CNS 12/2021

Opinion in relation to the consultation made by a public sector foundation on the development of a mathematical model for the prevention of epidemics and pandemics

A letter from the Foundation (...) is presented to the Catalan Data Protection Authority in which it is requested that the Authority issue an opinion on the development of a mathematical model for the prevention of epidemics and pandemics based on the use of data provided by certain public entities.

Specifically, it is considered whether the mechanisms implemented by these entities adequately guarantee the anonymization process of the set of data that each of them provides to the Foundation in order to be able to develop the aforementioned mathematical model.

The consultation is accompanied by the document "BIG DATA PROJECT FOR THE PREVENTION OF EPIDEMICS AND PANDEMICS".

Analyzed the query, given the current applicable regulations and in accordance with the report of the Area of Technology and Information Security of the Authority and the report of the Legal Advice I issue the following opinion:

I

(...)

II

The Foundation states in its consultation that, together with other entities, it is carrying out a "Project" which consists in the development of a mathematical model for the prevention of epidemics and pandemics based on the use of data provided by certain public entities.

According to the document "BIG DATA PROJECT FOR THE PREVENTION OF EPIDEMICS AND PANDEMICS", attached to the consultation, it is, in particular, to develop a mathematical model that analyzes the possible correlations between the mobility of citizens and the ratio of infection, enabling the quantitative incidence of the mobility of citizens with the incidence of epidemics and meteorological conditions.

This mathematical model will be made available to the research teams of the scientific community and the public health and safety managers of the Public Administrations participating in the Project.

The foundation states in said document that it is not, in any case, a traceability project that allows decisions to be made about specific individuals or the processing of personal data that identify natural persons, given that the personal data will have been anonymized and aggregated at source by the entities participating in the Project.

In this sense, it is pointed out that the mathematical model will be nourished by the following data:

- Data from the Public Data Analytics Program for Health Research and Innovation (“PADRIS”), managed by the Health Quality and Assessment Agency of Catalonia (AQuAS).
- Mobility data in Spain during the period of the COVID-19 pandemic, published by the Ministerio de Transportes, Movilidad y Agenda Urbana (MITMA).
- Open data from the National Institute of Statistics (INE). •

Environmental data, both meteorological and pollutant data, provided by the University of Girona.

The Foundation has announced that the following entities participate in the "Project":

a) Germans Trias i Pujol Hospital and the Fight Against AIDS Foundation, which provide scientific knowledge (epidemiology). b) The Center of Innovation for Data tech and Artificial Intelligence (CDIAI), which will develop the pandemic model for future prevention.

c) The Barcelona Supercomputing Center (BSC-CNS), which will develop the database model to analyze and monitor cases related to the pandemic.

d) The Rovira i Virgili University (URV), which will evaluate the mathematical model. e) The Global Health Institute of Barcelona (ISGlobal), which provides scientific advice on health matter f)

EURECAT, technical manager of the Project, which provides resources and facilities.

It also points out that the "Project" is framed in the provisions of Recommendation (EU) 2020/518 of the European Commission of April 8, 2020, relating to a set of common instruments of the Union for the use of technology and data in order to combat and overcome the COVID-19 crisis, in particular with respect to mobile applications and the use of anonymized mobility data.

In addition to all of this, the Foundation requests this Authority's assessment of the adequacy of the data anonymization procedures mentioned at the outset, this is from the entities that are responsible, for in order to guarantee that the mathematical model object of the "Project" can be developed without generating risks for the privacy of natural persons.

III

Point out that the principles and guarantees of data protection do not apply to anonymous information, that is to say, to that information that has lost all direct or indirect connection with the natural person - or that has no longer had it since of its obtaining-, so that the affected person is no longer identifiable without disproportionate efforts.

