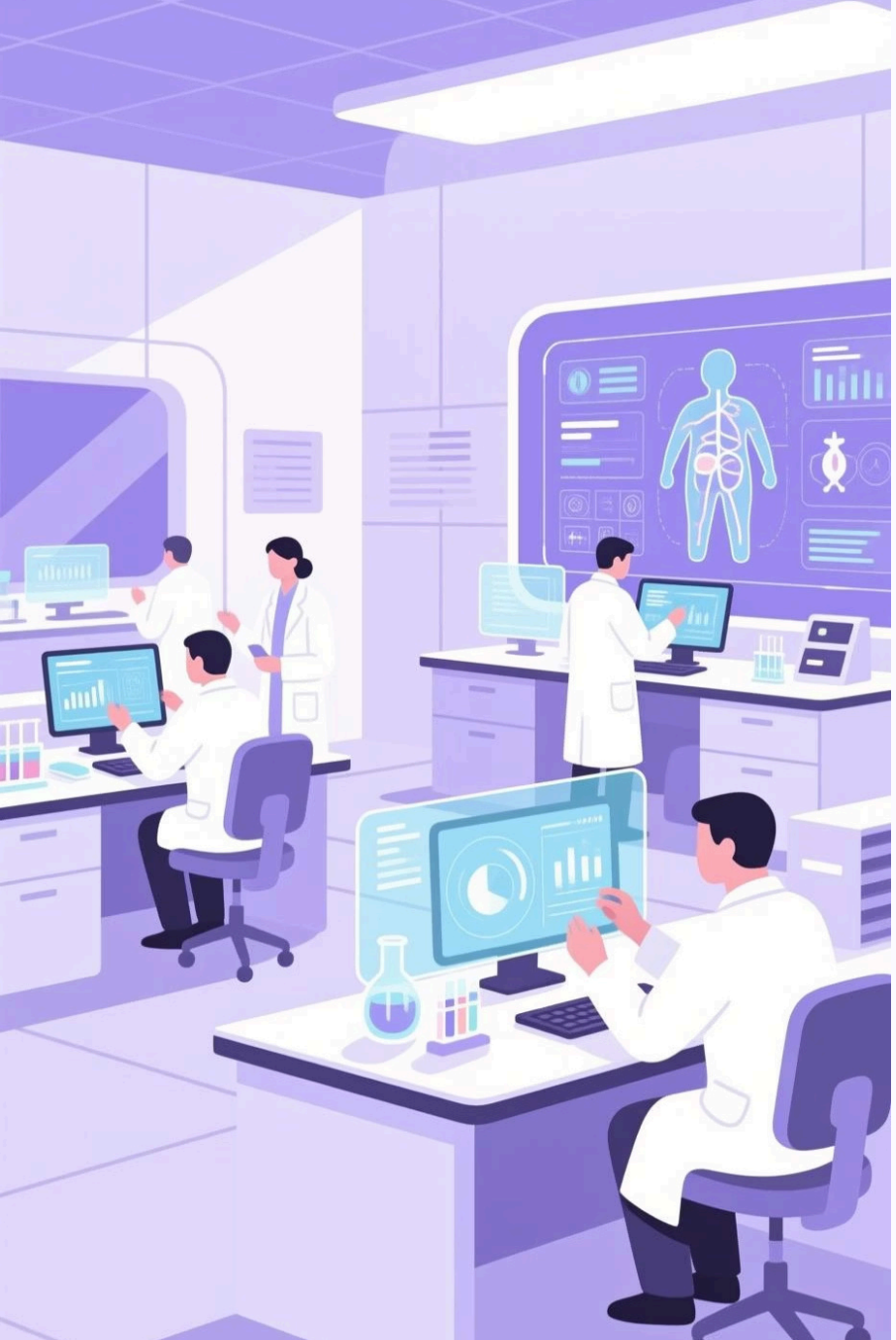


# Analyse Comparative de la Recherche en Oncologie

## Réalité Clinique vs. Mortalité Mondiale

Pipeline ETL complète

Focus sur cinq types de cancers majeurs : Poumon, Sein, Pancréas, Leucémie et Prostate.

