Comments on the definition of personal information and on the (re)use of personal information in anonymous or pseudonymised form in the proposed general data protection regulation.

Guido van 't Noordende
Informatics Institute, University of Amsterdam[1]
Amsterdam, April 30, 2013

This report gives high-level notes on current issues with the proposed general Data Protection Regulation (DPR). The comments in this report are founded in long-standing research on data anonymisation (up to 30 years back, e.g., [Dalenius]), as well as on more recent reports that confirm that microdata data can be re-identified easily in practice. Where possible, I try to describe things in an intuitive way. I will also make some notes on protecting medical information specifically.
The final section of this report points out specific problems with concrete definitions of *anonymous data* and *pseudonymous data* based on currently proposed amendments to the DPR.
This report focuses primarily on *microdata*: data or records that belong to a single individual, for example, a table with columns where each row contains attributes that belong to an individual.

Summary of main points

- Definitions of anonymous or pseudonymous data are unnecessary if a good definition of personal data exists. Currently proposed definitions of anonymous and pseudonymous data focus narrowly on removing directly identifiable features from microdata, thus ignoring significant risks of re-identifyability of the remaining data.
- Current Article 83 creates an exemption from DPR rules that allows usage of data for "historic, statistical, and scientific research" even if identifiable, without consent. This is too lenient, particularly given the broad application of Article 83 within the DPR.
- Medical information looses its special protection in the DPR (compared to 95/46/EC) under the original definition of Article 81 in the commission's proposal. A consent requirement should be included in Article 81(2). Also, Article 83 should not permit for processing of special categories of information for historical, statistical and scientific research without explicit consent.
- Re-identification mechanisms have been described for years, and it has been shown that combination of "anonymised" (or pseudonymised) data with other (background) information is straightforward. If due to new DPR regulations more and more "anonymised" micro-information becomes available, this problem is exacerbated, as re-identifcation and/or linkage of this information with other information becomes even more straightforward.
- Pseudonymised information, due to its *inherent* property of linkability and longitudinal stability (i.e., being the same over time), may bring additional risks compared to using anonymous data. Pseudonymisation is usable as a tool for securing information during processing, but should *not* be used as a means to escape DPR rules such as accountability and transparency.

1. On defining personal data

The definition of the data protection directive 95/46/EC refers to personal data as *"any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity"* [95/46/EC].

1 The views expressed in this report are those of the author and do not necessarily reflect those of the UvA.

The standing definition of personal data is thus data which can be *directly or indirectly related to a data subject.* The proposed DPR uses a similar definition. In practice, the interpretation of what constitutes (in)directly identifiable information, is problematic. Recently, amendments for introducing definitions of anonymous and pseudonymous data were proposed that further complicate the discussion and may, as a side-effect, weaken the DPR in unforeseen ways, as it impacts the definition of what constitutes identifiable information.
A single definition of what consitutes personal information is preferable, as this avoids inconsistencies or loopholes regarding the legal regime that applies to processing of personal data.

2. Anonymous data and re-identifyability – background and introduction

Anonymous data is not, or should not be, directly or indirectly be identifiable. As a minimum, to make data anonymous, all direct identifiers (such as names, address/date of birth combinations, social security numbers) must be removed; this is called *de-identification.* However, de-identification by itself does not make information anonymous. The resulting data may still *indirectly relate* to the data subject.

Microdata is data where records that belong to a single person are not aggregated or combined with data of other people but kept as separate records in the data set. Microdata can contain 1) direct identifiers, 2) *pseudonyms* that replace directly identifiable information but that can be "mapped back" to the data subject using some table kept separately, or 3) it can be "anonymised", meaning all directly and indirectly identifiable information gets removed from the record, and it contains no pseudonyms. To an outsider that does not know the mapping table pseudonymous data may appear as anonymous (more information on pseudonyms in the next section). Anonymised microdata may not contain features that allow for direct or indirect identification of the data subject.

