UNIVERSITÉ DE GENÈVE

CENTRE UNIVERSITAIRE D'INFORMATIQUE

Comparaisons de techniques de générations de résumés par clustering

Auteur
Nikita MISSIRI

Superviseurs
Prof. DI MARZO
SERUGENDO Giovanna
Mme FRIHA Lamia
M. HUGENTOBLER Alain

September 8, 2023

Contents

1 Introduction

1.1 Contexte du travail

Ce travail s'inscrit dans un projet avec l'équipe LiteRev de EpiGraphHub. LiteRev est un service numérique, permettant à un chercheur scientifique de gagner en temps pour sa revue littéraire. Plus précisément, le chercheur entre des mots-clés et le service fournit une liste d'articles scientifiques qui correspondent à ces mots-clés. Le chercheur a la possibilité ensuite d'affiner sa recherche, en sélectionnant un des articles fournis. Le service va alors fournir une plus petite liste d'articles, qui sont similaires au papier scientifique sélectionné. Ce service permet ainsi au chercheur d'avoir que les articles qui sont pertinents pour sa recherche. [1] Néanmoins, la liste d'articles retournés peut être très grande, et cela peut donc prendre beaucoup de temps pour parcourir tous ces résumés d'articles. C'est pourquoi il y est nécessaire de développer un service de génération automatique de résumé pour la liste d'articles sélectionnés par le chercheur. C'est dans cette optique, qu'intervient mon travail.

1.2 Objectif du travail

L'objectif de ce travail de recherche est donc de développer une nouvelle fonctionnalité pour le service LiteRev permettant de générer un résumé pour un ensemble d'articles scientifiques. Cette fonctionnalité sera disponible pour les utilisateurs Premium du service LiteRev. Le cas d'utilisation illustré ci-dessous en jaune représente mon travail.

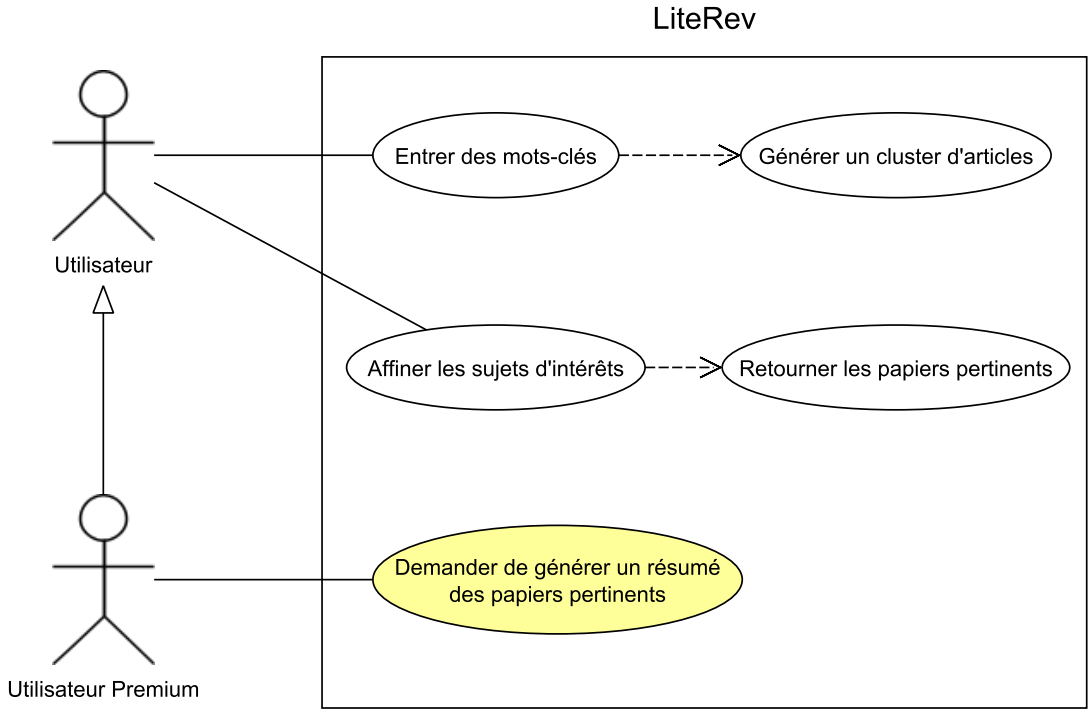

Figure 1: Cas d'utilisation du service LiteRev

De plus, étant donné le grand nombre de méthodes existantes pour pouvoir générer automatiquement un résumé à partir d'un ensemble d'abstracts, l'objectif de ce travail est également de comparer les différentes méthodes existantes de générations de résumés.
Dans le section 2, nous allons explorer les différentes outils existantes nous permettant de générer un nouveau résumé. Dans la section 3, nous expliquerons comment ces différentes technologies sont utilisés pour atteindre notre objectif. La section 4 va présenter les résultats obtenus à partir d'une combinaison de ces technologies, et enfin la section 5 discutera de ces résultats, ainsi que des limitations des méthodes appliqués et des pistes pour de potentiels travaux futurs.

2 Revue de littérature

Il existe beaucoup de recherches sur la génération automatique de résumés. Les deux principales modèles de résumés sont les résumés extractifs et les résumés abstractifs. [2]
Le premier modèle consiste à extraire les phrases les plus importantes d'un ensemble de phrases. Beaucoup de recherches ont été menées dans ce domaine. Plusieurs méthodes ont été développés pour obtenir ces phrases, notamment l'analyse statistique, les algorithmes de graphes, les algorithmes de clusterisation, les algorithmes utilisant de la logique floue ou encore du machine learning supervisé et non-supervisé. [3] Certains papiers utilisent une combinaison de ces technologies pour atteindre leurs objectifs. C'est le cas, par exemple de Ferreira et al. [4], qui ont construits un algorithme de clusterisation des phrases contenus dans un ensemble de documents disponible sur Internet, en se basant sur un modèle de graphes, qui prend en compte les similiarités statistiques entre phrases, ainsi que le traitement linguistique. Cela afin de relever les aspects importants mentionnés dans les documents et d'éviter de la redondance et de la diversité.
Avec le modèle abstractif, on construit de nouvelles phrases à partir des phrases existantes. C'est un domaine dans lequel il n'y a pas encore eu beaucoup de recherches, car il nécessite le développement des technologies du deep learning et de réseaux neuronales. Aliakbarpour et al. [5] proposent un modèle de résumé abstractif, qui utilise une combinaison de réseaux neuronales convolutionnelles, ainsi que de longues mémoires à court-terme avec une attention auxiliaire dans l'encoder pour augmenter la saillance et la cohérence des résumés générés.
Il existe également des modèles de résumés hybrides, qui incorporent une méthode extractive, puis une méthode abstractive, afin de générer des résumés de meilleures qualités, qui se rapprochent plus vers ce que produirait un expert humain. En effet, les humains identifient d'abord les phrases importantes contenus dans un ensemble de documents et introduisent ensuite de nouvelles phrases grammaticales, permettant de mieux résumer les concepts contenus dans les articles. Singh et al. [6] proposent une méthodologie qui intègre les méthodes abstractives et extractives par une approche par pipeline pour produire un résumé concis et par la suite générer un titre.
Pour notre problématique, le type de récapitulation, qui serait la plus approprié est un résumé générique, car on n'interagit pas avec l'utilisateur pour connaître les spécificités. Le résumé est également informatif, afin que l'utilisateur n'a pas à devoir chercher de l'information supplémentaire et que le résumé gère plusieurs

documents, étant donné que nous travaillons avec un ensemble d'abstracts. On décide également de produire un résumé extractif, puisque c'est une méthode simple pour générer de nouveaux résumés.
Pour identifier les phrases les plus importantes dans un ensemble de documents, beaucoup de recherches font recours à un système de score. Ceci est une technique, qui a prouvé de son efficacité.
Parmi les méthodes extractives existantes, la méthode extractive basé sur les clusters a été utilisé, puisque l'équipe LiteRev dispose déjà de technologies de clustering avec lesquelles nous pouvons faire des comparaisons. Dans les sous-sections suivantes, on introduira les outils qui ont été considérés pour ce travail.

2.1 Term Frequency - Inverse Document Frequency (TF-IDF)

Cet algorithme est très répandu dans le domaine du traitement de langages naturels, *Natural Language Processing (NLP)*, avec du machine learning, et permet dans notre cas de déterminer les termes qui sont les plus importants parmi un ensemble de documents.
TF-IDF est une matrice de dimensions $mots \times documents$, où pour un terme t_i dans un document d_i, d'un ensemble de documents D, les valeurs TF-IDF se calculent de la manière suivante:

$$TF(t_i, d_i) = \frac{f_{d_i}(t)}{\max_{w \in d_i} f_{d_i}(w)}$$

$$IDF(t_i, D) = \log(\frac{|D|}{|\{d \in D : t_i \in d}})$$

$$TF - IDF(t_i, d_i, D) = TF(t_i, d_i) \cdot IDF(t_i, D)$$

où $|D|$ est le nombre total de documents dans un ensemble, w est le nombre de mots totales.
La fréquence du terme t_i dans un document d_i se calcule en comptant le nombre d'occurrence du terme dans le document d_i divisé par le nombre total de termes présents dans ce même document. La fréquence inverse du document se calcule en comptant le nombre de documents dans lequel se trouve le terme t_i au moins une fois et divisant par le nombre total de documents compris dans notre ensemble $|D|$. Le résultat de cette fraction est ensuite inversé, puis son logarithme est calculé. Finalement, le résultat des deux fréquences sont multipliés. [7]
Une fois que nous avons calculé la matrice TF-IDF, il est nécessaire de réduire les dimensions de la matrice pour effectuer d'autres traitements, tels que la création de clusters. Pour effectuer la réduction des dimensions de la matrice, deux méthodes différentes ont été appliqués: Singular-Value Decomposition (SVD) et Uniform Manifold Approximation and Projection (UMAP). La première est une méthode linéaire, la deuxième est une méthode non-linéaire.

2.2 Singular-Value Decomposition (SVD)

Cette méthode de calcul algébrique linéaire est très répandu dans le domaine du NLP. C'est une méthode d'analyse sémantique latente (*Latent Semantic Analysis*), où une analyse des relations entre un ensemble de documents et ses termes est

faite. La réduction d'une matrice de dimensions $m \times n$ à une matrice de dimensions $k \times n$ se fait de la manière suivante:

$$A_{m \times n} = U_{m \times m} \times \Sigma_{m \times n} \times V^T_{n \times n}$$

où A, U, Σ, et V^T sont des matrices.

A est ici la matrice de départ, qui est connu. Les autres matrices sont quant à elles inconnus. Pour obtenir les valeurs de la matrice U, il faut considérer la propriété suivante: les colonnes de U sont les vecteurs propres orthonormales du produit $A \times A^T$. Il faut donc calculer le produit de A avec sa transposée, avant de calculer ses vecteurs propres. Le calcul de V est similaire: les colonnes de V sont les vecteurs propres orthonormales du produit $A^T \times A$. Cette fois-ci, on multiplie la matrice transposée par l'originale. Et finalement, Σ est la matrice diagonale contenant les racines carrées des valeurs propres de U ou V dans un ordre décroissant. Une fois les vecteurs propres calculés, nous pouvons obtenir les valeurs propres de ces vecteurs propres et les introduire sur la diagonale de la matrice Σ.

Une fois les quatre matrices connus, la matrice réduite $A'_{k \times n}$ se calcule par le produit $U_{k \times k} \times \Sigma_{k \times n}$, où on prend les k premières valeurs propres de Σ [8].

2.3 *Uniform Manifold Approximation and Projection (UMAP)*

UMAP est une technique de réduction de dimensions matricielles, principalement utilisé pour la visualisation des données. Il est similaire à l'algorithme t-SNE. De manière générale, l'algorithme effectue une réduction non-linéaire des dimensions. L'algorithme se base sur trois hypothèses à propos des données. Les données sont uniformément distribuées sur un collecteur Riemannien. Le métrique Riemannien est localement constant, ou est approximé. Le collecteur est localement lié. À partir de ces trois hypothèses, il est possible de modéliser le collecteur avec une structure topologique flou. L'intégration est trouvé en cherchant une projection des données dans une dimension inférieure, qui est le plus proche de manière équivalente à une structure topologique flou.[9]

Il existe plusieurs techniques de clustering, qui prennent en entrée un ensemble de points avec des coordonnées à 2 dimensions. Nous allons nous focalisé sur deux techniques bien répandus dans le domaine des NLP: k-Means et HDBSCAN.

2.4 *k-Means*

L'algorithme de clusterisation de k-Means est un des algorithmes de machine learning non-supervisé les plus simples. L'algorithme permet de partitionner un ensemble de points en k clusters. Le nombre k représente le nombre de centroids, points fictifs représentant le centre de clusters. Ces centroids permettent d'attribuer un point à un cluster.

Pour cela, nous choisissons des coordonnées aléatoires pour chacune des k centroids et allouons les points au cluster le plus proche en calculant la distance euclidienne. On optimise la position des centroids et recommençons les opérations, jusqu'à ce que les clusters formés ne changent plus. [10]

Le problème est qu'il faut spécifier le nombre k. Pour déterminer le nombre k optimal, il existe deux méthodes: *Elbow Method* et *Silhouette Method*. La première méthode parcourt les valeurs k définit et calcule la somme des carrés des erreurs dans un cluster, *Within-Cluster-Sum of Squared Errors (WSS)*. Les valeurs obtenus sont ensuite tracés, et la valeur pour laquelle la pente change est définit comme la valeur optimale de k. Cette méthode fonctionne que pour une petite quantité de points, ce qui n'est pas le cas de notre exemple. La deuxième méthode calcule la valeur silhouette pour chaque valeur de k, qui se calcule de la manière suivante:

$$s_i = \begin{cases} \frac{b_i - a_i}{\max\{a_i, b_i\}} \; si \; |C_i| > 1 \\ 0 \; si \; |C_i| = 1 \end{cases} \tag{1}$$

où s_i est la valeur silhouette d'un point i, a_i est la mesure de similarité entre le point i et le cluster C_i ou en d'autres termes la distance moyenne entre i et les autres points du même cluster, b_i est la mesure de dissimilarité entre le point i et le cluster C_i.
À noter que le nombre s_i est défini à 0, si le nombre de clusters est égal à 1, puisque nous ne pouvons faire de comparaisons pour un seul cluster. La valeur maximale obtenu pour s_i correspond à la valeur optimale de k. [11]

2.5 HDBSCAN

HDBSCAN est un algorithme hiérarchique de clustering, qui est une extension de DBSCAN. Cet algorithme utilise une technique pour extraire les clusters plats, mesurés en fonction de la stabilité des clusters [12]. Pour cela, l'algorithme s'aide d'un arbre hiérarchique.
Cette méthode considère des hyper-paramètres, tels que la taille minimale des clusters, l'échantillon minimale, qui donne une mesure de la conservativité des clusters, la méthode de sélection, où nous spécifions si nous souhaitons récupérer les feuilles de l'arbre hiérarchique ou si nous récupérons également les documents contenus dans les branches par *EOM (Excess of Mass)*. La méthode de sélection par les feuilles est plus approprié, s'il y a une trop grande différence de tailles entre les clusters.