This is clear from recital 26 of Regulation (EU) 2016/679, of the Parliament and of the European Council, of April 27, 2016, General Data Protection Regulation (hereinafter, RGPD):

> "The principles of data protection must be applied to all information relating to an identified or identifiable natural person. Pseudonymized personal data, which could be attributed to a natural person through the use of additional information, must be considered information about an identifiable natural person. To determine whether a natural person is identifiable, all means, such as identification, that can reasonably be used by the data controller or any other person to directly or indirectly identify the natural person must be taken into account. To determine whether there is a reasonable probability that means will be used to identify a natural person, all objective factors must be taken into account, such as the costs and time required for identification, taking into account both the

> technology available at the time of treatment and technological advances. Therefore, the principles of data protection should not be applied to anonymous information, that is, information that is not related to an identified or identifiable natural person, nor to data converted into anonymous data in such a way that the interested party is not identifiable, or to be Consequently, this Regulation does not affect the treatment of said anonymous information, including for statistical or research purposes."

The consultation asks whether the anonymization techniques applied by the entities that will provide the data that must serve as the basis for the development of the mathematical model by the Foundation and the other entities participating in the "Project" guarantee that we are facing anonymized data processing.

It should be clarified that any anonymization procedure, applied to personal data, must aim to destroy the link or nexus between the personal data and the affected natural person, to whom this information refers. The aim is that the affected person cannot be identified by third parties without disproportionate effort.

While this link between the data and the natural person to which it refers can be reconstructed in a relatively simple way - in this sense, it is necessary to consider all the objective factors, such as the costs and time required for identification, having taking into account both the technology available at the time of treatment and technological advances -, it cannot be considered that the information has been subject to an appropriate anonymization procedure and will remain subject to the principles and obligations of the data protection regulations.

Agree that the examination of the problem referred to in this consultation is carried out, immediately, taking into account the information described in the document "BIG DATA PROJECT FOR THE PREVENTION OF EPIDEMICS AND PANDEMICS", to which has mentioned in the previous foundation, as well as the documentation that is cited in this same document.

Before that, however, it should be pointed out, given the terms in which the query is formulated, that this opinion focuses on examining the quality of the anonymization of the data at source and the methodology of the "Project" in relation to this data, for the purposes of 'assess the risks of possible subsequent re-identification of natural persons.

As this Authority has agreed in previous opinions (for example, CNS 10/2016, CNS 52/2015, CNS 29/2015, CNS 20/2015 or CNS 34/2014, available on the website https://apdcat .gencat.cat/ca/inici), in the environment of big data and with the possibilities offered by techniques such as data mining and artificial intelligence - concepts to which the "Project" refers (section 4 of the document attached)-, the crossing of information obtained from various sources, even if it has been anonymized, may end up making a person identifiable.

That is to say, depending on the volume of data that, in the context of this "Project", is made available to the bodies that participate, and according to the form in which it is offered, the possibility that the combination of this information obtained from various sources may end up making specific people identifiable should not be ruled out (from general traits, the number of individuals at the intersection of all of them decreases until specific people are identified).

In this same sense, the Article 29 Working Group (hereinafter, GTA29) has pronounced in its Opinion 5/2014 on anonymization techniques. In this opinion, to which we refer, the following is agreed:

> "(...) those responsible for the treatment must be aware that an anonymized set of data can still carry residual risks for the interested parties.
> Effectively, on the one hand, anonymization and re-identification are fields of

> active research in which new discoveries are regularly published and, on the other hand, even anonymized data, such as statistics, can be used to enrich the existing profiles of people, with the consequent creation of new data protection problems. In short, anonymization should not be seen as a sporadic procedure, and those responsible for data processing must regularly evaluate existing risks.”

As WG29 points out, the risk of re-identification is inherent in any anonymization technique, so the privacy and right to data protection of the owner could be compromised, even though the data has been anonymized.

For this reason, it is necessary to always carry out an analysis of possible risks of re-identification and, in view of the result obtained, articulate the necessary measures to mitigate the probability of them materializing, even anticipating reactive measures to mitigate the possible damage that could be caused to a natural person if said re-identification took place. These measures or guarantees must be higher in those cases in which special categories of data are treated (as is the case in the present case), given that the risk is greater in view of the greater impact that this re-identification would represent, if materialized, on the rights and freedoms of the people affected.

