Métadonnées et protection de la sphère privée
Exemple d'un protocole

Sandrine Estoppey Younes – IUMSP

Métadonnées et protection de la sphère privée
Exemple d'un protocole

SKIPOGH (Swiss Kidney Project on Genes in Hypertension)

SNF – PI Prof Murielle Bochud, IUMSP

Cohorte multicentrique (Ge, Vd, Be) – familles de volontaires sains

Baseline (2009-2012) N= 1150 + suivi à trois ans (2012-2016)

- 20 publications
- 10 manuscrits en préparation
- 15 collaborations nationales et internationales (données, biobanque, génétique)

→ Partage de données

<u>Information/Consentement:</u>

- Recherche et analyses génétiques en lien avec fonction rénale et maladies cardiovasculaires

- Réutilisation des données et échantillons dans la recherche biomédicale du même domaine

- Réutilisation pour projets ultérieurs non encore définis actuellement (consentement élargi)

Données

- Hétérogènes (sources, formats)
- Massives (cliniques, auto rapportées, génomiques, épigénomiques, métabolomiques, images)

Problèmes liés aux données

- Transferts sécurisés, intégration des données sources
- Gestion des données massives hétérogènes
- Partage (intra-, inter-institutionnel, plus large)

Sécurité (vol)
Confidentialité

Volonté d'assurer le respect de la sphère privée

- Manque d'outils/compétences

Demandes de collaboration

- Recherche sur la sécurité des données en lien avec la santé (EPFL, UNIL, HESSO, SDSC),
- Besoin de sets des données 'réelles' (données hospitalières inaccessibles)
- Besoin de données non biomédicales (métadonnées), p.ex localisation

Loi / éthique

Codage + dé-identification

- Retrait de tous les identifiants (Noms, numéros, adresses, dates, etc…)

→ Données normalement inaccessibles car permettant justement la ré-identification du participant

Consentement

- Projets non couverts par le consentement
- Difficulté d'obtenir un consentement éclairé sur un sujet délicat et complexe

Dilemme

Partager des données confidentielles non partageables pour permettre de la recherche visant à mieux les sécuriser

Recherche sur la sécurité des données

- En périphérie de la recherche sur les données de santé
- Recherche sur la qualité (processus).
- Limites du domaine de la LRH ? → compétences CE ?
- Soumise à la loi sur la protection des données

Buts

Compréhension et limitation des risques de ré-identification (robustesse du codage, déductions possibles par inférence)

Dévelopement d'outils pour la protection des données dans le cadre du partage des données biomédicales (recherche, dossier informatisé du patient, intégration des données de routine)

Suivre évolution des technologies servant à la collecte des données (apps, smartphones), aussi intéressantes pour la recherche.

Le processus de dépersonnalisation consiste à retirer (définitivement) les identifiants directs et indirects, de façon à rendre très difficile l'identification d'un individu par une combinaison de données.

Supprimer - retirer:

- noms et initiales : patients, proches, médecins, partenaires
- Indications géographiques plus petites qu'un canton : rue, ville, lieu, NPA.
- Dates (sauf l'année) directement liées à un individu, dates de naissance, admission/sortie, décès, visite
-Numéros uniques : téléphone, fax, email, URL, IP, AVS, dossier, assuré, etc..
- Images (photos visages, marques distinctives, tatouages)

<u>Plusieurs niveaux de dé-identification selon travail sur données et selon intervenant (autorisation)</u>

Structure données SKIPOGH - Codage

Données sources (clinique, labos, Q auto-rapportés, etc…)

Codage 1: retrait identifiants principaux (création fichier clé de codage sécurisé)
- Noms, adresses, numéros

Bases données 1 , Bases données 2 , etc… - travail interne

Codage 2: retrait identifiants potentiels nécessaires au travail interne (datamanagement/analyses)
- Dates précises (selon base concernée)

Codage 3: analyses génétiques
- Double codage

Extractions bases données → partage interne vs externe

Structure données SKIPOGH - Codage

Données sources (clinique, labos, Q auto-rapportés, etc…)

Codage 1: retrait identifiants principaux
(création fichier clé de codage sécurisé)
- Noms, adresses, numéros

Bases données 1 , Bases données 2 , etc… travail interne

Perte d'information utile à la recherche
géolocalisation(études pollution),
source des données (statistiques qualité)

