unisanté

Anonymiser ou désidentifier ses données de recherche

Comment protéger ses participant-es

Céline Racine – Spécialiste en gestion des données – Unisanté
Udd.data@unisante.ch

Le plan

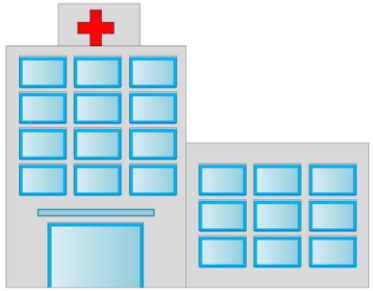
- Les définitions
  - Cadre légal
  - Anonymisation ou désidentification
  - Risques de réidentification
- Vos questions
- Conclusion

Cadre légal

La protection des données

Cambridge Analytica

UNDERGROUND

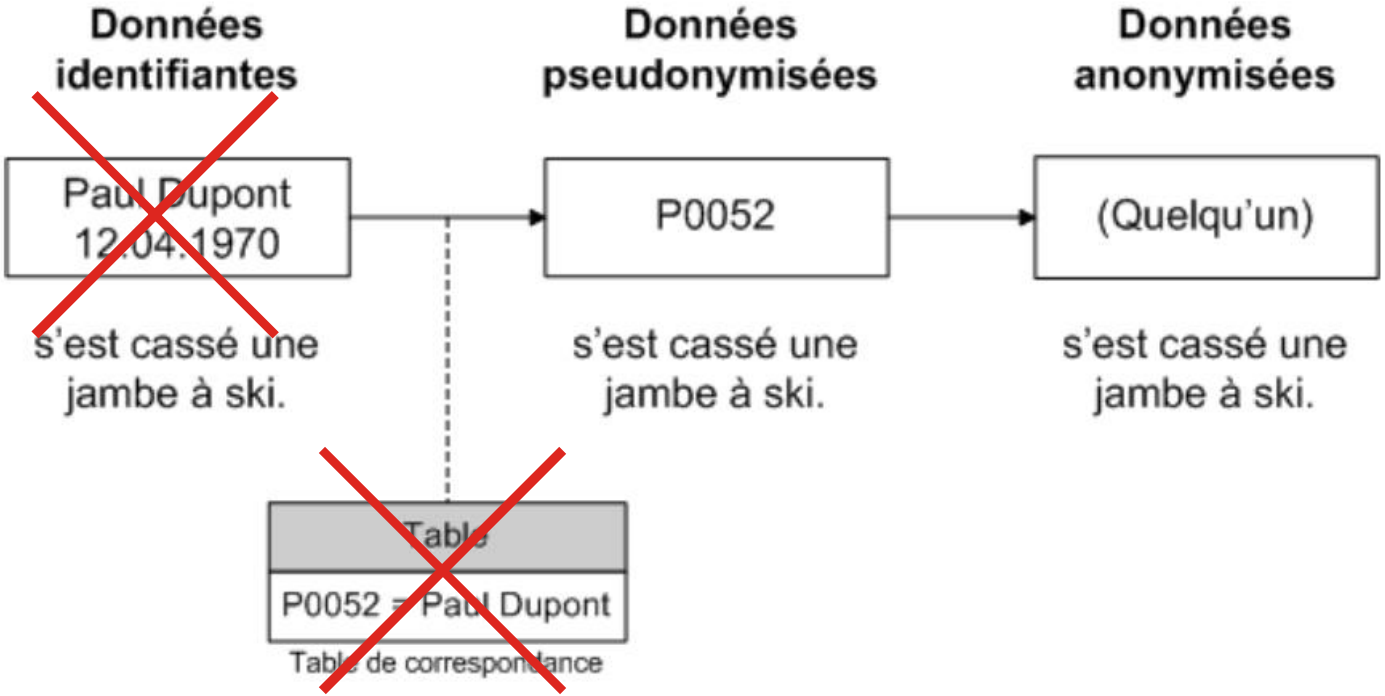


Cadre légal et consentement

- Loi sur la recherche sur l'être humain (LRH)
  - Information éclairée écrite
  - Consentement écrit obligatoire (2 feuilles séparées : participation et réutilisation)
- Loi fédérale sur la protection des données (LPD)
  - Information éclairée
  - Consentement explicite obligatoire pour le partage des données
- Hors-LPD
  - Les données ne sont plus relatives à une personne, le consentement n'est pas obligatoire

Anonyme ou non ?