This identification and analysis of the risk of re-identification should be understood in the present case as an activity framed within the data protection impact assessment (PIA) referred to in article 35 of the RGPD.

The RGPD requires an impact assessment on privacy "when it is likely that a type of treatment, in particular if it uses new technologies, by its nature, scope, context or purposes, entails a high risk for the rights and freedoms of physical persons"
(article 35.1). And it expressly mentions as a case in which an impact assessment will need to be carried out, the large-scale processing of special categories of data, such as, for example, health data (art. 35.2.b)) or the systematic and comprehensive assessment that allow the creation of profiles

In relation to this impact assessment, the LOPDGDD lists, in its article 28.2, some cases in which the existence of a high risk for the rights and freedoms of people is considered likely, among which "when produces a massive treatment that involves a large number of affected or entails the collection of a large amount of personal data" (letter d).

Although, as has been said, the data protection regulations do not apply to the treatment of anonymous data and therefore a priori the performance of a PIA would not be required in this case, given that it is a procedure that seeks to identify and control the risks to the rights and freedoms of people associated with the processing of data (in this case, of anonymized data) and that, as we have seen, the risk of re-identification is inherent in any anonymization technique, the fact that the examined "Project" is based on the combined use of anonymized data highlights the convenience, at least, of carrying it out in part (a complete process should not necessarily be carried out) of a PIA that allows to measure, evaluate and manage the risk of re-identification.

For these purposes, it may be of interest to consult the "Guide on impact assessment relating to data protection in the RGPD", available on the Authority's website.

In any case, in order to answer this query, each of the sets of data that will be made available to the Foundation, and to the rest of the participants, for the development of the "Project", for the purposes of determining the quality of the anonymization offered at the origin and being able to determine, based on this, if there is a risk of ending up re-identifying the affected people.

Environmental data from the University of Girona has been excluded from this analysis, given that this type of information cannot be classified as personal data, in accordance with article 4.1) of the RGPD.

IV

Evaluation of the level of anonymization of health data

The data that the Department of Health will contribute to the "Project" come from the PADRIS program, managed by AQuAS.

Specifically, the following data will be available:

• Basic Health Area (ABS). • Age range. • Date of first positive PCR. • Number of hospitalizations. • Average days of hospitalization. • Number of ICU admissions. • Average days of admission in ICU. • Hospital mortality. • Non-hospital mortality at 30 and 60 days. • Non-hospital mortality after more than 60 days.

The specified health data show that the information provided for the implementation of the "Project" includes different information on people who have suffered from COVID-19, which is provided in an aggregated manner depending on the ABS and the rank of age

However, special attention must be paid to the "Date of first positive PCR" field since, as described, it is a field that would contain information relating to a specific person. Therefore, in areas with a small incidence, it cannot be ruled out that a person can end up being identified based on their age and the fact that they have been diagnosed positive.

Also the fields "number of hospitalizations" and "number of ICU admissions", which, as described, would seem to refer to a specific person.

The rest of the fields are aggregated data (averages and totals) at the ABS level and age range. The quality of the anonymization of these fields depends on the number of people contributing to each of the groups determined by the ABS and the age range.

In this sense, the time period covered by each age range is unknown, but it is important, in order to reduce the risk of re-identification, that the age ranges, or even the time ranges in terms of the date of the first positive PCR, are sufficiently wide, especially in those areas where the number of incidents is lower.

In general, given the size of the ABS (between 2100 and 50000 people approx.) it could be thought that the aggregation is large enough and that it is difficult to obtain information on a specific person (see in this regard, the population data of reference to the CatSalut Central Population Register, available on the website https://catsalut.gencat.cat/ca/inici/). However, bearing in mind that the aggregation is also done by age and that the incidence of COVID-19 is very variable, it cannot be ruled out that some of these aggregations refer to a small number of people.

The key factor in determining whether the aggregations provide sufficient protection is the "Number of hospitalizations" field. When the value of this field is small, the effect of each person on the other fields will be very high, and therefore it might be possible to get information from a specific person.