3 Méthodologie

3.1 Pré-traitement de texte

Avant de faire quelconque traitement sur le texte, il faut regrouper les termes identiques ensemble. Pour cela, il faut parcourir les abstracts et ensuite transformer les lettres en minuscules, retirer les caractères spéciaux et la ponctuation, fractionner les phrases en tokens pour n'avoir que des mots, et ensuite lémmatiser ces mots, afin de regrouper toutes les formes d'un mot en un seul terme.
De plus, certaines phrases ont été exclus en identifiant par des expressions régulières, si la phrase contient un nombre suivi d'un élément de recherche, tels que les participants, les patients, les enfants. En effet, sachant que les articles scientifiques avec lesquelles on travaille sont à des fins médicales, il a été jugé que les détails concernant les participants ne devraient pas ressortir dans le résumé final.

Exemple de Pré-traitement:

Avant: Integrating Enhanced HIV Pre-exposure Prophylaxis Into a Sexually Transmitted Infection Clinic in Lilongwe: Protocol for a Prospective Cohort Study.

Après: integration enhancement hiv prophylaxis sexually transmission infection clinic lilongwe protocol prospective cohort study

Le pré-traitement est fait séparément de la génération de résumé, afin d'améliorer l'efficacité en temps. Il faudra donc stocker les résultats du pré-traitement dans les méta-données des articles: plus précisément le pré-traitement des titres, pré-traitement des phrases dans les abstractes des articles et le pré-traitement des abstracts eux-mêmes.
Une explication plus approfondi de la méthode avec un pseudo-code, ainsi qu'un exemple comme démonstration de lemmatisation est présente en annexe à la section 7.2.

3.2 Déroulement de la méthode utilisée

Afin de réaliser la génération automatique d'un nouveau résumé à partir d'un ensemble de documents on suit la procédure suivante:

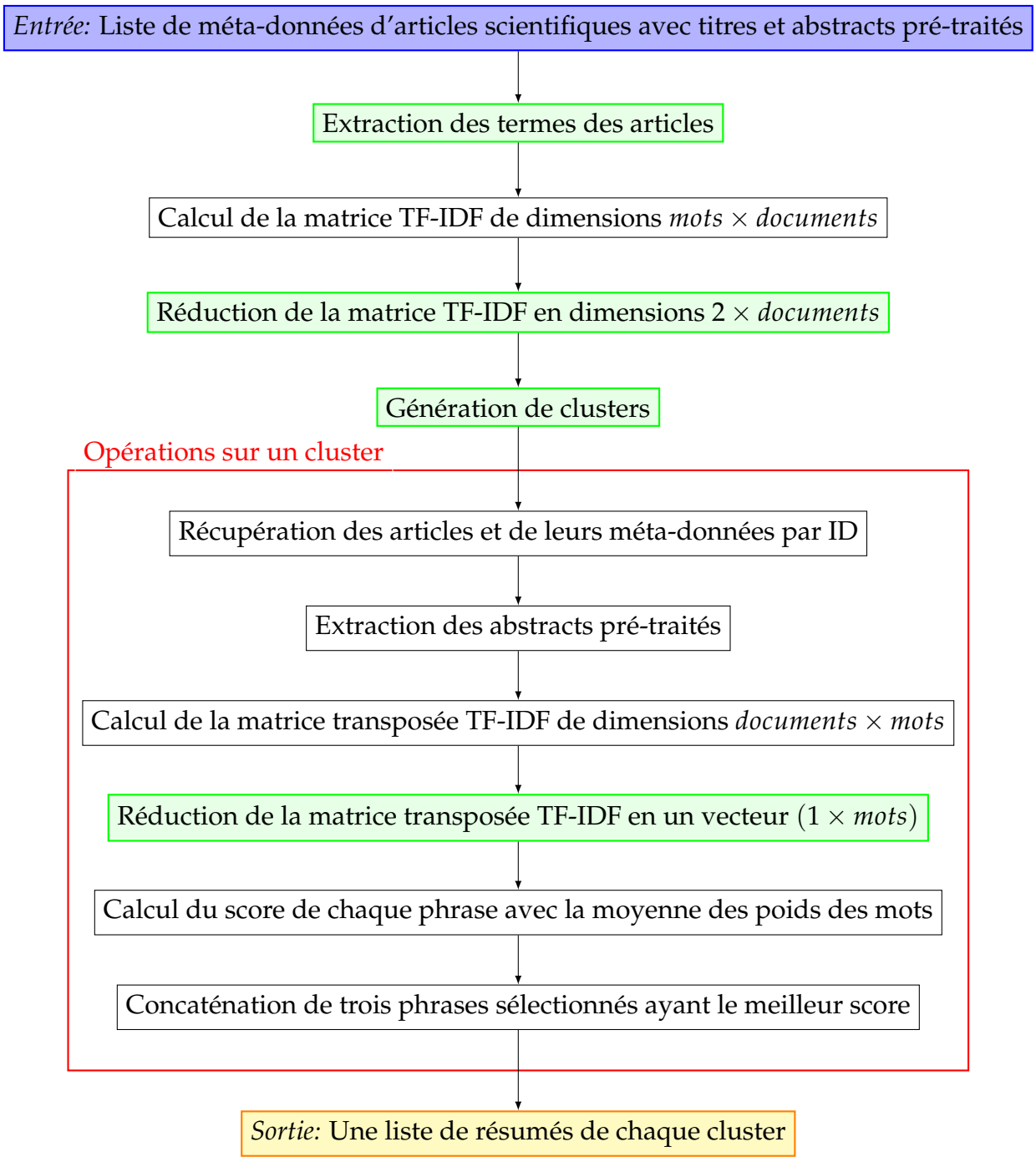

Figure 2: Processus de génération de résumé

L'algorithme de génération de résumés doit prendre une liste d'articles scientifiques en entrée, avec leurs méta-données, y compris les titres et les abstracts, ainsi que les titres et les abstracts, qui ont été préalablement pré-traités. La première étape est d'extraire les termes d'articles, avec lesquelles nous voulons regrouper. Nous avons la possibilité d'effectuer des clusters par titres d'articles ou des clusters par les abstracts. On procède ensuite au calcul de la matrice TF-IDF. Nous avons également la possibilité ensuite de réduire cette matrice, soit de manière linéaire avec SVD, soit de manière non-linéaire avec UMAP. Cette matrice permet ensuite de générer de nouveaux clusters. La génération peut être faite soit avec k-Means, soit avec HDBSCAN. Une fois que les clusters ont été construits, on parcourt chaque cluster pour récupérer les abstracts pré-traités

contenus dans le cluster et en calculer la matrice TF-IDF. On est ensuite intéressé aux mots et non aux documents. Il faut alors prendre la transposée de la matrice. Pour générer une pondération des mots-clés, il faut réduire cette matrice en un vecteur. On peut soit le faire en prenant la moyenne en fonction des documents présent dans le cluster, soit en réduisant linéairement avec SVD. Une fois qu'on a la pondération de ces mots, on procède au calcul du score de chaque phrase pré-traité de la manière suivante:

$$s_i = \frac{1}{n} \sum_{j}^{n} w_j$$

où s_i est le poids d'une phrase contenu dans un cluster, w_i est le poids d'un mot, calculé dans le vecteur de mots et n est le nombre total de mots contenus dans la phrase s_i.
On additionne donc le poids des termes dans une phrase et on divise par le nombre de mots présents dans la phrase.

Nous avons donc une combinaison de techniques permettant de générer des résultats différents. Les combinaisons que nous allons explorer sont les suivantes:

1. Clusters par titres, SVD et k-Means, pondération par la moyenne
2. Clusters par titres, SVD et k-Means, pondération par SVD
3. Clusters par abstracts, SVD et k-Means, pondération par la moyenne
4. Clusters par abstracts, SVD et k-Means, pondération par SVD
5. Clusters par titres, UMAP et HDBSCAN, pondération par la moyenne
6. Clusters par titres, UMAP et HDBSCAN, pondération par SVD
7. Clusters par abstracts, UMAP et HDBSCAN, pondération par la moyenne
8. Clusters par abstracts, UMAP et HDBSCAN, pondération par SVD

3.3 Évaluation des clusters

Il est difficile de mener un jugement sur la qualité des résumés. Il faudrait plutôt comparer l'équilibre des clusters. Pour ce faire, il existe deux méthodes, qui sont appliqués dans cette recherche:

3.3.1. *Écart-type*
Pour comparer les combinaisons d'outils les plus appropriés, cette méthode propose de calculer la distance entre clusters formées à partir d'une même combinaison, en calculant l'écart-type des poids maximales de chaque cluster.

$$\sigma^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

où X est une variable aléatoire qui représente le poids maximal d'un des clusters, $\bar{X}$ représente la moyenne de tous les poids maximales de tous les clusters et n est le nombre de clusters.

Avec cette méthode, nous pouvons identifier si les clusters ont bien été formées. En effet, si la valeur de l'écart-type est grande, et donc si les poids maximales diffèrent et que les mêmes mots-clés apparaissent parmi les meilleurs termes pour chaque cluster, nous pouvons déduire que les clusters ne sont pas équilibrés. Cette méthode ne compare que les poids maximales, mais ne compare pas les autres valeurs du vecteur de pondération des mots. C'est pourquoi, une seconde méthode d'évaluation a été utilisée.

3.3.2. *Similarité Cosinus*

Pour cette méthode, nous ne prenons pas une valeur de chaque cluster, mais bien le vecteur entier. Pour faire des comparaisons appropriés, on compare les vecteurs de chaque cluster entre elles, en les plaçant dans un espace vectorielle de dimension n, où n est le nombre de termes qui existe dans toute l'ensemble de documents D, et en calculant la similarité cosinus, qui consiste à calculer l'angle entre les deux vecteurs. La formule pour le faire est la suivante:

$$cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||}$$

où A et B sont deux vecteurs.

Ici, nous calculons le produit scalaire entre les deux vecteurs au numérateur, et au dénominateur on calcule la magnitude entre les deux vecteurs. Plus deux vecteurs sont distincts, plus la valeur de similarité cosinus s'approche de 0. Et à l'inverse, plus deux vecteurs sont proches, plus la valeur de similarité cosinus se rapproche de 1. En effet, deux vecteurs sont similaires, si pour les mêmes termes, les fréquences sont identiques. [13]

Pour l'évaluation, nous souhaitons obtenir les clusters qui n'ont que peu de similarités entre elles. Pour faire la comparaison entre les combinaisons d'outils, nous calculons $k = \frac{n \cdot (n-1)}{2}$ similarités cosinus entre les n clusters qui ont été formés, et ces valeurs sont ensuite agrégés en un seul, en calculant la moyenne des k valeurs. Plus les combinaisons de technologies ont une valeur proche de 1, plus les clusters seront équilibrés.

4 Résultats

4.1 Données de test

L'algorithme prend en entrée un taux de compression, qui détermine le nombre de clusters qui seront formés, et donc le nombre de résumés qui seront générés. L'algorithme HDBSCAN ne propose pas de paramètres pour le nombre de clusters, qui peuvent être générés, il y a eu donc un tâtonnement des hyperparamètres, afin d'arriver à un nombre de cluster quasiment identique qu'avec la méthode k-Means.

Pour cet exercice, le taux de compression a été fixé à 8. Les articles scientifiques portent sur le VIH en Afrique sub-saharienne.

Exemple d'un article:

> **We describe tuberculosis (TB) in children living with HIV (CLHIV) eligible for HIV treatment in South Africa to highlight opportunities to**

prevent TB.

We analyzed additional data from our original study of CLHIV who were 0˜12 years old and due to start HIV treatment in five health facilities in Eastern Cape Province from 2012 to 2015 and assessed characteristics associated with existing and new TB.
Of 397 enrolled children, 114 (28.7%) had existing TB. Children with a higher measure of household income had higher odds of existing TB. CD4+ cell count $<$ 350 cells/μl and malnutrition were also associated with existing TB. There were 5.2 new cases of TB for every 100 child-years. New TB was 4.7 times more likely for children with delayed HIV treatment start, 1.8 times more likely for children with malnutrition and 2.3 times more likely for children who did not get cotrimoxazole. Among 362 children with data, 8.6% received treatment to prevent TB.
Among these CLHIV, existing and new TB were common. Early HIV treatment, cotrimoxazole and addressing malnutrition may prevent TB in these children.

Il est d'ailleurs intéressant de relever que tous les articles scientifiques suivent tous une structure similaire: une contextualisation de la recherche, les objectifs de la recherche, les méthodes utilisés pour réaliser la recherche, les résultats obtenus et les conclusions qui ont sont retenus.

4.2 Illustrations des méthodes de clustering

Dans cette section, nous allons analyser de manière visuelle les clusters qui ont été formées.

