



Cancer Research Analytics

Analyse Comparative de la Recherche en Oncologie

Projet d'Analyse de Données | Décembre 2025 - Janvier 2026

Problématique

"L'investissement en recherche est-il corrélé à la gravité réelle des maladies ?"

Cette étude révèle des disparités critiques dans l'allocation des ressources de recherche oncologique, créant un décalage majeur entre l'impact sanitaire réel et l'effort de recherche déployé.

Constat Principal

Le cancer du poumon reçoit 3,7 fois moins de budget par mort que la leucémie, malgré une mortalité 6 fois supérieure.

| Cancer | Budget 2023 | Mortalité 2022 | \$ par mort |
|-------------------|-------------|----------------|-------------|
| Breast Cancer | \$542M | 666,103 | \$814 |
| Leukemia | \$270.6M | 305,405 | \$886 |
| Prostate Cancer | \$258.6M | 397,430 | \$651 |
| Pancreatic Cancer | \$246M | 467,409 | \$526 |
| Lung Cancer | \$435M | 1,817,469 | \$239 ⚠ |

Table des Matières

I. Introduction

- [Vue d'ensemble](#)
- [Objectifs du projet](#)
- [Types de cancer analysés](#)

II. Méthodologie

4. [Sources de données](#)
5. [Architecture du projet](#)
6. [Pipeline de traitement \(ETL\)](#)

III. Implémentation Technique

7. [Scripts de collecte](#)
8. [Nettoyage et transformation](#)
9. [Stockage et persistance](#)
10. [Dashboard interactif](#)

IV. Résultats et Analyse

11. [KPIs calculés](#)
12. [Analyse des biais](#)
13. [Résultats clés](#)

V. Conclusion

14. [Réponse à la problématique](#)
 15. [Technologies utilisées](#)
 16. [Pistes d'amélioration](#)
-

I. Vue d'Ensemble

Objectifs du Projet

Ce projet vise à révéler les disparités dans la recherche sur le cancer en croisant plusieurs dimensions :

- **Mortalité mondiale** (OMS 2022)
- **Financement public** (Budget NCI 2023)
- **Activité scientifique** (Publications PubMed 2024)
- **Essais cliniques** (ClinicalTrials.gov - 5000+ essais)
- **Visibilité médiatique** (Google Trends)

Chiffres Clés

-  **5000+ essais cliniques** analysés
-  **800+ mots-clés géographiques** pour classification
-  **8 régions du monde** cartographiées
-  **10 KPIs** calculés
-  **8 tables de données** structurées
-  **15+ visualisations** interactives

Types de Cancer Analysés

1. **Lung Cancer** (Cancer du poumon) - Mortalité : 1,8M décès/an
2. **Breast Cancer** (Cancer du sein) - Mortalité : 666K décès/an
3. **Pancreatic Cancer** (Cancer du pancréas) - Mortalité : 467K décès/an
4. **Prostate Cancer** (Cancer de la prostate) - Mortalité : 397K décès/an
5. **Leukemia** (Leucémie) - Mortalité : 305K décès/an

Ces 5 cancers représentent une part importante de la mortalité mondiale et des investissements en recherche.

II. Sources de Données

1. Organisation Mondiale de la Santé (OMS)

- **Dataset** : GLOBOCAN 2022 - Mortalité mondiale par cancer
- **Format** : CSV
- **Données** : Nombre de décès par type de cancer, toutes régions confondues

2. National Cancer Institute (NCI)

- **Dataset** : Budget de recherche 2023
- **Source** : <https://www.cancer.gov/about-nci/budget/fact-book/data/research-funding>
- **Données** : Financement alloué par type de cancer en millions USD

3. ClinicalTrials.gov

- **Méthode** : Web scraping avec Selenium
- **Données collectées** :
 - ID de l'essai clinique (NCT)
 - Titre de l'essai
 - Sponsor (institution/entreprise)
 - Statut (Recruiting, Completed, Terminated, etc.)

- **Volume** : 30 pages par cancer (5000 essais au total)

4. PubMed (NCBI)

- **Méthode** : API publique (E-utilities)
- **Données** : Nombre de publications scientifiques par cancer (2024)
- **Requête** : Recherche dans Title/Abstract + filtre année 2024

5. Google Trends

- **Méthode** : Export CSV manuel
 - **Données** : Score d'intérêt médiatique par pays et par cancer
-

Pipeline de Traitement (ETL)

1. Scraping ClinicalTrials.gov

Script : `scripts/1_Scrapping.py`

Objectif : Collecter les données des essais cliniques pour chaque type de cancer.

