A Study on the Re-Identifiability of Dutch Citizens

Matthijs R. Koot, Guido van 't Noordende, and Cees de Laat

University of Amsterdam, Informatics Institute,
Science Park 107, 1098 XG Amsterdam, Netherlands
{koot,noordende,delaat}@uva.nl

Updated December 16th 2012

Abstract. This paper analyses the re-identifiability of Dutch citizens by various demographics. Our analysis is based on registry office data of 2.7 million Dutch citizens, ∼16% of the total population. We provide overall statistics on re-identifiability for a range of quasi-identifiers, and present an in-depth analysis of quasi-identifiers found in two de-identified data sets. We found that 67.0% of the sampled population is unambiguously identifiable by date of birth and four-digit postal code alone, and that 99.4% is unambiguously identifiable if date of birth, full postal code and gender are known. Furthermore, two quasi-identifiers we examined from real-life data sets turn out to unambiguously identify a small fraction of the sampled population. As far as we are aware, this is the first re-identifiability assessment of Dutch citizens that uses authoritative registry office data.

Key words: re-identification, data anonymity

1 Introduction

These days, large amounts of data about citizens are stored in various data sets, spread over databases managed by different organisations all around the world. Data about individual citizens drives policy research on all sorts of topics: finances, health and public administration, to name a few. Using personally identifiable information outside the purpose for which it was originally collected is prohibited in general by EU directive 95/46/EC on data protection. De-identification techniques are often used to remove identifying information from data sets, while attempting to retain as much useful information as possible, for example to still allow (statistical) analysis involving demographics.

Most data sets can therefore not be called completely anonymised, even if they are claimed to be; especially for *microdata*, i.e., data consisting of entries that map to single persons, but from which identifying parts are removed, a risk exists that entries can be de-anonymised when sufficient additional information is available. Our research deals with the question of which pieces of partially identifying information can, when combined, lead to re-identification. Such a

combination of partially identifying information is called a quasi-identifier. This paper uses real registry office data of citizens in the Netherlands, to experimentally assess the re-identifiability of Dutch citizens using quasi-identifiers found in real-world data sets.

A seminal work on re-identification is due to Latanya Sweeney [14]. Using 1990 U.S. Census summary data, she established that 87% of the US population was uniquely identifiable by a quasi-identifier (QID) composed of three demographic variables [13, 14]:

Definition 1. $QID_{example}$ = { *Date-of-Birth + gender + 5-digit ZIP* }

In Massachusetts (U.S.) the Group Insurance Commission provides and administers health insurance to state employees. Sweeney legitimately obtained a de-identified data set containing medical information about Massachusetts' employees from them, including details about ethnicity, medical diagnoses and medication [14]. The data set contained the variables described in $QID_{example}$. Sweeney also legitimately obtained the identified 1997 voter registration list from the city of Cambridge, Massachusetts, which contained the same variables. By linking both data sets, it turned out to be possible to re-identify medical records, including records related to Massachusetts' governor of that time.

Sweeney proposed k-anonymity, a test asserting that for each value of a quasi-identifier in a data set, at least k records must exist with that same value and be indistinguishable from each other. This introduces a minimal level of uncertainty in re-identification: assuming no additional information is available, each record may belong to any of at least k individuals.

We analyze the (re-)identifiability of Dutch citizens by looking at demographic characteristics such as postal code and (part of the) date of birth. By 'citizen' we refer to a person who is registered as an inhabitant of the Netherlands. We examine the re-identifiability only in the context of linking the data sets that are described in this paper, and not using any additional outside information. For this paper, we limit ourselves to quasi-identifiers that we believe are most likely to be found in (identified) data sets elsewhere, based on commonly collected demographics. Regarding two real-life data sets, we only provide an assessment *of two specific quasi-identifiers*; other quasi-identifiers exist in those data sets, e.g. involving ethnicity and marital status, which are not discussed in this paper.

