Narration de données en santé publique: cas de la tuberculose au Gabon

Raymond Ondzigue Mbenga*,** Veronika Peralta* Edgard Brice Ngoungou** Sydney Maghendji Nzondo** Thomas Devogele*

*LIFAT, Université de Tours, 3 Place Jean Jaurès, 41000 Blois, France
raymond.ondziguembenga@etu.univ-tours.fr,veronika.peralta@univ-tours.fr,
thomas.devogele@univ-tours.fr
**DEBIM-UREMCSE, Université des Sciences de la Santé, BP:18 231, Gabon
sydneymaghendji@yahoo.fr, ngoungou2001@yahoo.fr

Résumé. La narration de données est l'activité consistant à élaborer des récits étayés par des faits extraits de l'analyse des données, à l'aide de visualisations interactives. Elle facilite la communication des résultats d'analyses par des moyens visuels à un public cible. Malgré son utilité reconnue en santé publique, la narration de données est généralement limitée à la transmission de recommandations de traitement pour éduquer le grand public. Dans cet article, nous présentons un retour d'expérience sur l'élaboration d'une narration de données sur la tuberculose dans un pays d'Afrique subsaharienne, le Gabon.

1 Introduction

Une narration de données est définie comme une composition structurée de messages qui transmettent des trouvailles sur les données et sont généralement communiqués par des moyens visuels afin de faciliter leur réception par un public ciblé (Outa et al., 2020). En outre, elle peut être considérée comme une séquence ordonnée d'étapes, chacune pouvant contenir des mots, des graphiques, des cartes, des sons, des vidéos ou toute combinaison de ces éléments, et reposant sur des données (Kosara et Mackinlay, 2013).

Malgré leur utilité reconnue en santé publique (Bouman, 2017), où les informations sur les problèmes de santé publique doivent concurrencer avec des milliers d'autres messages de communication (infodémies), les narrations de données sont généralement limités à la transmission de recommandations de traitement et de retours de patients pour éduquer le grand public. En effet, peu de travaux visent à transmettre des résultats scientifiques à un public d'experts pour rendre compte de la situation sanitaire et des résultats des politiques mises en œuvre, et plus généralement, pour aider à la prise de décision. Aussi, ces narrations de données sont élaborées selon des processus ad hoc. D'une manière générale, il y a un manque de processus et de lignes directrices méthodologiques pour la narration de données en santé publique.

Dans cet article, nous présentons un retour d'expérience sur l'élaboration d'une narration de données sur la tuberculose [1] (TB) au Gabon. Plus précisément, nous décrivons les phases du processus, les aspects méthodologiques clés, la prise en compte des particularités du domaine épidémiologique et les leçons apprises. Il s'agit d'une version résumée d'un article qui a été accepté à la conférence DARLI-AP 2022 (Raymond Ondzigue et al., 2022).

2 Etat de l'art

Dans la littérature, de nombreux auteurs ont proposé des processus de narration de données classiques et en intelligence épidémique. Pour ce qui est du processus de narration de données classique, (Chen et al., 2018) proposent un processus en trois phases (i) l'analyse visuelle qui nécessite de voir tous les aspects des données complexes, d'explorer leurs interrelations, (ii) la synthèse des données, au cours de laquelle l'analyste rassemble et organise les éléments d'information à communiquer, facilitant ainsi la présentation des résultats de l'analyse visuelle dans un récit convaincant et (iii) le storytelling, qui vise à transmettre uniquement les trouvailles extraites de l'analyse des données, présentées de manière simple et facilement compréhensible. Plus récemment, dans (Outa et al., 2020), les auteurs ont proposé un processus en quatre phases (i) définition des objectifs, (ii) exploration des données, (iii) organisation du récit, et (iv) présentation.