For example, if the number of hospitalizations is 1, the average number of days of hospitalization corresponds to the number of days that person has been hospitalized. Similarly, if the number of hospitalizations is 2, anyone who knew the data of one of the hospitalized could find out the data of the other.

In order to avoid these risks, you should choose not to use data when it refers to a small group of people.

To point out, at this point, that point 7 of section 5.4 "Privacy guarantees of personal data" of the PADRIS Report - document to which the query refers - foresees, as one of the mechanisms that AQuAS will use to anonymize in origin the PADRIS data, which "will apply a specific anonymization process for each work."

Given this, it cannot be ruled out that some specific type of anonymization has been carried out (beyond the aggregation by ABS and age range) in relation to the data provided for the development of the "Project". However, at the time of making this opinion, this fact is not known.

V

Evaluation of the level of anonymization of mobility data

The mobility data that will be used in the "Project" are open data provided by MITMA (https://www.mitma.gob.es/ministerio/covid-19/evolucion-movilidad-big-data/opendata mobilidad), which refer to mobility in Spain during the period of the COVID-19 pandemic.

According to the document "Analysis of mobility in Spain with Big Data technology during the state of alarm for the management of the COVID-19 crisis", to which the query refers, the main source of this data is anonymized records from mobile phone networks, which would have been combined with other data sources to generate origin-destination matrices and other indicators of mobility and population presence (data on land use, from the Municipal Register of Inhabitants and information from the transport network), all of them anonymised and aggrega

This same document describes the data used, the methodology and data analysis algorithms, as well as the indicators generated.

In essence, as the query shows, the mobility data correspond to the number of trips between different mobility areas that correspond to municipalities, districts and groups of these (in the case of Catalonia, a total of 442 areas have been considered). The population within these areas is, in general, greater than 5000 inhabitants and, in no case, less than 1000 inhabitants. Data on the number of trips and total kilometers aggregated are published by:

- Origin area.
- Destination area.
- Origin activity (home, work, others).
- Destination activity (home, work, others).
- Time (0, 1, ... , 23).
- Distance (ranges: 0.5-2km, 2-5km, 5-10km, 10-50km, 50-100km and >100km).

As described, in the event that there is a stop longer than one hour, the journey is already counted as a different journey.

It should be noted that, given that the data is based on information reported by a telephone company, to obtain the indicators of interest, it is necessary to extrapolate them to the population as a whole. This is done by expanding the data according to the Population Register of the INE and the penetration of the telephone operator in each area. In other words, the data on which the aggregations are based (number of trips and total kilometers) are calculated partly from real data and partly from estimates made from the census and telephone operator pe

There is a risk that if a person had a very unique movement pattern, this would allow their movements to be tracked.

As an example, in a case where it is known that a person has made an unusual journey between two Catalan towns, with a distance between them of about 300 km., on the same day, in which the person who made it states that they did not make stops, we can end up verifying whether or not the journey without stops has really been made based on the information provided (it should be borne in mind that the information provided includes the time of departure and arrival , and it also lets you know that no stops have been made for more than an hour).

In any case, note that, observing the data published by MITMA, it can be seen that whenever data on the number of trips between two areas is provided in the open, this is greater than 1,000, so the aforementioned risk would be greatly mitigated .

However, in the methodological description of the construction of the mobility data contemplated in the cited document, there is no reference to any anonymization mechanism that only allows mobility data to be given when the number of trips between two areas is large enough, so it cannot be stated that the publication of the data is always carried out with this guarantee.

VI

Evaluation of the anonymization of socio-demographic data

The set of INE data, published as open data, that will be used for the "Project" are the following:

- Average income per person (in euros). Average of the years 2015, 2016 and 2017. Variable observed at census section level.
- 2011 unemployment rate (in percentages). Variable observed at census section level.
- Percentage of population aged 65 or over in 2019. Variable observed at census section level.
- Percentage of foreigners from countries with a medium and low human development index according to the United Nations Development Program (2019). Variable observed at census section level.
- Infrastructure. Percentage of homes with less than 45 m² of living space 2011 Variable observed at census section level. •

Population density in 2019 (in inhabitants/km²). Variable observed at ABS level.

