SMALL CELL RISK ASSESSMENT

Pseudonymised Personal Data on Mental Health

PURPOSE: SMALL CELL RISKS ANALYSIS FOR Pseudonymised Personal Data on Mental Health

SMALL CELL RISK ASSESSMENT FORM

Project title
It is important to ensure that the title of the study is clear, easy to understand and accurately reflects the main purpose/focus of the project.
Pseudonymised Personal Data on Mental Health (GEPSEUDONIMISEERDE PERSOONSGEGEVENS DIE DE GEZONDHEID BETREFFEN AFKOMSTIG VAN ZORGINSTELLINGEN ACTIEF IN DE GEESTELIJKE GEZONDHEIDSZORG AAN DE ONDERZOEKSGROEPEN CENTRUM VOOR HUISARTSENGENEESKUNDE EN COLLABORATIVE ANTWERP PSYCHIATRIC RESEARCH INITIATIVE VAN DE UNIVERSITEIT ANTWERPEN IN HET KADER VAN EEN PROJECT MET ALS DOELSTELLING DE GEESTELIJKE VOLKSGEZONDHEID TE VERBETEREN)

Applicant	
Institution	**Centrum voor Huisartsengeneeskunde en Collaborative Antwerp Psychiatric Research Initiative van de Universiteit Antwerpen**
Address	Gouverneur Kinsbergencentrum, room 00.56 Doornstraat 331 - 2610 Wilrijk - Belgium
Principal investigator	Prof. Dr. Kris Van den Broeck

Disclosure risk assessor	
Institution	P95 CV
Address	Koning Leopold III laan 1 3001 Heverlee BELGIUM
Assessors (contact details)	Tom De Smedt (data analyst): tom.desmedt@p-95.com Ward Schrooten (MD): ward.schrooten@p-95.com

I. DESCRIPTION OF THE DATA USE (filled by applicant)

*** Should be aligned with the authorization request**

A.1. Motivation of the data request (Max 500 words)
Provide the necessary background information and key references, providing evidence that the applicants know the relevant scientific literature. Clearly describe the reasons for data collection (mention legal obligations if any), the research questions and their relevance for policy making and science. Provide a concise overview of the objectives, methods and data analysis for the proposed research.
Data on mental health is currently collected by many different health care facilities. In order to get a better picture of the care needs, a centralized data repository combining all the fragmented data from the different health care facilities is deemed necessary.

A.2. Objective(s)
Describe the objectives ordered from most to least important in sufficient detail, allowing assessment as to whether the data collection and intended analyses meet the objectives.
A centralized data repository on mental health combining the data from different health care facilities.

A.3. Target population
Describe the reference or target population, its key features and size.
All patients treated by health care facilities on mental health on 1/1/202 with updates every three months..

A.4. Population intended to be covered by the new data collection
Describe the population that will be covered by the data collection and its intended size. Include some consideration of whether the sample-size/power will be sufficient to meet the scientific objectives of the project.
Same as A.3.

A.5. Study design
Describe design characteristics that might be important for the small cell risk analysis.
Observational study

A.6. Variables

Give an overview of the key variables (or groups) of variables of the study

Dataset	Variables in code list
Gegevens setting	Erkenningsnunner
Afdelingsgegevens	Nummer
	Doelpopulatie
Patiëntgegevens	(gepseudonimiseerde) INS
	Geslacht
	Geboortejaar
	Nationaliteit (in klassen)
	NIS-code woonplaats
	Burgerlijke staat
	Kinderen ten laste
	Werksituatie (beroepsstat
	Opleidingsniveau (niveau l
	Type woonplaats (leefmilie
	Verzekerbaarheid code pat
	Hospitalisatieverzekering
Gegevens m.b.t. (deel)opname	Identificatie van de opnam
	(gepseudonimiseerde) RIZI
	Erkenningsnummer verwij
	Nummer afdeling indien ir
	Datum opname
	(gepseudonimiseerde) RIZI
	Kenletter bedtype
	Hoeveelste opname (op de
	Datum ontslag
	Omkadering na ontslag
	Erkenningsnummer van se
	Nummer afdeling waar pat
Gegevens m.b.t. behandeling (per deelopname)	Aanmeldingsklacht (vrije t
	Wijze van opname
	DSM-IV-diagnoses op As 1
	DSM5 data indien beschikl
	ICD9/10 indien beschikba
	Aantal contacten met vers
	Suïcidescreening
	Andere schalen indien bes
	Start vrijheidsbeperkende
	Einde vrijheidsbeperkende
	Reden van ontslag
	Wijze van ontslag
Medicatievoorschriften	Datum
	Naam medicatie
	CNK-code
Somatische parameters (bij opname)	Gewicht
	Lengte
	Bloeddruk
	Buikomtrek
	HDL-cholesterol
	Roken
	Middelengebruik

A.7. Data/statistical analyses planned
An overview of the data management and data analysis to be performed should be covered in this section. Applicants should ensure that analytical methods proposed are consistent with the objectives of the project and the data collected.
Statistical analysis aimed at providing insights in the current demand and offerings of mental health care.

A.8. Plans for disseminating and communicating study results, including target audience
Describe the way the results will be disseminated and the intended target audience.
The dataset as described here will only become available to researchers. Only aggregated results without the risk of identification will be shared with the wider public.

