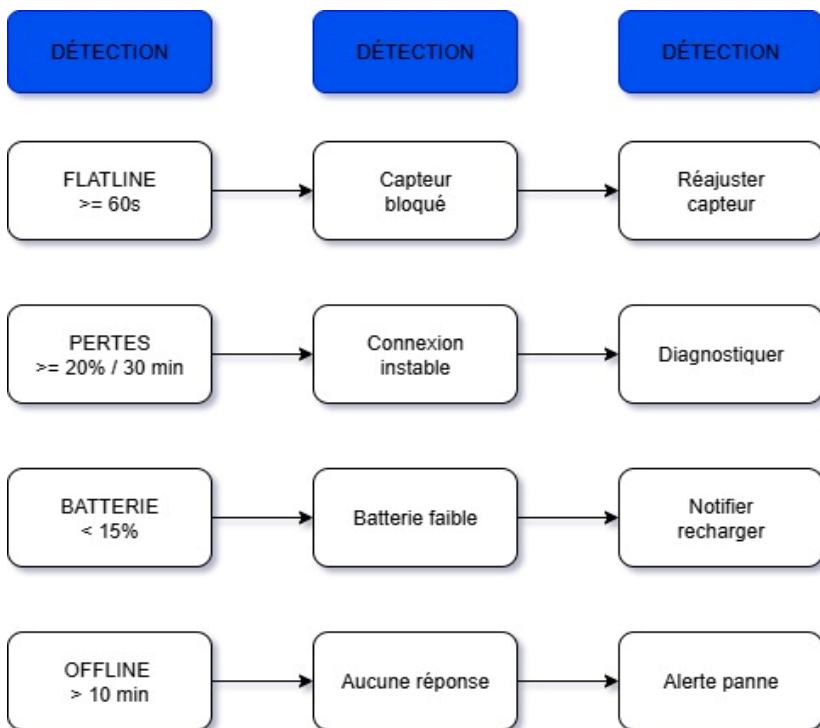


**amo**  
Less fear, more life

## **7. Détection d'anomalies capteur & pannes**

- Flatline  $\geq 60$  s; pertes  $> 20\% / 5$  min; batterie  $< 15\%$ ; heartbeat device 60 s (offline  $> 10$  min).
- Alerta panne :  $< 1$  min (local),  $< 5$  min (cloud) ; faux positifs pannes  $< 5\%$ .

Condition détectée	Cause probable	Action à prendre
<b>Flatline ≥ 60 s</b>	Capteur bloqué, bug logiciel, obstruction physique (bracelet mal placé)	Vérification automatique → si persiste, notifier utilisateur de réajuster/repositionner le capteur
<b>Pertes &gt; 20 % / 5 min</b>	Problème de connexion BLE/Wi-Fi, interférences, capteur surchargé	Diagnostiquer connectivité → relancer session, optimiser fréquence d'envoi
<b>Batterie &lt; 15 %</b>	Batterie proche de la coupure	Notification immédiate pour recharge + log événement batterie
<b>Pas de heartbeat depuis 10 min</b>	Capteur hors ligne, panne matérielle, batterie vide, hors portée réseau	Alerte au cloud + tentative de reconnexion automatique
<b>Alerte panne (local) &lt; 1 min</b>	Panne détectée sur le device/app passerelle	Notification locale instantanée à l'utilisateur et/ou médecin
<b>Alerte panne (cloud) &lt; 5 min</b>	Panne détectée par surveillance serveur	Alerte envoyée par SMS/email/push au référent
<b>Faux positifs pannes &lt; 5 %</b>	Bruit de données ou seuils mal calibrés	Ajuster algorithmes et filtres pour réduire erreurs d'alerte



## **8. Mémoire tampon locale & synchronisation différée**

- a. Buffer bracelet 30-60min (AES-128), cache mobile >= 24h**
  - **Bracelet** : s'il perd le lien BLE/réseau, il **met en file d'attente** les mesures pendant **30–60 min** dans sa mémoire interne, **chiffrées en AES-128** (confidentialité même si on vole le bracelet).
  - **Téléphone (app)** : il garde une **copie locale chiffrée** dans une base **SQLite** pendant **≥ 24 h**.
  - **Objectif** : ne perdre aucune donnée si l'utilisateur n'a pas de réseau ( métro, zone blanche). À la reconnexion, tout repart.
  