This paper is structured as follows: section 2 describes our approach; section 3 lists the results; section 4 describes related work and section 5 discusses the results.

2 Background

The Netherlands consists of 12 provinces and 441 municipalities of varying size [5]. A municipality is an administrative region that typically spans several villages or cities. Municipal registry offices are the official record-keepers of persons residing in the Netherlands, and maintain identified data about them.

De-identified data about individual citizens is available in number of research databases. To illustrate our analysis we picked two, which we describe below. In section 3 we assess, amongst others, re-identifiability of entries in these data sets.

2.1 Example Data Sets

The Dutch *National Medical Registration* (LMR) is a data collection program established in 1963, in which hospitals in the Netherlands participate by periodically sending in copies of medical and administrative information about hospital admissions and day care treatment. Example purposes of the LMR are the analysis of the effects of treatment, performance comparison between hospitals, and epidemiological studies. The LMR is currently managed by the Dutch Hospital Data foundation[1]. Statistics Netherlands, the Dutch organisation for conducting statistical studies on behalf of the Dutch government, also receives annual copies of the LMR for research purposes [6]. External researchers can currently request access to the records collected during 2005 and 2007 [2, 4]. These data sets contain only records about Dutch citizens; records about other patients are omitted. Each record in the LMR describes the hospital admission or day care treatment of a single individual, and multiple records may be present per individual. The 2005 and 2007 data sets each contain approximately 2.5 million records.

The Dutch *Welfare Fraud Statistics* (BFS) data set at Statistics Netherlands contains records about investigations on suspected welfare fraud of Dutch citizens [3]. Each record in the data set relates to a single, completed investigation, and multiple records may be present per person. The information in the data set is provided by municipalities. Between 2002 and 2007, the average number of records (cases) per year was 38,161[2]. The BFS data set contains different information at a different granularity than the LMR data set, which is the reason we selected it as a second example. For example, the LMR data set contains information about postal code, whereas BFS does not.

Re-identified records from the BFS data set could be abused to embarrass or discriminate citizens that have been subject of fraud investigation. Similarly, re-identified records from the LMR data set could be abused to embarrass or discriminate people based on medical history or medical conditions, potentially negatively impacting job or insurance prospects. Such consequences are at the disposal of the person possessing the (re-)identified records.

2.2 Approach and Terminology

A data set containing information about persons is said to be *de-identified* if direct identifiers like social security number, phone numbers, names and house numbers are omitted. A *quasi-identifier* is a variable or combination of variables

[1] `http://www.dutchhospitaldata.nl`
[2] Source: http://statline.cbs.nl

which, although perhaps not intended or expected to identify individuals, can in practice be used for that purpose.

A quasi-identifier may unambiguously identify a single individual, or reduce the number of possibilities to some small set of k individuals, the *anonymity set* [12]. A de-identified data set containing one or more quasi-identifiers can be *re-identified* by linking records to an *identified* data set containing the same quasi-identifying variable(s).

We assessed the (re-)identifiability of Dutch citizens by using quasi-identifiers composed of information about postal code, date of birth and gender information. We used registry office data of approximately 2.7 million persons, ∼16% of the total population, obtained from 15 of 441 Dutch municipalities. The 15 municipalities and number of citizens are shown in table 1. The sample contains small, mid-size and large municipalities. Although this selection is not random (selected by size) or necessarily representative for the whole population, we considered the selection appropriate for our analysis, since it enables us to assess whether differences in re-identifiability are observable for small municipalities compared to large municipalities that contain a city, for example. The municipalities chosen are spread over the country, such that there is no obvious bias due to geographical location of the municipalities in the countries - although the largest cities, Amsterdam, Rotterdam, and Den Haag, are located in the west of the Netherlands which is known as the most densely populated area of the Netherlands, called the "Randstad".

We requested a (nameless) listing of gender, full postal code and full date of birth of all citizens of 30 municipalities, and eventually obtained records of 15 municipalities, totalling approximately 2.7 million citizens. The remainder of this paper is based on analysis of this data. We distinctly discuss data only at municipal level; i.e. 'Amsterdam' refers to the *municipality of Amsterdam* rather than the *city of Amsterdam*.