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

# Architecture du Projet

## Collecte

Scraping (ClinicalTrials.gov), API (PubMed), Open Data (OMS)

## Transformation

Script Python, Nettoyage, Jointure (Pandas)

## Stockage

Base de données Cloud (Supabase / PostgreSQL)

## Visualisation

Notebook Jupyter (Matplotlib)

# Scraping De ClinicalTrials.gov

## Défi Technique

### Problème

- Sructures dynamiques

### Solution : Tout scrapper !

Selenium interagit comme un utilisateur réel, on aspireTOUT le texte brut pour contourner les erreurs de sélecteurs.

### Performance

Pagination automatisée sur 30 pages par cancer, avec 10 essais par page, permettant la récupération de plus de 1500 essais cliniques.

## Robustesse

### Auto-Healing

Vérifie à chaque action que le navigateur répond

```
    self.driver.current_url
except:
    print("\n🚨 ALERTE : Le navigateur ne répond plus ! Redémarrage d'urgence...")
```

### Sauvegarde Incrémentale

Enregistrement progressif des données pour minimiser la perte et optimiser le temps de traitement.

```
if compteur_total % 10 == 0:
    self.sauvegarder()
```

### Contournement des Blocages

Imitation du temps de lecture d'un humaine

```
time.sleep(2)
```

# Enrichissement des Données : API & Open Data OMS

Scripts `2_ApiSearch.py` & `3_Nettoyage.py`

Pour contextualiser les essais cliniques, nous avons enrichi notre dataset avec des informations provenant de sources académiques et de santé publique.

1

## Intégration API PubMed

Interrogation de l'API NCBI/PubMed pour récupérer le volume de publications scientifiques 2024, reflétant l'intérêt académique pour chaque cancer.

2

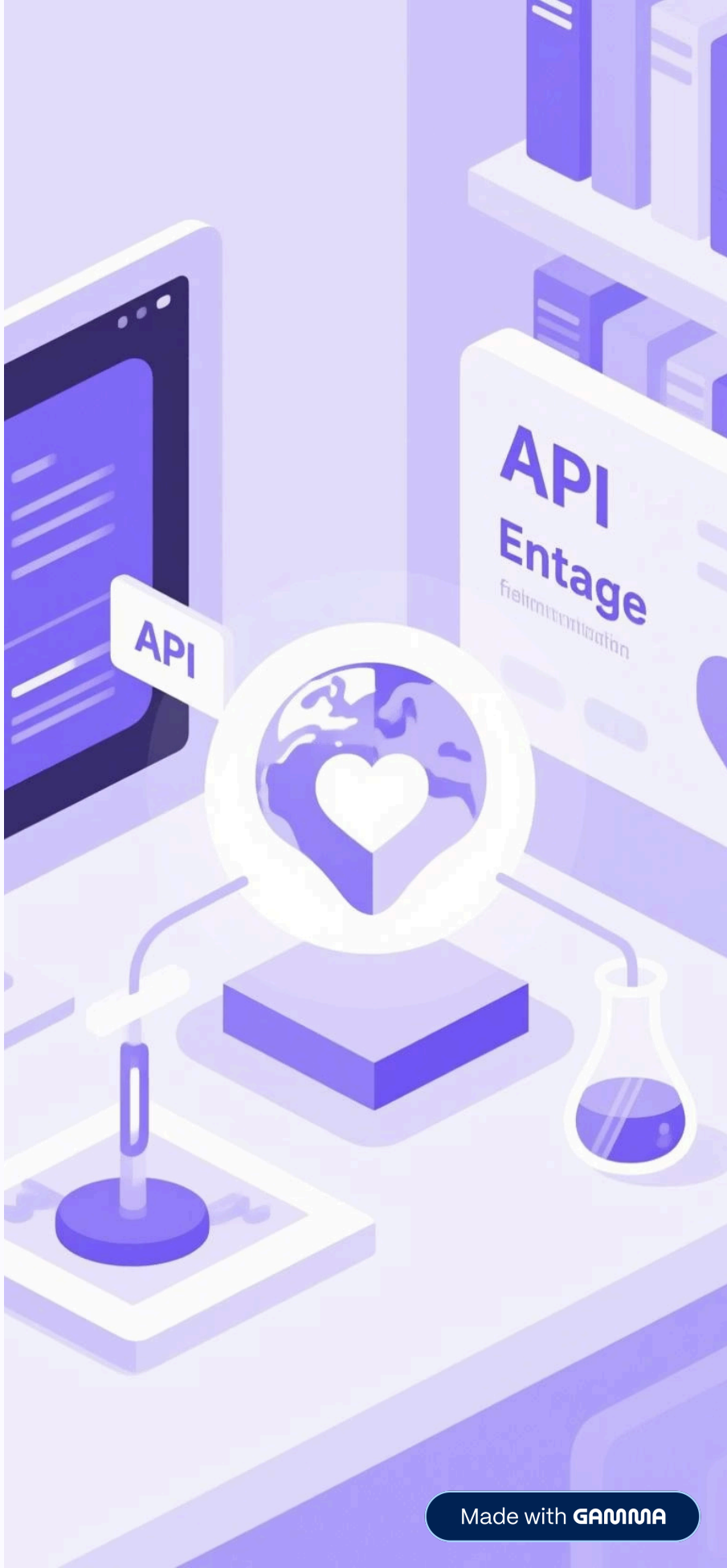
## Données de Mortalité GLOBOCAN 2022 (OMS)

