# NORDCAN – protecting confidentiality in tables and graphs of cancer statistics in the Nordic countries

Version 1.0

NORDCAN, Desember 2020

# CONTENTS

# 1 Aim

The aim of this document is to:

1) Give a theoretical background to maximise the statistical use of the aggregate counts and statistics in NORDCAN while minimising the risk of unauthorized disclosure of personal health information
2) Advise data managers in the Nordic countries on how to minimise the risk of unauthorized disclosure of personal data, especially concerning health
3) Build a foundation for automated data processing in NORDCAN

## 1.1 Unauthorized disclosure of personal health information

Unauthorized disclosure of personal information/personal health information occurs if the output contains sufficient detail to identify an individual and get additional information about them, like for instance information about their personal health. Both the published information by itself and the published information combined with other, public available information must be taken into account.

## 1.2 Personal data

“Personal data” means any information relating to an identified or identifiable natural person (“data subject”); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. (1)

## 1.3 Data concerning health

“Data concerning health” means personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status. In NORDCAN-documents, this is used synonymous with personal (health) information. (1)

## 1.4 Anonymous data

Data is anonymous if it is no longer possible, with the tools that can reasonably be expected to be used, to identify individuals in the data set. Anonymous data falls out of the scope of GDPR. (2)

# 2 Background

## 2.1 NORDCAN data management and data flow

Until 2018, each of the Nordic countries transferred individual, de-identified data concerning cancer diagnosis to the NORDCAN-secretariat in Denmark where the data were quality checked, enriched with the NORDCAN entity code and anonymized before aggregate counts and statistics were transferred to IARC. In IARC, the aggregate counts and statistics were stored in a database, building the backbone of the NORDCAN web tool for displaying statistics on cancer incidence, mortality, prevalence and survival.

From 2020, all data processing of personal data within NORDCAN is carried out within the safe boundaries of each of the Nordic cancer registries. No personal data in NORDCAN should be shared or transferred outside of the secure systems in each of the registries unless the parties agree in written and the provisions of the GDPR are fulfilled. The NORDCAN IT-group has developed a series of R-modules to be run within each registry. Output of the R-modules is aggregate counts and statistics to be used by the NORDCAN web tool in IARC.

Within each registry, standard procedures entail a high security level including for instance access control, role-based access, encryption and de-identification (pseudonymisation of patient ID by a key or algorithm). Aggregate counts and statistics prepared for IARC contain no link back to the original data.

The above mentioned change of data management and data flow in NORDCAN is an important step to minimise the risk of unauthorized disclosure of personal information. However, we also need to assess the risk of unauthorized disclosure of personal information derived from the aggregate counts and statistics.

## 2.2 Main risks of unauthorized disclosure in NORDCAN aggregate counts and statistics

The main issue for NORDCAN aggregate counts and statistics is to ensure that the information transferred out of each country can be regarded as anonymous data. There are three main scenarios for disclosure risk:

1) **The risk of identifying oneself through counts and statistics in NORDCAN**
2) **The risk of identifying other persons through counts and statistics in NORDCAN**
3) **The risk of gaining additional information about identified persons through counts and statistics in NORDCAN**

This document will give a theoretical view on these risks and the means to reduce such risks, and the *NORDCAN risk assessment* will address the assessed risk of these scenarios in NORDCAN.

## 2.3 Disadvantages of not publishing NORDCAN aggregate counts and statistics

Although the risks of unauthorized disclosure of personal information is one to be taken seriously, we must not forget the disadvantages tied to scenarios where such aggregate counts and statistics are *not* published. The main scenarios for this could be:

1) **Breach of data accessibility guidelines in the Nordic countries and the FAIR[1] principles**

[1] FAIR = Findable, Accessible, Interoperable, Reusable - https://www.go-fair.org/fair-principles/

2) **Less data and lower usability for prevention and for surveillance and comparison of the cancer trends in the Nordic countries**
3) **Less data easily available for research**

It would be considered both unethical and a breach of the purpose limitation principle in Article 5 of the GDPR (3) to store data that cannot or is not used for the purpose it was collected. The main purpose of all cancer registries is contribute to the research, hereunder cancer epidemiology and cancer trends.

The *NORDCAN risk assessment* will also take these disadvantages into account.

