SMALL CELL RISK ASSESSMENT

INTEGO

PURPOSE: SMALL CELL RISKS ANALYSIS FOR SCIENSANO@INTEGO

SMALL CELL RISK ASSESSMENT FORM
FOR DATA USE WITHIN THE HEALTHDATA.BE PLATFORM
V0.0

Project title
Système d'enregistrement général et continu de la morbidité, Intego ('Integrated Computerized Network').

Applicant	
Institution	KU Leuven
Address	Kapucijnenvoer 33 blok j B-3000 Leuven
Principal investigator (contact details)	Bert Vaes bert.vaes@kuleuven.be

Disclosure risk assessor	
Institution	P-95
Address	Koning Leopold III laan 1 3001 Heverlee BELGIUM
Assessors (contact details)	Estelle Méroc (data analyst) ; estelle.meroc@p-95.com: Kaatje Bollaerts (statistician); kaatje.bollaerts@p-95.com Chukwuemeka Onwuchekwa (MD); chukwuemeka.onwuchekwa@p-95.com

Signature	
Applicant: Name: Bert Vaes Date: 10 October 2022 Signature: DocuSigned by: 29725C3C07724DC...	Assessor: Name: Estelle Méroc Date: 10 October 2022 Signature: DocuSigned by: Estelle Méroc 20D67DB5392B473...

I. DESCRIPTION OF THE DATA USE

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

1. Data use scenario
Data use scenario: ☒ New data collection……..SECTION A ❑ Changes to existing data collection ……SECTION B ❑ Re-use of existing data……SECTION C ❑ Publication of private and/or public reports……SECTION D

SECTION A: New data collection

A.1. Motivation of the data request
INTEGO (Integrated Computerized Network) is a general practice (GP) registration network and database in which coded personal data has been registered since 1994 concerning the health of patients from the participating GP practices. INTEGO aims to develop an increasingly large and longitudinal database of diseases, which can provide data on the incidence and prevalence of diseases in Flanders, on medication, on laboratory results, on clinical characteristics and on specific background characteristics of patients. INTEGO is a computerized GP registration network that continuously collects epidemiological data and allows a wide range of epidemiological and operational research and supports data-driven policy.

A.2. Objective(s)
The main objectives of INTEGO are: • Epidemiological research: thanks to the longitudinal data, it is not only possible to map the incidence and prevalence of disorders for a certain year, but also to study the trend over different years. Questions that can be solved in this way are e.g., which disorders are most seen by general practitioners? What conditions does the doctor see in 2022 that were rare 20 years ago? Which one does he see less now? Is there an increase in the number of depressions? Are more antidepressants being prescribed? • Monitoring the quality of care: INTEGO makes it possible to monitor the impact of, for example, care processes or to test antibiotic prescriptions against quality indicators. The GP practices are given feedback about their quality of registration in the EHR and their quality of care. The registration network can also be used to set up cluster randomized trials, to measure the impact of certain interventions in general practice.

- Developing new hypotheses: INTEGO allows observational research, which means we cannot investigate causality. But it is perfectly possible to develop new hypotheses (e.g., the link between blood pressure and kidney function or the association between environmental pollution and health).
- Develop prediction models: Intego allows to validate existing prediction models (e.g., cardiovascular risk) and to develop new prediction models, based on the longitudinal data.

A.3. Target population

The target population consists of patients visiting the GP. Currently, the Intego population is representative of the Flemish population according to age and gender.

A.4. Population intended to be covered by the new data collection

The population covered by the data collection is the same as the target population.

A.5. Study design

The INTEGO project is a GP morbidity registry.

Often case-control designs are used to perform the analyses.

Also, the case-crossover design will be used to study the association between health and environmental factors, for instance.

A.6. Variables

Information on the following groups of variables is collected:

- Demographic data
- Diagnosis
- Laboratory results
- Vaccines
- Allergies – intolerances
- Medication prescriptions
- Clinical parameters
- Therapeutic history
- Family history
- Care trajectories
- Absence certificates
- Physiotherapy prescriptions
- Radiology prescriptions

A.7. Data/statistical analyses planned
In Intego, we perform retrospective analyses. These analyses, in general, encompass: - Trend analyses for prevalence and incidence using joinpoint regression analysis - Regression analysis - Survival statistics, using for instance cox regression analysis - Risk prediction modelling - Spatio-temporal analyses

A.8. Plans for disseminating and communicating study results, including target audience
The coded personal data is accessible to researchers working at the Academic Centre of General Practice (ACHG) of the KU Leuven, and researchers outside KU Leuven who are working on Intego through interuniversity collaboration. Aggregated level data is accessible to: - ACHG – KU Leuven - Participating GP centres through the feedback reports on registration quality and quality of care - Broad audience through the Intego website (www.intego.be) - Sponsors, partners and federal and regional ministries of health