Et si je supprime l'identité ?



Source : https://www.edoeb.admin.ch/edoeb/fr/home/datenschutz/internet_technologie/informationssicherheit.html

Catégorisation des identifiants

- Directs
  - Réidentification certaine (nom, numéro AVS)
- Forts
  - Réidentification très probable en croisant des données (adresse IP, maladie rare)
- Faible
  - Réidentification probable en croisant beaucoup de données (genre, date de naissance, lieu de travail)

En réalité

Anonymisation stricte
(= Anonymisation)

- Irréversible
- IMPOSSIBLE de réidentifier un individu
- Partage en Open Access

Anonymisation relative
(= Désidentification)

- Risques de réidentification
- Niveau de protection
- Partage en accès restreint

Stratégie de protection des données

- Avant collecte
  - Information éclairée aux participant-es
  - Consentement
- Collecte
  - Minimisation de données
  - Anonymisation par design
- Fin du projet
  - Traitement des données
  - Partage en accès restreint

Terminologie

	Identifiées / identifiables	**Codées**	**Désidentifiées (faible)**	**Désidentifiées (fort)**	**Anonymes**
Identifiants directs (HIPAA)	Intact	Transformés ou supprimés	Transformés ou supprimés	Transformés ou supprimés	Supprimés
Identifiants indirects	Intact	Intact	Intact	Transformés ou supprimés	Fortement transformés ou supprimés
Table de correspondance	-	Intacte	Supprimée	Supprimée	Supprimée

La désidentification est réalisée au cas par cas. Son objectif est de trouver le bon équilibre entre utilité du set de données et protection des participant-e-s.

Les données anonymes ne peuvent plus être reliées à un être humain et sortent du cadre légal de la LRH et LPD. L'impossibilité de ré-identification doit être garantie (expl : differential privacy) ou les données doivent être fortement agrégées.