Research stemming years back [Dalenius], up to more recenly [Sweeney, Koot, Narayanan and Shmatikov] demonstrated that it is often straightforward to re-identify anonymous (or pseudonymised) microdata.
Recent work in the Netherlands shows that approximately 99.7% of the Dutch population is uniquely identifiable using a combination of 6-digit postal code, gender, and date of birth. Reducing this information to only contain a 4-digit postal code, leads to a percentage of 67%. Using a better "de-identification" method that removes also the day of birth, decreases the percentage to 4.7% of the population – but with 79.1% of the records still being re-identifiable to a group of 10 people or less. It is clear to see that when information is added to a data set containing a 4-digit postal code, gender, month and year of birth, it becomes straightforward to disentangle the individuals from the remaining group of 10 people. Experimental work confirms this [Koot 2012].
Earlier work underpins these results. Sweeney experimentally established that about 87% of the US population is uniquely identifiable using a combination of three demographic variables. She showed that using a (legally available) "anonymised" microdata set containing medical historic information of Massachusetts' employees, combined with an identified voter list of Cambridge, Massachusetts, she could re-identify many records in the medical data set: she was able to (correctly) associate de-identified medical information with a Massachusetts governor, among others. Narayan and Shmatikov empirically confirmed in 2007 that "anonymity" in microdata sets quickly erodes to zero when adding (background) information to it, particularly when the microdata contains sparse (rare) information. In practice, many microdata sets indeed contain sparse (rare) information [Narayan and Shmatikov].

Koot's work, among that of other scientists, shows that already a part of the postal code, the date of birth and gender information are sufficient to uniquely or next to uniquely distinguish individual records that can be uniquely mapped to data subjects in a data set. From this it is easy to see how

adding information to a record makes the record straightforwardly de-anonymisable. The amount of background information that is available for re-identification is increasing over time, also in the public domain when more "anonymised" microdata becomes available.

Simply put: *the more information you add to an "anonymous" record, or the more "anonymised" records you combine, the higher the probability that the record contains uniquely discernable records that can be de-identified using exernal, often publically known background information.*

The larger a (micro)data set becomes, the more the anonymity of the subjects in the dataset decreases. The risks of re-identification are exacerbated when the amount of publically available "background information" increases. More and more microdata sets are released these days, ranging from Twitter "traces" to research data sets released under "open data" policies. The work by Narayan and Shmatikov convincingly showed, using real-world information obtained from the Netflix movie rating database, that combinations of de-identified microdata with other publically available information lead to finding movie preferences of individuals; the Netflix database, thus, should not have been regarded anonymous.

The related work discussed above makes clear that anonymity is *not about what information is removed from a data set.* Anonymisation is *about what information is kept in.*

3. On pseudonyms

Amendments to the DPR propose to introduce the notion of pseudonymised data [DPR Am's, 2013].

A *pseudonym* is a random, cryptographic, or other number or identifier that is used to replace real, direct identifiers in a dataset. A pseudonym can be linked to the data subject, or to other pseudonymised data of the same subject, by the party who generates the pseudonym or the party who has the linking codes. Pseudonymous data are, by definition, *microdata.*

An inherent *goal* of using pseudonyms is that microdata can be linked. Also, using pseudonyms, microdata becomes trackable over time.

The intent of pseudonyms is that linkage of microdatasets is possible: linkage back to the subject, or linkage of separately collected datasets that use the same pseudonyms or pseudonyms that can be combined. Trusted Third Parties (TTPs) are in the business of issuing pseudonyms and "relinking" data sets that were often collected for different purposes into a single combined data set[2]. Relinking of data sets with (initially) different pseudonyms is becoming commonplace in, for example, healthcare research [Mondriaan].

Because professionally created pseudonyms are reproducible and stable over time, pseudonymised microdata is often linkable over a long period of time. Pseudonyms are derived from, for example, name, date of birth address, and/or social security number, with a secret code mixed in by a trusted third party for security (to avoid trivial re-identification).

Pseudonyms carry additional risks compared to strictly "anonymous" information. As an example, pseudonymised data sets used in the Netherlands[3] contain data originating from many medical sources, including all hospitals in the Netherlands. The pseudonyms are stable and map to the same person,

2 See e.g., healthcare TTP ZorgTTP, http://www.zorgttp.nl/ (in Dutch).

3 See e.g., the Dutch DBC information system http://www.dbcinformatiesysteem.nl/

irrespective of where it was created (a secret created by a TTP is mixed in to avoid trivial "guessing" attacks). Different medical episodes of a person are thus connected to the same pseudonym, creating a large, longitudinal microdataset of a given person in pseudonymised form.