- b. Reprise par lots avec backoff ; clés idempotentes ; déduplication côté API**
  - **Reprise par lots** : quand le réseau revient, l'app envoie par paquets (ex. 500 records) plutôt qu'un par un => plus rapide et stable.
  - **Backoff** : si l'API répond "trop chargé"/erreur, on réessaie plus tard en espaçant les tentatives (ex. 1s, 2s, 4s, 8s...) pour ne pas saturer le serveur.
  - **Déduplication côté API** : l'API garde une table de reçus (message\_id, device\_id, ts\_reçu).
  - **Objectif** : fiabilité et exact-once même avec réseaux capricieux.
  
- c. Mode offline : règles locales (SpO2 < 88%, T° > 38,5°C) + protocole d'urgence**
  - **Sans internet**, l'app mobile applique des règles vitales en local :
    - o SpO2 < 88% => alerte immédiate (risque hypoxie).
    - o T° > 38,5°C => suspicion infection/fièvre => une alerte.
  - **Protocole d'urgence affiché** : comme notification, **instructions claires** (boire, s'hydrater, antalgiques selon protocole, contacter centre, numéro d'urgence, localisation si possible).
  - **Objectif** : protéger le patient même hors ligne ; l'alerte ne dépendra pas du cloud.

## **9. Module d'explicabilité ML**

### **A. Modèles interprétables (Logistic/GBT) + SHAP local**

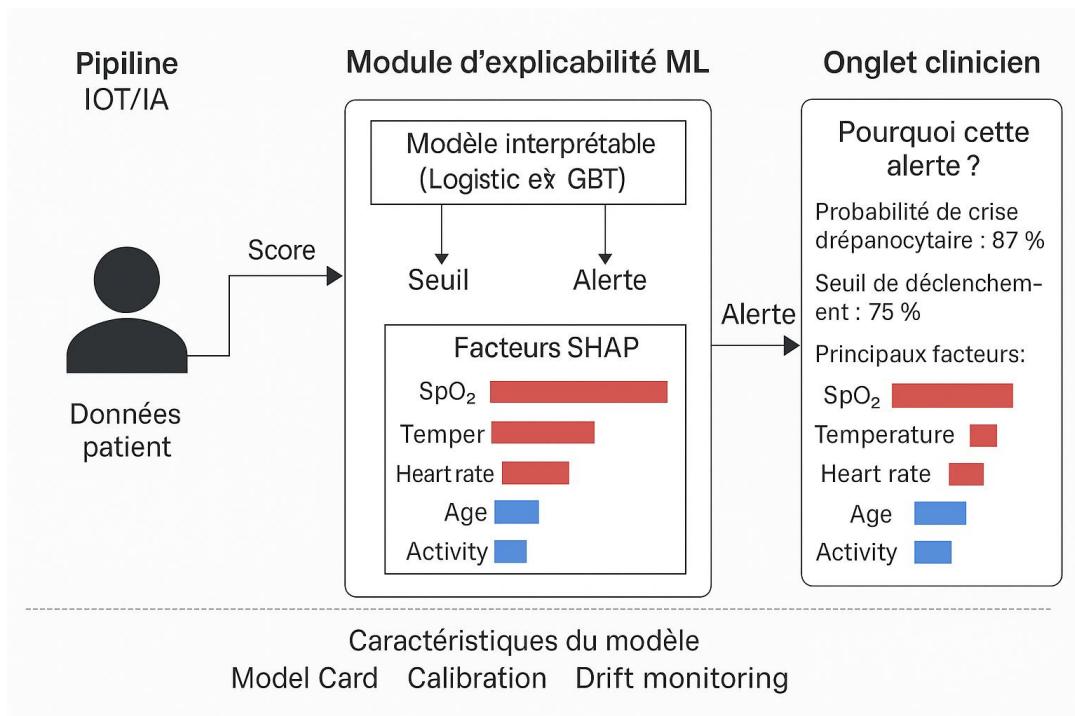
- a. Logistic Regression ou Gradient Boosted Trees (GBT)**
  - Ce sont des modèles ML plus simples à **expliquer** que des réseaux neuronaux profonds.
- b. SHAP local** (Shapley Additive exPlanations)
  - Méthode issue de la théorie des jeux qui explique pour chaque prédiction quels facteurs ont le plus contribué au résultat.

### B. Chaque alerte inclut score, seuil et top facteurs.

- **Score** : probabilité estimée par le modèle (exemp.  $0,87 = 87\%$  de risque de crise).
- **Seuil** : limite fixée pour déclencher l'alerte (exemp. alerte si score  $\geq 0,75$ ).
- **Top facteurs** : les variables les plus influentes pour cette décision (exemp. SpO<sub>2</sub>, température, rythme cardiaque).
- **Objectif** : donner un résumé **quantitatif et qualitatif** à chaque alerte.

