

SMALL CELL RISK ASSESSMENT

**Electronic Laboratory Medicine ordering with evidence-based
Order sets in primary care study – ELMO**

PURPOSE: SMALL CELL RISKS ANALYSIS FOR THE ELMO STUDY

SMALL CELL RISK ASSESSMENT FORM
FOR DATA USE WITHIN THE HEALTHDATA.BE PLATFORM
V1.0 (final)

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.

Electronic Laboratory Medicine ordering with evidence-based Order sets in primary care


Applicant

Institution	Catholic University Leuven, Academic centre for general practice
Address	Kapucijnenvoer 33 blok j, 3000 Leuven
Principal investigator	Prof Bert Aertgeerts bert.aertgeerts@kuleuven.be

Disclosure risk assessor

Institution	P95
Address	Koning Leopold III laan 1 3001 Heverlee BELGIUM
Assessors	Kaatje Bollaerts (statistician) : kaatje.bollaerts@p-95.com Margarita Riera (MD) : margarita.riera@p-95.com Maria Alexandridou (data analyst): maria.alexandridou@p-95.com

Signature

<u>Applicant:</u>	<u>Assessor:</u>
Name:	Name: Kaatje Bollaerts
Date:	Date: 20/01/2020
Signature:	 Signature:

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

* Should be aligned with the authorization request

1. Data use scenario

Data use scenario:

- ☒ New data collection.....SECTION A
- ☐ Changes to existing data collectionSECTION 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 (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.

In 2015, the minister of Health commissioned the Belgian Health Care Knowledge Centre (KCE) to start a programme to improve the efficiency and effectiveness of the Belgian health care system. The proposal "The effect of evidence-based order sets within a computerised physician order entry (CPOE) system on the quantity and quality of laboratory test ordering in family practice: a cluster randomised trial" was selected and financed by KCE. Order sets are a form of clinical decision support systems (CDSSs), where a limited set of evidence-based tests are proposed for a series of indications, integrated in a CPOE. This study aims to evaluate the effect of order sets on the quality and quantity of laboratory tests orders by physicians and, moreover, to evaluate the effect of order sets on diagnostic error and explore the effect on downstream or cascade activities. To achieve these aims, electronic health record (EHR)-based data is collected from all patients who are treated by a primary care practice (PCP) affiliated to one of three collaborating laboratories; Medisch Centrum Huisartsen (MCH), Algemeen Medisch Laboratorium (AML) or Anacura. A cluster randomized trial will be planned with four levels of clustering comprising the tests, patients, treating physician and the PCP.

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.

The project has two main objectives:

- To evaluate the effect of order sets on the quality and quantity of laboratory test orders by physicians for 17 common indications

235856003 - Disorder of liver (disorder)
 236071009 - Chronic diarrhea (disorder)
 264580006 - Thyroid dysfunction (disorder)
 271737000 - Anemia (disorder)
 38341003 - Hypertensive disorder, systemic arterial (disorder)
 394659003 - Acute coronary syndrome (disorder)
 409966000 - Acute diarrhea (disorder)
 44054006- Diabetes mellitus type 2 (disorder) |
 49601007 - Disorder of the cardiovascular system (disorder)
 59282003 - Pulmonary embolism (disorder)
 69896004 - Rheumatoid arthritis (disorder)
 709044004 - Chronic kidney disease (disorder)
 8098009 - Sexually transmitted infectious disease (disorder)
 84114007 - Heart failure (disorder)
 84229001 - Fatigue (finding)
 90560007 - Gout (disorder)

- To evaluate the effect of order sets on diagnostic error and explore the effect on downstream or cascade activities

A.3. Target population

Describe the reference or target population, its key features and size.

The data collected from electronic health records (eHR) concerns patients from Belgian PCPs that are affiliated to one of three collaborating laboratories; namely MCH, AML or Anacura.

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.

Inclusion criteria

PCPs will be considered eligible if all the physicians active in the practice agree to be involved in the study. All physicians will be considered eligible if they:


- Collaborate with either one of three collaborating laboratories (MCH, Anacura or AML)
- Agree to use the online CPOE for their laboratory test orders
- Use a computerised eHR for patient care
- Have little or no experience in the use of order sets within a CPOE
- Agree to the terms in the clinical study agreement

Exclusion criteria

Physicians who have experience in the use of order sets within a CPOE.
--

A.5. Study design
<i>Describe design characteristics that might be important for the small cell risk analysis.</i>
Cluster randomized study.