Consider that a researcher knows that the prime minister was admitted for a broken toe in hospital A on 11 December 2012 around noon. Matching this information to a single entry in the pseudonymised dataset, it is straightforward to find all other treatment episodes matching the, thusly re-identified, prime minister by finding all records with the same pseudonym. This may include things that took place years ago, some of which could be of a very private nature[4].

Bearing the longitudinal uniqueness and linkability of pseudonymous data in mind, it is important to regard pseudonymised data as personal data in general, given that information from different sources may be linked. Currently, linkage of pseudonymous data collected in different contexts typically happens without consent. Thus data subjects are unaware that their data are being recombined into increasingly large microdata sets, while this data – as more data is linked – is becoming increasingly re-identifiable, detailed, and more sensitive as this happens.

Effectively, pseudonymous data sets should ideally be viewed as *personal* information, to ensure that appropriate DPR rights such as transparency are applied to (combined) data sets.

Countries increasingly adopt "open data" policies where "anonymised" (research) data is being disclosed to the general public. This is problematic with any microdata, but certainly so with pseudonymised data. If the DPR does not take special measures to guard pseudonymised data, or if it does not define "anonymised" data well, the outlined "prime minister example" may become a reality to be encountered time and again in many contexts.

4. How <u>not</u> to define anonymous or pseudonymised data

A narrow definition of a pseudonym or of pseudonymised data that concentrates on removing only a few "identifying" attributes of a data set does little or nothing to protect the subject. Whether a dataset is re-identifiable depends not on what is removed from it (in terms of directly identifiable information) but on what remains in it. In most if not all cases, pseudonymised data should be considered identifiable information.
Given the pseudonymisation procedure's explicit intent of making data linkable, providing for the possibility of longitudinal linking over time, the processing of pseudonymous information should be regarded as the processing of <u>identifiable information.</u>

Note that usage of pseudonyms as a security measure *within* a given context or system, *keeping other security measures and DPR constraints in place*, is a good idea. Pseudonymisation should however <u>not</u> be seen as a way to "anonymise" information aiming to escape DPR constraints such as purpose limitation, appropriate protection, consent, etc. These constraints are certainly also needed for pseudonymised and often even for "anonymised" information as well.

If the DPR defines anonymised data narrowly, focusing only on what information has to be *removed* to obtain "anonymity" instead of looking at the probability that remaining information can be re-identified, this will lead citizens to *very dangerous territory* in the digital domain.

---

4 If a few entries match, little additional information may be needed to select the correct match from the different options; an option is to pick up the phone and make a call to ask about another distinguishing episode.

Were “anonymous” or “anonymised” information - in a narrow sense of the word – exempted from the strict constraints and regulations of the DPR, this “anonymous” or “anonymised” information[5] will increasingly and without any possibility for control by the data subject find its way into the public domain, where it is no longer appropriately protected and where it can be processed and used by anyone. This in turn increases the threat to privacy, as availability of many sets of microdata may facilitate re-identification of these data sets by simply exploiting the overlap in them.

After removal of some directly identifiable information, the remaining data can often still trivially be linked to other data sets and often to the data subject. Given current knowledge on how easy it is to combine/re-identify sets of microdata, one should *by default* consider microdata as personal information. Narrow definitions of what consitutes anonymised data (e.g., focusing on a minimal set of items that must be removed from a data set) should be avoided.

It also follows from the above that anonymised/pseudonymised data should be appropriately protected, and be subject to the same DPR constraints as other personally identifiable data in general. This should include, whenever possible, transparency, removal rights, accountability, etc.

5. Possible societal effects of using weak, narrow definitions of anonymous and pseudonymous data

When permitting “anonymous” or “pseudonymous” microdata to escape the strict requirements of the DPR for handling personal information, this in itself can thus lead to anonymised information to become less “anonymous”, even less than it already was. Intuitively, one can foresee how automated processing would be able to link the many anonymous data sets out there in a highly precise manner, simply by matching (“stitching together”) rare/sparse patterns that exist in the different microdata sets, pseudonymised or non-pseudonymised.

It has been convincingly shown that it is straightforward to link information from microdata sets to other data sets with microdata given sufficient background knowledge. If in the future a lot of microdata were in the public domain as the result of too narrow and weak a definition of anonymous or pseudonymised information , also more and more “background knowlege” would accumulate in the public domain that could be used to link and re-identify data of individuals available through public, private, or stolen microdata sets.