Risques de réidentification

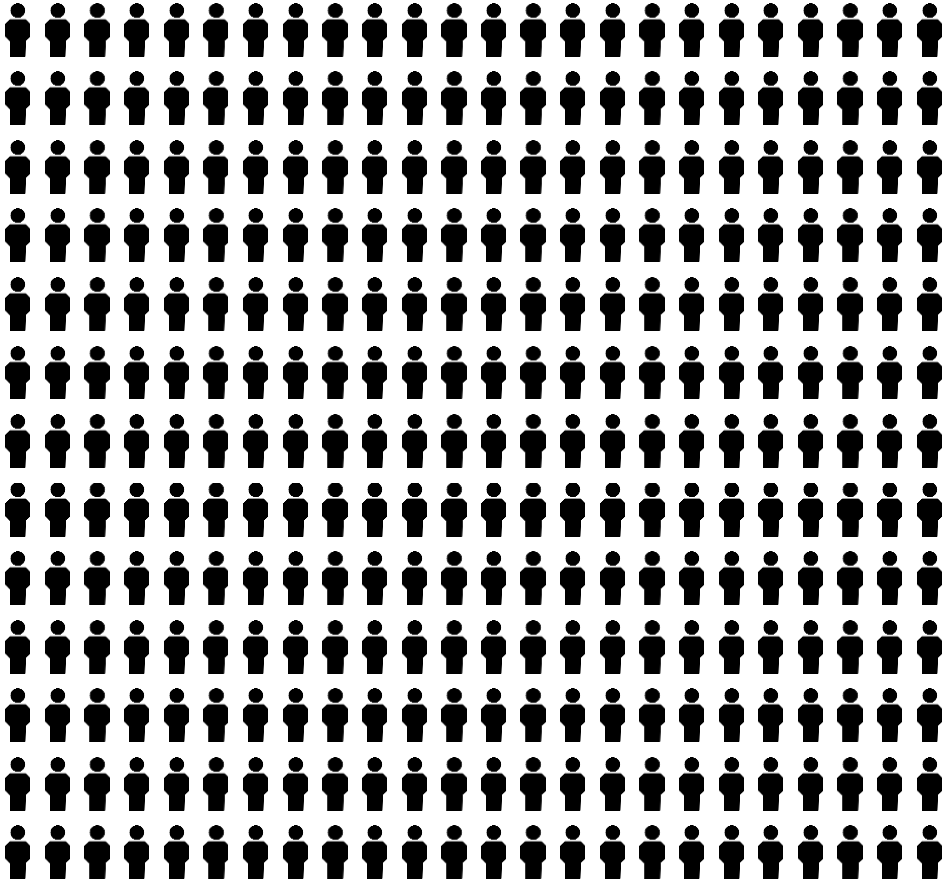
unisanté

Image by Gordon Johnson from Pixabay

Risque 1: Individualisation

Possibilité d'isoler 1 individu dans le set de données

- Attention particulière:
  - Petite population
  - Caractéristiques originales



unisanté

Risque 2: Corrélation

Possibilité de ré-identifier un individu en croisant des données

- Attention particulière:
  - Identifiants forts et faibles
  - Données codées/pseudonymisées

ID	0000515
Tél.	4.89.33
Institution	Unisanté

ID	9745831
Tél.	4.89.33
Personne	Céline Racine

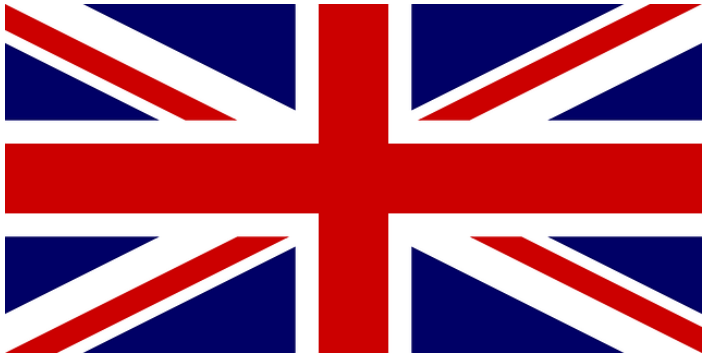
ID 0000515 = Céline Racine

Exemples de sources de données

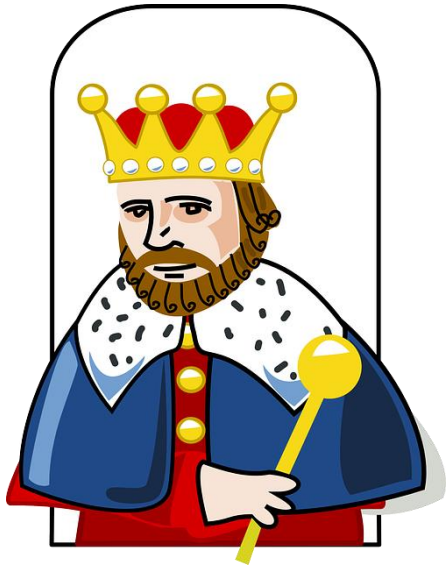
- List broker
  - https://shop.kbdata.ch/consumer/selection
- Dépôts de données
- Réseaux sociaux
- Entreprises
- Darknet

Risque 3:

Inférence



Possibilité de deviner l'identité d'un individu sans information supplémentaire



- Attention particulière:
  - Contexte
  - Informations publiques

Risques de réidentification

- S’il y a un risque, ce n’est pas anonyme

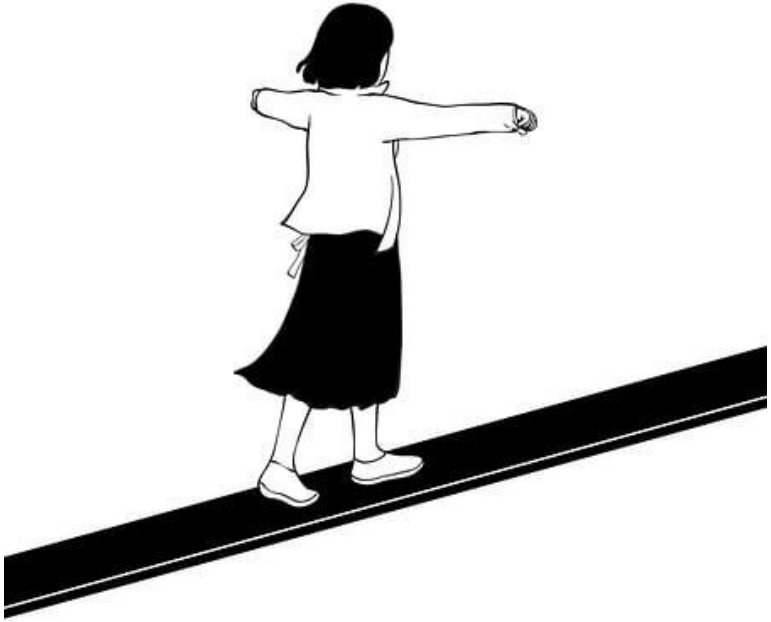
- Attention particulièrement à
  - Données sensibles
  - Petite population
  - Caractéristiques originales / données socio-démographiques
  - Destinataire des données