Importation des statistiques de mortalité de l'Organisation Mondiale de la Santé pour évaluer la gravité réelle de chaque cancer.

3

## Table Unifiée

Création d'une table finale reliant l'Offre Clinique (Scraping), l'Intérêt Académique (API) et la Gravité (OMS).



# Pipeline de nettoyage : Sources multiples vers données exploitables

## Sources de données intégrées

### Données OMS

Mortalité par cancer (2022) - nettoyage et tri par décès décroissants pour établir la gravité.

### API PubMed

Publications scientifiques 2024 par type de cancer - indicateur de l'intérêt académique.

### NCI Budget

Financement de la recherche américaine 2023 - extrait et converti en USD pour une base comparable.

### Google Trends

Visibilité médiatique par pays - scores d'intérêt moyen pour la perception publique.

## Transformations effectuées

### Mapping des Nomenclatures

Uniformisation des terminologies (ex: "Trachea bronchus and lung" → "Lung Cancer").

### Fusion des Datasets

Regroupement des données via des colonnes clés communes pour une vue intégrée.

### Calcul de Métriques

Définition de nouvelles métriques, comme les publications pour 1000 décès.

### Nettoyage des Formats

Standardisation des formats monétaires ($, virgules) en valeurs numériques flottantes.

# Garantir la fiabilité : Nettoyage des 5000+ essais cliniques

## 1 Extraction et Nettoyage des Identifiants

- Utilisation de Regex pour isoler les **NCT IDs** des URLs.
- Parsing intelligent des titres pour standardiser les informations d'étude.

## 2 Déduplication Massive

Mise en œuvre d'un processus strict basé sur l'**ID d'Essai Clinique** unique, garantissant un dataset sans doublons et la conservation de la première occurrence.

## 3 Géolocalisation Précise des Essais

- Fonctionnalité d'extraction pays/région via une base de plus de 800 mots-clés.
- Classification en 8 grandes zones géographiques pour analyser la répartition des efforts de recherche par type de cancer.


# Stockage Cloud : Persistance des Données avec Supabase

Script `4_Supabase.py`

Supabase, une plateforme "open source Firebase alternative", a été choisie pour héberger notre base de données PostgreSQL dans le cloud, offrant scalabilité et accessibilité.

### PostgreSQL Cloud

Utilisation d'une instance PostgreSQL gérée par Supabase pour une solution de base de données robuste et évolutive.

### Chargement par Batchs

Chargement des données par paquets de 1000 lignes pour optimiser les performances et minimiser la charge sur la base.