### C. Drift monitoring

- **Surveillance de la dérive** : détection quand le modèle voit des données différentes de son entraînement (changeement population, saison, capteurs).
- **But** : déclencher une ré-évaluation ou un ré-entraînement.



### 10. Livrables attendus

- Schémas data
- Pipelines de données opérationnels pour ingestion, transformation et stockage
- Modèle ML entraînés et déployés
- Datasets synthétiques + scénarios de test.
- API ML pour intégration avec mobile et web
- Tableaux de bord et monitoring en temps réel
- Documentation technique complète.
-

## **11. Planification (Agile – 6 sprints)**

Sprint	Semaine(s)	Objectifs principaux	Livrable / Jalon
<b>S1 – S5 Pré-projet</b>	Semaine 1-5	<ul style="list-style-type: none"> <li>Cadrage du projet</li> <li>Choix des technologies</li> <li>Création des comptes Cloud</li> <li>Mise en place sécurité de base</li> </ul>	—
<b>S6 – Fondations Cloud &amp; ingestion</b>	Semaine 6	<ul style="list-style-type: none"> <li>Création infra Cloud (VPC, buckets, secrets, IAM).</li> <li>Mise en place topics Kafka/Kinesis.</li> <li>Déploiement API Gateway.</li> </ul>	<b>Jalon 1</b> : premier flux IoT → Data Lake / TimeSeries (moins important)
<b>S7 – Pipelines &amp; qualité</b>	Semaine 7	<ul style="list-style-type: none"> <li>ETL/ELT bruts vers zones bronze/silver.</li> <li>Tests de qualité des données</li> <li>Définition et validation des schémas</li> </ul>	<b>Jalon 2</b> : dashboard technique (moins important) ingestion + qualité
<b>S8-S9 – ML v1 &amp; alertes</b>	Semaines 8-9	<ul style="list-style-type: none"> <li>Mise en place Feature Store</li> <li>Développement premiers modèles (baselines + règles)</li> <li>Création endpoints</li> <li>Moteur d'alertes et notifications</li> </ul>	<b>Jalon 3</b> : alerte temps réel sur jeu de données simulées
<b>S10 – Portails &amp; BI</b>	Semaine 10	<ul style="list-style-type: none"> <li>Développement tableaux de bord (soignant / admin) (je ne sais pas encore, si rôle des dev)</li> <li>Système d'exports (collaboration avec dev)</li> <li>Gestion des rôles &amp; permissions (collaboration avec dev)</li> </ul>	<b>Jalon 4</b> : démonstration clinique interne (UX soignant)
<b>S11 – MLOps &amp; sécurité avancée</b>	Semaine 11	<ul style="list-style-type: none"> <li>Registry modèles</li> <li>CI/CD pour modèles ML</li> <li>Drift monitoring</li> <li>PIA/DPIA</li> <li>Journal de sécurité</li> </ul>	<b>Jalon 5</b> : décision go/no-go pilote terrain
<b>S12-S13 – Buffer / Pilote</b>	Semaines 12-13	<ul style="list-style-type: none"> <li>Observabilité complète</li> <li>Calibrage des seuils</li> <li>Collecte retours utilisateurs</li> <li>Documentation finale</li> </ul>	<b>Jalon 6</b> : pilote validé + plan d'industrialisation

## 12. Analyse des risques et contraintes :

Risque	Prob.	Impact	Mitigation	Owner
Données insuffisantes/bruitées	Moyen	Haute	Collecte pilote, règles qualité, imputation	Data
Modèle ML non fiable	Moyen	Haute	Baseline simple, features robustes, validation clinique	Data
Instabilité IoT/connectivité	Haute	Moyen	Buffering edge, retry, SMS fallback	IoT
Coûts cloud	Moyen	Moyen	Compression, tiering S3, autoscaling, budgets/alerts	Dev & Data
Sécurité/Conformité	Large	Haute	Chiffrement, IAM least privilege, DPIA, audit	Dev & Data
Adoption utilisateur	Moyen	Moyen	UX simple, co-design avec patients/soignant, support et information.	UI & Dev