A.9. Codebook
The assessor should have access to the codebook, listing all variables that will be collected, the variable name, short description, variable type (binary, categorical, continuous) and possible values (in case of a categorical variable) or value range (in case of a continuous variable).
The codebook can be found at:  Codelist_UA_MentalHealth_15022021

II. SMALL CELL RISK ASSESSMENT (filled by assessor)

1. Identify direct identifiers, indirect identifiers and sensitive information
Complete the codebook by indicating whether variables are direct identifiers, indirect identifiers or contain sensitive information.
File name:  *Classification of variables: Ward Schrooten, MD*

2. Disclosure risk assessment based on direct identifiers
Identify the direct identifiers. These are variables that unambiguously identify units of observation (e.g. names, addresses, phone numbers, social insurance numbers).
There is no direct identifier in the datasets used for the current small cell risk analysis.

3. Disclosure risk assessment based on indirect identifiers
Assess the disclosure risk based on indirect identifiers. These are variables that –in combination with other indirect identifiers – can be used to disclose the identity of individuals or institutions.
Gender, weight and height variables are classified as indirect identifiers. When both weight and height are included, guaranteeing subject anonymity will be very hard. The qualitative assessment is:

Dataset	Variables in code list	Unique values	Combined unique values	Comments
Gegevens setting	Erkenningsnunner	1	1	Onbekend
Afdelingsgegevens	Nummer	1	1	Onbekend
	Doelpopulatie	1	1	Onbekend
Patiëntgegevens	Geslacht	2	2	Man, vrouw
	Geboortejaar	64	128	1940 tot 2002 (ouder in 1 klasse)
	Nationaliteit (in klassen)	10	1280	Geschatte aantal nationaliteten
	NIS-code woonplaats	589	753920	Aantal verschillende NIS-codes in België
	Burgerlijke staat	5	3769600	ongehuwd, gehuwd, gescheiden, verweduwd, wettelijk samenwonend
	Kinderen ten laste	6	22617600	0, 1, 2, 3, 4, 5, +5
	Type woonplaats (leefmilieu voor opname)	8	180940800	Zie beschrijving
Gegevens m.b.t. (deel)opname	Identificatie van de opname (opnamenummer)	1	180940800	Onbekend
	Erkenningsnummer verwijzende organisatie	1	180940800	Onbekend
	Nummer afdeling indien interne verwijzing	1	180940800	Onbekend
	Datum opname	1	180940800	
	Hoeveelste opname (op deze afdeling)	10	1809408000	Geschat maximaal aantal opnames
	Datum ontslag	1	1809408000	
	Erkenningsnummer van setting waar patiënt na ontslag naar wordt verwezen	1	1809408000	
	Nummer afdeling waar patiënt na ontslag naar wordt verwezen	1	1809408000	
Somatische parameters (bij opname)	Gewicht	1301	2,35404E+12	40 tot 170kg per 0.1kg
	Lengte	61	1,43596E+14	150 tot 210cm per 1cm
	Buikomtrek	30	4,30789E+15	Geschatte buikomtrek range in cm

The number of unique combinations is very high. Where the unique values is set to 1, the actual number of unique values is currently unknown. It is clear that the high number of indirect identifiers will lead to a very high risk of identification.

4. Impact of potential disclosure

Assess the impact of potential disclosure based on the sensitive variables.

A potential identity disclosure is particularly problematic if sensitive information is revealed. If a patient is a sample unique, then sensitive information is revealed upon identity disclosure.

The following variables are classified as sensitive:

Dataset	Variables in code list
Afdelingsgegevens	Nummer
	Doelpopulatie
Patiëntgegevens	Nationaliteit (in klassen)
	Burgerlijke staat
	Kinderen ten laste
	Werksituatie (beroepsstatus bij opname)
	Opleidingsniveau (niveau laatst beëindigd onderwijs)
	Type woonplaats (leefmilieu voor opname)
Gegevens m.b.t. (deel)opname	Erkenningsnummer verwijzende organisatie
	Hoeveelste opname (op deze afdeling)
	Omkadering na ontslag
	Erkenningsnummer van setting waar patiënt na ontslag naar wordt verwezen
	Nummer afdeling waar patiënt na ontslag naar wordt verwezen
Gegevens m.b.t. behandeling (per deelopname)	Aanmeldingsklacht (vrije tekst)
	Wijze van opname
	DSM-IV-diagnoses op As 1 t.e.m. 5 (met data van toekenning)
	DSM5 data indien beschikbaar
	ICD9/10 indien beschikbaar
	Suïcidescreening
	Andere schalen indien beschikbaar (Honos, …)
	Start vrijheidsbeperkende maatregelen
	Einde vrijheidsbeperkende maatregelen
	Reden van ontslag
	Wijze van ontslag
Medicatievoorschriften	Naam medicatie
	CNK-code
Somatische parameters (bij opname)	Bloeddruk
	HDL-cholesterol
	Roken
	Middelengebruik

If BMI is added to the dataset instead of height and weight, this would be an additional sensitive variable.

5. Recommended disclosure control strategies

Based on the analysis done, we would like to suggest the following actions:

- Consider whether defining the specific research questions upfront is possible and have different datasets per research question. That would potentially allow the number of variables to be drastically decreased.
- If the above suggestion is not possible, we would strongly recommend the following:
    - Where possible move from continuous variables (such as age, weight, etc.) to categorical variables
    - Remove weight and height and use BMI
    - Consider whether pseudonimization of ID variables is possible
    - Implement very strong technical and organisational measures to ensure data protection (f.e. encrypted data transfers, only store and analyze data on a server with strict access control, etc.)
    - Implement a very strict data access policy
    - Implement a very strict data retention policy
    - Carefully check the results to ensure no personal data is shared