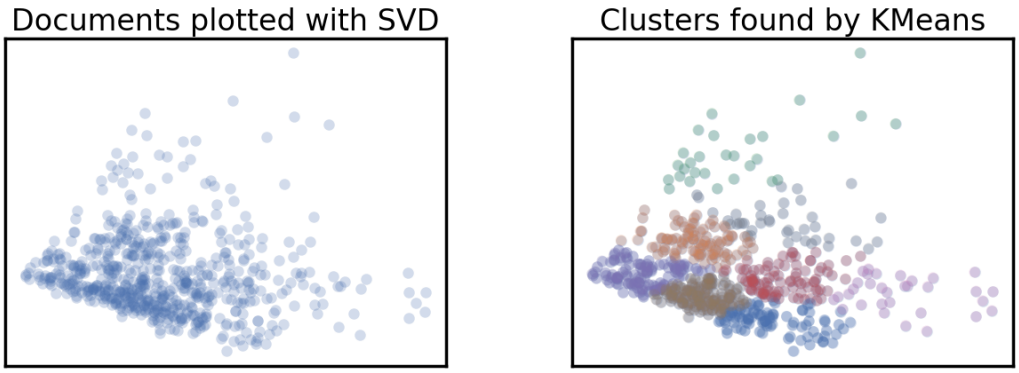

Figure 3: Représentation des titres avec SVD, clustérisés avec k-Means

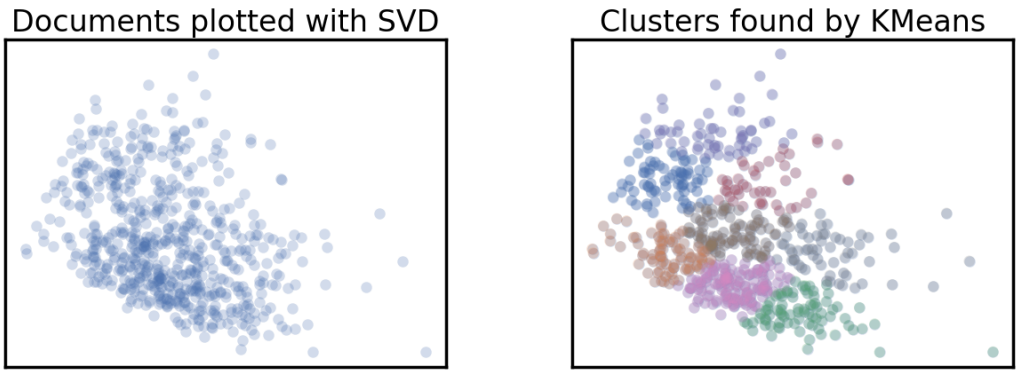

Figure 4: Représentation des abstracts avec SVD, clustérisés avec k-Means

En observant les représentation des titres et des abstracts à l'aide de SVD, nous constatons qu'une grande partie des titres sont condensés à un seul endroit, tandis que les autres titres sont beaucoup plus éparpillés. Concernant les abstracts, nous voyons qu'ils sont regroupés dans un espace. Dans les deux cas, nous constatons des valeurs aberrantes. Nous pouvons également constater que les clusters formés par l'algorithme k-Means représentent des clusters de tailles équivalents.

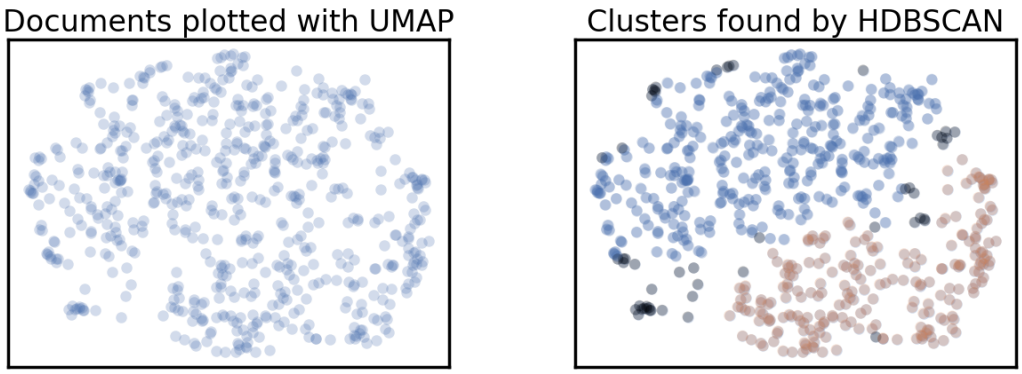

Figure 5: Représentation des titres avec UMAP, clustérisés avec HDBSCAN

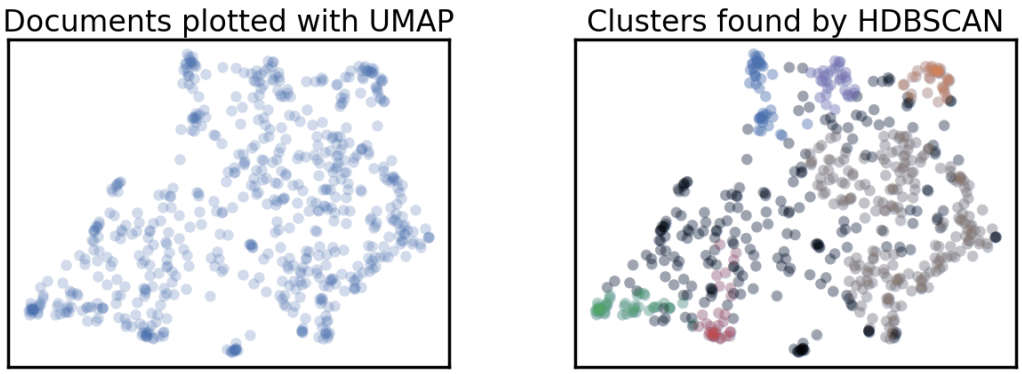

Figure 6: Représentation des abstracts avec UMAP, clustérisés avec HDBSCAN

Concernant les représentations des titres et des abstracts à l'aide de UMAP, nous constatons que les données sont éparpillés beaucoup plus uniformément, que dans le cas de SVD. Le clustering par les titres sont mêmes plus uniformes que le clustering par les abstracts. Nous pouvons tout de même observer que l'algorithme de clustering HDBSCAN ne sélectionne que des zones condensés de documents.

4.3 Comparaison de la pondération des mots-clés des clusters générés

Dans cette partie, pour chaque combinaison de technologies, nous allons comparer la pondération des mots-clés entre chaque cluster. Cela sera fait en calculant l'écart-type et la similarité cosinus. Illustrés dans les tables ci-dessous sont la pondération des six meilleurs mots-clés pour chaque cluster. Un tableau représente qu'une seule combinaison de technologies.

Résultats avec les clusters par titres, réduction SVD, clusters k-Means et pondération par moyenne

Cluster n°1	**Poids**	**Cluster n°2**	**Poids**	**Cluster n°3**	**Poids**	**Cluster n°4**	**Poids**
hiv	**0.1012**	infection	**0.1145**	hiv	**0.0793**	hiv	**0.0693**
arty	**0.0778**	hiv	**0.085**	test	**0.0410**	infection	**0.0558**
test	**0.0677**	cell	**0.0644**	infection	**0.0405**	cell	**0.0297**
ci	**0.0476**	viral	**0.0498**	arty	**0.0313**	response	**0.0287**
infant	**0.0474**	acuteness	**0.0484**	usage	**0.0302**	association	**0.0272**
treatment	**0.0431**	earliness	**0.0469**	treatment	**0.0274**	study	**0.0257**
Cluster n°5	**Poids**	**Cluster n°6**	**Poids**	**Cluster n°7**	**Poids**	**Cluster n°8**	**Poids**
hiv	**0.1126**	hiv	**0.1498**	infection	**0.0615**	infection	**0.0682**
test	**0.0632**	test	**0.1065**	hiv	**0.0519**	hiv	**0.0647**
ahi	**0.0631**	infection	**0.0969**	subtype	**0.0462**	cell	**0.0422**
infection	**0.0631**	ahi	**0.0730**	viral	**0.0445**	virus	**0.0390**
earliness	**0.0427**	diagnosis	**0.0590**	virus	**0.0441**	neutralization	**0.0370**
detection	**0.0409**	acuteness	**0.0568**	antibody	**0.0418**	viral	**0.0368**

Table 1: Pondération des mots-clés par clusters

L'écart-type est de $\sigma^2 = 0.0304$.
La moyenne des similarités cosinus est de $cos(\theta) = 0.9803$.

Résultats avec les clusters par titres, réduction SVD, clusters k-Means et pondération par SVD

Cluster n°1	**Poids**	**Cluster n°2**	**Poids**	**Cluster n°3**	**Poids**	**Cluster n°4**	**Poids**
hiv	**1.0160**	infection	**0.7548**	infection	**0.3701**	infection	**0.8178**
test	**0.5243**	hiv	**0.6549**	viral	**0.2879**	hiv	**0.6753**
infection	**0.491**	acuteness	**0.3655**	neutralization	**0.2873**	cell	**0.5347**
arty	**0.4266**	viral	**0.3069**	antibody	**0.2866**	response	**0.4286**
usage	**0.3701**	cell	**0.3051**	subtype	**0.2734**	viral	**0.4168**
treatment	**0.3442**	earliness	**0.2770**	virus	**0.2570**	neutralization	**0.3947**
Cluster n°5	**Poids**	**Cluster n°6**	**Poids**	**Cluster n°7**	**Poids**	**Cluster n°8**	**Poids**
hiv	**0.9594**	hiv	**0.8567**	hiv	**0.8307**	hiv	**1.1343**
arty	**0.7332**	test	**0.7193**	infection	**0.6820**	test	**0.6887**
test	**0.6056**	infection	**0.5158**	cell	**0.3552**	infection	**0.6356**
infant	**0.4607**	ahi	**0.4254**	response	**0.3412**	ahi	**0.6178**
ci	**0.4429**	acuteness	**0.3455**	association	**0.3410**	detection	**0.4819**
diagnosis	**0.3706**	detection	**0.3435**	transmission	**0.2898**	earliness	**0.4549**

Table 2: Pondération des mots-clés par clusters

L'écart-type est de $\sigma^2 = 0.227$.
La moyenne des similarités cosinus est de $cos(\theta) = 0.98$.

Résultats avec les clusters par abstracts, réduction SVD, clusters k-Means et pondération par moyenne

Cluster n°1	**Poids**	**Cluster n°2**	**Poids**	**Cluster n°3**	**Poids**	**Cluster n°4**	**Poids**
infection	**0.0542**	hiv	**0.0832**	hiv	**0.0736**	infection	**0.1166**
hiv	**0.0500**	infection	**0.0585**	test	**0.0431**	hiv	**0.0988**
cell	**0.0464**	test	**0.0373**	arty	**0.0411**	viral	**0.0559**
virus	**0.0396**	blood	**0.0330**	ahi	**0.0403**	cell	**0.0501**
subtype	**0.0392**	donor	**0.0305**	diagnosis	**0.0330**	association	**0.0475**
viral	**0.0364**	usage	**0.0288**	ci	**0.0325**	acuteness	**0.0472**
Cluster n°5	**Poids**	**Cluster n°6**	**Poids**	**Cluster n°7**	**Poids**	**Cluster n°8**	**Poids**
hiv	**0.0493**	hiv	**0.1765**	infection	**0.0948**	hiv	**0.1367**
infection	**0.0304**	infection	**0.1164**	neutralization	**0.0725**	test	**0.0829**
patience	**0.0277**	test	**0.0910**	cell	**0.0687**	arty	**0.0762**
usage	**0.0273**	acuteness	**0.0637**	antibody	**0.0649**	infant	**0.0539**
study	**0.0252**	earliness	**0.0552**	response	**0.0630**	diagnosis	**0.0505**
arty	**0.0223**	detection	**0.0550**	viral	**0.0474**	ci	**0.0503**

Table 3: Pondération des mots-clés par clusters

L'écart-type est de $\sigma^2 = 0.0433$.
La moyenne des similarités cosinus est de $cos(\theta) = 0.9764$.

Résultats avec les clusters par abstracts, réduction SVD, clusters k-Means et pondération par SVD

Cluster n°1	**Poids**	**Cluster n°2**	**Poids**	**Cluster n°3**	**Poids**	**Cluster n°4**	**Poids**
hiv	**0.8401**	infection	**0.7669**	hiv	**1.2551**	hiv	**0.8052**
ahi	**0.5561**	neutralization	**0.6174**	test	**0.7549**	infection	**0.5699**
test	**0.4986**	response	**0.5450**	arty	**0.6592**	blood	**0.4078**
arty	**0.4882**	antibody	**0.5251**	infant	**0.4846**	test	**0.4074**
diagnosis	**0.3952**	cell	**0.5221**	diagnosis	**0.4696**	donor	**0.3966**
ci	**0.3883**	viral	**0.3819**	ci	**0.4590**	donation	**0.3289**
Cluster n°5	**Poids**	**Cluster n°6**	**Poids**	**Cluster n°7**	**Poids**	**Cluster n°8**	**Poids**
hiv	**0.4921**	infection	**0.4545**	hiv	**1.378**	infection	**0.8055**
infection	**0.3044**	hiv	**0.4340**	infection	**0.8950**	hiv	**0.7096**
patience	**0.2949**	cell	**0.4128**	test	**0.7452**	viral	**0.3876**
usage	**0.2589**	virus	**0.3499**	acuteness	**0.5323**	cell	**0.3697**
study	**0.2536**	subtype	**0.3430**	detection	**0.4543**	acuteness	**0.3339**
node	**0.2356**	sequence	**0.2944**	earliness	**0.4510**	association	**0.3325**

Table 4: Pondération des mots-clés par clusters

L'écart-type est de $\sigma^2 = 0.3248$.
La moyenne des similarités cosinus est de $cos(\theta) = 0.976$.