A weak regime for the treatment of “anonymised” or “pseudomised” information that makes this data escape the DPR definition of personal information could cause a self-enforcing detrimental effect to individual citizens' security and well-being; effectively, it can void digital privacy.

6. The specific importance of keeping medical information under strict DPR data protection regime

Under Directive 95/46/EC, medical information has special protection. Whereas in some cases usage of information for statistical purposes is allowed given this information is anonymised, consent of the data subject is required in in nearly all cases, including medical research. The basis for this is founded in medical secrecy which in turn ensures individuals' trustful access to healthcare. Research in the US has shown that access to healthcare decreases to a significant degree when people have no confidence in medical privacy, leading to adverse public health effects [HHS, Forrester].

5 This equally applies to “pseudonymous” or “pseudonymised” data

The DPR uses very broad exempts for using information for "historical, statistical and scientific research" in anonymised form, or even in identifiable form. The current definition of Article 83 is very broad, and makes it easy to process personal information for secondary use without consent[6], even in identifiable or easily identifiable form. This is worrisome; use of indirectly identifiable information, but certainly of directly identifiable information should be forbidden, except in cases where consent is given. Article 83 should be modified to include a consent requirement for secondary use of identifiable information.

If amendments with narrow definitions of anonymous and pseudonyms data find their way into Article 83, this will have comparable detrimental effect. As stated above, Article 83 should *not* allow processing of directly or indirectly identifiable personal information for historical statistical and scientific purposes without consent. It should also not refer to definitions of "pseudonyms", "pseudonymous" or "anonymous" that do not take into account the probability of re-identifiability of microdata sets.

Particularly worrisome is that Article 81, which concerns health information, creates a broad exemption for secondary use of medical information without consent by referring to Article 83. This should not be the case. Given the broad defintion of historical, statistical, and scientific research processing without consent in Art. 83, at the very least Article 81 *must* place delimiters around this exemption to ensure that consent, purpose limitation, appropriate protection, and accounting measures stay in place for medical information. *This must certainly be ensured for medical information.*

If the DPR becomes as lenient with secondary use of medical information as it appears to become for other information, European citizens will face a future where not only "regular" information is easily available by linking public, weakly anonymised or pseudonymised data sets, but where their medical information can be linked to this information as well, and in the public domain -- in the same straightforward way as illustrated for the Netflix database in 2007 [Narayanan and Shmatikov].

Summarising: Article 83 in particular is currently much too lenient on the use of identifiable information for historical, statistical, and scientific research purposes when anonymisation is difficult[7]. Referring to Article 83 directly and without limitation from Article 81(2), is extremely risky as it provides an "escape route" from the strict constraints of the DPR for processing sensitive medical information to processing in hardly or not even anonymised form.

In contrast to some other parts of the world, Europe has traditionally done well in protecting medical information in data protection directive 95/46/EC, and it has a good starting position for maintaining protection of medical information (and the important benefits for public health and individual access to healthcare that result from protecting medical confidentiality), including providing the data subject appropriate rights related to processing of his or her medical information. Europe would make a grave mistake were it to remove the existing strong protection for our most sensitive information from the DPR, as seems to be proposed now. It is imperative that, if nothing else, Article 81 is amended to take this into account.

---

6 Note that such secondary use may even include recombination of data sets, exacerbating the problem as outlined above.

7 If a weak definition of anonymous and pseudonymous is included in the DPR, this may have the same effect as the broad definitions that exist in Article 83.

7. Final notes, examples, and specific recommendations.

This section contains some notes on the wording of elected (non-exhaustive) amendments as examples of how to evaluate proposed amendments to the DPR.

The most important point of this report is that a *narrow definition* of what identification means, either in the definition of `data subject' or by definitions of pseudonym or anonymous should be avoided. Whether data is to be considered anonymous is not determined by what is *removed* from the data, but by what is *left in,* and on how easy that can be *combined* with other data sources[8] .

In this view, **Amendment 726** describing 'pseudonymised data' is detrimental, as it states that data becomes pseudonymous when *personal identifiers have been removed.* This definition focuses on what is *removed*, not on what *remains,* which is the critical part. Amendment 726 makes it possible to focus on a strict defintion of what "personal identifiers" are, and, once so decided, stop having to worry about the re-identification risks that remain with the information that is left over[9].