### Gestion d'Erreurs & Logs

Implémentation d'une gestion d'erreurs détaillée et de logs pour assurer l'intégrité du processus de chargement et le débogage facile.

# Indicateurs Clés (KPI)

Objectif: Quantifier et comparer l'effort de recherche, le financement et la visibilité des cancers au regard de leur impact sanitaire.

## Essais Cliniques par Type de Cancer

Nombre d'essais recensés par cancer, identifiant les domaines de recherche prioritaires.

| cancer | trials_count |
|---|---|
| Pancreatic Cancer | 300 |
| Lung Cancer | 300 |
| Prostate Cancer | 300 |
| Leukemia | 300 |
| Breast Cancer | 299 |

## Budget de Recherche (NCI)

Analyse des budgets moyens, min. et max. par cancer, révélant la répartition des financements.

| avg_budget_mı | min_budget_mı | max_budget_mı |
|---|---|---|
| 350.44 | 246 | 542 |

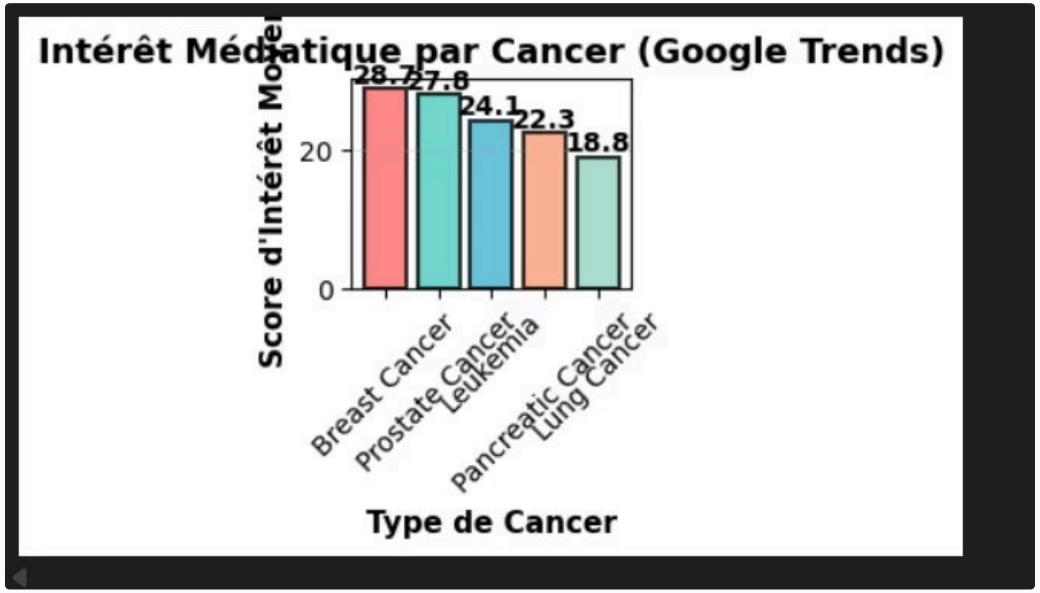
## Indicateur de Priorité

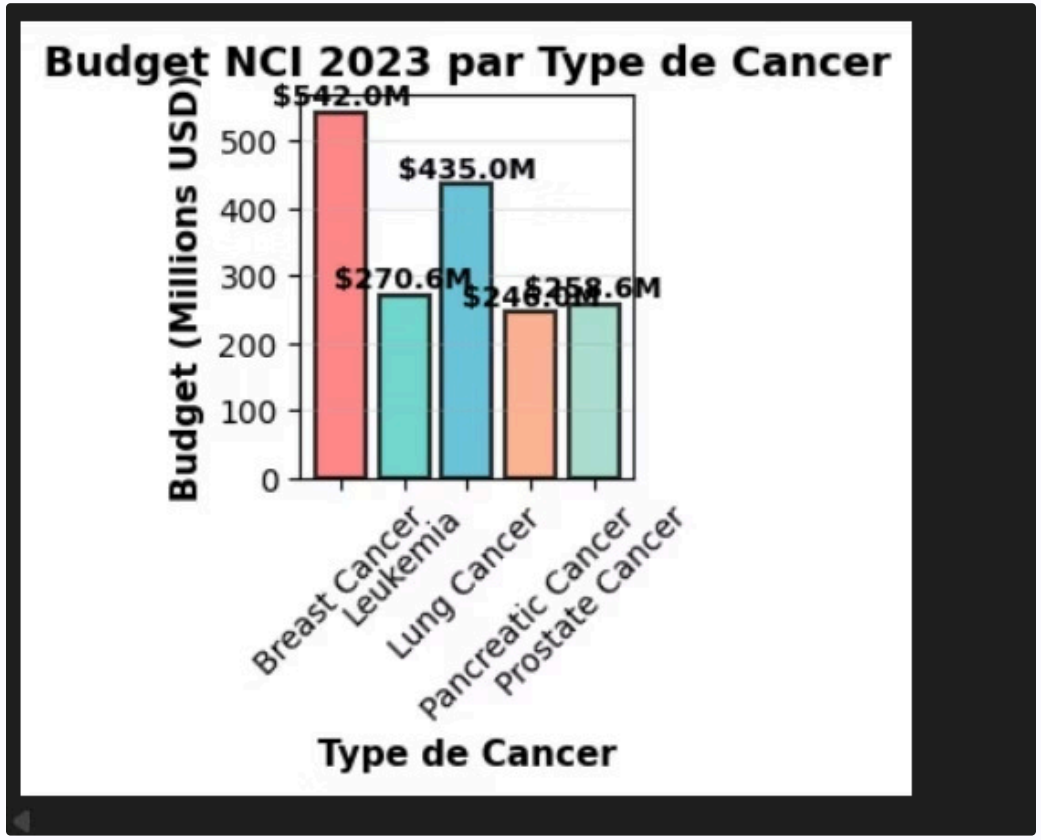
Ratio **mortalité / publications scientifiques** : une valeur élevée suggère un cancer sous-recherché.

| cancer | mortality_20: | publications_ | deaths_per_publicatio |
|---|---|---|---|
| Lung Cancer | 1817469 | 17299 | 105.0620845135556969 |
| Pancreatic Cancer | 467409 | 4665 | 100.1948553054662379 |
| Prostate Cancer | 397430 | 9628 | 41.2785625259659327 |
| Leukemia | 305405 | 9988 | 30.5771926311573889 |
| Breast Cancer | 666103 | 24582 | 27.0971849320641120 |

# Intérêt Médiatique et Budget de Recherche

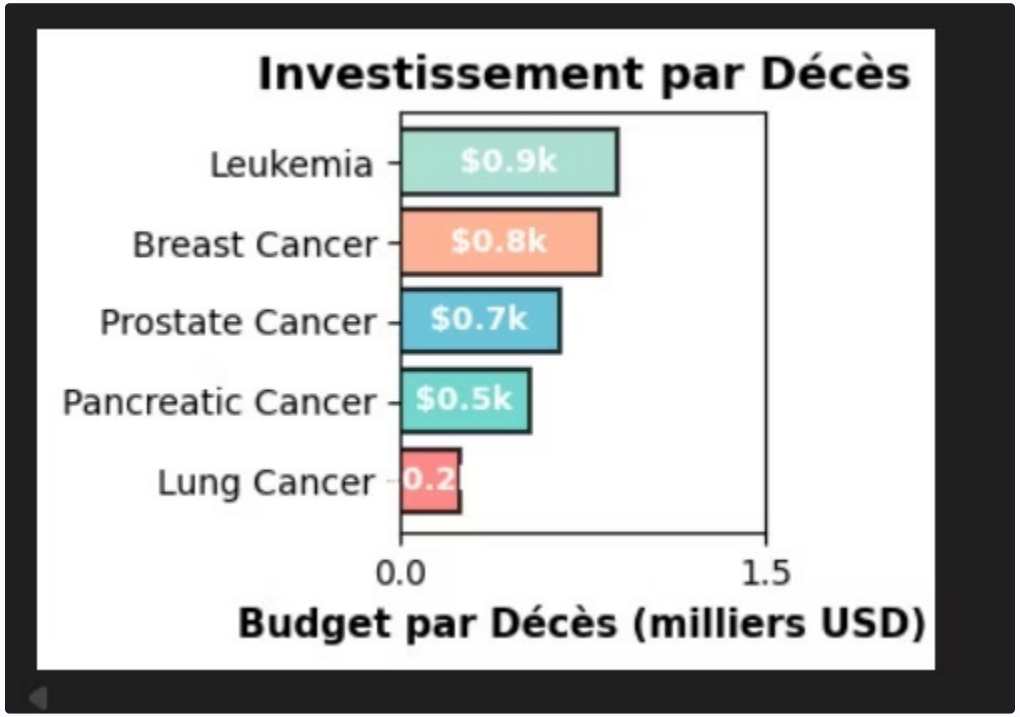
## Analyse comparative des investissements