With the exception of population density, all other variables are observed at census section level.

The consultation explains how the Statistics, Econometrics and Health Research Group ("GRECS") of the University of Girona has calculated the value at ABS level for the variable corresponding to population density. In this sense, it is pointed out that using the population of each of the census sections as weights (the source of the total population of the census section and of the population of the census section by sex was INE (2020b)), calculated the weighted average of the values of the census tracts that make up the ABS to obtain its va

Considering the information available, it can be said that in this case there would be no risk of obtaining personal information from this data.

VII

In view of the considerations made in the previous sections, it must be concluded that:

The three data sets examined (health data, mobility data and socio-economic data) present the data in an aggregated form. Although this significantly reduces the risk of revealing personal information, that is, the possibility of re-identifying those affected. Although a priori it cannot be ruled out that anonymization is effective depending on the levels of aggregation that are finally applied to each of the subsets of information, except in the case of socio-economic data, the information provided does not offer sufficient guarantees that the information is offered with sufficiently broad levels of aggregation.

In the case of health data, it cannot be ruled out that the sample of individuals consisted of a single person (or a very small number of people). If so, the published data would be easily associated with a specific person. Although the PADRIS report refers to the use of specific anonymization methods in each project, given that this information is not available in relation to the present "Project" it is not possible to rule out the possibility of aggregations on small groups of individuals.

Although some of the information provided could be understood, at the level of the study, as not particularly critical (average days of hospitalization, average days in the ICU, etc.), it must be taken into account that it is in any case data relating to health (article 9 RGPD), and which could allow to know if a certain person has passed the disease, or even to reconstruct some kind of sequence regarding the concatenation of contagions in a certain area, for which which requires extreme guarantees in its treatment, even when it has been anonymized.

The aforementioned risk cannot be ruled out in the case of mobility data either, given that, although it has been observed in some of the published data files that travel data is only offered in the open when the aggregation corresponds to more than 1000, this observation is not fixed in the anonymization methodology made by MITMA.

All in all, given the detected risks of re-identification due to correlation with other sets of data, in order to be able to carry out the "Project" it would be necessary to adopt the appropriate measures to mitigate the possible consequences that could arise in the event that to materialize the re-identification of natural persons.

In this sense, it would be necessary to consider whether the development of the mathematical model to which the query refers requires the inexcusable availability of all the data examined (PADRIS, Open data mobility and sociodemographics).

The necessary application of the principle of minimization contained in article 5.1.c) RGPD (treating the minimum essential personal information) is key when processing personal data, but also when processing anonymized data.

The processing of data that may be excessive or unnecessary can considerably increase the risk of re-identification, therefore, from the data protection point of view, it is always necessary to limit the processed data to the minimum necessary to achieve the intended purpose, both from the point of view of population volume (number of records) and analyzed data (processed attributes).

It would also be necessary to maximize the level of aggregation and in the face of small or extremely small samples of individuals, mask the data relating to these groups or choose to eliminate them.

In turn, it would be necessary to offer additional guarantees in order to preserve the rights of those affected, which would imply a series of commitments by all participants in the "Project

a) Duty of confidentiality.
b) Maintain anonymity, that is to say, do not take actions to re-identify. c) Link the information delivered exclusively to the "Project", not allocating it to other purposes. d) Immediately notify the entities that provide the data of any suspicion of re-identification (for these purposes, it would be advisable to establish a protocol to be able to carry out this communication in an agile and secure way).
e) Establish a maximum retention period for the data and destroy them once they are no longer needed by the "Project" (which would need to be proven to the entities that provide the data).

In accordance with the considerations made so far in relation to the query raised, the following are made,

Conclusions

Based on the information available, it cannot be ruled out that from the crossing of the anonymized data referred to in the query in order to develop a mathematical model for the prevention of epidemics and pandemics, specific people could end up being identified , so it is necessary to measure, evaluate and manage this risk of re-identification by adopting the appropriate measures, which have been referred to, to reduce the probability of re-ide

Barcelona, March 23, 2021