En intelligence épidémique, les processus proposés par divers auteurs (Che et Desenclos, 2002; Noah, 2006; Kaiser et al., 2006; Astagneau et Ancelle, 2011; Thacker et al., 2012; Eilstein et al., 2012) s'appuient sur le processus recommandé par l'Organisation Mondiale de la Santé (OMS, 2014) pour la surveillance des maladies. Il s'agit d'un processus en 5 phases (i) détection, qui consiste à sélectionner les sources de données et à collecter les données, (ii) triage, qui est décomposée en deux sous-phases l'analyse des données (qualité des données et épidémiologie descriptive et analytique) et l'interprétation des données (évaluation qualitative de la signification des résultats), (iii) vérification, qui consiste à confirmer l'authenticité et la conformité des résultats et de leurs caractéristiques, généralement par recoupement à l'aide d'autres sources fiables, (iv) évaluation du risque épidémique, qui implique de déterminer le niveau de risque pour la santé humaine et les mesures de contrôle potentielles qui peuvent être mises en œuvre, et (v) communication, qui concerne la communication des indicateurs à différents publics. Dans ce processus, en dehors des phases vérification et évaluation du risque épidémique, les phases détection, triage et communication correspondent respectivement aux phases d'exploration des données et présentation d'un processus de narration de données classique.

3 Production de la narration

Le processus de production d'une narration de données se trouve à la croisée de plusieurs domaines, notamment l'extraction et traitement des données, l'analyse des données, la visualisation, la communication, entre autres. Malgré de nombreuses contributions dans chacun de

1. La tuberculose est une maladie infectieuse causée par une mycobactérie, Mycobacterium tuberculosis, qui atteint le plus souvent les poumons (tuberculose pulmonaire) mais qui peut atteindre d'autres organes (tuberculose extra-pulmonaire). En 2019, elle a infecté plus de 10 millions de personnes dans le monde.

ces domaines, peu de travaux proposent des méthodologies globales, décrivant l'ensemble du processus de narration, mais ils s'accordent sur les principales phases (Lee et al., 2015; Chen et al., 2018; Outa et al., 2020) : (i) définition des objectifs, (ii) exploration des données, (iii) organisation du récit et (iv) présentation. Nous nous sommes inspirés de ces travaux, et avons enrichi la phase exploration des données du processus avec des phases particulières (vérification et évaluation du risque épidémique) issues du processus d'Intelligence Epidémique (IE) de l'Organisation Mondiale de la Santé (OMS, 2014). Aux sous-sections ci-après, nous décrivons les principales phases de la production de la narration de données sur la TB.

3.1 Définition des objectifs et questions analytiques

A cette phase, nous avons défini l'objectif de la narration de données et l'avons décliné en questions analytiques. L'objectif était de décrire le profil épidémiologique la tuberculose dans la région sanitaire Libreville-Owendo-Akanda du Gabon. En affinant cet objectif et en menant des entretiens avec divers responsables de la lutte antituberculeuse, nous avons obtenu la liste questions analytiques suivantes :

— Q1 : Quelles sont les caractéristiques épidémiologiques de la tuberculose ? Nous cherchons à décrire le profil (fréquence, variations) de la tuberculose en fonction des caractéristiques des patients tuberculeux ;
— Q2 : Quelle est la répartition spatiale et temporelle des patients ? L'identification des zones les plus touchées est essentielle pour cibler les actions de riposte.

3.2 Exploration des données

Pour l'exploration des données, nous avons procédé en plusieurs étapes : collecte de données, prétraitement, analyse, interprétation, vérification, évaluation du risque épidémique et finalement formulation des messages. Nous avons collecté les données cliniques et sociodémographiques de 7968 patients tuberculeux qui ont été pris en charge à l'hôpital spécialisé Nkembo de Libreville entre 2016 et 2018. Ensuite, ces données ont été prétraitées (correction des incohérences, suppression des doublons, etc.) avant d'être stockées dans un entrepôt de données spatiales sous PostGIS. Les patients ont été décrits en fonction des dimensions temps (2016 à 2018), forme clinique (pulmonaire, extra-pulmonaire, multirésistante ou inconnue), âge (organisé en deux niveaux : âge et groupe d'âge), genre, statut professionnel, lieu de résidence (organisé en deux niveaux : quartier et arrondissement), statut VIH et résultat thérapeutique.

Nous avons ensuite réalisé des multiples requêtes et validations afin de répondre aux questions analytiques. Nous illustrons les étapes d'analyse, d'interprétation et de vérification à travers de l'une des études réalisées.