Résultats avec les clusters par titres, réduction UMAP, clusters HDBSCAN et pondération par moyenne

Cluster n°1	**Poids**	**Cluster n°2**	**Poids**	**Cluster n°3**	**Poids**	**Cluster n°4**	**Poids**
hiv	**0.1243**	hiv	**0.0883**	infection	**0.0695**	hiv	**0.0759**
test	**0.0852**	tb	**0.0766**	hiv	**0.0684**	infection	**0.0594**
infection	**0.0834**	cell	**0.0676**	arty	**0.0541**	infant	**0.0585**
detection	**0.0717**	infection	**0.0589**	viral	**0.0495**	cell	**0.0575**
assay	**0.0633**	arty	**0.0575**	response	**0.0466**	child	**0.0475**
blood	**0.0603**	ci	**0.0484**	infant	**0.0411**	association	**0.0424**
Cluster n°5	**Poids**	**Cluster n°6**	**Poids**	**Cluster n°7**	**Poids**	**Cluster n°8**	**Poids**
hiv	**0.1079**	hiv	**0.1031**	hiv	**0.0916**	hiv	**0.1143**
infection	**0.0722**	infection	**0.0702**	infection	**0.0602**	infection	**0.0525**
test	**0.0580**	prep	**0.0500**	test	**0.0468**	test	**0.0499**
ahi	**0.0541**	transmission	**0.0423**	woman	**0.0362**	neutralization	**0.0409**
antibody	**0.0370**	model	**0.0420**	earliness	**0.0307**	diagnosis	**0.0365**
acuteness	**0.0359**	test	**0.0403**	diagnosis	**0.0287**	patience	**0.0353**

Table 5: Pondération des mots-clés par clusters

L'écart-type est de $\sigma^2 = 0.0189$.
La moyenne des similarités cosinus est de $cos(\theta) = 0.9871$.

Résultats avec les clusters par titres, réduction UMAP, clusters HDBSCAN et pondération par SVD

Cluster n°1	**Poids**	**Cluster n°2**	**Poids**	**Cluster n°3**	**Poids**	**Cluster n°4**	**Poids**
hiv	**0.7722**	cell	**0.7498**	infection	**0.4782**	hiv	**0.8204**
infection	**0.6998**	infection	**0.5384**	viral	**0.4381**	infection	**0.4718**
blood	**0.5172**	hiv	**0.3594**	virus	**0.3403**	arty	**0.4052**
cell	**0.4034**	viral	**0.2446**	response	**0.3062**	ahi	**0.3470**
donor	**0.3959**	level	**0.2257**	subtype	**0.2942**	transmission	**0.2926**
association	**0.3147**	acuteness	**0.1948**	hiv	**0.2773**	resistant	**0.2906**
Cluster n°5	**Poids**	**Cluster n°6**	**Poids**	**Cluster n°7**	**Poids**	**-**	**-**
hiv	**0.8800**	hiv	**0.7532**	hiv	**0.6032**	-	**-**
test	**0.7622**	tb	**0.5525**	test	**0.4331**	-	**-**
diagnosis	**0.4392**	test	**0.4558**	arty	**0.3964**	-	**-**
infection	**0.4362**	arty	**0.4190**	diagnosis	**0.3096**	-	**-**
ahi	**0.3501**	mortality	**0.3807**	ahi	**0.2731**	-	**-**
intervention	**0.3131**	ci	**0.3435**	infection	**0.2595**	-	**-**

Table 6: Pondération des mots-clés par clusters

L'écart-type est de $\sigma^2 = 0.1369$.
La moyenne des similarités cosinus est de $cos(\theta) = 0.9899$.

Résultats avec les clusters par abstracts, réduction UMAP, clusters HDBSCAN et pondération par moyenne

Cluster n°1	**Poids**	**Cluster n°2**	**Poids**	**Cluster n°3**	**Poids**	**Cluster n°4**	**Poids**
response	**0.0727**	hiv	**0.1329**	hiv	**0.0893**	infection	**0.0898**
infection	**0.0724**	infection	**0.0896**	pregnant	**0.0635**	hiv	**0.0760**
subtype	**0.0660**	test	**0.0840**	mortality	**0.0556**	incidence	**0.0693**
superinfection	**0.0625**	detection	**0.0594**	initiation	**0.0537**	bed	**0.0690**
hiv	**0.0605**	acuteness	**0.0561**	arty	**0.0496**	estimation	**0.0603**
virus	**0.0530**	earliness	**0.0450**	child	**0.0492**	viral	**0.0526**
Cluster n°5	**Poids**	**Cluster n°6**	**Poids**	**Cluster n°7**	**Poids**	**-**	**-**
hiv	**0.0928**	arty	**0.1110**	ahi	**0.1277**	-	**-**
infection	**0.0799**	hiv	**0.0919**	hiv	**0.1010**	-	**-**
donor	**0.0798**	test	**0.0631**	risk	**0.0600**	-	**-**
blood	**0.0763**	care	**0.0548**	infection	**0.0519**	-	**-**
viral	**0.0506**	hivst	**0.0477**	aehi	**0.0510**	-	**-**
estimation	**0.0447**	health	**0.0380**	test	**0.0504**	-	**-**

Table 7: Pondération des mots-clés par clusters

L'écart-type est de $\sigma^2 = 0.0222$.
La moyenne des similarités cosinus est de $cos(\theta) = 0.987$.

Résultats avec les clusters par abstracts, réduction UMAP, clusters HDBSCAN et pondération par SVD

Cluster n°1	**Poids**	**Cluster n°2**	**Poids**	**Cluster n°3**	**Poids**	**Cluster n°4**	**Poids**
tb	**0.7922**	hiv	**0.7126**	hiv	**0.7495**	infant	**0.5762**
arty	**0.4561**	test	**0.6732**	infection	**0.6044**	hiv	**0.5503**
resistant	**0.4094**	infection	**0.5596**	pregnant	**0.5719**	arty	**0.5424**
ci	**0.3815**	detection	**0.5519**	response	**0.4913**	infection	**0.3762**
treatment	**0.3681**	acuteness	**0.5349**	association	**0.3460**	initiation	**0.3494**
hiv	**0.3314**	combo	**0.4194**	viral	**0.3299**	week	**0.3088**
Cluster n°5	**Poids**	**Cluster n°6**	**Poids**	**Cluster n°7**	**Poids**	**-**	**-**
cell	**0.8284**	hiv	**0.7337**	neutralization	**0.8937**	-	**-**
infection	**0.6602**	ahi	**0.5758**	antibody	**0.5900**	-	**-**
hiv	**0.5897**	infection	**0.5351**	infection	**0.4108**	-	**-**
mortality	**0.4622**	diagnosis	**0.3369**	vaccine	**0.3730**	-	**-**
arty	**0.4468**	acuteness	**0.3350**	epitope	**0.2961**	-	**-**
child	**0.4061**	transmission	**0.3192**	envelope	**0.2877**	-	**-**

Table 8: Pondération des mots-clés par clusters

L'écart-type est de $\sigma^2 = 0.1002$.
La moyenne des similarités cosinus est de $cos(\theta) = 0.9856$.

5 Conclusion

5.1 Analyse des résultats

En classant les combinaisons de technologies par la valeur de l'écart-type puis de la similarité cosinus, nous obtenons les classements suivants:

5.1.1. *Écart-type*

1. Clusters par titres, UMAP et HDBSCAN, pondération par la moyenne ($\sigma^2 = 0.0189$)
2. Clusters par abstracts, UMAP et HDBSCAN, pondération par la moyenne ($\sigma^2 = 0.0222$)
3. Clusters par titres, SVD et k-Means, pondération par la moyenne ($\sigma^2 = 0.0304$)
4. Clusters par abstracts, SVD et k-Means, pondération par la moyenne ($\sigma^2 = 0.0433$)
5. Clusters par abstracts, UMAP et HDBSCAN, pondération par SVD ($\sigma^2 = 0.1002$)
6. Clusters par titres, UMAP et HDBSCAN, pondération par SVD ($\sigma^2 = 0.1369$)

7. Clusters par titres, SVD et k-Means, pondération par SVD ($\sigma^2 = 0.2270$)
8. Clusters par abstracts, SVD et k-Means, pondération par SVD ($\sigma^2 = 0.3248$)

Nous pouvons constater dans le cas des écarts-types, que les valeurs les plus bas sont enregistrés pour les combinaisons d'outils où la méthode de pondérations des mots-clés a été la moyenne. Nous pouvons également constater que dans la grande partie des cas, la clusterisation faite par UMAP et HDBSCAN produit des clusters plus équilibrés que lorsqu'elle a été faite par SVD et k-Means. Nous pouvons également relever que le clustering par titres est plus efficace que par les abstracts. Nous pouvons conclure à partir des calculs de l'écart-type, que la meilleure combinaison de méthodes appropriée pour notre problématique est de faire un clustering par titres, avec les algorithmes UMAP et HDBSCAN, et d'ensuite faire la pondération des mots-clés par la moyenne.

5.1.2. *Similarité cosinus*

1. Clusters par titres, UMAP et HDBSCAN, pondération par SVD ($cos(\theta) = 0.9899$)
2. Clusters par titres, UMAP et HDBSCAN, pondération par la moyenne ($cos(\theta) = 0.9871$)
3. Clusters par abstracts, UMAP et HDBSCAN, pondération par la moyenne ($cos(\theta) = 0.9870$)
4. Clusters par abstracts, UMAP et HDBSCAN, pondération par SVD ($cos(\theta) = 0.9856$)
5. Clusters par titres, SVD et k-Means, pondération par la moyenne ($cos(\theta) = 0.9803$)
6. Clusters par titres, SVD et k-Means, pondération par SVD ($cos(\theta) = 0.9800$)
7. Clusters par abstracts, SVD et k-Means, pondération par SVD ($cos(\theta) = 0.9760$)
8. Clusters par abstracts, SVD et k-Means, pondération par la moyenne ($cos(\theta) = 0.9764$)

Concernant la similarité cosinus, nous pouvons constater que les valeurs sont relativement proches les unes des autres, et qu'elles sont toutes proches de la valeur 1. Cela signifie donc que toutes les clusters ont beaucoup de choses en commun entre elles. Toutefois, nous constatons que les combinaisons qui se rapprochent de la valeur de 1 pour la similarité cosinus sont les clusters formés avec les algorithmes UMAP et HDBSCAN. Nous pouvons également remarquer que les clusters par titres sont plus similaires entre eux, que les clusters par abstracts. Il n'y a pas de différence entre les deux méthodes de pondération des mots-clés dans chacune des clusters (moyenne ou SVD). Nous pouvons donc également conclure avec la similarité cosinus, que la meilleure méthode pour générer des clusters équilibrés serait de clusteriser les titres avec les algorithmes UMAP et HDBSCAN. Rien ne peut toutefois être dit sur la méthode de pondération des mots.

5.2 Limitations

Malgré le fait d'avoir généré des clusters et un résumé à partir de chaque cluster, il y a des limitations qui ont été observés. Les algorithmes UMAP et HDBSCAN n'ont pas été explorés en profondeur. Il est possible que s'il y avait eu la possibilité d'explorer un peu plus les paramètres de ces méthodes, nous arrivions à des résultats plus fructueuses.
De plus, l'algorithme de génération de résumés ne distingue pas les phrases, qui parlent de la méthodologie, des résultats ou des conclusions dans l'article. Il était possible avec l'aide de la librairie d'expressions régulières de filtrer les phrases qui mentionnent des détails sur les participants, mais il n'est pas possible de déterminer la nature de la phrase. Si cela avait été possible, les clusters pourraient être construits plus intelligemment en regroupant les articles utilisant des méthodes similaires, ou arrivant à des conclusions identiques. Cela aurait donné des résultats plus intéressants. Finalement, ceci est une génération de résumés extractive. Il n'y a donc pas de relations sémantiques entre les phrases sélectionnés.

5.3 Travaux futurs

Afin de surmonter ces limites, une exploration approfondi des paramètres de l'algorithme de clustering HDBSCAN est nécessaire pour améliorer les résultats. Il serait intéressant aussi de comparer cette technologie avec d'autres existantes. Il serait également bénéfique de développer un algorithme de machine learning, se reposant sur les grands modèles de langages (LLM) et s'entraînant sur les résumés d'articles, est proposé pour identifier les phrases parlant d'une méthodologie, de résultats ou de conclusions, voire identifier les méthodes de recherche appliqués et les conclusions de la recherche. Avec ces informations, nous pourrions créer des clusters d'articles regroupant les mêmes méthodes, ou les mêmes conclusions. Cela nous permettraient idéalement d'obtenir des phrases qui parlent tous de la même idée et nous pouvons alors résumer un cluster à une seule phrase. L'algorithme de machine learning permettrait également de sélectionner les phrases plus intelligemment, en considérant que les phrases de conclusions, sachant qu'ils sont les plus pertinents pour le chercheur. Finalement, une fois les phrases sélectionnés, l'algorithme applique la méthode abstractive de génération de résumés, en réarrangeant les phrases sélectionnés et en ajoutant des phrases supplémentaires, afin d'introduire des relations sémantiques entre les phrases sélectionnés. Un moyen à court terme qui peut être utilisé est de demander à intelligence artificielle existante, tel que ChatGPT, de générer un résumé à partir des phrases sélectionnés. Cela a été fait dans le cadre de ce travail dans l'exemple qui suit et sert en tant que piste pour les travaux futurs

To study the structure of human immunodeficiency virus (HIV)-1 drug resistance (DR) in patients with newly diagnosed infection. Blood donations in South Africa are tested for HIV RNA using individual donation NAT (ID-NAT), allowing detection and rapid antiretroviral therapy (ART) of acute HIV infections. Broadly neutralizing antibodies (bNAbs) for HIV-1 prevention or cure strategies must inhibit transmitted/founder and reservoir viruses. Multi-assay algorithms (MAAs) are used to estimate population-level HIV incidence and identify individuals with recent infection. Information on treatment failure (TF) in People living with HIV in a data-poor setting is necessary to counter the epidemic of TF with first-line combined antiretroviral therapies (cART) in sub-Saharan Africa (SSA). Declines in HIV incidence have been slower than expected during the roll-out of antiretroviral treatment (ART) services in sub-Saharan African populations suffering generalized epidemics. Pre-exposure prophylaxis (PrEP) reduces HIV acquisition risk by >90% and is a critical lever to reduce HIV incidence.

To reduce HIV incidence, it is critical to study the structure of drug resistance in newly diagnosed patients and develop prevention strategies using broadly neutralizing antibodies (bNAbs) that can inhibit transmitted/founder and reservoir viruses. Multi-assay algorithms (MAAs) help estimate population-level HIV incidence and identify recent infections, while monitoring treatment failure (TF) is necessary to counter the epidemic of TF with first-line combined antiretroviral therapies (cART) in sub-Saharan Africa (SSA). Despite the roll-out of antiretroviral treatment (ART) services, declines in HIV incidence have been slower than expected in SSA populations, highlighting the importance of interventions like pre-exposure prophylaxis (PrEP) which can reduce HIV acquisition risk by >90%.