## 13. Gouvernance & RACI



Domaine	Responsable (R)	Appui (A)	Consulté (C)	Informé (I)
Architecture Cloud & Data	Data/Cloud Lead (toi)	CTO	Dev Back	Équipe
Pipelines & Qualité	Big Data	Data/Cloud Lead	Dev Mobile, IoT	-
Modèles ML & MLOps	Big Data – ML	Big Data - ML	Médecins	-
Portails & APIs	Dev Back/Web	Big Data - ML	UI/UX, Médecins	Équipe
Sécurité/Conformité		Big Data - ML	Juridique	Équipe
Pilote terrain	-	Équipe	Médecins, AS communautaires	Stakeholders

## 14. Proposition de valeur et impact

- Prévention des crises (alertes précoces) => moins d'hospitalisations, meilleure qualité de vie.
- Accessibilité en zone à faible connectivité (buffer offline, règles locales)

## **15. Ciblage & déploiement**

- Pilote : 2 centres (Douala et Yaoundé), 300 patients.
- Passage à l'échelle : 2000 puis 5000 patients (capacités infra prévues)

## **16. Modèle économique & soutenabilité**

- Freemium social (alertes vitales, journal) + Pack Pro.
- 

## **17. Sécurité & conformité (santé)**

<b>Exigence</b>	<b>Mesure technique/organisationnelle</b>
<b>Chiffrement en transit/au repos</b>	TLS 1.2+, AWS KMS, disques chiffrés
<b>Gestion des accès</b>	IAM RBAC/ABAC, Secrets Manager, moindre privilège
<b>Traçabilité</b>	CloudWatch/CloudTrail, journaux d'accès et d'usage
<b>Privacy</b>	Consentement explicite, anonymisation/pseudo-anonymisation
<b>DPIA/PIA &amp; registre</b>	Évaluation d'impact, registre des traitements RGPD

## **18. Stack (référence possible)**

- **IoT:** BLE → mobile gateway, MQTT/HTTP.
- **Cloud:** AWS (API Gateway, IoT Core/Kinesis, S3, Glue, Lambda, ECS/EKS, SageMaker, Athena, Redshift, CloudWatch, SNS).
- **Data:** Kafka/Flink/Spark, DBT, TimescaleDB/InfluxDB, DynamoDB/MongoDB, QuickSight.
- **ML:** PyTorch/Scikit, Feast/Feature Store, MLflow/Registry, Great Expectations, Evidently AI (drift).
- **Sécurité :** KMS, WAF, IAM, Secrets Manager.

## 19. Seuil

Paramètre	0–1 an	2–5 ans	6–12 ans	13–18 ans	19–40 ans	41–60 ans	61–80 ans
Température corporelle (°C)	≥ 37.8 (alerte)	≥ 38.0 (alerte)	≥ 37.8 (alerte plus précoce, immunité ↓)				
Fréquence cardiaque (bpm)	> 160 ou < 90	> 140 ou < 80	> 120 ou < 70	> 110 ou < 60	> 100 ou < 55	> 100 ou < 55	> 95 ou < 50
Fréquence respiratoire (/min)	> 50 ou < 20	> 40 ou < 18	> 30 ou < 16	> 25 ou < 12	> 22 ou < 10	> 22 ou < 10	> 20 ou < 10
SpO <sub>2</sub> (%)	< 94 %	< 94 %	< 93 %	< 93 %	< 92 %	< 92 %	< 92 % (alerte si < 93)
Hydratation (L/jour)	< 0.7 L	< 1.0 L	< 1.2 L	< 1.5 L	< 2.0 L	< 1.8 L	< 1.5 L (surveillance ↑)
Indice de chaleur (°C ressenti)	> 35	> 36	> 37	> 38	> 40	> 38	> 37

Paramètre	Vert (OK)	Jaune (Surv.)	Orange (Action)	Rouge (Urgence)	Durée avant alerte	Notes & personnalisation
Température corporelle (°C)	36,1–37,5	37,6–37,9 (≥15 min)	38,0–38,4 (≥30 min) ou 35,0–35,4 (≥15 min)	≥ 38,5 à tout moment ; < 35,0	Immédiat si ≥38,5 ; sinon selon paliers	Seuil ≥38,5°C = urgence pour SCD (fièvre) ; confirmer avec thermomètre buccal/tympanique fiable.
SpO <sub>2</sub> (%)	≥ 94% ou ≥ (Baseline–2)	90–93% ou ↓ 3–4 pts vs baseline (≥3 min)	88–89% ou ↓ ≥5 pts vs baseline (≥2 min)	< 88% (≥1 min) ou chute rapide + dyspnée/cyanose	1–3 min (anti-bruit)	Configurer Baseline_SpO <sub>2</sub> patient. Si inconnue, utiliser 94% comme repère symptômes respiratoires.
Fréquence cardiaque au repos (bpm – adulte)	60–100 (50–100 si sportif)	101–120 (≥10 min) ou 50–59 avec symptômes	>120 (≥5 min) ou <50 avec symptômes	>140 (≥3 min) + douleur thoracique/syncopes ; <40	5–10 min (selon profil)	Adapter par âge/entraînement. Vérifier contexte (stress, ...)
Fréquence respiratoire au repos (cycles/min – adulte)	12–20	21–24 (≥10 min)	25–30 (≥5 min)	>30 ou <10 + détresse/altération de conscience	5–10 min	Associer à SpO <sub>2</sub> et température; risque d'ACS si fièvre + toux