Amendment 729 seems to do better than 726. This amendment states that the information should *not be attributable to the data subject*, without defining how this attribution takes place. This way, the "escape" from defining anonymity (as observed by an outsider who does not have access to the linking codes) using a narrow, technocratic interpretation of identifiable information, cannot occur. Controllers should always evaluate whether a data set is sufficiently anonymous at the time of its construction, use, or release, considering the microdata in this set[10]. Note that the reference to technical controls in amendment 729 seems superfluous, and also the definition is not self-contained, as amendment 729 does not define what a pseudonym is.

We can conceive a compromise amendment, combining parts of 726 and 729, for example:

"Pseudonymised data" means any personal data that has been altered so that it cannot be attributed directly or indirectly to a data subject, but where a new (context-dependent) identifier is inserted in the data that allows linkage back to the data subject or to other pseudonymised data by anyone who has access to the linking codes."

Note that, in our opinion, a defintion of anonymous data or pseusonymous data is in fact unnecessary: the key importance of a definition of personal data is for the DPR to determine its scope precisely, that is, to what (directly or indirectly) identifiable personal information it applies. Pseudonymous data should in general be considered identifiable information, and by using a proper definition of personal information it is clear when this does and does not apply.

As a final example, note that a legal difficulty may arise in a case where two controllers independently determine that a pseudonymised data set is sufficiently "anonymous" for it to be released. When two independent controllers generate data with matching pseudonyms for the same data subject, each controller assessing the respective data *they* release as anonymous, the combined data set may not be

8 I use examples of pseudonymous data here, but comparable considerations apply to amendments with definitions of anonymous data.

9 Note: Amendment 728 defines 'pseudonym' and not 'pseudonymous data', so we do not include 728 here.

10 Decisions on whether to process information in this form thus have to be made based on an explicit risk assessement. Only if assessed in a well-documented way that information in a data set is not identifiable in any way, either directly or indirectly, should data be declared anonymous or pseudonymous in the sense that it can be released from the reach and application of the constraints of the DPR.

anonymous. This is another example of complexity which is introduced by having separate (narrow) definitions of anonymous or pseudonymous data. Note also the critical role that trusted third parties would and should have if a pseudonymous data definition is included in the DPR. Such technical complexities can be avoided by not including a definition of pseudonymous data in the DPR, instead focusing on a proper, unambiguous definition of personal data.

Amendment 714: the definition of 'data subject'. The addition of the words “or singled out” as proposed in this amendment, is appropriate. This closely relates to the point made earlier that microdata (certainly with a certain granularity or when combined with other microdata) should be considered identifiable information.

Further, without the direct interpretation as “singling out” relating to microdata, it is of key importance to recognise that when a specific individual (data subject) can be singled out using given information (irrespective of the method of identification, and even when not *directly* identified), this information must be considered personal information. Singling out relates to profiling, which is beyond my expertise, but it is important to note that the wording “singling out” is consistent with the way that personal information is viewed in this article. “Singled out” is an appropriate addition in that it avoids a narrow interpretation of personal data, including being coupled to a physical identity alone.

Article 81(2): the reference to the “conditions and safeguards” in Article 83 is misleading. There are hardly any safeguards in Article 83, it is simply a very broad exemption to the DPR that allows any processing that falls under the way too broad definition of historic, statistical and scientific research to escape the controls of the DPR – even if this involves processing of directly identifiable information.

If this broad scope remains in place for Art. 83, then for Article 81 at least a consent requirement for the use of directly or indirectly identifiable medical information for secondary (i.e., historic, statistical, or scientific) use *must* be ensured.

A good approach to alleviate the problems with Article 81 are the amendments proposed by Albrecht [Albrecht 2012]: amendment 327 is essential as it reinstates the consent requirement for secondary use of medical information. Amendment 326 is important as it restricts the use of directly identifiable information when not needed. Amendment 334 ensures that special data (including health information) as defined in Article 8/9 are excluded from the overly broad exemptions created in Article 83.

Note: Article 83 states: “[if] these purposes cannot be otherwise fulfilled...” and “as long as these purposes can be fulfilled in this manner”. The definition of Article 83 is much too lenient. It is much too easy to allow for usage of identifiable information. Identifiable information should only be used with consent. I recommend adding a consent requirement to Art. 83 by adding an Art. 83(4)[11].