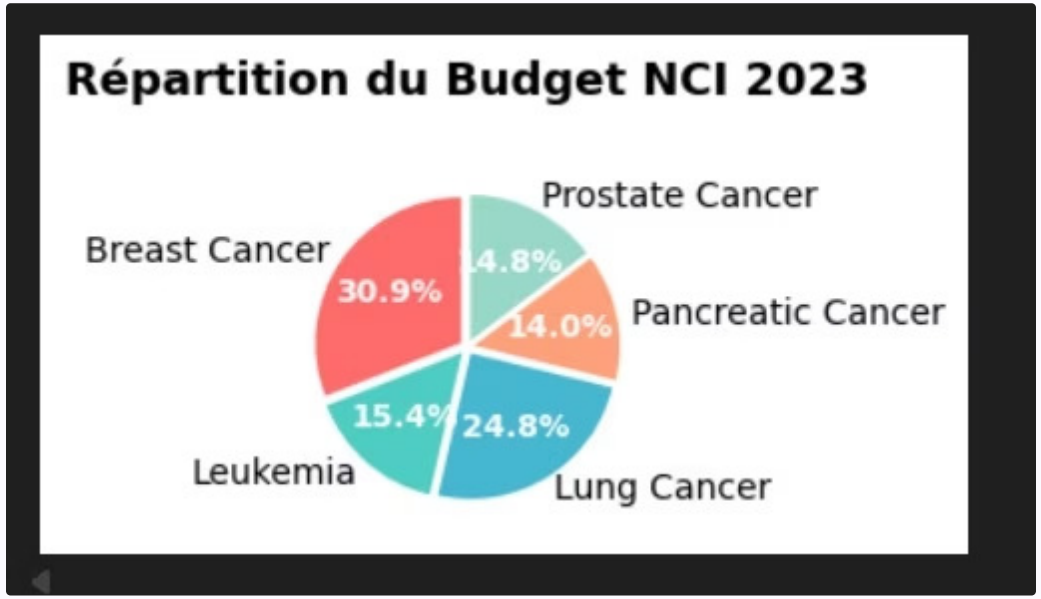


# Répartition des Ressources et Efficacité

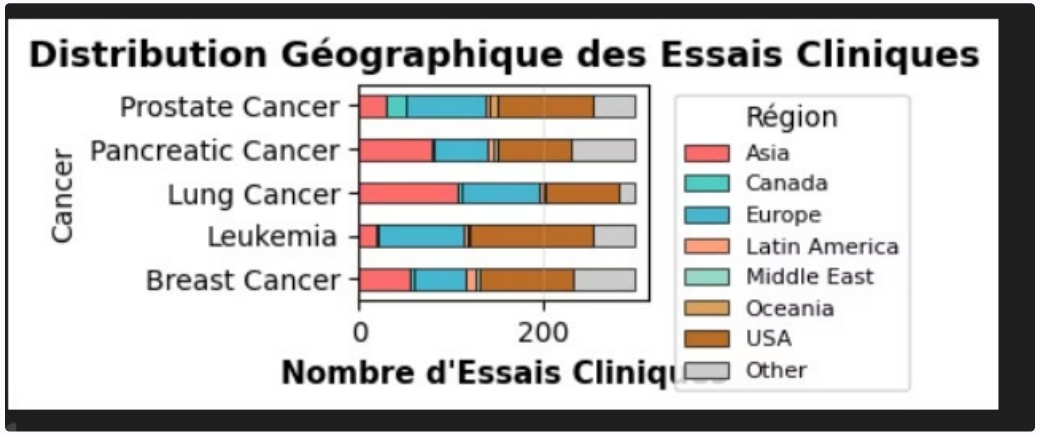
## Budget par décès et allocation NCI

