quinten health

Générer des données de patients synthétiques grâce au machine learning: accélérer les études en vie réelle (RWE)

Margaux Törnqvist* , Louise Dry[1], Gaëtan Pinon[1], Antoine Movschin[1]

[1]Quinten Health – 8 rue Vernier, Paris – France

20 juin 2024

Une méthode de génération de données synthétiques de patients pour accélerer la recherche médicale

Questions de recherche

RWDs

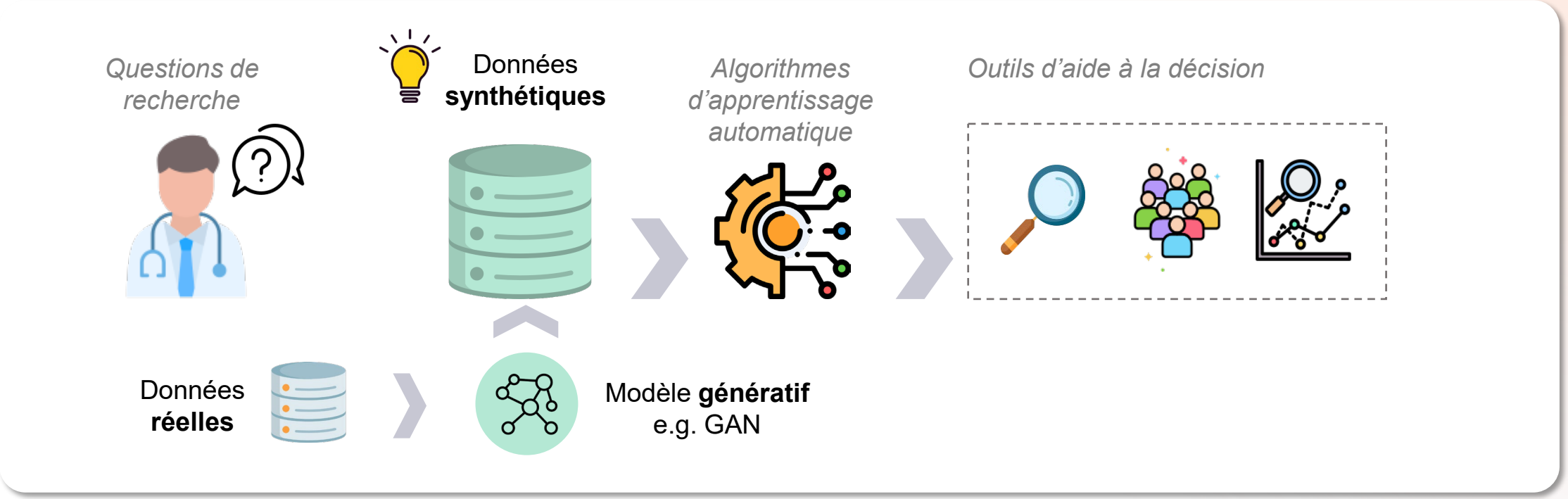
Algorithmes d'apprentissage automatique

Outils d'aide à la décision

Confidentialité
Coût
Délais
Accessibilité

RWD: Real World Data

Une méthode de génération de données synthétiques de patients pour accélerer la recherche médicale



OBJECTIF: Evaluer la capacité des GANs à générer des données tabulaires synthétiques

GAN: Réseau adversaire génératif

Méthode pour évaluer le machine learning dans la génération de données synthétiques de patients

CREATION DE COHORTE

- MIMIC-III[1]
- Soins intensifs
- U.S.
- 2001 - 2012
- 44k patients
- Démographie
  Diagnostiques(ICD-9-CM)
  Mesures de laboratoires

EXTRACTION DES DONNEES

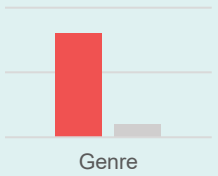
Sélection de variables:

- 45 ans
- Diabète
- Septicémie
- Hb 11 mg/L

MODELISATION

Entraînement de deux GANs:

CTGAN[2]:



CTABGAN[3]:

EVALUATION

Evaluation de la fidelité et confidentialité:

Scores

- ✓ 0.90
- ✓ 0.94

Patients réels

Patients synthétiques

quinten health

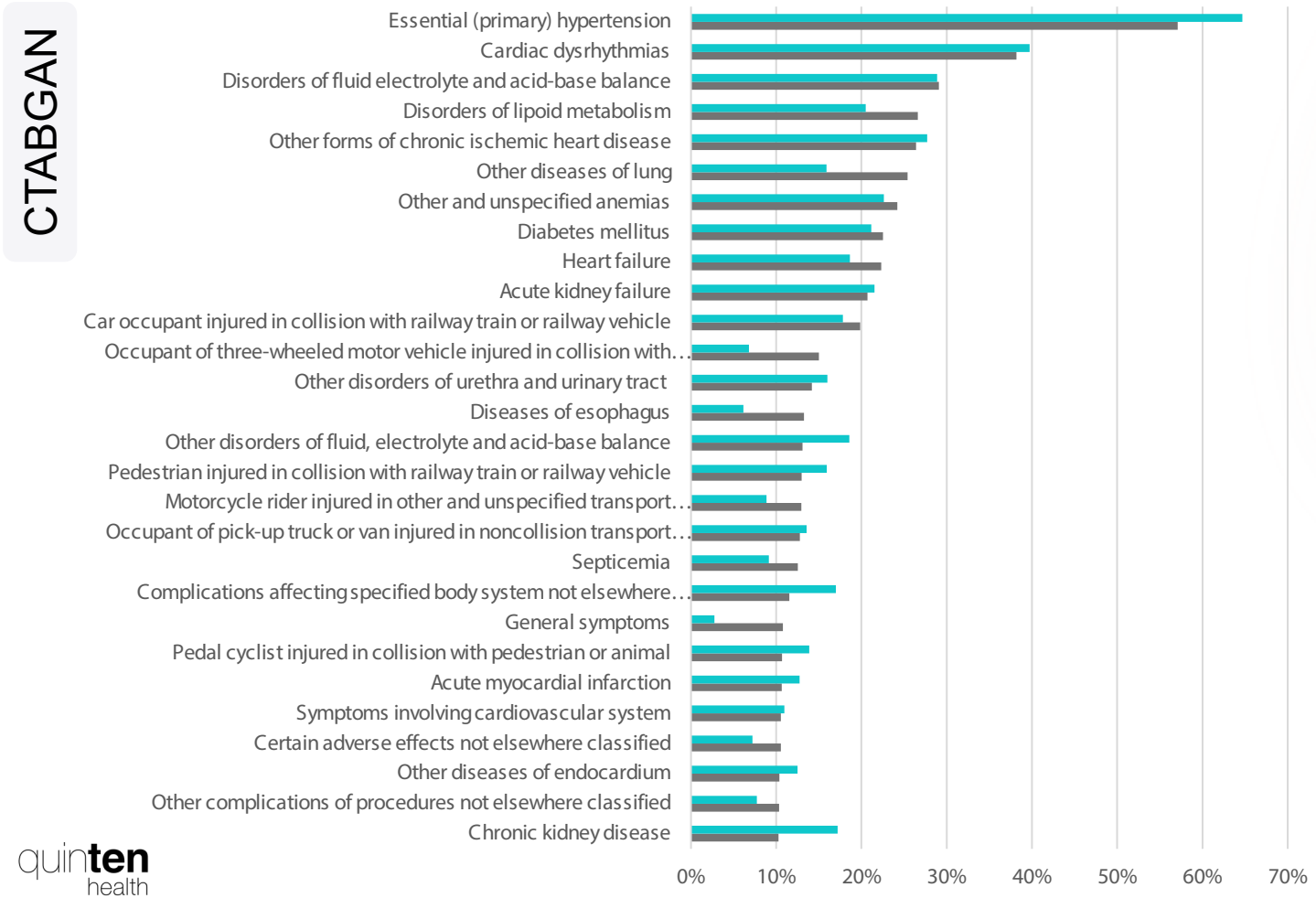
[1]Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).

[2]Xu et al., Modeling Tabular data using Conditional GAN, arXiv:1907.00503 (2019).

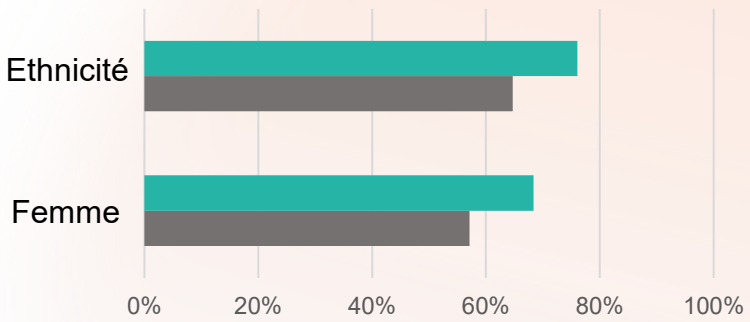
[3]Zhao et al.,CTAB-GAN: Effectice Table Data Synthetizing, arXiv:2102.08369 (2021).