Vos questions sur l’anonymisation

Utilité des données

- Les données doivent être utile pour la recherche

- Stratégie de désidentification
  - Consentement
  - Traitement des données
  - Restriction d’accès



unisanté

Utilité d’un texte

« Mon nom est Céline Racine. Je suis enseignante à l’école primaire de Charrat (VS). Je suis des classes de 3-4 HARMOS. Dans une de mes classes, 2 élèves ont des besoins spéciaux, en effet, l’une est atteinte d’un TDAH et l’autre est en chaise roulante. Cela me fait de la peine, je fais énormément d’effort pour les intégrer mais sans moyen et sans formation spécialisée, je me sens surchargée et suis à la limite du burnout.»

Utilité d'un texte

« Mon nom est Céline Racine. Je suis enseignante à l'école primaire de Charrat (VS). Je suis des classes de 3-4 HARMOS. Dans une de mes classes, 2 élèves ont des besoins spéciaux, en effet, l'une est atteinte d'un TDAH et l'autre est en chaise roulante. Cela me fait de la peine, je fais énormément d'effort pour les intégrer mais sans moyen et sans formation spécialisée, je me sens surchargée et suis à la limite du burnout.»

Puis-je modifier ?

Je m’intéresse aux enfants ou à la structure des classes

- Modifier les identifiants indirects impactent ma recherche

Je m’intéresse aux adultes

- Je peux modifier les identifiants indirects sans grand impact sur le cœur de ma recherche

Conditions d’accès

Open Access

- Prend du temps
- Connaissance des données
- Suppression ou modification des identifiants directs et indirects
- Données agrégées

Accès restreint

- Permet d’appliquer une désidentification plutôt qu’anonymisation
- Assure un contrôle des données

Guide FORS

- Qualitative data anonymisation: theoretical and practical considerations for anonymising interview transcripts (FORS Guide N°20), Alexandra Stam, Pablo Diaz
- Quantitative data anonymisation: practical guidance for anonymising social science data (FORS Guide N°23), Brian Kleiner, Marieke Heers

Les techniques d’anonymisation

Catégorisation des identifiants

- Directs
  - Réidentification certaine (nom, numéro AVS)
- Forts
  - Réidentification très probable en croisant des données (adresse IP, maladie rare)
- Faible
  - Réidentification probable en croisant beaucoup de données (genre, date de naissance, lieu de travail)

Norme HIPAA

1. Nom
2. Adresse (y compris les subdivisions plus petites que les États, telles qu'une adresse postale, une ville, un comté ou un code postal)
3. Toutes les dates (à l'exclusion des années) qui sont directement liées à une personne, y compris l'anniversaire, la date d'admission ou de sortie, la date de décès ou l'âge exact des personnes âgées de plus de 89 ans
4. Numéro de fax
5. Numéro de téléphone
6. Adresse email
7. Numéro de dossier médical
8. Numéro de sécurité sociale
9. Numéro de bénéficiaire d’assurance
10. Numéro de compte
11. Numéro de certificat / permis de conduire
12. Identifiants de véhicule, numéros de série ou numéros de plaque d'immatriculation
13. Identifiants d'appareils ou numéros de série
14. URL Web
15. Adresse IP
16. Identifiants biométriques tels que les empreintes digitales ou les empreintes vocales
17. Photos identifiantes (visage, tatouages…)
18. Tout autre numéro d'identification, caractéristique ou code unique