We primarily focus on quasi-identifiers that match the LMR and BFS examples in this paper. The results, however, apply to *any* data set that contains these quasi-identifiers. We did not attempt to obtain access to data from the example data sets, since for our purposes it suffices to know which possible quasi-identifying variables they contain, and this information is available from public documents [2–4].

2.3 Data Quality

Data from municipal registry offices is relied upon during transactions between the Dutch government and its citizens, including the process of passport issuance. Registry office data is not free of error: data may be inconsistent with reality due to e.g. failure of citizens to report changes timely and truthfully, typographical errors and software errors [10]. The registry offices are required to undergo a periodical audit, which includes an integrity check of a random sample of the electronic person records. Each record from that sample is matched against other official files associated with the person whom the record is about, such as birth certificates. Each variable containing an incorrect value is counted as a single

Table 1. Municipalities included in our study (ordered by size)

Municipality	# of citizens
Amsterdam	766,656
Rotterdam	591,046
Den Haag	487,582
Utrecht	305,845
Nijmegen	161,882
Enschede	156,761
Arnhem	147,091
Overbetuwe	45,548
Geldermalsen	26,097
Diemen	24,679
Reimerswaal	21,457
Enkhuizen	18,158
Simpelveld	11,019
Millingen a/d Rijn	5,915
Terschelling	4,751
TOTAL:	2,774,476

error, and the maximum allowed rate for errors in 'essential' fields like DoB and postal code is 1% of the sample set size: to pass the test, a 100-record sample cannot contain more than 1 error in essential fields. The sample size depends on the municipality size. During the 2002-2005 audit cycle, 339 of the 370 (92%) audited municipalities passed this test [10]. This suggests that Dutch registry offices are generally a reliable source of data. During our own data sanity checks we removed 11 records containing a postal code from outside the sampled municipalities, as those records would have caused false outliers[3]; the remainder of the records passed all sanity checks.

2.4 Postal Codes in the Netherlands

In the Netherlands, a postal code consist of a four-digit number and a two-character extension — e.g. "1098 XG", the postal code of our institution. The four-digit number is referred to as '4-Position PostalCode' (PC4), and corresponds to exactly one town (city, village). A town may be divided into multiple PC4-regions: for example, our data contains eighty different PC4-regions for the city of Amsterdam, "1098" being one of them.

The two-character extension indicates a street, but often also a specific odd or even range of house numbers *within* that street. The full postal code is referred to as '6-Position PostalCode' (PC6). A combination of full (PC6) postal code and house or P.O. box number uniquely indicates a postal delivery address in the Netherlands.

[3] These cases may be related to moving citizens, e.g. pending handover of data between municipalities.

3 Results

This section describes the results of our analysis. Section 3.1 describes an overall analysis of our input data. From the result data it becomes clear which combinations of variables can be used to single out individuals or small groups of citizens, and which combinations pose less of a privacy risk in that sense. Section 3.2 describes the potential re-identifiability of citizens in the LMR data set. Section 3.3 analyses the potential re-identifiability of citizens in the BFS data set. Throughout this paper, we use the following notations: *QID*=Quasi-IDentifier, *DoB*=Date of Birth, *YoB*=Year of Birth, *MoB*=Month of Birth.

By 'quasi-identifier' we refer to abstract variables, by 'quasi-identifier value' to a valuation of those variables. We use rounded values for the sake of readability. For each quasi-identifier, we counted the number of different (distinct) values in the data — this is the number of anonymity sets; the number of people sharing a specific quasi-identifier value represents the anonymity set size.