Des prévalences de caractéristiques cliniques relativement fidèles aux données d'origine

Prévalence des diagnostics des données réelles vs. synthétique



Démographie

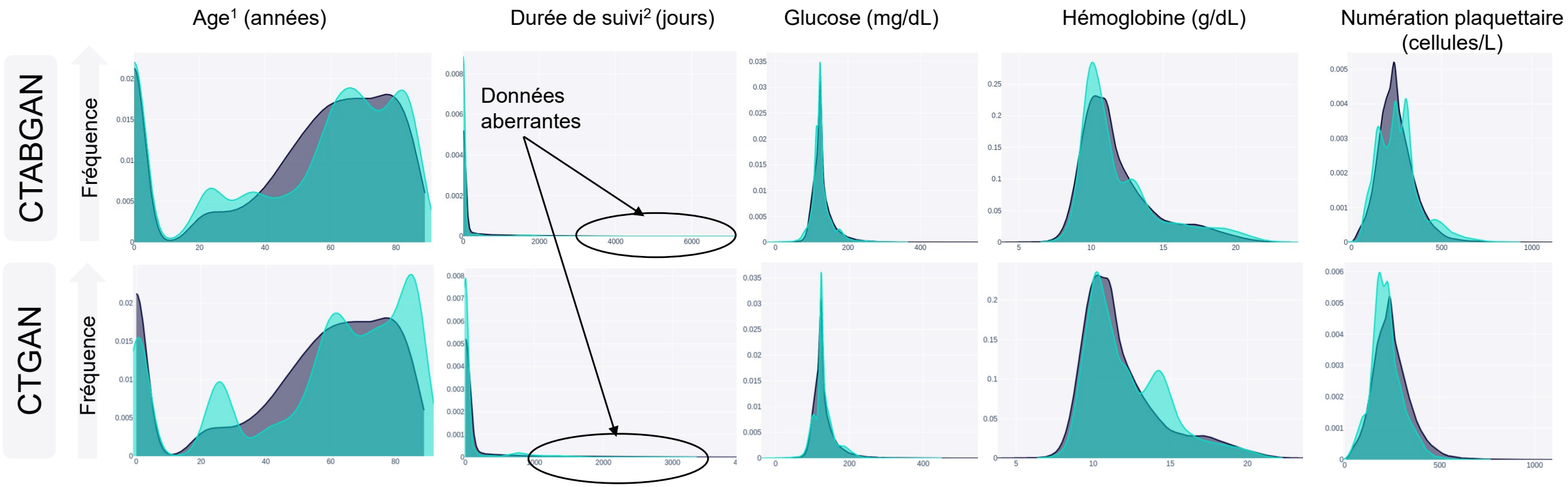


Patients réels

Patients synthétiques

quinten health

Des distributions continues fidèles aux données d'origine



Patients réels

Patients synthétiques

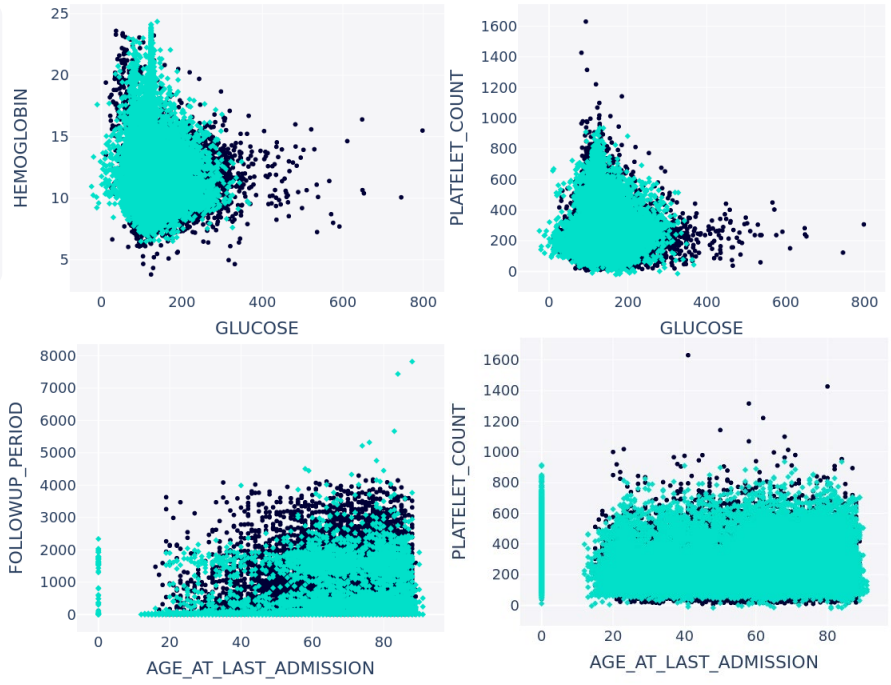
quinten health

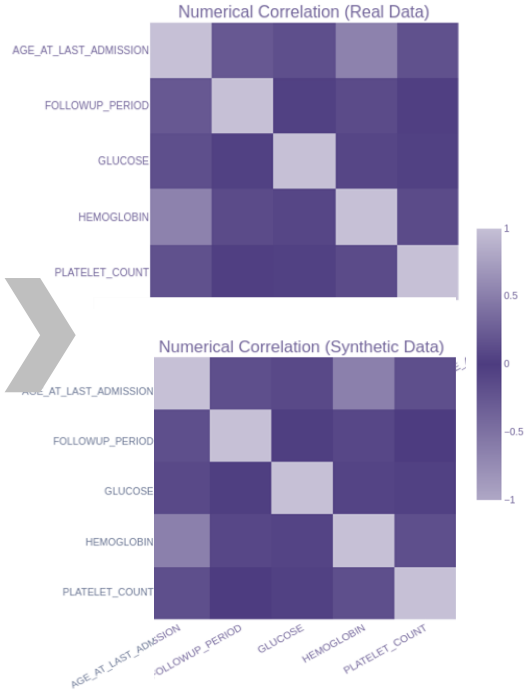
[1]Age à la dernière admission aux urgences

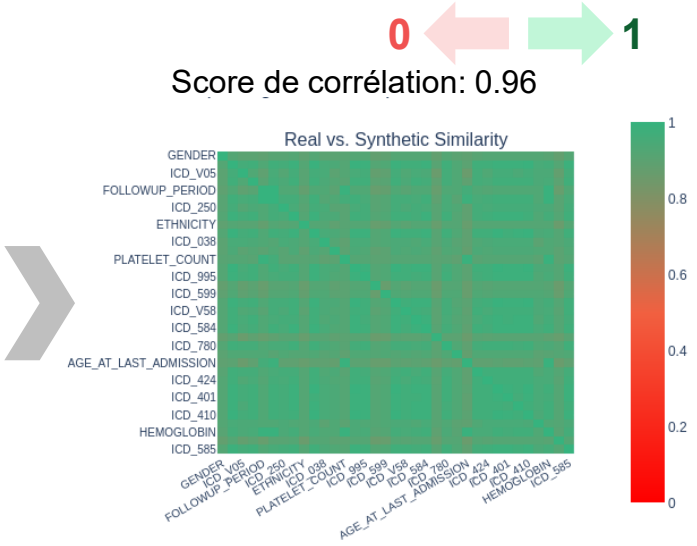
[2]Durée entre la première et derniere admissions aux urgences en jours. Un patient ayant une unique admission aura une durée de suivi de 0 jours.

Des corrélations préservées entre l'ensemble des variables

CTABGAN







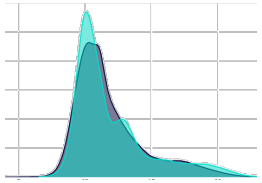
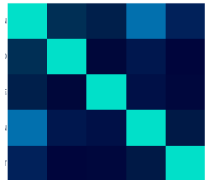
- CTABGAN préserve les correlations entre variables continues mais aussi discrètes
- MAIS génère des points aberrants pour les patients nouveaux nés

Patients réels

Patients synthétiques

quinten health

Génération de données réalistes préservant la confidentialité, un équilibre

	Fidélité		Confidentialité	
	Distribution[1]	Corrélation[2]	Nouveaux patients (%)	Score de distance[3]
Model				
CTGAN	0.90	0.93	**100%**	**0.92**
CTABGAN	**0.94**	**0.96**	96%	0.85