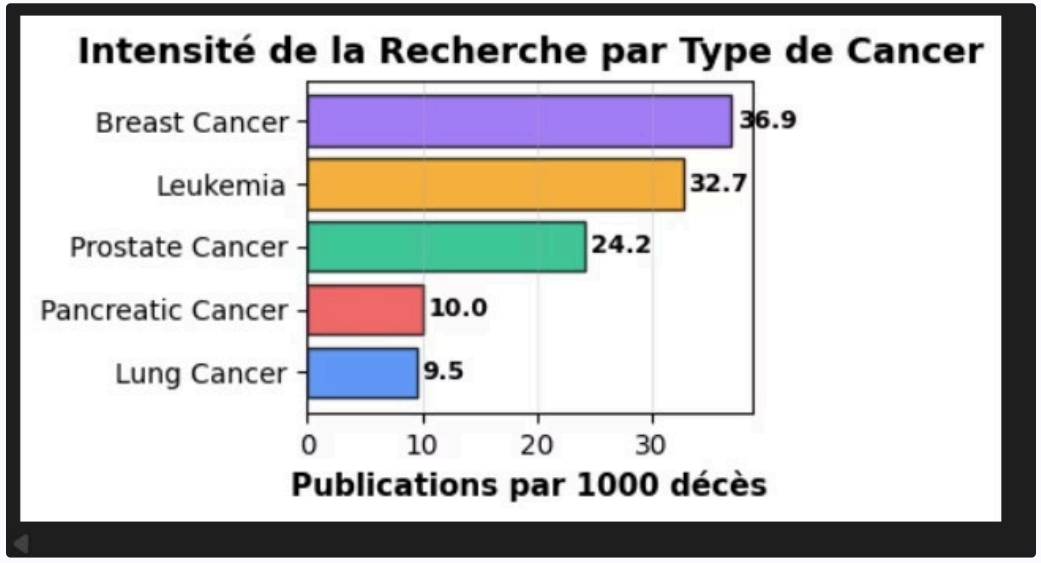


# Distribution Géographique et Intensité de Recherche

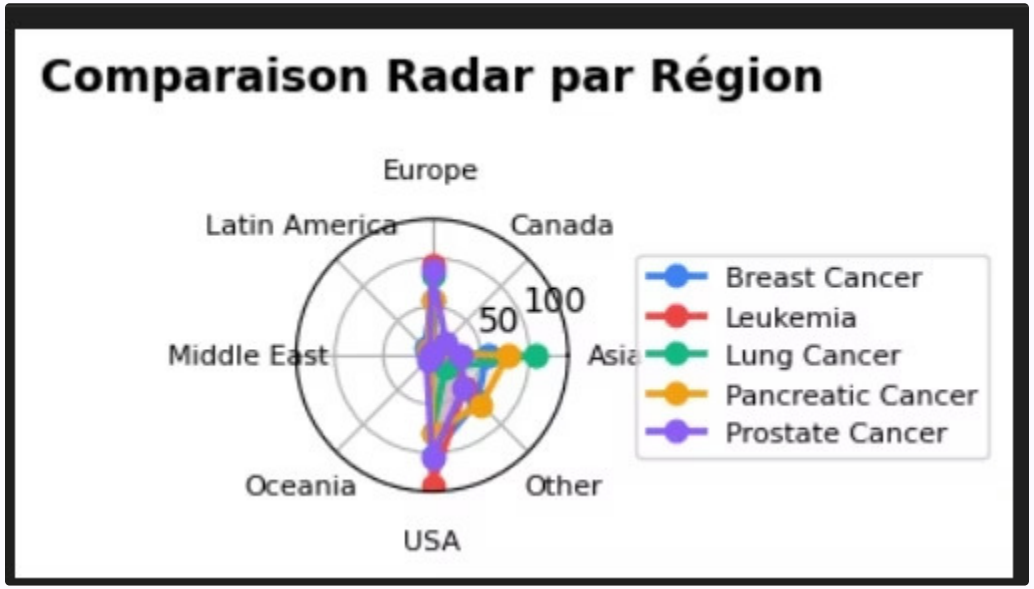
## Analyse régionale et effort académique