## 2.4 NORDCAN risk assessment

A detailed risk assessment for NORDCAN can be found in the document *NORDCAN risk assessment.*

# 3 Alternative methods

Could NORDCAN have chosen an alternative method, based on newer technology, to keep a high level of detailed statistics while at the same time reducing the risk of unlawful disclosure even more?

As sharing, and also reuse, of data is both an increasing demand and an increasing risk, many efforts are put together over the last couple of years to maximise the statistical/analytical use of data while still protecting the confidentiality.

We will mention a few of the methods below. At present time, however, our evaluation is that while these methods might be promising for the future, NORDCAN needs to have an immediate, short-time solution to put into practice now.

## 3.1 Synthetic data

Synthetic data are real data which are sent through an anonymisation/synthetisation algorithm which, in short, throws the data points around so that the statistical characteristics of the dataset is still the same, but no records are identical to the original data set, hereby making identity disclosure and attribute disclosure much more difficult.

A synthetic dataset was made in 2020 to be used within the NORDCAN R-development process, so that the modules for quality assurance, enrichment and aggregation can be thoroughly tested. While this is helpful in avoiding the sharing of de-identified data and ensuring we don't do testing on real data, we have yet to assess whether this might be a solution for the future NORDCAN. Since most of the data management in NORDCAN happens within each country, the number of categories in NORDCAN is quite few and the synthetic data set is supposed to have close to identical statistical characteristics to the original dataset, there might be little to gain from going through a process of synthetisation of data.

## 3.2 Distributed analysis

Distributed analysis, or distributed learning, is a way of sharing statistical models and model parameters instead of sharing sensitive data. It is a step away from the traditional organisation of research projects where data are sent to one physical location and all analysis are done there. In distributed analysis, all data are kept safe within the organisation and only the result of an analysis is shared with others.

There are many advantages to distributed analysis, but the basic notion that the NORDCAN results still have to be anonymous to be shared with IARC and online is not solved through this method. NORDCAN will most likely explore this method in the future, but as it requires a lot of initial processes in each organisation regarding risk analysis, documentation, set up of servers and software etc., this is not a solution to be implemented in NORDCAN at present time.

## 3.3 Homomorphic encryption

Homomorphic encryption is a form of encryption that allows computation on the encrypted data, generating an encrypted result which, when decrypted, matches the result of the operations as if they had been performed on the unencrypted data.

This is a promising direction for protecting confidentiality, but the technology is still in its infancy and is presently not readily implemented. Massive computing power is needed and the processes are time consuming, making current schemes not viable for online use. Handling of encryption keys, for decrypting results, is also a not straight forward in an online tool.

# 4 Theoretical background for risk assessment and strategies for disclosure control

Most of the theoretical background in this chapter is taken from the British Office for National Statistics, "Policy on protecting confidentiality in tables of birth and death statistics". (4)

## 4.1 Risk assessment

When assessing the risk, a key point is that small numbers, even unique cases, are not necessarily disclosive.

What needs to be assessed is:

- Can any individual be identified from the table, with any degree of certainty?
- If an individual can be identified, is any new information revealed about them?
- If an individual can be identified, is any information revealed about other persons connected with them?
- Is the disclosed information sensitive?
- Is there a possible harm to the individual by the disclosure?
- Who will benefit from a disclosure?

All health data are considered to be sensitive.

There is usually not an unauthorized disclosure when:

- A person can be identified (recognition), but nothing new can be learnt about them
- Persons can identify themselves (self-recognition), but others cannot identify them
- A person believes he/she can be identified or can recognise someone else, but there is no certainty that this is the case

The distinction between recognition of an individual's identity without revealing new or sensitive information about them ("simple recognition") and revealing new, sensitive information about the person ("attribute disclosure") is important.