Codage 2: retrait identifiants potentiels
nécessaire au travail interne
datamanagement/analyses)
Dates précises (selon base concernée)

Codage 3: analyses génétiques
- Double codage

Bases données extractions → partage interne vs externe

SKIPOGH Metadata and patient's privacy

Comment les métadonnées d'information de santé peuvent affecter le respect de la vie privée

- Zone d'interaction où la dé-identification classique empêche la recherche (données inaccessibles →perte d'utilité)
- Tests d'algorithmes de sécurisation de données (modèle k-anonymity)
- Trouver un compromis utilité / sécurité

<u>Risques étudiés</u>

Inférences possibles avec données librement accessibles

→ (pages jaunes/blanches, réseaux sociaux, recensement)

Exemple: localisation médecin traitant

[1] El Emam Khaled, Rodgers Sam, Malin Bradley. Anonymising and sharing individual patient data BMJ 2015;
[2] Sariyar, Murat, and Irene Schlünder. "Reconsidering Anonymization-Related Concepts and the Term "Identification" Against the Backdrop of the European Legal Framework." *Biopreservation and Biobanking* (2016).
[3] Data Protection Directive 95/46/EC, http://ec.europa.eu/justice/policies/privacy/docs/95-46-ce/dir1995-46_part1_en.pdf
[4] The Health Insurance Portability and Accountability Act of 1996 (HIPAA)

Localisation de la source de données peut être déduite

1- Signature numérique de la source (pour vérifier que la source est le véritable médecin / l'institution médicale)

2-L'adresse IP de la source des données[1]

http://whatismyipaddress.com/ip/128.179.165.48

128.179.165.48 Lookup IP Address

General IP Information

IP: 128.179.165.48
Decimal: 2159256880
Hostname: tsf-460-wpa-5-048.epfl.ch
ASN: 559
ISP: Ecole Polytechnique Federale de Lausanne
Organization: Ecole Polytechnique Federale de Lausanne
Services: None detected
Type: Broadband
Assignment: Static IP
Blacklist: Blacklist Check

Geolocation Information

Country: Switzerland
State/Region: Vaud
City: Ecublens
Latitude: 46.5185219 (46° 31′ 6.68″ N)
Longitude: 6.5619088 (6° 33′ 42.87″ E)

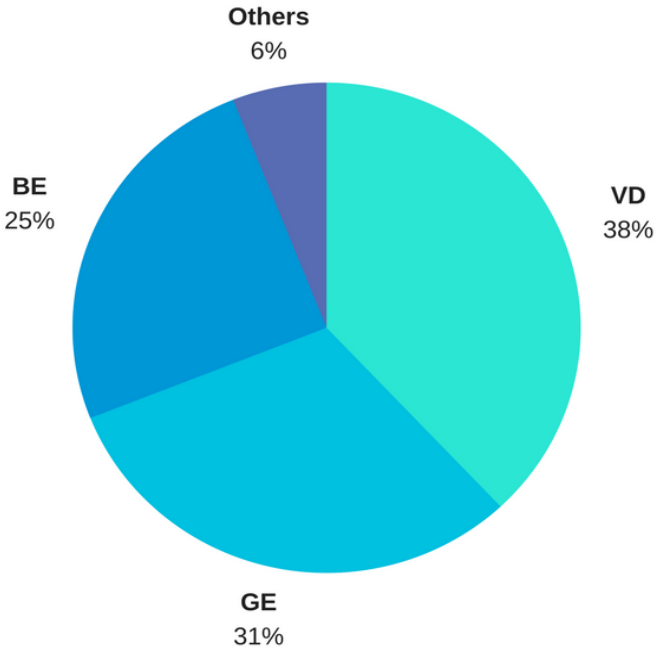
Don't want this known? Hide your IP details
Location not accurate? Update your location

Geolocation Map

[1] Muir, James A., and Paul C. Van Oorschot. "Internet geolocation: Evasion and counterevasion." *Acm computing surveys (csur)* 42.1 (2009): 4.