II. SMALL CELL RISK ASSESSMENT

1. Identify direct identifiers, indirect identifiers and sensitive information
For the SCRA, we use the patient's identifier and the indirect identifiers. A list of the indirect identifiers and sensitive variables is provided in Codelist_INTEGO.xlsx file.  Codelist_INTEGO.xlsx *File name: Codelist_INTEGO.xlsx* *Classification of variables: Chukwuemeka Onwuchekwa, MD*

2. Disclosure risk assessment based on direct identifiers
There are no direct identifiers in the data.

3. Disclosure risk assessment based on indirect identifiers

Given that sample uniques (i.e. patients with a unique pattern of indirect identifiers) are more likely to be identified, one way to assess disclosure risk is to calculate the number of subjects in the sample having the same distinct pattern of indirect identifiers. This approach is called k-anonymity. It is typically required that each pattern of indirect identifiers has at least 3 sample records (k = 3) to ensure confidentiality.

The variables classified as indirect identifiers are summarized in **Table 1a** and **Table1b**.

Table 1a. Indirect identifiers (Main INTEGO datasets)

INDIRECT IDENTIFIER	DESCRIPTION
CD_PAT_SEX	Sex
CD_PAT_PLC_RESDC	Postal code of residence
NR_PAT_BIRTHY	Year of birth
DT_PAT_DOD	Date of death
CD_NATLTY	Nationality
TX_CIV_STA	Civil status
TX_PROF	Profession
CD_RIZIV_TREAT_PHYS_SPE	3 last digits RIZIV/INAMI-RIZIV/INAMI-code of the treating physician

Table 1b. Indirect identifiers (2017 INTEGO datasets)

INDIRECT IDENTIFIER	DESCRIPTION
CD_PAT_SEX	Sex
TX_PAT_BIRTHY	Year of birth

K-anonymity

As displayed in **Table 2a,** the risk of identity disclosure in the Main INTEGO datasets based on the indirect identifiers is fairly high, with 32% and 44% of the 747,705 samples violating 2- and

3-anonymity, respectively. On the other hand, the risk of disclosure in the 2017 INTEGO datasets is almost null (3-anonymity<0.01%) (**Table 2b**).

Table 2a. K-anonymity Main datasets: based on the indirect identifiers: sex, postal code of residence, year of birth, date of death, Nationality, Civil status, Profession and RIZIV/INAMI-code of the treating physician (set 1)

	Sample	
K-anonymity	Nr violations	%violations
2	239,276	32
3	331,078	44

Table 2b. K-anonymity 2017 datasets: based on the indirect identifiers: sex and year of birth

	Sample	
K-anonymity	Nr violations	%violations
2	1	<0.01
3	7	<0.01

To find guidance on how to improve the k-anonymity in the Main INTEGO datasets, we calculated k-anonymity using a leave one-out procedure, excluding one variable at a time (**Table 3**). The variables that yielded the highest reductions in 2-anonymity were CD_PAT_PLC_RESDC and NR_PAT_BIRTHY (reductions of 25%).

Table 3. 2-anonymity based on leave-on-out procedure, (set 1).

Variables	Nr violations	%reduction in violations
All	239,276	NA
Excl. Sex	181,932	8
Excl. Postal code	48,934	25
Excl. Year of birth	53,171	25
Excl. Date of death	229,569	1
Excl. Nationality	185,351	7
Excl. Civil status	214,109	3
Excl. Profession	189,996	7
Excl. Treating physician	168,272	9

Excl. – excluding NA – not applicable

After discussing with the INTEGO researchers, we agreed to keep the CD_PAT_PLC_RESDC and NR_PAT_BIRTHY which are key variables in the project, and, instead, to drop the three variables CD_NATLTY, TX_CIV_STA and TX_PROF. Secondly, it was decided to pseudonymize the CD_RIZIV_TREAT_PHYS_SPE. As shown in **Table 4,** these two steps would lead to a final 3-anomymity of 8%.

Table 4. Changes in K-anonymity for the different data management steps.