## 4.2 General practical advice

- Consider a 0, 1 or 2 in any table cell as having potential disclosure risk
- Take into account row and column totals, whether presented in the table or not, when assessing the risk
- Take into account the underlying population total and geographical area
- A table with many dimensions, variables or small counts is likely to be riskier than one with fewer
- Overlapping geographical areas and rolling multi-year aggregates can give rise to disclosure through differencing
- A rate or other statistics may be disclosive if it is possible to infer that the underlying counts contain small numbers, even if the numbers are not published

## 4.3 Types of disclosure risk

There are several types of disclosure risks:

| Type of risk | Description | Risk level | Comment |
|---|---|---|---|
| **Identity disclosure** | The risk of identifying oneself of another person in the dataset | Relatively low, but be aware | Identification itself poses a relatively low risk, but the identification might lead to a discovery of rareness or uniqueness for oneself. |
| **Individual attribute disclosure** | The uncovering of new/not previously known information about a person through the use of published data | High | Identification (identity disclosure) is a necessary precondition for this to occur. |
| **Group attribute disclosure / profiling** | Learning a new attribute about an identifiable group or learning a group does not have a particular attribute | Evaluation needed | This is a much neglected threat, and does not require individual identification. |
| **Within group disclosure** | The uncovering of new/not previously known information about one or several individuals from a group where all persons fall into two categories and one of the categories only contains 1 individual | High | This is a combination of both identity and attribute disclosure types. |
| **Disclosure by differencing** | The use of two or more overlapping tables and subtraction to gather additional information about the differences between them | High | This is one of the possible disclosure scenarios if other tables exist on the same group of patients. |
| **Perception of disclosure risk** | The perception by individuals or the general public that because there are small counts, no/insufficient protection has been applied | Low | This is a low-risk scenario, but might discredit the data/the organisation even if there is no unlawful disclosure. |

## 4.4 Strategies for disclosure control (SDC)

If the risk assessment, after taking into account technological developments, existing published data, sensitivity of the data, potential harm and benefits for the individual, benefits for public health and disease control and the development in treatment, reveals a too high risk of unauthorized disclosure of personal information, one should take use of one or more strategies for disclosure control.

Some strategies can be defined as “perturbative” – meaning that they “disturb” or alter the underlying data – for instance through approximation, and some can be defined as “non-

perturbative" – non-disturbing – where the original data might be hidden but not disturbed or altered. There are advantages and disadvantages to all methods, so a thorough evaluation of which method(s) best fit the data is necessary.

## 4.5 Non-perturbative methods

The most common non-perturbative method is to redesign the table, collapsing categories to reduce the sparsity of the table. For instance:

- Collapse categories, merge categories of low counts with neighbouring categories
- Collapse breakdowns in the table, for instance combine data years, combine sexes or use broader age groups
- Collapse/recode selected categories or breakdowns, for instance broadening the youngest and the oldest age groups.
- Anonymise/recode categories, for instance replace named geographical areas with a breakdown of another geography-related characteristic, such as deprivation quintile
- Split a multi-dimension table into two or more separate tables

## 4.6 Perturbative methods

The most common perturbative methods are:

- Suppression of data by replacing low counts by a symbol (such as "-"). To ensure that the value cannot be discovered through differencing, additional cells also need to be suppressed (secondary suppression)
- Rounding cell counts. The most common is to round small numbers to either 5 or 10, but it is also possible to round for instance rates to a lesser level of precision by having fewer decimal places

# 5 Literature


(1) EU. **Article 4, GDPR - Definitions** https://gdpr-info.eu/art-4-gdpr/

(2) A guide to the anonymisation of personal data
https://www.datatilsynet.no/en/regulations-and-tools/reports-on-specific-subjects/anonymisation/

(3) EU. **Article 5, GDPR – Principles relating to processing of personal data** https://gdpr-info.eu/art-5-gdpr/

(4) Policy on protection confidentiality in tables of birth and death statistics
(https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyonprotectingconfidentialityintablesofbirthanddeathstatistics)

(5) Statistical disclosure control
https://www.wiley.com/en-no/Statistical+Disclosure+Control-p-9781119978152

(6) EU Data protection rules
https://ec.europa.eu/info/law/law-topic/data-protection_en

(7) EU GDPR Recital 26
https://www.privacy-regulation.eu/en/recital-26-GDPR.htm

(8) STINE DPIA and Risk- and vulnerability analysis
Documents describing the Cancer Registry of Norway Statistics Bank. Can be obtained by sending a request to the Cancer Registry of Norway

(9) Synthetic data
https://en.wikipedia.org/wiki/Synthetic_data

(10) Distributed learning
https://vantage6.ai/

(11) Homomorphic encryption
https://en.wikipedia.org/wiki/Homomorphic_encryption