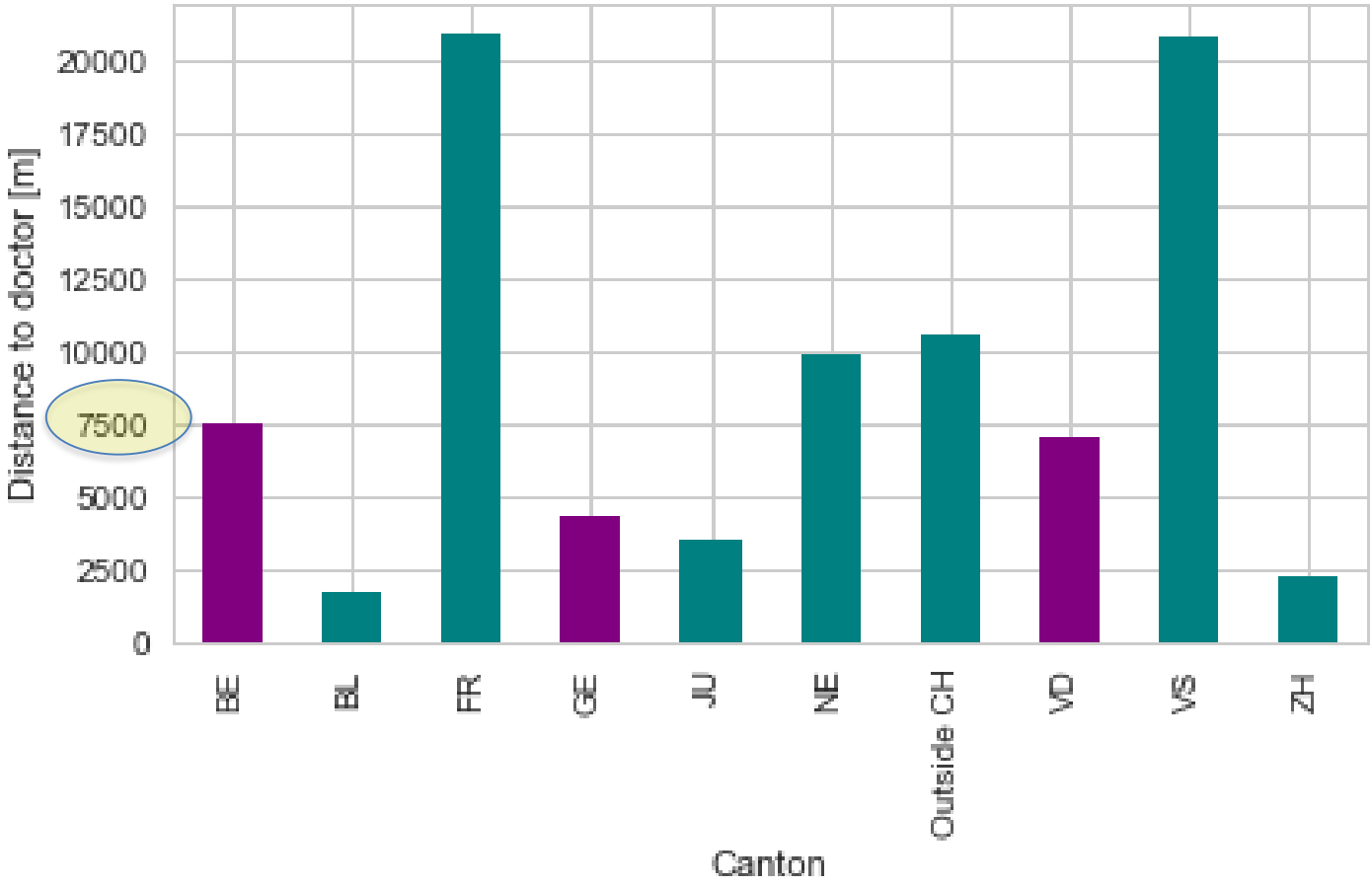
Données extraites de SKIPOGH

- 800 enregistrements (participants)
- Utilisés: adresse (rue+NPA), genre, âge du participant, adresse et nom du médecin.
- Dés-identification: pas de nom de patient, “mélange ” des attributs (taille, genre, âge)