6 Remerciements

Je tiens à remercier mes superviseurs Prof. DI MARZO SERUGENDO Giovanna, Mme FRIHA Lamia et M. HUGENTOBLER Alain, ainsi qu'à toute l'équipe LitRev pour permettre la réalisation de ce travail.

References

[1] Erol Orel et al. *LiteRev, an Automation Tool to Support Literature Reviews: A Case Study on Acute and Early HIV Infection in Sub-Saharan Africa*. Pages: 2023.02.20.23286179. Feb. 21, 2023. DOI: 10.1101/2023.02.20.23286179. URL: https://www.medrxiv.org/content/10.1101/2023.02.20.23286179v1 (visited on 03/20/2023).

[2] Mahak Gambhir and Vishal Gupta. "Recent automatic text summarization techniques: a survey". In: *Artificial Intelligence Review* 47.1 (Jan. 1, 2017), pp. 1–66. ISSN: 1573-7462. DOI: 10.1007/s10462-016-9475-9. URL: https://doi.org/10.1007/s10462-016-9475-9 (visited on 05/16/2023).

[3] Avaneesh Kumar Yadav et al. "State-of-the-art approach to extractive text summarization: a comprehensive review". In: *Multimedia Tools and Applications* (Feb. 16, 2023). ISSN: 1573-7721. DOI: 10.1007/s11042-023-14613-9. URL: https://doi.org/10.1007/s11042-023-14613-9 (visited on 05/15/2023).

[4] Rafael Ferreira et al. "A multi-document summarization system based on statistics and linguistic treatment". In: *Expert Systems with Applications* 41.13 (Oct. 1, 2014), pp. 5780–5787. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2014.03.023. URL: https://www.sciencedirect.com/science/article/pii/S0957417414001523 (visited on 06/05/2023).

[5] Hassan Aliakbarpour, Mohammad Taghi Manzuri, and Amir Masoud Rahmani. "Improving the readability and saliency of abstractive text summarization using combination of deep neural networks equipped with auxiliary attention mechanism". In: *The Journal of Supercomputing* 78.2 (Feb. 1, 2022), pp. 2528–2555. ISSN: 1573-0484. DOI: 10.1007/s11227-021-03950-x. URL: https://doi.org/10.1007/s11227-021-03950-x (visited on 05/30/2023).

[6] Rajeev Kumar Singh et al. "SHEG: summarization and headline generation of news articles using deep learning". In: *Neural Computing and Applications* 33.8 (Apr. 1, 2021), pp. 3251–3265. ISSN: 1433-3058. DOI: 10.1007/s00521-020-05188-9. URL: https://doi.org/10.1007/s00521-020-05188-9 (visited on 06/05/2023).

[7] Kinder Chen. *Introduction to Natural Language Processing — TF-IDF*. Medium. May 24, 2021. URL: https://kinder-chen.medium.com/introduction-to-natural-language-processing-tf-idf-1507e907c19 (visited on 09/04/2023).

[8] Kirk Baker. "Singular value decomposition tutorial". In: *The Ohio State University* 24 (2005), p. 22.

[9] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. Sept. 17, 2020. DOI: 10.48550/arXiv.1802.03426. arXiv: 1802.03426[cs,stat]. URL: http://arxiv.org/abs/1802.03426 (visited on 09/04/2023).

[10] Education Ecosystem (LEDU). *Understanding K-means Clustering in Machine Learning*. Medium. Sept. 12, 2018. URL: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1 (visited on 09/04/2023).

[11] Khyati Mahendru. *How to determine the optimal K for K-Means?* Analytics Vidhya. June 17, 2019. URL: https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb (visited on 09/04/2023).

[12] Leland McInnes, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering". In: *Journal of Open Source Software* 2.11 (Mar. 21, 2017), p. 205. ISSN: 2475-9066. DOI: 10.21105/joss.00205. URL: https://joss.theoj.org/papers/10.21105/joss.00205 (visited on 08/28/2023).

[13] *Cosine Similarity - an overview | ScienceDirect Topics*. URL: https://www.sciencedirect.com/topics/computer-science/cosine-similarity (visited on 09/08/2023).

7 Annexes

7.1 Résumés produits pour chaque méthode

Chaque paragraphe représente un résumé d'articles contenus dans les clusters qui ont été identifiés dans chacune des méthodes.

Clusters par titres, réduction SVD, clusters k-Means et pondération par moyenne
Wealth-related inequality in early uptake of HIV testing was measured using the Erreygers concentration index (CI) further adjusted for inequality introduced by predicted healthcare need (ie, need-standardised). To assess the performance of rapid HIV tests in comparison to a laboratory-based HIV ELISA test for determining HIV-exposure and excluding HIV infection during infancy. The timing of initiation of ART after commencing TB therapy was not significantly associated with increased mortality or survival.

Primary HIV infection (PHI) and subsequent chronic infection alter B-cell compartment. Compared with women with BED *geq* 0.8/CD4 *geq* 350 (typical of HIV-1 chronic patients) there was insufficient evidence to conclude that women presenting with BED $<$ 0.8/CD4 *geq* 350 (typical of recent infections) were more likely to transmit in utero [adjusted odds ratio (aOR) = 1.37, 96% confidence interval (CI) 0.90-2.08, P = 0.14], whereas women with BED $<$ 0.8/CD4 200-349 (possibly recently infected patients) had a 2.57 (95% CI 1.39-4.77, P-value $<$ 0.01) odds of transmitting in utero. Haemotologic parameters were assessed before infection and at regular intervals in the first twelve months of HIV infection.

Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised. Reactivity with all HIV genotypes was 100%. HIV self-screening (HIVSS) is acceptable to adults, but there is limited data on HIVSS feasibility in community programmes.

There was a trend towards lower levels of CD161++CD8+ T cells in HIV-negative individuals with active and latent TB. These cells were dramatically increased in chronic HIV infection. SP-D levels are increased in lung fluid from AIDS patients but not in patients with early HIV infection.

The Defer/Test strategy averted the most HIV infections. Because window-period donations are the most important source of residual HIV contamination and arise from incident infections, research to develop risk factor exclusion strategies must focus on predictors of HIV seroconversion. Individuals with acute (preseroconversion) HIV infection (AHI) are important in the spread of HIV.

From July 2005 to June 2006, women were offered HIV testing following group information and education on HIV and STDs in the clinic waiting area. The overall HIV prevalence was 37.3% [95% confidence interval (CI) 34.3–41.3]. HIV transmission risk is higher during acute and early HIV infection than it is during chronic infection, but the contribution of early infection to the spread of HIV is controversial.

Establishment of persistent human immunodeficiency virus type 1 (HIV-1) reser-

voirs occurs early in infection, and biomarkers of infected CD4+ T cells during acute infection are poorly defined. Parotid lymph nodes reflect these HIV-related changes. The focus of infection with human immunodeficiency virus (HIV) is the lymphatic system and this results in a specific HIV-related pathology in the parotid.

Sensitivity, specificity, predictive values, and agreements were compared among the EIA kits using PCR-confirmed HIV-positive and negative samples. A single viral variant is transmitted in the majority of HIV infections. During acute HIV infection, HIV actively replicates but seroconversion has not yet occurred.

*Clusters par titres, réduction SVD, clusters k-Means et pondération par SVD*Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised. Reactivity with all HIV genotypes was 100%. HIV self-screening (HIVSS) is acceptable to adults, but there is limited data on HIVSS feasibility in community programmes.

Acute HIV infection lasts approximately 3 weeks and early HIV infection, which includes acute HIV infection, lasts approximately 7 weeks. Persons with acute HIV infection (AHI) are highly infectious and responsible for a disproportionate share of incident infections. During a South African cervical cancer screening trial, 5595 women 35-65 years of age were followed up for 36 months; 577 women were HIV positive at enrollment, HIV seroconversion occurred in 123 women, and 4895 women remained HIV negative throughout.

To determine the availability of the gp120 V3 loop for neutralizing antibody binding on SHIV-89.6 and KB9 virions, we have constructed immunogenic C4-V3 peptides from these SHIVs and induced anti-V3 antibodies in guinea pigs and rhesus monkeys. In this study, we assessed the autologous and heterologous neutralization responses of 14 HIV-1 subtype C-infected individuals, using envelope clones obtained within the first 2 months postinfection. Establishment of persistent human immunodeficiency virus type 1 (HIV-1) reservoirs occurs early in infection, and biomarkers of infected CD4+ T cells during acute infection are poorly defined.

Sensitivity, specificity, predictive values, and agreements were compared among the EIA kits using PCR-confirmed HIV-positive and negative samples. A single viral variant is transmitted in the majority of HIV infections. Matched preinfection and postinfection samples were available from 13 individuals.

Patients who had never tested for HIV or tested negative over one year prior to recruitment were enrolled between May 2008 and March 2010. Wealth-related inequality in early uptake of HIV testing was measured using the Erreygers concentration index (CI) further adjusted for inequality introduced by predicted healthcare need (ie, need-standardised). To assess the performance of rapid HIV tests in comparison to a laboratory-based HIV ELISA test for determining HIV-exposure and excluding HIV infection during infancy.

From July 2005 to June 2006, women were offered HIV testing following group information and education on HIV and STDs in the clinic waiting area. The overall HIV prevalence was 37.3% [95% confidence interval (CI) 34.3–41.3]. To identify, diagnose and counsel patients with acute HIV infection (AHI) during routine HIV testing in South Africa.

There was a trend towards lower levels of CD161++CD8+ T cells in HIV-negative individuals with active and latent TB. These cells were dramatically increased in chronic HIV infection. SP-D levels are increased in lung fluid from AIDS patients but not in patients with early HIV infection.

The Defer/Test strategy averted the most HIV infections. Because window-period donations are the most important source of residual HIV contamination and arise from incident infections, research to develop risk factor exclusion strategies must focus on predictors of HIV seroconversion. Fourth-generation HIV assays detect both antigen and antibody, facilitating detection of acute/early HIV infection.

Clusters par abstracts, réduction SVD, clusters k-Means et pondération par moyenne These cells were dramatically increased in chronic HIV infection. Sustained viremia after acute HIV infection is associated with profound CD4 We have studied the coreceptor requirements of 12 primary HIV-1 O-type isolates from individuals with different clinical symptoms.

The Defer/Test strategy averted the most HIV infections. It seems likely that these latter subjects lead to spread of HIV. Acute HIV infection (prior to antibody seroconversion) represents a high-risk window for HIV transmission.

Reactivity with all HIV genotypes was 100%. Mobile HIV screening may facilitate early HIV diagnosis. Wealth-related inequality in early uptake of HIV testing was measured using the Erreygers concentration index (CI) further adjusted for inequality introduced by predicted healthcare need (ie, need-standardised).

Sensitivity, specificity, predictive values, and agreements were compared among the EIA kits using PCR-confirmed HIV-positive and negative samples. There was a trend towards lower levels of CD161++CD8+ T cells in HIV-negative individuals with active and latent TB. Fourth-generation HIV assays detect both antigen and antibody, facilitating detection of acute/early HIV infection.

Data from a prospective cohort study conducted during 1989-1990 of HIV serology and from a retrospective review of laboratory records of 727 patients presenting for superficial lymph node biopsy at the University Teaching Hospital in Lusaka, Zambia, were analyzed to determine the relative significance of HIV-associated lymphadenopathy among patients undergoing lymph node biopsy. HIV voluntary counselling and testing (VCT) is important for prevention, detection and treatment of HIV infection. A higher AUC statistic confirmed the superior predictive performance of the PwD assay for the three cut-offs.

Because window-period donations are the most important source of residual HIV contamination and arise from incident infections, research to develop risk fac-

tor exclusion strategies must focus on predictors of HIV seroconversion. HIV transmission risk is higher during acute and early HIV infection than it is during chronic infection, but the contribution of early infection to the spread of HIV is controversial. Acute HIV infection lasts approximately 3 weeks and early HIV infection, which includes acute HIV infection, lasts approximately 7 weeks.

Matched preinfection and postinfection samples were available from 13 individuals. TRIM5alphahu levels did not change significantly after infection. Primary HIV infection (PHI) and subsequent chronic infection alter B-cell compartment.

Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised. The overall HIV prevalence was 37.3% [95% confidence interval (CI) 34.3–41.3]. HIV self-screening (HIVSS) is acceptable to adults, but there is limited data on HIVSS feasibility in community programmes.

Clusters par abstracts, réduction SVD, clusters k-Means et pondération par SVD Reactivity with all HIV genotypes was 100%. Mobile HIV screening may facilitate early HIV diagnosis. Wealth-related inequality in early uptake of HIV testing was measured using the Erreygers concentration index (CI) further adjusted for inequality introduced by predicted healthcare need (ie, need-standardised).

Matched preinfection and postinfection samples were available from 13 individuals. TRIM5alphahu levels did not change significantly after infection. To determine the availability of the gp120 V3 loop for neutralizing antibody binding on SHIV-89.6 and KB9 virions, we have constructed immunogenic C4-V3 peptides from these SHIVs and induced anti-V3 antibodies in guinea pigs and rhesus monkeys.

Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised. The overall HIV prevalence was 37.3% [95% confidence interval (CI) 34.3–41.3]. HIV self-screening (HIVSS) is acceptable to adults, but there is limited data on HIVSS feasibility in community programmes.

The Defer/Test strategy averted the most HIV infections. It seems likely that these latter subjects lead to spread of HIV. Acute HIV infection (prior to antibody seroconversion) represents a high-risk window for HIV transmission.

Data from a prospective cohort study conducted during 1989-1990 of HIV serology and from a retrospective review of laboratory records of 727 patients presenting for superficial lymph node biopsy at the University Teaching Hospital in Lusaka, Zambia, were analyzed to determine the relative significance of HIV-associated lymphadenopathy among patients undergoing lymph node biopsy. HIV voluntary counselling and testing (VCT) is important for prevention, detection and treatment of HIV infection. HIV serology was tested in 22 children and was positive in eight, including four of 14 with tuberculous lymphadenitis.