Défis techniques surmontés :

1. Structures dynamiques

- Problème : Les sélecteurs CSS changent fréquemment
- Solution : Aspiration de TOUT le texte brut pour contourner les erreurs

2. Auto-Healing

- Vérifie à chaque action que le navigateur répond
- Redémarrage automatique en cas de crash

3. Sauvegarde incrémentale

- Enregistrement tous les 10 essais pour minimiser les pertes

4. Contournement des blocages

- `time.sleep(2)` entre requêtes pour imiter un comportement humain

Performance : Plus de 5000 essais cliniques collectés.

2. Recherche API PubMed

Script : `scripts/2_ApiSearch.py`

Objectif : Récupérer le nombre de publications scientifiques par cancer en 2024.

Fonctionnement :

```
term = '"Lung Cancer" [Title/Abstract] AND 2024 [Date - Publication]"
```

Utilise l'API NCBI E-utilities pour compter les publications par type de cancer.

3. Nettoyage et Enrichissement

Script : `scripts/3_Nettoyage.py`

Objectif : Nettoyer, normaliser et enrichir toutes les données collectées.

3.1 Nettoyage mortalité OMS

- Garde colonnes `Label` et `Mortality`
- Supprime lignes vides
- Trie par mortalité décroissante

3.2 Enrichissement PubMed

- Mapping entre noms PubMed et OMS
- Calcul : **Publications par 1000 décès**

3.3 Nettoyage essais cliniques

- Extraction ID NCT depuis URL
- Déduplication (5000+ → données uniques)
- Compte des essais par cancer

3.4 Géolocalisation Précise

Base de 800+ mots-clés pour classifier les essais par région :

USA (300+ mots-clés) :

- Hôpitaux : Johns Hopkins, MD Anderson, Mayo Clinic...
- Universités : Harvard, Yale, Stanford...

- Pharma : Pfizer, Merck, Johnson & Johnson...

Europe (250+ mots-clés) :

- Hôpitaux : NHS UK, Institut Curie, Charité Berlin...
- Pharma : Roche, Novartis, AstraZeneca...

Asia (150+ mots-clés) :

- Hôpitaux : Peking Union, Samsung Medical Center...
- Pharma : Takeda, Daiichi Sankyo...

Classification en 8 régions : USA, Europe, Asia, Canada, Latin America, Middle East, Oceania, Other

4. Upload vers Supabase

Script : `scripts/4_Supabase.py`

Objectif : Migrer les données nettoyées vers une base PostgreSQL cloud.

Fonctionnement :

- Upload par batchs de 1000 enregistrements
- API REST Supabase avec authentification
- Gestion d'erreurs et logs détaillés

Tables créées :

- `cancer_mortality`
 - `research_vs_mortality`
 - `clinical_trials`
 - `geography_count`
 - `google_trends`
 - `nci_budget`
-

5. Orchestration Automatique

Script : `main.py`

Orchestre l'exécution séquentielle de tous les scripts avec :

- Logs colorés et détaillés
- Vérification de l'existence des fichiers générés

- Gestion d'erreurs avec traceback
 - Résumé final du pipeline
-

Dashboard Interactif

Script : `dashboard.py`

Fonctionnalités

Filtres Dynamiques

- Sélection multi-cancer
- Filtre par statut d'essai clinique
- Ajustement hauteur des graphiques

KPIs en Temps Réel

- Total des essais cliniques
- Budget moyen NCI
- Cancer le plus mortel
- Intérêt médiatique moyen

4 Onglets d'Analyse

1. Recherche & Mortalité

- Corrélation mortalité vs publications
- Publications par 1000 décès
- Research Gap (décès/publication)

2. Budget NCI

- Répartition du budget 2023
- Budget par type de cancer
- Budget par décès

3. Essais Cliniques

- Distribution géographique (stacked bar)
- Total par cancer
- Heatmap cancer × région

4. Tendances Média

- Intérêt médiatique (Google Trends)
- Répartition attention médiatique
- Visibilité vs gravité sanitaire

Design

- Style sobre et professionnel
 - Palette cohérente (bleu #2d5a7b)
 - Typographie soignée (Georgia, Arial)
 - CSS personnalisé pour Streamlit
-

KPIs Calculés

KPI1 : Nombre total d'essais cliniques

```
SELECT COUNT(*) AS total_trials
FROM clinical_trials;
```

KPI2 : Essais par type de cancer

```
SELECT cancer, COUNT(*) AS trials_count
FROM clinical_trials
GROUP BY cancer
ORDER BY trials_count DESC;
```

KPI4 : Statistiques budget NCI

```
SELECT
    AVG(budget_2023_million_usd) AS avg_budget,
    MIN(budget_2023_million_usd) AS min_budget,
    MAX(budget_2023_million_usd) AS max_budget
FROM nci_budget;
```

KPI8 : Research Gap

```
SELECT
    cancer,
    mortality_2022,
    publications_2024,
    (mortality_2022 * 1.0 / NULLIF(publications_2024, 0))
        AS deaths_per_publication
FROM research_vs_mortality
ORDER BY deaths_per_publication DESC;
```

(Voir *scripts/KPI.sql* pour les 10 KPIs complets)

Analyse des Biais Structurels

1. Biais Médiatique

Le cancer du sein bénéficie d'une triple surreprésentation :

- Intérêt médiatique : **24,1** (Google Trends)
- Budget : **\$542M** (30,9% du budget NCI)
- Intensité recherche : **36,9** publications/1000 décès

Explication : Décennies de mobilisation associative (Ruban Rose, Octobre Rose).