Indice de chaleur (°C) – si humidité disponible	< 32	32–40 ( $\geq 30$ min)	41–53 ( $\geq 15$ min)	$\geq 54$ (exposition même brève)	15–30 min	Basé sur catégories NWS. Ajouter notifications d'hydratation
Température ambiante (°C) – si HI indisponible	18–28	29–31 ( $\geq 30$ min)	$\geq 32$ ( $\geq 20$ min)	$\geq 36$ ( $\geq 10$ min) + symptômes de coup de chaleur	10–30 min	Privilégier l'Indice de chaleur quand possible (plus précis)
Hydratation (auto-déclarée / suivi prise d'eau)	$\geq 150$ – $250$ mL/heure en période chaude	< 500 mL sur 4 h avec HI $\geq 32$ ou activité ↑	Aucune prise sur $\geq 6$ h + HI $\geq 32$ ou urines foncées	Confusion, syncope, vomissements incoercibles	Fenêtres glissantes 4–6 h	Objectif à personnaliser avec clinicien; viser urines claires. Intervalle de temps entre les prises d'eau

Fréquence respiratoire :



Tranche d'âge	Normale (au repos)	Alerte basse (Bradypnée)	Alerte haute (Tachypnée)
Nouveau-né (0–1 an)	30 – 60	< 30	> 60
Nourrisson (1–5 ans)	20 – 30	< 20	> 40
Enfant (6–12 ans)	18 – 25	< 15	> 30
Adolescent (13–17)	12 – 20	< 12	> 25
Adulte (18–59)	12 – 20	< 10	> 24
Sénior (60–80)	12 – 25	< 10	> 26

Fréquence\_cardiaque :

Tranche d'âge	Fréquence cardiaque normale (bpm)
Nouveau-né (0–1 an)	100 – 180
Nourrisson (1–5 ans)	90 – 140
Enfant (6–12 ans)	75 – 120
Adolescent (13–17)	60 – 100
Adulte (18–59 ans)	60 – 100
Sénior (60–80 ans)	60 – 100 (parfois 55–90)

Tranche d'âge	Normale	Alerte basse (brady)	Alerte haute (tachy)	Critique immédiate
Nouveau-né (0–11 m)	100–180	< 90	> 180	< 80 ou > 200
Nourrisson (1–5 a)	90–140	< 80	> 150	< 70 ou > 180
Enfant (6–12 a)	75–120	< 60	> 130	< 50 ou > 160
Adolescent (13–17)	60–100	< 50	> 120	< 45 ou > 150
Adulte (18–59)	60–100	< 50*	> 100	< 40 ou > 150
Sénior (60–80)	60–100	< 50	> 110	< 40 ou > 150

#### ◆ Combinaisons physiologiques critiques (alerte immédiate)

Certaines **combinaisons** sont plus dangereuses que les anomalies isolées :

##### 1. Fièvre + SpO<sub>2</sub> basse

- Température  $\geq 38^{\circ}\text{C}$  ET SpO<sub>2</sub>  $\leq 93\%$   
→ Suspicion infection + hypoxie → URGENCE

##### 2. Tachycardie + Hypotension

- FC > 120 (adulte) **ET TAS < 100 mmHg**  
→ Déshydratation sévère ou choc hémodynamique

### 3. **Douleur EVA ≥ 7 + SpO<sub>2</sub> ≤ 92 %**

→ Crise vaso-occlusive sévère → oxygénation + antalgiques forts requis

### 4. **Déshydratation + Chaleur extrême**

- Apports < 1.5 L **ET indice de chaleur > 38 °C**  
→ Risque majeur de crise en extérieur

### 5. **Fièvre + Tachypnée**

- Température ≥ 38 °C **ET FR > 25/min**  
→ Syndrome thoracique aigu suspecté (urgence vitale)