These cells were dramatically increased in chronic HIV infection. We have stud-

ied the coreceptor requirements of 12 primary HIV-1 O-type isolates from individuals with different clinical symptoms. HIV-1 and related viruses require coreceptors, in addition to CD4, to infect target cells.

Because window-period donations are the most important source of residual HIV contamination and arise from incident infections, research to develop risk factor exclusion strategies must focus on predictors of HIV seroconversion. HIV transmission risk is higher during acute and early HIV infection than it is during chronic infection, but the contribution of early infection to the spread of HIV is controversial. Acute HIV infection lasts approximately 3 weeks and early HIV infection, which includes acute HIV infection, lasts approximately 7 weeks.

Sensitivity, specificity, predictive values, and agreements were compared among the EIA kits using PCR-confirmed HIV-positive and negative samples. There was a trend towards lower levels of CD161++CD8+ T cells in HIV-negative individuals with active and latent TB. Fourth-generation HIV assays detect both antigen and antibody, facilitating detection of acute/early HIV infection.

Clusters par titres, réduction UMAP, clusters HDBSCAN et pondération par moyenne
Fourth-generation HIV assays detect both antigen and antibody, facilitating detection of acute/early HIV infection. Most point-of-care HIV assays have poor sensitivity to diagnose acute HIV infection as they only detect antibodies against HIV-1 and HIV-2 (HIV-1/2). Reactivity with all HIV genotypes was 100%.

Acute and chronic HIV were associated with lower frequencies of CD161++CD8+ T cells, which did not correlate with CD4 count or HIV viral load. CD45RA(+) CD4(+) T cells were depleted during the CDC-B stage. To estimate the effect of ART on TB incidence while accounting for time-dependent confounders affected by exposure, a Cox proportional hazards marginal structural model was used.

Acute HIV infection visits were defined as those up to 3 months prior to and including the visit at which HIV DNA was first detected. Differences in the prevalence of symptoms at acute infection versus noninfection visits were determined overall and were stratified by age at infection (<2 months vs. >or=2 months). The cost of diagnosing HIV infection earlier in infancy was measured.

These cells were dramatically increased in chronic HIV infection. Sustained viremia after acute HIV infection is associated with profound CD4 Here, we assessed CD8+ T cell functional evolution from primary to chronic HIV infection.

Individuals with acute (preseroconversion) HIV infection (AHI) are important in the spread of HIV. Acute HIV infection (prior to antibody seroconversion) represents a high-risk window for HIV transmission. During a South African cervical cancer screening trial, 5595 women 35-65 years of age were followed up for 36 months; 577 women were HIV positive at enrollment, HIV seroconversion occurred in 123 women, and 4895 women remained HIV negative throughout.

HIV and other sexually transmitted infections (STIs) often co-occur. Primary HIV infection among women primarily drives the pediatric HIV epidemic. Pre-

exposure prophylaxis (PrEP) reduces HIV acquisition risk by >90% and is a critical lever to reduce HIV incidence.

The overall HIV prevalence was 37.3% [95% confidence interval (CI) 34.3–41.3]. From July 2005 to June 2006, women were offered HIV testing following group information and education on HIV and STDs in the clinic waiting area. Late diagnosis of HIV infection is a major challenge to the scale-up of HIV prevention and treatment.

Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised. HIV self-screening (HIVSS) is acceptable to adults, but there is limited data on HIVSS feasibility in community programmes. In sub-Saharan Africa, most people with HIV do not know they are infected.

Clusters par titres, réduction UMAP, clusters HDBSCAN et pondération par SVD
The Defer/Test strategy averted the most HIV infections. Sensitivity, specificity, predictive values, and agreements were compared among the EIA kits using PCR-confirmed HIV-positive and negative samples. Because window-period donations are the most important source of residual HIV contamination and arise from incident infections, research to develop risk factor exclusion strategies must focus on predictors of HIV seroconversion.

The peak frequency of Gag-responsive IFN-γ(+)CD4(+) T cells was observed at a median of 28 d (interquartile range: 21-81 d) post-Fiebig I/II staging, whereas Gag-specific IFN-γ(+)CD8(+) T cell responses peaked at a median of 253 d (interquartile range: 136-401 d) and showed a significant biphasic expansion. Prior to the disappearance of Gag-responsive Ki67(+)CD4(+) T cells, these cells positively correlated (p = 0.00038) with viremia, indicating that early Gag-responsive CD4 events are shaped by viral burden. 80 individuals were recruited for cross-sectional analysis: controls (n = 18), latent MTB infection (LTBI) only (n = 16), pulmonary tuberculosis (TB) only (n = 8), HIV only (n = 13), HIV and LTBI co-infection (n = 15) and HIV and TB co-infection (n = 10). Hormonal contraception was not associated with either the HIV-1 plasma setpoint or cervical loads during early infection. Cervical loads were significantly higher (0.7-1.1 log10 copies/swab) during acute infection than subsequently. N. gonorrhoeae coinfection was present during HIV acquisition in 6 out of 35 (17%), and was associated with an increased breadth and magnitude of systemic HIV-specific CD8 T-cell responses, using both interferon-gamma gamma and MIP-1 beta as an output.

Individuals with acute (preseroconversion) HIV infection (AHI) are important in the spread of HIV. The stages were early HIV disease, late HIV disease and AIDS. Serum specimens negative for HIV antibodies were screened by HIV RNA PCR using a highly specific pooling/resolution testing algorithm.

Patients who had never tested for HIV or tested negative over one year prior to recruitment were enrolled between May 2008 and March 2010. Community-based HIV testing services (HTS) can contribute to increased testing coverage and early HIV diagnosis, with HIV self-testing (HIVST) strategies showing promise.

HIV testing is the first step to stop transmission.

From July 2005 to June 2006, women were offered HIV testing following group information and education on HIV and STDs in the clinic waiting area. Wealth-related inequality in early uptake of HIV testing was measured using the Erreygers concentration index (CI) further adjusted for inequality introduced by predicted healthcare need (ie, need-standardised). Mobile HIV screening may facilitate early HIV diagnosis.

Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised. The overall HIV prevalence was 37.3% [95% confidence interval (CI) 34.3–41.3]. HIV self-screening (HIVSS) is acceptable to adults, but there is limited data on HIVSS feasibility in community programmes.

*Clusters par abstracts, réduction UMAP, clusters HDBSCAN et pondération par moyenne*The proportions of subtype A, D, or recombinants showed no significant increasing or decreasing trend over this time period (p=0.51). We used log-binomial regression to compare ARS symptom prevalence among those with subtype A vs. C or D, adjusting for sex, time since enrolment, and enrolment viral load. We characterized CD8(+) T-cell responses in 20 acutely infected, antiretroviral-naive individuals with HIV-1 subtype C infection using the interferon- enzyme-linked immunosorbent spot assay.

The overall HIV prevalence was 37.3% [95% confidence interval (CI) 34.3–41.3]. Fourth-generation HIV assays detect both antigen and antibody, facilitating detection of acute/early HIV infection. HIV transmission risk is higher during acute and early HIV infection than it is during chronic infection, but the contribution of early infection to the spread of HIV is controversial.

There is a risk of mother-to-child transmission of HIV (MTCT) during pregnancy and breastfeeding. In all, 140 children were recruited. When treatment was initiated at a CD4

Compared with women with BED $\geq$ 0.8/CD4 $\geq$ 350 (typical of HIV-1 chronic patients) there was insufficient evidence to conclude that women presenting with BED $<$ 0.8/CD4 $\geq$ 350 (typical of recent infections) were more likely to transmit in utero [adjusted odds ratio (aOR) = 1.37, 96% confidence interval (CI) 0.90-2.08, P = 0.14], whereas women with BED $<$ 0.8/CD4 200-349 (possibly recently infected patients) had a 2.57 (95% CI 1.39-4.77, P-value $<$ 0.01) odds of transmitting in utero. The BED assay was developed to estimate the proportion of recent HIV infections in a population. A single viral variant is transmitted in the majority of HIV infections.

The Defer/Test strategy averted the most HIV infections. Because window-period donations are the most important source of residual HIV contamination and arise from incident infections, research to develop risk factor exclusion strategies must focus on predictors of HIV seroconversion. We conducted a convenience sampling comparing the seroprevalence of infectious agents (HIV, HBsAg, HCV and

syphilis) in deferred versus accepted blood donors after medical selection.

In total, 353 initiated ART with median (IQR) 97.9 (60.5, 384.5) days from estimated seroconversion; 253/353 early ART, 100 deferred ART. Infants with HIV $<$ 12 weeks old with CD4% $\geq$ 25% were randomized in the CHER trial to early limited ART for 40 or 96 weeks (ART-40 W, ART-96 W), or deferred ART (ART-Def). From 2011 onward, the proportion of patients entering ART with advanced HIV disease has remained relatively unchanged.

Individuals with acute (preseroconversion) HIV infection (AHI) are important in the spread of HIV. A model-based score was assigned to each predictor to create a risk score for every woman. The unchanged global HIV incidence may be related to ignoring acute HIV infection (AHI).

*Clusters par abstracts, réduction UMAP, clusters HDBSCAN et pondération par SVD*To estimate the effect of ART on TB incidence while accounting for time-dependent confounders affected by exposure, a Cox proportional hazards marginal structural model was used. During follow-up, 242 (66.5%) children initiated ART and 81 (22.3%) developed TB. The pooled incident TB among adult HIV infected patients in Ethiopia was 16.58% (95% CI; 13.25-19.91%).

Most point-of-care HIV assays have poor sensitivity to diagnose acute HIV infection as they only detect antibodies against HIV-1 and HIV-2 (HIV-1/2). Development of a test to detect acute infection at the point-of-care is urgent. Real-time pooled RNA testing for the detection of acute HIV infection is feasible in resource-limited settings.

Prevention of acute HIV infections in pregnancy is required to achieve elimination of pediatric HIV. There is a risk of mother-to-child transmission of HIV (MTCT) during pregnancy and breastfeeding. Public health programs need to continue to reinforce prevention strategies and HIV retesting during pregnancy.

Acute HIV infection visits were defined as those up to 3 months prior to and including the visit at which HIV DNA was first detected. HIV is a major contributor to infant mortality. Infants received sdNVP and zidovudine (ZDV) for 1 week.

These cells were dramatically increased in chronic HIV infection. By age 1 year, an estimated 35.2% infected and 4.9% uninfected children will have died; by 2 years of age, 52.5% and 7.6% will have died, respectively. The frequency of CD161++CD8+ T cells was assessed prior to and during antiretroviral therapy (ART) in 14 HIV-positive patients.

Diagnosis of acute HIV infection (AHI) presents an opportunity to prevent HIV transmission during a highly infectious period. Acute HIV infection lasts approximately 3 weeks and early HIV infection, which includes acute HIV infection, lasts approximately 7 weeks. The unchanged global HIV incidence may be related to ignoring acute HIV infection (AHI).

To determine the availability of the gp120 V3 loop for neutralizing antibody binding on SHIV-89.6 and KB9 virions, we have constructed immunogenic C4-V3 peptides from these SHIVs and induced anti-V3 antibodies in guinea pigs and rhesus monkeys. Despite high neutralizing potential and cross-reactivity, anti-V3 antibodies do not contribute to autologous neutralization. The magnitude of this response was associated with shorter V1-to-V5 envelope lengths and fewer glycosylation sites, particularly in the V1-V2 region.

7.2 *Méthode avancée de lemmatisation*

Algorithme 1 : Lemmatisation

Entrées : un mot w // *'transmitted'*
Sorties : un nouveau mot w' // *'transmission'*
début
  $w_l \leftarrow w$ lemmatisé en fonction de sa forme; // $w_l \leftarrow$ *transmit* ;
  $W \leftarrow$ lemmas de WordNet pour w_l; // [*'convey'.v.03.transmit'*, *'impart.v.03.transmit'*, *'air.v.03.transmit'*, *'transmit.v.04.transmit'*] ;
  NomsCommuns $\leftarrow$ { } ;
  pour chaque *lemma* $\in W$ **faire**
    si *lemma est un nom commun* ***et*** *lemma se termine par -ion, -ment, -ity, -x, -k, -lt, -ence, -ness, -ant, -ure, -ude, -is, -ood, -end, -age* ***et non*** *lemma se termine par -hood, -ist, -r* **alors**
      ajouter *lemma* à *NomsCommuns*
    fin
    pour chaque *LemmaRelié* $\in$ *formes sémantiquement liés au lemma* **faire**
      // *pour 'convey'.v.03.transmit': ['familial.s.02.transmissible', 'catching.s.01.transmissible', 'infection.n.04.transmission', 'vector.n.03.transmitter'] ;*
      // *pour 'impart.v.03.transmit': ['transmission.n.02.transmission', 'transmission.n.01.transmittal'] ;*
      // *pour 'air.v.03.transmit': ['transmission.n.02.transmission', 'transmitter.n.03.transmitter'] ;*
      // *pour 'transmit.v.04.transmit': ['transmission.n.01.transmission', 'sender.n.01.transmitter', 'transmission.n.01.transmitting', 'transmission.n.01.transmittal'] ;*
      si *LemmaRelié est un nom commun* ***et*** *LemmaRelié se termine par -ion, -ment, -ity, -x, -k, -lt, -ence, -ness, -ant, -ure, -ude, -is, -ood, -end, -age* ***et non*** *LemmaRelié se termine par -hood, -ist, -r* **alors**
        ajouter *LemmaRelié* à *NomsCommuns*
      fin
    fin
  fin
  NomsCommunsList $\leftarrow$ conversion de *NomsCommuns* en liste; // [*'transmission'*] ;
  si *NomsCommunsList est vide* **alors**
    retourner w_l
  sinon
    retourner premier élément de *NomsCommunsList* lemmatisé
  fin
fin

Algorithme 2 : Pré-traitement de texte

Entrées : une liste de documents D avec leurs méta-données, une liste de mots vides M