	0 ⟵ ⟶ 1	0 ⟵ ⟶ 1		0 ⟵ ⟶ 1

- CTABGAN génère des données synthétiques plus fidèles aux données d'origine que CTGAN
- MAIS CTGAN préserve mieux la confidentialité des données

[1]Distance shape score: Average between KSComplement and TVComplement

[2]Distribution correlation score: Average between contingency and correlation similarity

[3]DCR: Distance to Closest Record computed on 10% of the synthetic data for computational reasons

quinten health

Discussion

DES RESULTATS ENCOURAGEANTS

- Les GANs produisent des données synthétiques fidèles à la réalité et préservant la confidentialité
- MAIS des ressources computationnelles et données importantes, difficulté de convergence

UNE EVALUATION DEPENDANTE DE L'USAGE DES DONNEES

- Un niveau et quantité de preuve dépendant du critère de jugement
- Un compromis entre fidelité et confidentialité

POUR ALLER PLUS LOIN

- Evaluation sur d'autres bases de données de santé, de taille et nature différentes
- Métrique d'utilité adaptée à la question de recherche
- Comparaison à d'autres approches: Large Language Models

GAN: Réseaux adversaires génératifs

Thank you

Merci pour votre attention

Des questions ?

8, rue vernier
75017 Paris - France
+33 1 45 74 33 05

www.quinten-health.com

Confidential document – Copyright Quinten 20