2. La Grande Injustice : Cancer du Poumon

Triple handicap malgré la mortalité la plus élevée :

- Intérêt médiatique : **18,8** (le plus faible)
- Investissement : **\$0,2k par mort** (3,7x moins que leucémie)
- Intensité recherche : **9,5** publications/1000 décès

Ratio : 185 décès par publication - le cancer le plus sous-recherché.

3. Cancer Pancréatique : L'Oublié Mortel

Taux de survie à 5 ans < 10%, pourtant :

- Investissement : **\$0,5k par mort**
- Intensité recherche : **10,0** publications/1000 décès

- Ratio : **108 décès par publication**
-

4. Inégalités Géographiques

Concentration dans les pays riches :

- **85% de la population mondiale** exclue des essais
 - USA/Europe/Canada : ~80–90% des essais
 - Limite la généralisation des résultats
-

Résultats Clés

1. Disparité de Financement

| Cancer | Budget par décès |
|-------------------|------------------|
| Leukemia | \$886 |
| Breast Cancer | \$814 |
| Prostate Cancer | \$651 |
| Pancreatic Cancer | \$526 |
| Lung Cancer | \$239 ! |

2. Research Gap

| Cancer | Décès par publication |
|-------------------|------------------------------|
| Lung Cancer | 185 (sous-représenté) |
| Pancreatic Cancer | 108 |
| Prostate Cancer | 41 |
| Breast Cancer | 27 |
| Leukemia | 31 |

3. Distribution Géographique

- **USA** : Dominant (~60–70%)
- **Europe** : ~15–25%

- **Asie** : ~10-15%
 - **Autres** : <5%
-

Conclusion : Réponse à la Problématique

Question

"L'investissement en recherche est-il corrélé à la gravité réelle des maladies ?"

Réponse : **NON** 

L'analyse démontre l'existence de **biais structurels** créant des **inégalités majeures** :

1. Biais Médiatique

Attention disproportionnée au cancer du sein grâce à la mobilisation associative, créant un cercle vertueux qui ne reflète pas strictement la gravité sanitaire.

2. Biais Économique

Les cancers "rentables" (survivants à long terme) attirent plus d'investissement que les cancers rapidement létaux.

3. Biais Géographique

Concentration de la recherche dans les pays riches, excluant 85% de la population mondiale.

Message Clé

L'allocation des ressources de recherche est davantage influencée par la visibilité médiatique, les intérêts économiques et les inégalités géographiques que par l'impact sanitaire réel.

Cette étude appelle à une **réévaluation des priorités** basée sur des critères objectifs de santé publique.

Pistes d'Amélioration

1. Élargir le Périmètre

- Intégrer d'autres cancers (colorectal, ovaire, mélanome)

- Comparer avec les cancers rares

2. Analyser l'Impact sur les Patients Vivants

- Évaluer la prévalence (pas seulement mortalité)
- Intégrer DALY (Disability-Adjusted Life Years)

3. Intégrer les Coûts Sociétaux

- Coût des traitements
- Perte de productivité
- Charge économique totale

4. Suivi Temporel

- Automatisation du scraping périodique
 - Tracking des tendances émergentes
 - Modèles prédictifs (ML)
-

Documentation

Ce README est accompagné d'une **présentation complète** :

- [Analyse-Comparative-de-la-Recherche-en-Oncologie.pdf](#)

Le document contient :

- Visualisations des résultats
 - Graphiques comparatifs
 - Schémas d'architecture
 - Analyse des biais
-

Références

- **OMS GLOBOCAN 2022** : <https://gco.iarc.fr/>
 - **NCI Budget** : <https://www.cancer.gov/about-nci/budget>
 - **ClinicalTrials.gov** : <https://clinicaltrials.gov/>
 - **PubMed API** : <https://www.ncbi.nlm.nih.gov/home/develop/api/>
 - **Google Trends** : <https://trends.google.com/>
-

Projet réalisé dans le cadre d'une analyse de données en oncologie

Décembre 2025 - Janvier 2026

Pour sauver plus de vies, investissons là où l'impact est le plus grand.