Modifier des noms

- Suppression

- Pseudonymisation
  - Pseudo aléatoire : ID2866
  - Catégorisation : [responsable hiérarchique]
  - Alias / prénom fictif : Eléanore
    - Sociologie des prénoms : https://coulmont.com/bac/

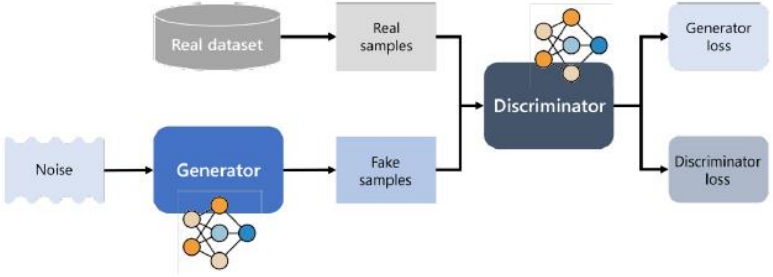
Bruit

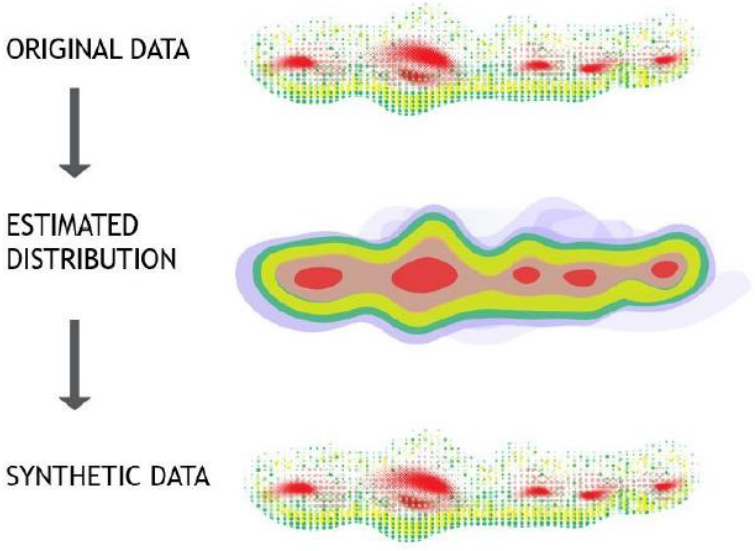
- Méthodes statistiques
- Données quantitatives
- Valeurs continues

- But : «noyer les données sensibles» en conservant les statistiques

Participant-e	Satisfaction
001	4.4
002	4.8
003	4.6
004	4.4
005	2
006	4
007	4.5

Synthetic data: a promising solution to alleviate the concerns on the privacy-utility trade-off





- Can address privacy concerns associated with real data
- Can address bias in real data with synthetic data diversification
- Can be a cost-effective approach for creating large datasets

Limites de l’anonymisation

Données médicales

- Essai clinique : non-anonymisable
  - Les identités identifiées doivent être conservées 10 à 15 ans (Oclin)

- Hors essai clinique : difficilement anonymisable
  - Trop d'identifiants indirect
  - Perte d'information trop importante

Données médicales

- Consentement
  - 1 consentement pour la participation
  - 1 consentement pour le partage et la réutilisation
- Le protocole explique le partage et la réutilisation des données

- Partage en accès restreint
- Réduction du risque = désidentification

Maladies rares et âges particuliers

- Généralisation des valeurs (catégorisation)

- Consentement éclairé des participant-es
- Restrictions d’accès

Personnalités publiques

- Retirer ou transformer un maximum d’identifiants
- Restreindre les informations contextuelles (métadonnées incluses)

- Recommandation : préférer un partage en accès restreint + conserver un risque

Vérifier l’anonymisation

Vérifier l’anonymisation

- Chaque projet est unique, pas de règles générales

- Prendre en compte
  - Risque pour les participant-es
  - Risques de réidentification (individualisation, inférence, corrélation)
  - Statistiques

Outils

- SPHN template : estimation de risque
- Outils statistiques
  - Amnesia
  - ARX
  - µ-ARGUS
  - SdcMicro

Conclusion

Stratégie de protection des données

- Avant collecte
  - Information éclairée aux participant-es
  - Consentement
- Collecte
  - Minimisation de données
  - Anonymisation par design
- Fin du projet
  - Traitement des données
  - Partage en accès restreint