Sorties : la liste de documents D' avec leurs méta-données et les titres et abstracts pré-traités

début

TitresPretraités ← [ ] ;
PhrasesPretraités ← [ ] ;
AbstractsPretraités ← [ ] ;
pour chaque *titre* $\in D[Titres]$, *abstract* $\in D[Abstracts]$ **faire**
PhrasesTitre ← *titres* décomposé par des points, les éléments sont transformés en minuscule ;
TitreMotsClés ← [ ] ;
pour chaque *mot* $\in$ *PhrasesTitre* **faire**
si *mot est alphanumérique* ***et*** *mot* $\notin M$ **alors**
ajouter Lemmatization(*mot*) à *TitreMotsClés* ;
fin
fin
joindre les éléments de la liste *TitreMotsClés* par un espace et les ajouter à *TitresPretraités* ;
PhrasesAbstract ← *abstract* décomposé par des points, les éléments sont transformés en minuscule ;
AbstractPhrasesClés ← [ ] ;
pour chaque *phrase* $\in$ *PhrasesAbstract* **faire**
si *phrase a plus de 6 mots* ***et non*** *phrase contient un nombre suivi de mots hommes, femmes, enfants, mères, patients, participants, etc.* ***et non*** *phrase contient les mots 'select', 'enroll', 'visit' ou 'of those'* **alors**
ajouter *phrase* à *AbstractPhrasesClés* ;
fin
fin
MotsClésParPhraseClé ← [ ] ;
AbstractPretraité ← [ ] ;
AbstractMotsClés ← [ ] ;
pour chaque *phrase* $\in$ *AbstractPhrasesClés* **faire**
pour chaque *mot* $\in$ *phrase* **faire**
si *mot est alphanumérique* ***et*** *mot* $\notin M$ **alors**
ajouter Lemmatization(*mot*) à *MotsClésParPhraseClé* ;
fin
fin
ajouter *MotsClésParPhraseClé* à *AbstractMotsClés* ;
joindre les éléments de la liste *MotsClésParPhraseClé* par un espace et les ajouter à *AbstractPretraité* ;
fin
ajouter *AbstractPretraité* à *AbstractsPretraités* ;
joindre les éléments de la liste *AbstractMotsClés* par un espace et les ajouter à *PhrasesPretraités* ;
fin
insérer de nouveaux attributs *TitresPretraités*, *PhrasesPretraités* et *AbstractsPretraités* dans la liste de documents D ;
retourner D' ;

fin

7.3 Résumés générés en utilisant les combinaisons de technologies

Temps pour faire le preprocessing: **14 minutes 55 secondes**

1. **Cluster avec les titres, réduction linéaire avec SVD, clustering avec k-Means, pondération des mots avec la moyenne**
   The cystine-cystine chemokine receptor 5 (CCR5) is the primary HIV co-receptor involved in the viral entry process into human cells. Pre-exposure prophylaxis (PrEP) reduces HIV acquisition risk by >90% and is a critical lever to reduce HIV incidence. The observation that HIV-1 subtype D progresses faster to disease than subtype A prompted us to examine cytokine levels early after infection within the predominant viral subtypes that circulate in Uganda and address the following research questions: (1) Do cytokine levels vary between subtypes A1 and D? Human immunodeficiency virus (HIV) and hepatitis B virus (HBV) are endemic in South Africa while hepatitis C virus (HCV) infection is rare. To study the structure of human immunodeficiency virus (HIV)-1 drug resistance (DR) in patients with newly diagnosed infection. Define the clinical presentation of acute human immunodeficiency virus infection (AHI) among men and women from 2 continents to create a clinical scoring algorithm.Comparison of incident sign and symptom between those with and without AHI.At-risk human immunodeficiency virus (HIV) negative men and women in Thailand, Kenya, Tanzania, and Uganda underwent twice-weekly testing for HIV. Blood donations in South Africa are tested for HIV RNA using individual donation NAT (ID-NAT), allowing detection and rapid antiretroviral therapy (ART) of acute HIV infections. Broadly neutralizing antibodies (bNAbs) for HIV-1 prevention or cure strategies must inhibit transmitted/founder and reservoir viruses.
   Execution time: 1.13 s.

2. **Cluster avec les titres, réduction linéaire avec SVD, clustering avec k-Means, pondération des mots avec SVD**
   There was a trend towards lower levels of CD161++CD8+ T cells in HIV-negative individuals with active and latent TB. The Defer/Test strategy averted the most HIV infections. Sensitivity, specificity, predictive values, and agreements were compared among the EIA kits using PCR-confirmed HIV-positive and negative samples. Establishment of persistent human immunodeficiency virus type 1 (HIV-1) reservoirs occurs early in infection, and biomarkers of infected CD4+ T cells during acute infection are poorly defined. Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised. Wealth-related inequality in early uptake of HIV testing was measured using the Erreygers concentration index (CI) further adjusted for inequality introduced by predicted healthcare need (ie, need-standardised). From July 2005 to June 2006, women were offered HIV testing following group information and education on HIV and STDs in the clinic waiting area. Acute HIV infection lasts approximately 3 weeks and early HIV infection, which includes acute HIV infection, lasts approximately 7 weeks.
   Execution time: 1.30 s.

3. **Cluster avec les titres, réduction linéaire avec SVD, clustering avec HDBSCAN, pondération des mots avec la moyenne**
   The observation that HIV-1 subtype D progresses faster to disease than subtype A prompted us to examine cytokine levels early after infection within the predominant viral subtypes that circulate in Uganda and address the following research questions: (1) Do cytokine levels vary between subtypes A1 and D? Broadly neutralizing antibodies (bNAbs) for HIV-1 prevention or cure strategies must inhibit transmitted/founder and reservoir viruses. The cystine-cystine chemokine receptor 5 (CCR5) is the primary HIV coreceptor involved in the viral entry process into human cells. Detection of acute and prevalent HIV infection using point-of-care nucleic acid amplification testing (POC-NAAT) among outpatients with symptoms compatible with acute HIV is critical to HIV prevention, but it is not clear if it is cost-effective compared with existing HIV testing strategies. HIV and AIDS continue to be major public health concerns globally. Post-partum loss to follow-up and lack of early HIV infant diagnosis (EID) can significantly affect the efficiency of programs for the prevention of mother-to-child transmission. We describe tuberculosis (TB) disease among antiretroviral treatment (ART) eligible children living with HIV (CLHIV) in South Africa to highlight TB prevention opportunities.
   Execution time: 0.58 s.

4. **Cluster avec les titres, réduction linéaire avec SVD, clustering avec HDBSCAN, pondération des mots avec SVD**
   Here, we assessed CD8+ T cell functional evolution from primary to chronic HIV infection. Sensitivity, specificity, predictive values, and agreements were compared among the EIA kits using PCR-confirmed HIV-positive and negative samples. The Defer/Test strategy averted the most HIV infections. The overall HIV prevalence was 37.3% [95% confidence interval (CI) 34.3–41.3]. Because window-period donations are the most important source of residual HIV contamination and arise from incident infections, research to develop risk factor exclusion strategies must focus on predictors of HIV seroconversion. Failing and non-failing patients had comparable median time [interquartile] on ART (69.5 [23.0-89.5] vs 64.0 [34.0-99.0] months; At decentralized urban settings, there is need for enhanced virological monitoring and adherence support. Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised.
   Execution time: 0.74 s.

5. **Cluster avec les titres, réduction non-linéaire avec UMAP, clustering avec k-Means, pondération des mots avec la moyenne**
   Detection of acute and prevalent HIV infection using point-of-care nucleic acid amplification testing (POC-NAAT) among outpatients with symptoms compatible with acute HIV is critical to HIV prevention, but it is not clear if it is cost-effective compared with existing HIV testing strategies. The observation that HIV-1 subtype D progresses faster to disease than subtype A prompted us to examine cytokine levels early after infection within the predominant viral subtypes that circulate in Uganda and address the following research questions: (1) Do cytokine levels vary between subtypes A1 and D? In South Africa, women continue to face a high burden of Hu-

man Immunodeficiency Virus (HIV) infection and the possible complications thereof during pregnancy. Broadly neutralizing antibodies (bNAbs) for HIV-1 prevention or cure strategies must inhibit transmitted/founder and reservoir viruses. Multi-assay algorithms (MAAs) are used to estimate population-level HIV incidence and identify individuals with recent infection. Pre-exposure prophylaxis (PrEP) reduces HIV acquisition risk by >90% and is a critical lever to reduce HIV incidence. Information on treatment failure (TF) in People living with HIV in a data-poor setting is necessary to counter the epidemic of TF with first-line combined antiretroviral therapies (cART) in sub-Saharan Africa (SSA). Blood donations in South Africa are tested for HIV RNA using individual donation NAT (ID-NAT), allowing detection and rapid antiretroviral therapy (ART) of acute HIV infections.
Execution time: 9.49 s.

6. **Cluster avec les titres, réduction non-linéaire avec UMAP, clustering avec k-Means, pondération des mots avec SVD**
These cells were dramatically increased in chronic HIV infection. Fourth-generation HIV assays detect both antigen and antibody, facilitating detection of acute/early HIV infection. HIV is a major contributor to infant mortality. The Defer/Test strategy averted the most HIV infections. Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised. From July 2005 to June 2006, women were offered HIV testing following group information and education on HIV and STDs in the clinic waiting area. To estimate the effect of ART on TB incidence while accounting for time-dependent confounders affected by exposure, a Cox proportional hazards marginal structural model was used. Despite high neutralizing potential and cross-reactivity, anti-V3 antibodies do not contribute to autologous neutralization.
Execution time: 5.09 s.

7. **Cluster avec les titres, réduction non-linéaire avec UMAP, clustering avec HDBSCAN, pondération des mots avec la moyenne**
To study the structure of human immunodeficiency virus (HIV)-1 drug resistance (DR) in patients with newly diagnosed infection. Blood donations in South Africa are tested for HIV RNA using individual donation NAT (ID-NAT), allowing detection and rapid antiretroviral therapy (ART) of acute HIV infections. Broadly neutralizing antibodies (bNAbs) for HIV-1 prevention or cure strategies must inhibit transmitted/founder and reservoir viruses. Multi-assay algorithms (MAAs) are used to estimate population-level HIV incidence and identify individuals with recent infection. Information on treatment failure (TF) in People living with HIV in a data-poor setting is necessary to counter the epidemic of TF with first-line combined antiretroviral therapies (cART) in sub-Saharan Africa (SSA). Declines in HIV incidence have been slower than expected during the roll-out of antiretroviral treatment (ART) services in sub-Saharan African populations suffering generalized epidemics. Pre-exposure prophylaxis (PrEP) reduces HIV acquisition risk by >90% and is a critical lever to reduce HIV incidence.
Execution time: 4.21 s.

8. **Cluster avec les titres, réduction non-linéaire avec UMAP, clustering avec HDBSCAN, pondération des mots avec SVD**
   Acute and chronic HIV were associated with lower frequencies of CD161++CD8+ T cells, which did not correlate with CD4 count or HIV viral load. By age 1 year, an estimated 35.2% infected and 4.9% uninfected children will have died; by 2 years of age, 52.5% and 7.6% will have died, respectively. Fourth-generation HIV assays detect both antigen and antibody, facilitating detection of acute/early HIV infection. Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised. The HPTN 071 (PopART) trial evaluated the impact of an HIV combination prevention package that included "universal testing and treatment" on HIV incidence in 21 communities in Zambia and South Africa during 2013-2018. Individuals with acute (preseroconversion) HIV infection (AHI) are important in the spread of HIV. We sought to study the survival of newborn children according to HIV status of the mother, that of the child and the timing of infection. Mobile HIV screening may facilitate early HIV diagnosis.
   Execution time: 4.31 s.

9. **Cluster avec les abstracts, réduction linéaire avec SVD, clustering avec k-Means, pondération des mots avec la moyenne**
   Multi-assay algorithms (MAAs) are used to estimate population-level HIV incidence and identify individuals with recent infection. Although certain individuals with HIV infection can stop antiretroviral therapy (ART) without viral load rebound, the mechanisms under-pinning 'post-treatment control' remain unclear. Detection of acute and prevalent HIV infection using point-of-care nucleic acid amplification testing (POC-NAAT) among outpatients with symptoms compatible with acute HIV is critical to HIV prevention, but it is not clear if it is cost-effective compared with existing HIV testing strategies. The observation that HIV-1 subtype D progresses faster to disease than subtype A prompted us to examine cytokine levels early after infection within the predominant viral subtypes that circulate in Uganda and address the following research questions: (1) Do cytokine levels vary between subtypes A1 and D? Pre-exposure prophylaxis (PrEP) reduces HIV acquisition risk by >90% and is a critical lever to reduce HIV incidence. To study the structure of human immunodeficiency virus (HIV)-1 drug resistance (DR) in patients with newly diagnosed infection. Broadly neutralizing antibodies (bNAbs) for HIV-1 prevention or cure strategies must inhibit transmitted/founder and reservoir viruses. Blood donations in South Africa are tested for HIV RNA using individual donation NAT (ID-NAT), allowing detection and rapid antiretroviral therapy (ART) of acute HIV infections.
   Execution time: 1.17 s.

10. **Cluster avec les abstracts, réduction linéaire avec SVD, clustering avec k-Means, pondération des mots avec SVD**
    Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised. The Defer/Test strategy averted the most HIV infections. Matched preinfection and postinfection samples were available from 13 individuals. Reactivity with all HIV genotypes was 100%. Sensitivity,

specificity, predictive values, and agreements were compared among the EIA kits using PCR-confirmed HIV-positive and negative samples. HIV transmission risk is higher during acute and early HIV infection than it is during chronic infection, but the contribution of early infection to the spread of HIV is controversial. These cells were dramatically increased in chronic HIV infection. Data from a prospective cohort study conducted during 1989-1990 of HIV serology and from a retrospective review of laboratory records of 727 patients presenting for superficial lymph node biopsy at the University Teaching Hospital in Lusaka, Zambia, were analyzed to determine the relative significance of HIV-associated lymphadenopathy among patients undergoing lymph node biopsy.
Execution time: 1.40 s.