95% des personnes consultent un médecin dans le canton où elles vivent

Distance entre le participant et son médecin traitant



Pas plus de 7,5 km dans la plupart des cas

Données en libre accès

Augmentation du risque de ré-identification

su-f-01.02.03.06

Population résidante permanent selon âge, par commune

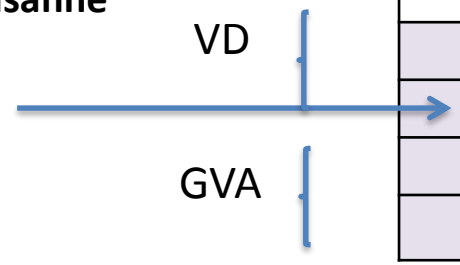
Région	Total	0	1	2	3	4	5	6	7	8	9	10	11
Suisse	8039060	79621	81185	82560	80900	80183	78980	78127	78023	78084	76848	77822	7832
- Zürich	1408575	15324	15350	15272	14820	14392	14034	13657	13466	13353	13015	13069	1270
>> Bezirk Affoltern	49446	514	546	545	585	590	513	607	536	581	513	534	57
.....0001 Aeugst am Albis	1955	21	18	19	24	19	20	19	13	15	22	27	1
.....0002 Affoltern am Albis	11276	107	129	118	122	109	99	114	104	107	117	103	11
.....0003 Bonstetten	5205	71	64	64	82	83	71	77	69	65	59	65	7
.....0004 Hausen am Albis	3376	27	28	35	31	48	40	48	40	44	37	38	3
.....0005 Hedingen	3511	32	36	29	30	37	36	49	47	55	43	43	4
.....0006 Kappel am Albis	922	6	16	14	11	19	9	19	9	12	13	9	1
.....0007 Knonau	1982	21	15	36	21	32	24	30	21	21	18	17	2
.....0008 Maschwanden	641	7	11	9	4	10	6	10	9	2	5	6	1
.....0009 Mettmenstetten	4420	48	42	43	49	41	44	57	45	58	39	47	5
.....0010 Obfelden	4833	45	54	51	73	54	42	61	55	65	49	62	6
.....0011 Ottenbach	2480	27	29	19	28	32	18	26	22	27	27	21	2
.....0012 Rifferswil	957	17	9	10	18	17	11	23	16	20	6	13	1
.....0013 Stallikon	3309	40	32	38	34	43	37	26	24	39	31	29	3
.....0014 Wettswil am Albis	4579	45	63	60	58	46	56	48	62	51	47	54	5
>> Bezirk Andelfingen	30038	292	298	290	335	307	331	341	316	341	329	341	33
.....0021 Adlikon	561	3	5	9	9	7	10	5	6	5	3	5	
.....0022 Benken (ZH)	823	10	8	6	4	8	7	7	8	6	12	9	1
.....0023 Berg am Irchel	587	6	2	5	5	5	2	7	4	3	1	6	
.....0024 Buch am Irchel	899	9	12	14	11	12	14	12	9	11	17	18	
.....0025 Dachsen	1960	18	19	17	26	21	30	24	16	29	23	27	2
.....0026 Dorf	634	1	4	1	4	6	7	7	5	13	9	10	

2016 2015 2014 2013 **2012** 2011 2010 +

https://www.bfs.admin.ch/bfs/en/home/statistics/population.assetdetail.3222001.html

D'après la localisation du médecin, même après brouillage de la localisation du participant, on peut repérer qu'un participant habite autour de Puidoux (n=2500)

Puidoux, 1070
15km from Lausanne



NPA du patient anonymisé	NPA du médecin/source
1007-1219	**1007**
1007-1219	**1070**
1007-1219	**1205**
1007-1219	**1208**

Ici, un médecin (travaillant à 1007 Lausanne), suit 23 participants
61% des participants allant chez lui vivent à 1007 Lausanne

NPA du patient	%
1005	7.69
1006	7.69
1007	**61.54**
1028	7.69
1053	7.69
1555	7.69

NPA du patient anonymisé	NPA du médecin/ source	NPA du médecin/source déduite
1007-1219	**1007**	**1007**
1007-1219	**1070**	**1070**
1007-1219	**1205**	**? (GVA)**
1007-1219	**1208**	**? (GVA)**

Contre-mesures

- k-anonymiser par la source de données / médecin

  (Utilité reduite)