Variables	K-anonymity	%violations (sample)
Set 1: sex, postal code of residence, year of birth, date of death, Nationality, Civil status, Profession and RIZIV/INAMI-code of the treating physician	2	32
	3	44
Set 2: sex, postal code of residence, year of birth, date of death, and RIZIV/INAMI-code of the treating physician [Drop Nationality, Civil status, Profession]	2	11
	3	17
Set 3: sex, postal code of residence, year of birth, and date of death [Pseudonymize Code of treating physician]	2	5
	3	8

4. Impact of potential disclosure

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. Sensitive information from patients that do not violate 2-anonymity might still be disclosed if all patients sharing the same pattern of identifying variables share the same sensitive information. We therefore calculated per distinct pattern of identifying variables, the proportion of patients with sensitive values for each sensitive variable in turn. If this proportion equals 1 or is very high (>0.95), sensitive information might get disclosed for all persons with that combination of identifying variables. We call this statistic M-sensitivity.

A subset (n=10) of the sensitive classified variables (from the main Intego datasets) used for the M-sensitivity is summarized in **Table 5**.

Table 5. Sensitive variables and sensitive values

Sensitive variable	Description of sensitive variable	Sensitive values
IDC_ALLER	Allergen ID	Any value except missing
TX_CRE_ELEM_TTL	Care element title (free text)	Any value except missing
TX_ICPC2	ICPC code of diagnosis (free text)	Any value except missing
TX_ICD10	ICD code of diagnosis (free text)	Any value except missing
TX_DIAGS	Diagnosis (free text)	Any value except missing
TX_MEDICT_ATC	Medication ATC code (free text)	Any value except missing
TX_PARAM_VAL	Parameter value (free text)	Any value except missing
IDC_VACCI	Vaccine ID	Any value except missing
TX_VACCI_ATC	Vaccine ATC code (free tect)	Any value except missing
TX_VACCI_CNK	Vaccine CNK code (free text)	Any value except missing

There are 36,779 unique patients (5%) that violate 2-anonymity, hence their sensitive information is revealed upon identity disclosure.

Using all the indirect identifiers we applied the M-sensitivity statistic. We used the Set 3 of indirect identifiers (**Table 4**).

Commented [BV1]: This will be added?

There are 47,827 distinct patterns of the identifying variables with the proportion of patients having a sensitive value being higher than 0.95. The patterns can be broken down as follows:

• For 7,946 (1%) patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding IDC_ALLER

• For 35 (<0.01%) patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TX_CRE_ELEM_TTL

• For 747 (0.1%) patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TX_ICD10

• For 5 (<0.01%) patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TX_ICPC2

• For 52 (0.01%) patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TX_DIAGS

• For 2,923 (0.4%) patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TX_MEDICT_ATC

• For 16,864 (2%) patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TX_PARAM_VAL

• For 26,122 (4%) patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding IDC_VACCI

• For 1,684 (0.2%) patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TX_VACCI_ATC

• For 23,257 (3%) patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TX_VACCI_CNK

Based on this impact assessment, we see that the sensitive variables are well protected. According to the M-sensitivity statistic, sensitive data on vaccination (IDC_VACCI and TX_VACCI_CNK) could potentially be revealed for respectively 4% and 3% of the patients, which is the most likely sensitive data that could be revealed and is low.

5. Recommended disclosure control strategies

In order to reduce the risk of patient's identity disclosure, we recommend in the main Intego datasets to:

1. Drop CD_NATLTY, TX_CIV_STA and TX_PROF (V_INTEGO_FPI_SCRA_20220901)
2. Pseudonymize CD_RIZIV_TREAT_PHYS_SPE (V_INTEGO_ABS_CERT_SCRA_20220901, V_INTEGO_ALLER_SCRA_20220901, V_INTEGO_CRE_APPR_SCRA_20220901, V_INTEGO_CRE_ELEM_SCRA_20220901, V_INTEGO_FAM_HISTY_SCRA_20220901, V_INTEGO_LRE_SCRA_20220901, V_INTEGO_MED_IMAG_SCRA_20220901, V_INTEGO_MED_PROC_SCRA_20220901, V_INTEGO_MEDICT_IN_SCRA_20220901, V_INTEGO_MEDICT_SCRA_20220901, V_INTEGO_PARAM_SCRA_20220901, V_INTEGO_PHYSIO_PR_SCRA_20220901, V_INTEGO_SOEP_RUL_SCRA_20220901, V_INTEGO_VACCI_SCRA_20220901)

With these mitigation measures, the proportion of patients violating 2-anonymity within the sample drops to 5% (n=36,779) which is completely acceptable. Moreover, the impact assessment showed that the sensitive variables were well protected.
No further actions are needed in order to reduce the impact of potential disclosure of any sensitive information. Data collections for future years do not require an additional SCRA, if the variables collected remain the same. In case new variables will be collected that are either sensitive or can be used as indirect identifiers, their impact on the SCRA will need to be assessed.