# Analyse Comparative Régionale et Corrélation Clé

## Disparités géographiques et relation mortalité-recherche



## 1. Biais Médiatique et "Effet de Visibilité"

Le cancer du sein bénéficie d'une **triple surreprésentation** :

- Intérêt médiatique élevé (24,1)
- Budget le plus important ($542M)
- Intensité de recherche la plus forte (36,9 publications/1000 décès)

## 2. La Grande Injustice : Le Cancer du Poumon

Malgré sa **mortalité la plus élevée**, le cancer du poumon souffre d'un **triple handicap** :

- Intérêt médiatique le plus faible (18,8)
- Investissement par décès dérisoire ($0,2k)
- Intensité de recherche la plus basse (9,5 publications/1000 décès)

## 3. Cancer Pancréatique : L'Oublié Mortel

Avec un **taux de survie à 5 ans inférieur à 10%**, c'est l'un des cancers les plus létaux. Pourtant :

- Investissement par décès modeste ($0,5k)
- Intensité de recherche faible (10,0 publications/1000 décès)

## 4. Inégalités Géographiques

La recherche clinique est massivement concentrée dans les pays à revenu élevé (USA, Europe, Canada), créant un **double problème** :

- Les populations des pays en développement n'accèdent pas aux innovations
- Les diversités génétiques et environnementales sont sous-étudiées, limitant la généralisation des résultats

# Conclusion : Réponse à la Problématique

L'analyse démontre l'existence de **multiples biais structurels** créant des inégalités majeures dans l'innovation médicale :

1

### 1. Biais Médiatique

Le cancer du sein reçoit une attention disproportionnée grâce à des décennies de mobilisation associative (Ruban Rose, Octobre Rose), créant un cercle vertueux de financement et de recherche qui ne reflète pas strictement sa gravité sanitaire relative.

2

### 2. Biais Économique

Les cancers "rentables" (avec des populations importantes de survivants à long terme nécessitant des traitements continus) attirent davantage l'investissement pharmaceutique que les cancers rapidement létaux comme le cancer pancréatique.

3

### 3. Biais Géographique

La concentration de la recherche dans les pays riches crée une "science à deux vitesses" excluant 85% de la population mondiale, perpétuant les inégalités sanitaires globales.

# Pistes d'Amélioration et Perspectives Futures

1

### 1. Élargir le Périmètre d'Analyse

Intégrer d'autres cancers (colorectal, ovaire, mélanome, etc.) pour une vision plus complète et identifier d'autres déséquilibres.

2

### 2. Analyser l'Impact sur les Patients Vivants

Évaluer le nombre de personnes vivantes touchées par chaque cancer, pas seulement la mortalité. Cela révélerait l'impact réel sur la qualité de vie et la charge de morbidité.

3

### 3. Intégrer les Coûts Sociétaux

Analyser le coût économique indirect (perte de productivité, soins palliatifs) pour une allocation plus équitable des ressources.

4

### 4. Suivi Temporel et Prédictif

Mettre en place une surveillance continue du pipeline pour détecter les tendances émergentes et anticiper les besoins futurs en recherche.