A.6. Variables
<i>Give an overview of the key variables (or groups) of variables of the study</i>
<p>Information on the following groups of variables is collected as part of the ELMO study:</p> <ul style="list-style-type: none"> - Patient identification number - Date of birth - Sex - Domiciliation (at the level of postal code) - Vital status: Alive/dead - If dead, date of death - CG1/CG2 (insurance status) - NIHDI-code of the physician - Date of the lab test - Total cost of the lab test panel - Study indication - Selected order set - Lab test: name, LOINC code, result, reference value(s), normal value

A.7. Data/statistical analyses planned
<p><i>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 consonant with the objectives of the project and the data collected.</i></p>
<p>Cluster randomized study. There are four levels of clusters comprising the tests, patients, treating physician and the PCP. All order sets logged between November 2017 and June 2018 will be analysed.</p>
A.8. Plans for disseminating and communicating study results, including target audience
<p><i>Describe the way the results will be disseminated and the intended target audience.</i></p>
<p>Results are disseminated in four ways:</p> <ol style="list-style-type: none"> 1) Individualized feedback reports to PCP and laboratories. 2) Global reports and cost-effectiveness reports to KCE and staff members of the centres. Centres will receive an individualized feedback report that allows them to compare their data to the average value of the other centres. 3) Scientific publications. 4) Global reports will be also become available to the public via healthstat.be and to sponsors, partners, federal and regional ministers of public health.
A.9. Codebook
<p><i>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).</i></p>
<p>The codebook can be found at Codelist_KCELABGP_LAB.xlsx</p> <div data-bbox="165 1397 373 1518">  <p>Codelist_KCELABGP_L AB.xlsx</p> </div>

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: Codelist_KCELABGP_LAB.xlsx

Classification of variables: Marga Riera, 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 are no direct identifiers in the data, patients receive an internal patient ID.

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.

For the small cell risk analysis (SCRA), we use the patient's identifier, the indirect identifiers, the sensitive variables and the date of laboratory test order. A list of the indirect identifiers is provided in Codelist_KCELABGP_LAB.xlsx file.

If there were multiple records for a patient, we used the latest socio-demographic information reported. If the latest record had missing information and the information was filled in previously, the information from the previous record was carried forward to the latest record.

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 1.

Table 1. Indirect identifiers

Indirect identifier	Description of indirect identifier
CD_DATA_PROV	NIHDI-code of the data provider
CD_PAT_PLC_RESDC	Postal code
CD_PAT_SEX	Sex
CD_RIZIV_TREAT_PHYS	NIHDI-code of the treating physician
DT_PAT_DOB	Date of birth
DT_PAT_DOD	Date of death
FL_PAT_DECEA	Deceased
TXT_DP_IDN_VAL	NIHDI-code of the data provider

Date of death was missing for all records. Deceased status was either missing or false. These variables were therefore not included in the current small cell risk analysis. TXT_DP_IDN_VAL is the same as the NIHDI-code of the data provider (CD_RIZIV_TREAT_PHYS), and is therefore omitted from the analysis.

K-anonymity

For this data, there are 6,940 unique patients out of a total population with approximately 13,000 patients, yielding a sampling fraction of 53.4%.

We checked 2-anonymity in our sample (Table 2) using the indirect identifiers listed in Table 1 with the exception of TXT_DP_IDN_VAL, DT_PAT_DOD and FL_PAT_DECEA.

Table 2. K-anonymity: based on the indirect identifiers: NIHDI-code of the data provider, Postal code, Sex, NIHDI-code of the treating physician, date of birth (set 1).

Sample

K-anonymity	Nr violations	% violations
2	6938	100
3	6940	100

To find guidance on how to improve the k-anonymity, we calculated k-anonymity using a leave-one-out procedure, excluding one variable at a time (Table 3). The variable that yielded the highest reductions in k-anonymity is date of birth (reduction 83.5%).

Table 3. 2-anonymity based on leave-on-out procedure, NIHDI-code of the data provider, Postal code, Sex, NIHDI-code of the treating physician, date of birth (set 1).

Variables	Number violations	% reduction in violations
All	6938	.
Excl. data provider	6938	0
Excl. treating physician	6900	0.5
Excl. postal code	6932	0.1
Excl. sex	6936	0
Excl. date of birth	1145	83.5

Excl. – excluding

As a first step to reduce 2-anonymity, date of birth was replaced by year of birth, leading to a reduction from 100% to 82.3%. Next, we suggest anonymizing treating physician (after which it no longer is an indirect identifier), leading to a reduction from 82.3% to 37.5%. Finally, the postal code was replaced by the province, leading to a reduction from 37.5% to 4.2%.

These different steps and corresponding reductions in k-anonymity are summarized in Table 4.