L'étude concerne la répartition géographique des patients. Nous avons réalisé des études univariées et bivariées. Les distributions spatiales et spatio-temporelles des cas par arrondissement et par quartier n'ont pas montré de corrélation, pourtant très courantes dans d'autres études d'épidémiologie spatiale. Afin d'aller plus loin, nous avons collecté des nouvelles données sur la typologie des quartiers et étudié la distribution, où les tendances sont plus marquées, plus de la moitié des cas provenant des quartiers précaires (p-value= 0,01). Une validation par rapport à l'état de l'art a montré l'accord avec autres études, par exemple (Engohan Alloghe et al., 2006).

Finalement, nous avons formulé des messages. Par exemple, l'étude géographique illustrée précédemment a conduit à deux messages : (i) il n'y a pas de corrélation spatiale ou spatio-temporelle des cas, (ii) l'évolution spatiale et temporelle est corrélée avec la typologie des quartiers (p-value= 0,01), les quartiers précaires et mixtes étant significativement plus touchés que les autres. La probabilité d'un risque d'épidémie de la tuberculose est donc beaucoup plus élevée dans ces quartiers.

Nous soulignons que pendant cette phase, plusieurs aller-retours ont été nécessaires afin de croiser nos résultats avec d'autres sources de données et de les comparer à l'état de l'art. Ces analyses ont aussi permis la formulation de nouvelles questions analytiques.

3.3 Structuration de la narration

A cette phase, les messages ont été assemblés et ordonnés dans une trame cohérente et logique afin de faciliter leur compréhension et d'attirer le public. Nous avons commencé par définir le public (experts en épidémiologie et en santé publique) et sélectionner les messages à transmettre à ce public. Ensuite, nous avons choisi la structure narrative et organisé les messages.

L'intrigue est organisée en 8 actes qui tiennent compte des messages clés à communiquer au public cible. Le premier acte introduit l'intention de l'étude, le deuxième présente les messages saillants et les 6 suivants se concentrent chacun sur une dimension décrivant les patients, respectivement, le statut professionnel, le genre, l'âge, le statut VIH et la géographie. Le dernier acte présente les conclusions et les recommandations. Cette structure facilite la compréhension du profil d'un patient, selon les différentes dimensions qui le composent. Il est prévu que le public regarde d'abord l'introduction et la présentation générale, mais qu'il puisse ensuite naviguer entre les actes suivants en fonction de ses besoins. Ce type d'interactivité est connue dans la narration moderne sous le nom de Martini-Glass (Segel et Heer, 2010).

3.4 Présentation

Cette phase concerne le choix de la représentation visuelle (par exemple, tableaux de bord interactifs, infographies, diaporama, vidéo) et la mise en place d'artefacts visuels (graphiques, couleurs, texte, etc.) pour raconter les messages en attirant l'attention du public. Nous avons conçu et mis en oeuvre deux versions de la narration des données avec un rendu visuel différent : (i) un récit interactif, composé de tableaux de bord interactifs interconnectés (un exemple de tableau de bord est donné en Figure 1) et (ii) une vidéo [2], capturant une navigation particulière à travers le récit interactif, avec des explications audio. Nous avons utilisé Tableau pour rendre la narration interactive et OBS Studio pour enregistrer la vidéo.

La figure 1 correspond au tableau de bord de l'acte II. Il présente (i) la distribution géographique du nombre des patients TB par quartier (carte), (ii) la répartition du nombre de patients par type de tuberculose (diagramme en secteurs), (iii) le nombre de patients selon le résultat thérapeutique (graphique en barres) et (iv) la distribution des patients selon le type de tuberculose et les résultats thérapeutiques (tableau croisé). En outre, ce tableau de bord donne accès à d'autres tableaux de bord dans lesquels d'autres visualisations (statut professionnel, genre, classe d'âge, statut VIH et géographiques) sont présentées.