Note that “historical, statistical, and scientific research information” (Article 83), which often concerns anonymised and pseudonymised information, is also exempted from the “right to be forgotten” (see Art. 17-3c). Again, an overly broad definition of data that falls outside the scope of the DPR involves a risk in that it removes citizens from the right to impose appropriate transparency and controls over this information.
As a final note, introducing the “option of broad consent” in amendment 2987 is not a good idea. It is easily conceivable that people consent to a particular research request once, but that does not mean they would agree to share information each and every time for any type of research automatically.

11 Some relevant suggestions for amending Art. 83 are given in [Korff].

Furthermore, transparency is important so that patients can be confident that medical confidentiality is upheld by their doctor. Finally, purpose binding should apply to any (reasonably) identifiable information, and particularly for medical information, as outlined in this report.

8. Conclusion.

Anonymisation and pseudonymisation should at best be considered weak protection mechanisms for personal information. Linking anonymous micro-information is often trivial.

The seemingly unnecessary act of defining anonymous data or pseudonymised data as separate terms in the DPR in itself – even when defined right[12] – creates a risk in that it creates an exception to the definition of data subject and directly or indirectly identifiable personal data. As a result, data may be released in easily re-identifiable form, again increasing the probability that other "anonymous" data is re-identified.

Care should be taken not to create broad exempts for processing of de-identified or identifiable data for "historical, statistical, and scientific research" without consent, particularly for medical data if not in all cases. Also, care should be taken to avoid overly narrow definitions of anonymous or pseudonymous data – if such definitions are introduced in the DPR at all.

An appropriate, technology-independent and thus time-resistant definition of personal data like 95/46/EC's *"any information relating to an identified or identifiable natural person ("data subject")" i*s and should remain the foundation of the DPR and should without ambiguity define the reach and applicability of the DPR.

Bibliography

[95/46/EC] Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Brussels, 1995.

[Albrecht 2012] J. Albrecht, Committee on Civil Liberties, Justice and Home Affairs , proposed amendments. Draft report 17.12.2012. (ref. PE501.927v02-00)

[Dalenius] T. Dalenius. "Finding a Needle In a Haystack or Identifying Anonymous Census Records", J. Official Statistics, Vol.2, No.3, 1986. pp. 329–336,

[DPR] European Comission: proposal for a regulation of the European Parliament and of the council on the protection of individual with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Brussels, 25.01.2012.

[DPR Am's 2013] J.P. Albrecht, Committee on Civil Liberties, Justice and Home Affairs. "Amendments on the proposal for a regulation of the European Parliament and of the council on the protection of individual with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)", Draft report, 04.03.2013.

[Forrester] Forrester research, Inc.: National Consumer Health Privacy Survey 2005, URL: http://www.chcf.org/publications/2005/11/national-consumer-health-privacy-survey-2005

12 If not defined right, it simply creates a direct escape from the reach of the DPR and a weakening of the (strong) definition of personal data.

[HHS] U.S. Dept health and human services (HHS), Fed. Reg. Vol 65 nr. 250, Dec. 28, 2000 pp. 82761-82810, http://aspe.hhs.gov/admnsimp/final/PvcFR07.txt

[Koot] M.R. Koot, G.J. Van 't Noordende, C.T.A.M. De Laat: “A study on the re-identifyability of Dutch citizens”, Workshop on hot topics in privacy enhancing technology (HotPETs), PETS, 2010.

[Koot 2012] M. Koot, “measuring and predicting anonymity”, PhD dissertation, University of Amsterdam, 2012

[Korff]: D. Korff “Comments on selected topics in the draft EU data protection regulation.” 2012 Available on: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2150151

[Mondriaan] TI-Pharma's (Dutch) database for medical (policy) research. See: http://www.tipharma.com/pharmaceutical-research-projects/drug-discovery-development-and-utilisation/mondriaan-project.html

[Narayanan and Shmatikov] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large datasets (how to break anonymity of the netflix prize dataset), U. Texas, 2007. Link: http://arxiv.org/abs/cs/0610105 . Also appeared in IEEE Security and Privacy Symposium, Oakland, 2008.

[Sweeney 2000]: L. Sweeney, “Uniqueness of simple demographics in the U.S. Population.” LIDAP-WP4, Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA: 2000.

[Sweeney 2001]: L. Sweeney: “Computational disclosure control: a primer on data privacy protection.” PhD dissertation, Massachusetts Institute of Technology, 2001.