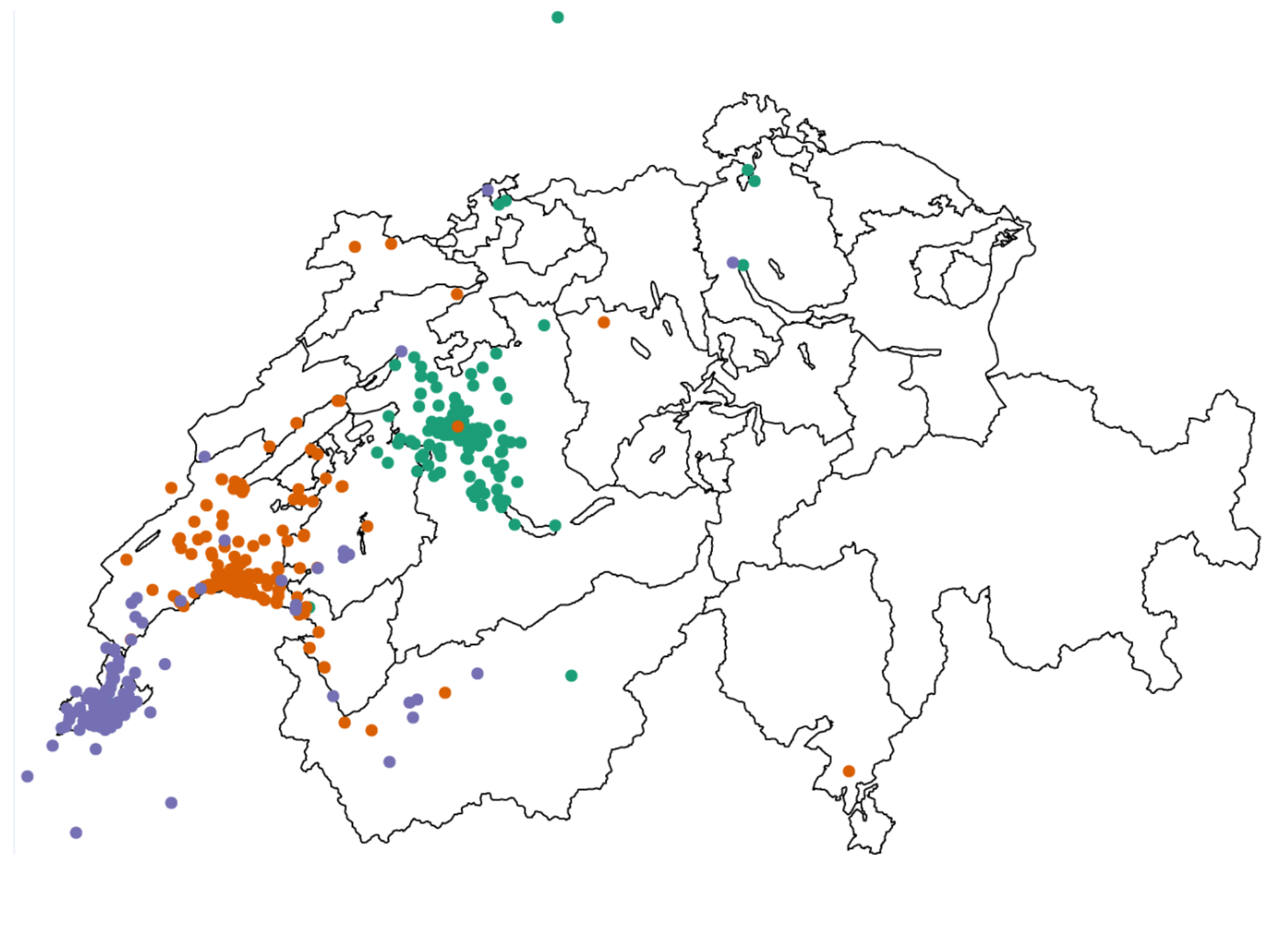
- “protéger” la localisation du médecin

  proxy approuvé, réseaux p2p, solution cryptographique (“anonymous credentials”)

  Besoin d’ algorithmes adaptatifs pour l’anonymization

<u>Besoins</u>

- Plus de données
- Données précises: numéro de la rue, données du participant non mélangées



Modèle k-anonymat

Modèle fréquemment utilisé en recherche clinique

- Définition de QID (quasi –identifiants): attributs pas directement des identifiants mais qui, combinés à des données d'accès libre, peuvent permettre d'inférer l'identité d'un individu.

- Créer du bruit autour des données sensibles pour qu'il y ait au moins k-enregistrements avec le même QID.
1 même QID est attribué à k individus→ probabilité de ré-identification = 1/k

- Choix du k important: permet d'équilibrer balance utilité / sécurité