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

Variables	K-anonymity	% violations (sample)
Set 1: NIHDI-code of the data provider, Postal code, Sex, NIHDI-code of the treating physician, date of birth	2	100
	3	100

Set 2: NIHDI-code of the data	2	82.3
provider, Postal code, Sex, NIHDI-code of the treating physician, year of birth	3	96.7
[date of birth to year of birth]	4	99.5
Set 3: NIHDI-code of the data	2	37.5
provider, Postal code, Sex, year of birth	3	55.9
[anonymize treating physician]	4	66.6
Set 4: NIHDI-code of the data	2	4.2
provider, province, Sex, year of birth	3	8.3
[postal code to province]	4	12.0

4. Impact of potential disclosure		
<i>Assess the impact of potential disclosure based on the sensitive variables.</i>		
<p>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.</p> <p>The variables classified as sensitive variables and their sensitive values are summarized in Table 5.</p>		
Table 5. Sensitive variables and sensitive values		
Sensitive variable	Description of sensitive variable	Sensitive values
CD_STDY_INDICATN	Study indication	Any value
TX_STDY_INDICATN_OTH	Study indication, if other	Any value
TX_LAB_TST_RESULT	Laboratory test result	Any abnormal value

To assess the impact of CD_STDY_INDICATN and TX_STDY_INDICATN_OTH, new variables have been created. CD_STDY_INDICATN, in combination with TX_STDY_INDICATN_OTH, has 19 different values (disorder of liver, chronic diarrhoea, thyroid dysfunction, etc.), each one of them being sensitive. Therefore, 19 new binary variables have been created (TRNSFM_study1-TRNSFM_study19) in order to calculate the proportion of patients for each sensitive value per distinct pattern of indirect identifiers. TRNSFM_study1 indicates whether the subject has had at least one test regarding disorder of liver, TRNSFM_study2 indicates whether the subject has had at least one test regarding chronic diarrhoea, and so on. The mapping of the new variables is given in *0928_Mapping reflists KCELAB_IP_MA.xlsx*. Finally, TX_LAB_TST_RESULT was transformed into a new variable (TRNSFM_result), which indicates whether the patient has had at least one abnormal lab test result.



0928_Mapping
reflists KCELAB_IP_MA

There are 469 distinct patterns of the identifying variables (set 4) with the proportion of patients having a sensitive value being higher than 0.95. The patterns can be broken down as follows:

- For 3023 patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding abnormal lab test results.
- For 9 patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TRNSFM_study13.
- For 2 patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TRNSFM_study14.
- For 12 patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TRNSFM_study15.
- For 33 patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TRNSFM_study17.
- For 149 patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TRNSFM_study18.
- For 10 patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TRNSFM_study19.
- For 6 patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TRNSFM_study3.
- For 7 patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TRNSFM_study4.
- For 9 patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TRNSFM_study5.

- For 27 patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TRNSFM_study8.
- For 9 patients, a distinct pattern of indirect identifiers might lead to a disclosure of sensitive information regarding TRNSFM_study9.

Based on the impact assessment, we conclude that all sensitive information regarding study indications is well protected. There are 3023 (43.6%) subjects whose information is not as well protected regarding abnormal lab test results. However, there is a strong research interest for the test results, which justifies keeping this variable.

5. Recommended disclosure control strategies

Recommendations: 26 September 2018

We recommend the following data transformations to minimize the risk of patient's identity disclosure:

1. Age group
2. Drop region
3. masking the NIHDl-code of the treating physician (creating an anonymous physician id)
4. dropping date
5. of birth, only keeping year of birth
6. recoding the place of residence to a larger geographical scale (province)
7. date of death (all missing) and diseased status (either false or missing) did not indicate any deaths and were not able to be assessed for this SCRA. In order to mitigate the disclosure risk in a future analysis, we suggest providing only the month and the year of the date of death.

After these data transformation steps, the percentage of patients violating 2-anonymity within the sample reduced from 100 to 4.2 (n = 292 patients). The percentage of patients violating 2-anonymity within the total patient population is expected to be 2.2 as the sampling fraction is estimated to be 53.4%.

Follow-up meeting: 20 January 2020

All data collections were finalized by the researchers and the following data pre-processing steps were carried out at the validation level of the Healthdata.be platform:

1. Data of birth was dropped and replaced by age (in years) at study start
2. Information on region of the patient was dropped
3. The NIHDl-code of the treating physician was masked
4. The date of death and diseased status were dropped as all patients were still alive at study end

It is therefore concluded that adequate data restrictions were implemented to protect the privacy of the patients while at the same time allowing for sufficient information to conduct high-quality research to support public health decision-making.