Conclusion

- Désidentifier : au cas par cas
- Garder une balance protection / utilité
- Attention au cadre légal et consentement

unisanté

Merci pour votre attention

Références

- BOUTET, Antoine, 2023. Données personnelles : rien à cacher, mais beaucoup à perdre. *The Conversation*. [en ligne]. 29 mars 2023. [Consulté le 2 mai 2023]. Disponible à l'adresse: http://theconversation.com/donnees-personnelles-rien-a-cacher-mais-beaucoup-a-perdre-201494
- CNIL, Commission nationale de l'informatique et des libertés, 2020. L'anonymisation de données personnelles. *cnil.fr*. [en ligne]. 19 mai 2020. [Consulté le 1 août 2022]. Disponible à l'adresse: https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles
- DATA PROTECTION COMMISSION, 2019. *Guidance Note: Guidance on Anonymisation and Pseudonymisation*. [en ligne]. Dublin: Data protection Commission. [Consulté le 2 février 2022]. Disponible à l'adresse: https://www.dataprotection.ie/sites/default/files/uploads/2020-09/190614%20Anonymisation%20and%20Pseudonymisation.pdf
- FINCH, Kelsey, 2016. A Visual Guide to Practical Data De-Identification. *https://fpf.org/*. [en ligne]. 25 avril 2016. [Consulté le 2 mai 2023]. Disponible à l'adresse: https://fpf.org/blog/a-visual-guide-to-practical-data-de-identification/
- How to GO FAIR, [sans date]. *GO FAIR*. [en ligne]. [Consulté le 18 juin 2022]. Disponible à l'adresse: https://www.go-fair.org/how-to-go-fair/
- JOHNSTON, Lisa (éd.), 2017. *Curating research data. Volume two: A handbook of current practices*. [en ligne]. Chicago, Illinois: Association of College and Research Libraries, a division of the American Library Association. ISBN 978-0-8389-8862-6. Disponible à l'adresse: https://renouvaud1.primo.exlibrisgroup.com/permalink/41BCULAUSA_LIB/1vikse1/alma991021107671202852
- JOTTERAND, Alexandre, 2022. Personal Data or Anonymous Data: where to draw the lines (and why)? *Jusletter*. [en ligne]. 15 août 2022. No. 1119. [Consulté le 24 août 2022]. Disponible à l'adresse: https://jusletter.weblaw.ch/juslissues/2022/1119/personal-data-or-ano_173939252d.html
- Kleiner, B. & Heers, M. (2024). Quantitative data anonymisation: practical guidance for anonymising sensitive social science data. FORS Guide, 23, Version 1.0, 1-17. https://doi:10.24449/FG-2024-00023

unisanté

Références

- NGUYEN, Benjamin et CASTELLUCCIA, Claude, 2020. Techniques d'anonymisation tabulaire : concepts et mise en oeuvre. *1024 : Bulletin de la Société Informatique de France*. 2020. No. 15, pp. 23. Disponible à l'adresse : https://hal.science/hal-02570847
- ROCHER, Luc, 2019. Données anonymes… bien trop faciles à identifier. *The Conversation*. [en ligne]. 17 septembre 2019. [Consulté le 13 décembre 2021]. Disponible à l'adresse: http://theconversation.com/donnees-anonymes-bien-trop-faciles-a-identifier-123157
- ROCHER, Luc, HENDRICKX, Julien M. et DE MONTJOYE, Yves-Alexandre, 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*. 23 juillet 2019. Vol. 10, no. 1, pp. 3069. DOI 10.1038/s41467-019-10933-3.
- SPRUMONT, Dominique, 2019. Protection des données, anonymisation et recherche. [en ligne]. Lunch LRH, Lausanne. 3 octobre 2019. [Consulté le 2 août 2022]. Disponible à l'adresse: https://static1.squarespace.com/static/60b94bed393f8064950b2821/t/616e9f1dce9f7f7a05927376/1634639647288/Presentation_Protection_des_donnees_anonymisation_et_recherche_191003.pdf
- STAM, A, DIAZ, P. (2023). Qualitative data anonymisation: theoretical and practical considerations for anonymising interview transcripts. FORS Guide No. 20, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. doi:10.24449/FG-2023-00020