11. **Cluster avec les abstracts, réduction linéaire avec SVD, clustering avec HDBSCAN, pondération des mots avec la moyenne**
The inflammasome pathway is an important arm of the innate immune system that provides antiviral immunity against many viruses. The cystine-cystine chemokine receptor 5 (CCR5) is the primary HIV co-receptor involved in the viral entry process into human cells. Multi-assay algorithms (MAAs) are used to estimate population-level HIV incidence and identify individuals with recent infection. Rapid initiation of antiretroviral therapy (ART) in early HIV infection is important to limit seeding of the viral reservoir. Human immunodeficiency virus (HIV)-1 genetic diversity increases during infection and can help infer the time elapsed since infection. Pre-exposure prophylaxis (PrEP) reduces HIV acquisition risk by >90% and is a critical lever to reduce HIV incidence.
Execution time: 0.91 s.

12. **Cluster avec les abstracts, réduction linéaire avec SVD, clustering avec HDBSCAN, pondération des mots avec SVD**
Sustained viremia after acute HIV infection is associated with profound CD4 These cells were dramatically increased in chronic HIV infection. Most point-of-care HIV assays have poor sensitivity to diagnose acute HIV infection as they only detect antibodies against HIV-1 and HIV-2 (HIV-1/2). The Defer/Test strategy averted the most HIV infections. Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised.
Execution time: 1.10 s.

13. **Cluster avec les abstracts, réduction non-linéaire avec UMAP, clustering avec k-Means, pondération des mots avec la moyenne**
Pre-exposure prophylaxis (PrEP) reduces HIV acquisition risk by >90% and is a critical lever to reduce HIV incidence. To study the structure of human immunodeficiency virus (HIV)-1 drug resistance (DR) in patients with newly diagnosed infection. Multi-assay algorithms (MAAs) are used to estimate population-level HIV incidence and identify individuals with recent infection. Detection of acute and prevalent HIV infection using point-of-care nucleic acid amplification testing (POC-NAAT) among outpatients with symptoms compatible with acute HIV is critical to HIV prevention, but it is not clear if it is cost-effective compared with existing HIV testing strategies. The cystine-cystine chemokine receptor 5 (CCR5) is the primary HIV

co-receptor involved in the viral entry process into human cells. Blood donations in South Africa are tested for HIV RNA using individual donation NAT (ID-NAT), allowing detection and rapid antiretroviral therapy (ART) of acute HIV infections. Broadly neutralizing antibodies (bNAbs) for HIV-1 prevention or cure strategies must inhibit transmitted/founder and reservoir viruses. Post-partum loss to follow-up and lack of early HIV infant diagnosis (EID) can significantly affect the efficiency of programs for the prevention of mother-to-child transmission.
Execution time: 4.43 s.

14. **Cluster avec les abstracts, réduction non-linéaire avec UMAP, clustering avec k-Means, pondération des mots avec SVD**
The proportions of subtype A, D, or recombinants showed no significant increasing or decreasing trend over this time period (p=0.51). Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised. These cells were dramatically increased in chronic HIV infection. To estimate the effect of ART on TB incidence while accounting for time-dependent confounders affected by exposure, a Cox proportional hazards marginal structural model was used. Fourth-generation HIV assays detect both antigen and antibody, facilitating detection of acute/early HIV infection. To determine the availability of the gp120 V3 loop for neutralizing antibody binding on SHIV-89.6 and KB9 virions, we have constructed immunogenic C4-V3 peptides from these SHIVs and induced anti-V3 antibodies in guinea pigs and rhesus monkeys. From July 2005 to June 2006, women were offered HIV testing following group information and education on HIV and STDs in the clinic waiting area. HIV transmission risk is higher during acute and early HIV infection than it is during chronic infection, but the contribution of early infection to the spread of HIV is controversial.
Execution time: 5.01 s.

15. **Cluster avec les abstracts, réduction non-linéaire avec UMAP, clustering avec HDBSCAN, pondération des mots avec la moyenne**
Multi-assay algorithms (MAAs) are used to estimate population-level HIV incidence and identify individuals with recent infection. Persons with acute HIV infection (AHI) are highly infectious and responsible for a disproportionate share of incident infections. HIV/Mycobacterium tuberculosis (Mtb) co-infected individuals have an increased risk of tuberculosis prior to loss of peripheral CD4 T cells, raising the possibility that HIV co-infection leads to CD4 T cell depletion in lung tissue before it is evident in blood. Pre-exposure prophylaxis (PrEP) reduces HIV acquisition risk by >90% and is a critical lever to reduce HIV incidence. Post-partum loss to follow-up and lack of early HIV infant diagnosis (EID) can significantly affect the efficiency of programs for the prevention of mother-to-child transmission. The inflammasome pathway is an important arm of the innate immune system that provides antiviral immunity against many viruses. Accessing family planning is a key investment in reducing the broader costs of health care and can reduce a significant proportion of maternal, infant, and childhood deaths.
Execution time: 4.40 s.

16. **Cluster avec les abstracts, réduction non-linéaire avec UMAP, clustering avec HDBSCAN, pondération des mots avec SVD**
Fourth-generation HIV assays detect both antigen and antibody, facilitating detection of acute/early HIV infection. Individuals with acute (preseroconversion) HIV infection (AHI) are important in the spread of HIV. These cells were dramatically increased in chronic HIV infection. Following an information session on HIVSS, interested participants were offered one of three methods of HIVSS testing: supervised, semi-supervised, and unsupervised. Data were collected by using pretested and structured extraction tool. Interrupting this mode of passage would provide protection for children.
Execution time: 5.07 s.

7.4 Résumé à partir de clusters fournis par l'équipe LiteRev

Immediate cART initiation significantly reduces risk of cancer. The dominant ileum B cell response was to Env gp41. The test group will receive training. The primary outcomes were HIV transmission at 1 week of age in the infant and maternal and infant safety. HIV contributes substantially to child mortality, but factors underlying these deaths are inadequately described.

Les 20 mots-clés les plus importants du Cluster n°1

woman	**0.054449**	disease	**0.024078**
patient	**0.049414**	prevalence	**0.023820**
risk	**0.035948**	month	**0.023242**
treatment	**0.032350**	diagnosis	**0.023158**
man	**0.030817**	sexual	**0.022565**
incidence	**0.030449**	acute	**0.022406**
year	**0.029035**	adult	**0.022080**
testing	**0.026083**	positive	**0.021920**
care	**0.025017**	test	**0.021780**
associate	**0.024771**	health	**0.021750**

Les 10 meilleurs phrases sélectionnés du Cluster n°1

Sentence	Weight
Immediate cART initiation significantly reduces risk of cancer.	**0.009412**
Early HIV testing is critical to prevention and timely treatment.	**0.007812**
To determine the incidence of HIV during pregnancy as defined by seroconversion using a repeat HIV rapid testing strategy during late pregnancy.	**0.006935**
The prevalence of late ART initiation was high.	**0.006802**
Early HIV diagnosis, enrollment on antiretroviral treatment, and isoniazid prophylaxis treatment should be considered to decrease the TB risk.	**0.006790**
A delay presentation for human immunodeficiency virus (HIV) patient's care (that is late engagement to HIV care due to delayed HIV testing or delayed linkage for HIV care after the diagnosis of HIV positive) is a critical step in the series of HIV patient care continuum.	**0.006688**
Genital ulcer disease (GUD) is a major risk factor for human immunodeficiency virus (HIV) transmission.	**0.006657**
Older age and baseline CD8 cell count were independent predictors of infection-unrelated cancer.	**0.006581**
These results highlight the substantial risk of transmission during acute HIV infection.	**0.006580**

Les 20 mots-clés les plus importants du Cluster n°2

cell	**0.085219**	antibody	**0.029279**
subtype	**0.058979**	resistance	**0.028890**
response	**0.053972**	plasma	**0.028825**
virus	**0.040775**	acute	**0.028537**
viral	**0.040014**	neutralize	**0.028087**
isolate	**0.039945**	level	**0.027754**
individual	**0.033167**	env	**0.027502**
sequence	**0.033069**	immune	**0.027150**
primary	**0.030091**	gag	**0.026829**
specific	**0.030033**	associate	**0.026642**

Les 10 meilleurs phrases sélectionnés du Cluster n°2

Sentence	Weight
The dominant ileum B cell response was to Env gp41.	**0.012922**
SIVmac grew equally well in both cell lines.	**0.011398**
Regulatory T cells (Tregs) have the potential to control systemic immune activation but also to suppress antigen specific T and B cell response.	**0.010153**
Primary HIV-1 drug resistance was low.	**0.009562**
HIV viral loads and peripheral blood CD4+ T cell counts were measured in all subjects.	**0.009487**
Here, we assessed CD8+ T cell functional evolution from primary to chronic HIV infection.	**0.009176**
ART was not associated with an increase in CD161++CD8+ T cell frequency.	**0.009165**
Higher set point viral load, lower early CD4+ cell count, and more-symptomatic acute HIV-1 illness each predicted death.	**0.008913**
The superinfected individual mounted a neutralizing antibody response to the primary TF virus, which remained TF-specific over time and even after superinfection, did not neutralize the superinfecting variant.	**0.008774**
Additionally, HIV-specific cytolytic CD4+ T cell responses in acute HIV infection are predictive of disease progression.	**0.008612**

Les 20 mots-clés les plus importants du Cluster n°3

test	**0.075072**	care	**0.033174**
ahi	**0.066854**	positive	**0.032452**
testing	**0.060238**	antibody	**0.031870**
acute	**0.039772**	risk	**0.031413**
assay	**0.039621**	patient	**0.030348**
sample	**0.039450**	incidence	**0.029061**
blood	**0.038846**	diagnosis	**0.028825**
rapid	**0.037243**	detect	**0.028673**
donor	**0.034822**	cost	**0.028433**
participant	**0.033388**	estimate	**0.028138**

Les 10 meilleurs phrases sélectionnés du Cluster n°3

Sentence	Weight
The test group will receive training.	**0.015217**
Mobilized participants received clinic-based rapid antibody testing and point-of-care HIV RNA testing.	**0.014584**
Available antibody testing cannot detect an acute HIV infection, but repeat testing after 2-4 weeks may detect seroconversion.	**0.013351**
Combo test results were reported as antigen positive, antibody positive, or both.	**0.013249**
Whole blood was used for Plasmodium falciparum rapid test determination at screening visit.	**0.013200**
The p24 ELISA antigen test remained positive at 5 pg/mL.	**0.012151**
Rapid testing was conducted with parallel testing in the clinic and serial testing in the center.	**0.012109**
Future directions for HIV testing include rapid testing technology and detection of acute HIV infection, self-testing expansion, and partner notification.	**0.011754**
Assays that detect p24 antigen reduce the diagnostic window period of HIV testing.	**0.011280**
Concordance of cobas HIV-1/2 Qual test with the comparator serological test and COBAS AmpliPrep/COBAS TaqMan test was $\geq$ 99.6% with all sample types.	**0.011270**

Les 20 mots-clés les plus importants du Cluster n°4

infant	**0.321222**	pmtct	**0.049729**
mother	**0.150459**	mtct	**0.049637**
week	**0.112531**	woman	**0.048541**
child	**0.082790**	birth	**0.045351**
transmission	**0.074193**	rate	**0.044278**
expose	**0.063175**	mortality	**0.042816**
month	**0.061769**	diagnosis	**0.039777**
maternal	**0.056052**	prophylaxis	**0.038537**
age	**0.052444**	test	**0.038504**
pcr	**0.050343**	receive	**0.038387**

Les 10 meilleurs phrases sélectionnés du Cluster n°4

Sentence	Weight
The primary outcomes were HIV transmission at 1 week of age in the infant and maternal and infant safety.	**0.048036**
Earlier diagnosis is necessary to reduce infant mortality.	**0.046573**
We measured MTCT prevalence at 4-12 weeks post-delivery and evaluated associations between infant HIV infection and maternal and infant characteristics including maternal treatment and infant prophylaxis.	**0.042277**
HIV is a major contributor to infant mortality.	**0.040948**
Early initiation of antiretroviral therapy reduces HIV-related infant mortality.	**0.040424**
Early initiation of antiretroviral therapy depends on an early infant diagnosis and is critical to reduce HIV-related infant mortality.	**0.039532**
Early infant diagnosis using HIV-RNA/PCR or HIV-DNA/PCR >6 weeks.	**0.032818**
A significant gap remains between the uptake of infant and maternal antiretroviral regimens and only a minority of HIV-exposed infants receives prophylaxis and safe infant feeding.	**0.029961**
Early HIV-1 diagnosis with antiretroviral therapy before symptomatic disease is critical for infant survival.	**0.029139**
Antiretroviral therapy is often initiated too late to impact early HIV-related infant mortality.	**0.029108**

Les 20 mots-clés les plus importants du Cluster n°5

child	**0.249055**	month	**0.046241**
mortality	**0.129734**	diagnosis	**0.043173**
year	**0.097707**	associate	**0.042511**
age	**0.075610**	tuberculosis	**0.036155**
patient	**0.064384**	die	**0.035661**
infect	**0.058968**	person	**0.034310**
death	**0.058359**	lftu	**0.033884**
treatment	**0.058161**	initiation	**0.033880**
clinical	**0.053035**	care	**0.032937**
stage	**0.047929**	predictor	**0.031883**

Les 10 meilleurs phrases sélectionnés du Cluster n°5

Sentence	Weight
HIV contributes substantially to child mortality, but factors underlying these deaths are inadequately described.	**0.023813**
All except one child were on antiretroviral treatment, 45% had commenced treatment $<$ 12 months of age.	**0.023143**
To determine the impact of HIV on child mortality and explore potential risk factors for mortality among HIV-infected and HIV-exposed uninfected children in a longitudinal cohort in rural Uganda.	**0.020463**
LTFU of HIV infected children was common with an incidence of 32.9 per 1000 child years and occurred early in treatment and risk factors included poverty, low caregiver education, male child and early HIV disease stage.	**0.019670**
Efforts should be intensified to prevent maternal to child transmission of HIV infection.	**0.019589**
We assessed overall mortality and stratified by year using random effects models.	**0.019496**
Increased mortality and attrition were also associated with advanced clinical stage, underweight and diagnosis of tuberculosis at programme entry.	**0.019482**
Intensified efforts to prevent mother-to-child transmission of HIV and ensure early HIV diagnosis and treatment are required to decrease child mortality caused by HIV in rural Africa.	**0.019027**
Early HIV testing and ART initiation is recommended to decrease mortality.	**0.017052**
Median age of antiretroviral treatment commencement was 3.9 years.	**0.016493**