In addition to mean values, we provide quartiles and min-max values to give an indication of how a quasi-identifier maps citizens in anonymity sets of rather diverse or rather similar size[4]. We chose quartiles as a means to indicate the value distribution while maintaining some brevity and readability of tables. Another choice could have been made (e.g., for deciles or percentiles), however, none has a definite advantage over the other. By using quartiles we can state properties of the distribution of anonymity set sizes such as "at most 25% of the anonymity sets are smaller than <1st quartile>" and "at most 50% of the anonymity sets are smaller than <median>".

3.1 Analysis over Aggregated Data

This section describes the results of an analysis of the combined data of the citizens of all municipalities listed in table 1. By including both small and large municipalities, covering the smallest villages (the smallest having two inhabitants) and largest cities (the largest having 684,926 inhabitants) in the Netherlands, the minimum and maximum anonymity set sizes represent the worst and best cases we expect to be found *anywhere* in the Netherlands. Furthermore, the statistics over the combined data indicate how strongly identifiable a quasi-identifier is for the overall population.

Throughout this paper, k denotes the anonymity set size; $k = 1$ means that some quasi-identifier value unambiguously identifies some individual, $k = 2$

[4] The lower (1st) quartile is the value separating the lower 25% of the values; the median value (2nd quartile) separates the higher half of the values from the lower half; the upper (3rd) quartile separates the higher 25% of the values. To illustrate: for both (100,100,100,100,100) and (1,1,1,1,496), the mean value is 100, while both sets are obviously very different. For the former set, all three quartiles are 100, as are both the minimum and maximum: this shows that the distribution is uniform. For the latter set of numbers, minimum value and all quartiles are 1, but the maximum value is 496: this shows that the distribution is skewed. Or, in our context, that the quasi-identifier maps citizens into anonymity sets of different sizes.

means that the value is shared by two individuals, and so on. Table 2 shows the statistical characteristics of anonymity set size k for various (potential) quasi-identifiers. The column *'# of sets'* contains the number of different values present in our data for a given quasi-identifier, i.e., the number of anonymity sets. Generally, the higher this number, the weaker privacy, because the anonymity sets will tend to be smaller in that case. The min/max values denote the size of the smallest and largest anonymity set.

Table 2. Anonymity set size k for various (potential) quasi-identifiers

Quasi-identifier:	# of sets	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
PC4	388	2	3,278	7,090	7,188	10,300	22,330
PC6	66,883	1	24	35	41	50	1,322
PC4+DoB	2,267,700	1	1	1	1	1	42
PC6+DoB	2,759,422	1	1	1	1	1	5
PC4+gender	776	1	1,652	3,536	3,594	5,151	11,730
PC6+gender	133,012	1	11	18	21	25	954
gender+YoB	221	1	5,219	14,570	12,550	19,740	25,580
gender+YoB+PC4	68,515	1	11	31	41	59	312
gender+YoB+MoB	2,699	1	397	1,177	1,028	1,594	2,326
gender+YoB+MoB+PC4[a]	635,679	1	2	3	4	6	40
gender+YoB+MoB+municipality[b]	34,790	1	6	18	80	96	733
gender+DoB	71,318	1	21	40	39	54	571
gender+DoB+PC4	2,488,828	1	1	1	1	1	22
gender+DoB+PC6	2,766,475	1	1	1	1	1	4
town+gender	134	1	222	1116	20,700	3259	347,100
town+YoB	5,642	1	6	29	492	101	14,270
town+YoB+MoB	49,207	1	2	5	56	20	1,262
town+DoB	463,134	1	1	2	6	7	419
town+YoB+gender	10,492	1	4	17	264	60	7,515
town+YoB+MoB+gender	83,172	1	1	3	33	14	695
town+DoB+gender	697,875	1	1	2	4	5	226

[a] QID_A, see section 3.2.
[b] QID_B, see section 3.3.

As an example, the median anonymity set size of PC6 is 35, the minimum size is 1 and the maximum size is 1,322. This means that at most half of the values for PC6 have anonymity sets of sizes between 1 and 35, and that the sizes of the anonymity sets in the upper half are between 35 and 1,322.