2. Une narration de données sur la tuberculose au Gabon : `https://www.youtube.com/watch?v=u_KoBWc_qJU&t=209s`

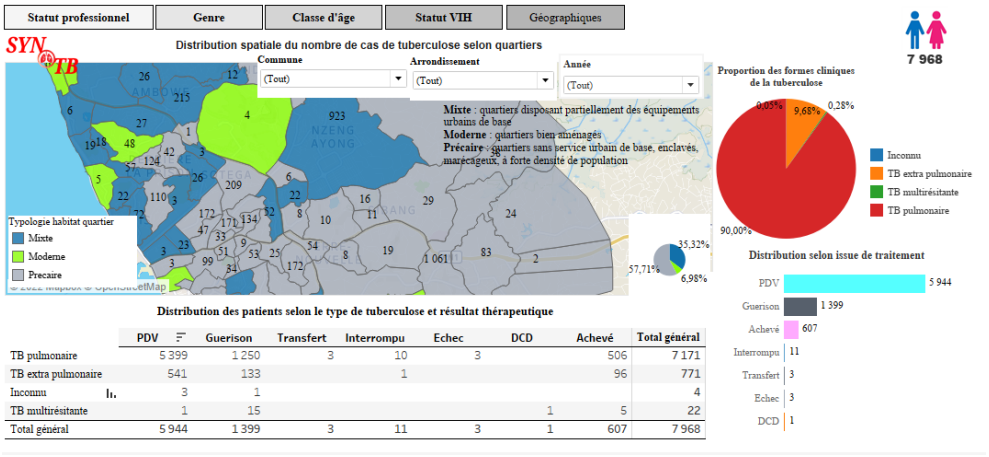

FIG. 1 – *Tableau de bord acte II*

4 Leçons apprises

Le processus d'élaboration est inspiré de modèles et de processus de l'état de l'art (Lee et al., 2015; Chen et al., 2018; Outa et al., 2020). Cependant, plusieurs particularités du contexte applicatif nous ont amenés à enrichir ce processus. Cette section présente les principaux enseignements tirés de cette adaptation.

L'analyse des données statistiques n'est pas suffisante pour la prise de décision en matière de santé publique. Une comparaison systématique avec l'état de l'art, par rapprochement des chiffres obtenus, est impérative pour discerner les phénomènes globaux des particularités régionales ou saisonnières. Ainsi, les décideurs peuvent juger quelles dimensions du profil des patients sont en accord avec la situation dans d'autres pays, pour lesquelles des actions communes peuvent être mises en place, et lesquelles concernent la population gabonaise. De même, les résultats obtenus doivent subir des tests approfondis afin de prouver leur valeur statistique. Le public cible étant majoritairement scientifique, ces résultats peuvent être communiqués dans le récit.

Contrairement aux récits saisonniers (fréquents dans le journalisme de données), dans les récits scientifiques, les questions analytiques ne sont pas toutes connues à l'avance. Au contraire, de nouvelles questions peuvent surgir au cours de l'analyse des données. Des itérations entre les phases de définition des objectifs et d'exploration des données sont souvent nécessaires. De nouvelles découvertes peuvent également avoir un impact sur les messages précédents et nécessiter une mise à jour.

La composante géographique est très importante pour évaluer l'étendue spatiale et spatio-temporelle des problèmes de santé. La restitution sous forme de cartes est à privilégier, mais aussi, les corrélations spatiales.

La narration des données doit permettre une navigation interactive entre les tableaux de bord. Il existe différents profils parmi les décideurs. D'une part, on retrouve des besoins variés en termes de dimensions et d'indicateurs étudiés, pour lesquels une organisation thématique

est parfaitement adaptée. D'autre part, les autorités sanitaires ont besoin d'une lecture plus complète et guidée du récit. Le défi consiste à trouver un bon équilibre pour le rendu, à la fois guidé et interactif.

5 Conclusion

Dans cet article, nous avons présenté un retour d'expérience sur l'élaboration d'une narration de données sur la tuberculose. A notre connaissance, c'est le premier travail qui décrit le processus complet de production, tout en tenant compte des particularités du domaine épidémiologique. Contrairement aux travaux qui visent le grand public, cette narration s'adresse aux experts en épidémiologie et en santé publique, qui sont des décideurs.

En plus des autorités gabonaises (responsables du programme national de lutte contre la tuberculose), la narration a été présentée à des experts de santé publique d'autres pays africains aux Journées Camerounaises d'Informatique Médicale (JCIM 2021), via une communication orale (Ondzigue Mbenga et al., 2021). Nous espérons que cette initiative servira à inspirer d'autres équipes pour reproduire l'expérience dans d'autres domaines de la santé, notamment pour faciliter la compréhension de la situation épidémiologique d'autres maladies infectieuses (choléra, ébola, etc.) qui sévissent encore en Afrique.

Références

Astagneau, P. et T. Ancelle (2011). *Surveillance épidémiologique : Principes, méthodes et applications en santé publique*. Lavoisier.

Bouman, M. (2017). Storytelling makes public health statistics more accessible. *European Journal of Public Health 27*(suppl_3).

Che, D. et J. Desenclos (2002). Detection systems for infectious diseases in france. *MEDECINE ET MALADIES INFECTIEUSES 32*(12), 704–716.

Chen, S., J. Li, G. Andrienko, N. Andrienko, Y. Wang, P. H. Nguyen, et C. Turkay (2018). Supporting story synthesis : Bridging the gap between visual analytics and storytelling. *IEEE transactions on visualization and computer graphics 26*(7), 2499–2516.

Eilstein, D., G. Salines, et J.-C. Desenclos (2012). Veille sanitaire : outils, fonctions, processus. *Revue d'épidémiologie et de Santé Publique 60*(5), 401–411.

Engohan Alloghe, E., M. Toung Mve, S. Ramarojoana, J. J. Iba Ba, et D. Nkoghe (2006). Epidemiologie de tuberculose infantile au centre antituberculeux de libreville de 1997–2001. *Med trop 66*, 469–471.

Kaiser, R., D. Coulombier, M. Baldari, D. Morgan, et C. Paquet (2006). What is epidemic intelligence, and how is it being improved in europe ? *Weekly releases (1997–2007) 11*(5), 2892.

Kosara, R. et J. Mackinlay (2013). Storytelling : The next step for visualization. *IEEE Computer*, 46.

Lee, B., N. H. Riche, P. Isenberg, et S. Carpendale (2015). More than telling a story : Transforming data into visually shared stories. *IEEE Computer Graphics and Applications 35*(5).

Noah, N. (2006). *Controlling communicable disease*. McGraw-Hill Education (UK).

OMS (2014). Early detection, assessment and response to acute public health events : implementation of early warning and response with a focus on event-based surveillance. Technical document.

Ondzigue Mbenga, R., V. Peralta, T. Devogele, S. Maghendji, et E. B. Ngoungou (2021). Processus de narration de données en intelligence épidémique avec application à la pandémie de tuberculose au gabon. In *8e Journées Camerounaises d'Informatique Médicale*.

Outa, F. E., M. Francia, P. Marcel, V. Peralta, et P. Vassiliadis (2020). Towards a conceptual model for data narratives. In *ER 2020, Vienna, Austria*.

Raymond Ondzigue, M., V. Peralta, T. Devogele, F. E. Outa, S. M. Nzondo, et E. B. Ngoungou (2022). A data narrative about tuberculosis pandemic in gabon. In M. Ramanath et T. Palpanas (Eds.), *Proceedings of the Workshops of the EDBT/ICDT 2022 Joint Conference, Edinburgh, UK, March 29, 2022*, Volume 3135 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Segel, E. et J. Heer (2010). Narrative visualization : Telling stories with data. *IEEE TVCG 16*(6).

Thacker, S. B., J. R. Qualters, L. M. Lee, C. for Disease Control, Prevention, et al. (2012). Public health surveillance in the united states : evolution and challenges. *MMWR Surveill Summ 61*(3), 3–9.

Summary

Data storytelling is the activity of developing narratives supported by facts extracted from data analysis, using interactive visualizations. It easily communicates the results of analyses through visual means to a target audience. Despite its usefulness in public health, data storytelling is generally limited to the delivery of treatment recommendations to educate the general public. In this article, we present feedback on the development of a tuberculosis data narrative in a sub-Saharan African country, Gabon.