Looking at the quartiles, it becomes clear that some quasi-identifiers are particularly strong, by which we mean that a large portion of the anonymity sets established by that quasi-identifier are of small size (e.g. $k = 1$ or $k \leq 5$). For example, for $\{PC4{+}DoB\}$, table 2 shows an anonymity set size of $k = 1$ for up to the 3rd quartile, meaning that 75% of the quasi-identifier values unambiguously identify a citizen. Looking at the lower quartiles, it also becomes clear that some

quasi-identifiers are weaker identifiers: for $\{PC4\}$, only at most 25% of the sets are of size $k \leq 3,278$; for $\{gender + YoB\}$, at most 25% of the sets are of size $k \leq 5,219$. Overall, it turns out that quasi-identifiers containing both PC4 or PC6, as well as date of birth, are most identifying.

We were surprised to find that PC4 postal codes exist which are shared by only two citizens: we had expected that PC4 codes always map to relatively large numbers of citizens. Upon closer inspection, it appears that the data is accurate: it represents the inhabitants of a new construction area in the harbour of Rotterdam. These pioneering citizens turn out to be unambiguously identifiable nation-wide by only their $\{PC4 + gender\}$ or $\{town + gender\}$ — albeit only until other citizens officially move in.

Table 2 also clearly shows that the two-character extension to the PC4 postal code, making PC6, strongly increases identifiability: the median anonymity set size for $\{PC4\}$ is 7,090, for $\{PC6\}$ only 35.

Table 3. Number of Dutch citizens per anonymity set size, for various quasi-identifiers

Quasi-identifier:	$k = 1$	$k \leq 5$	$k \leq 10$	$k \leq 50$	$k \leq 100$
PC4	0	9	19	345	996
PC6	429	6,109	25,103	1,459,939	2,354,255
PC4+DoB	1,861,081	2,754,465	2,765,932	2,774,476	-
PC6+DoB	2,744,653	2,774,476	-	-	-
PC4+gender	4	27	103	889	2,555
PC6+gender	1,854	31,262	184,803	2,342,242	2,629,017
gender+YoB	5	14	53	250	516
gender+YoB+PC4	4,160	28,206	71,948	942,306	2,076,880
gender+YoB+MoB	55	356	712	4,478	9,674
gender+YoB+MoB+PC4[a]	137,035	279,100	2,196,950	2,774,476	-
gender+YoB+MoB+municipality[b]	2,186	22,565	59,597	244,152	619,671
gender+DoB	2,014	14,506	40,322	1,392,622	2,725,472
gender+DoB+PC4	2,240,461	2,765,067	2,772,205	2,774,476	-
gender+DoB+PC6	2,758,578	2,774,476	-	-	-
town+gender	4	4	28	372	896
town+YoB	499	3,172	7,225	50,985	103,145
town+YoB+MoB	10,083	61,073	112,850	287,173	394,844
town+DoB	185,042	596,769	1,045,559	2,730,668	2,750,700
town+YoB+gender	1,153	7,195	16,333	102,018	150,135
town+YoB+MoB+gender	22,260	109,126	170,351	398,601	826,744
town+DoB+gender	288,409	1,029,601	1,813,559	2,750,669	2,764,050

[a] QID_A, see section 3.2.
[b] QID_B, see section 3.3.

Whereas table 2 focusses on the size distribution of the anonymity sets, table 3 shows the actual number of *citizens* found in those anonymity sets. The larger the value in columns '$k = 1$', '$k \leq 5$' and possibly '$k \leq 10$', the larger

the portion of the population that is covered by anonymity sets of those (small) sizes and the stronger the quasi-identifier identifies citizens. The numbers confirm that $\{PC6 + DoB\}$ is a strong identifier, because here nearly all citizens have $k = 1$; $\{PC6\}$ alone is not a strong identifier, because only a very small portion of the citizens have $k \leq 10$ (compared to $k \leq 50$). We also included columns for a few larger set sizes ($k \leq 50$ and $k \leq 100$) for illustrative purposes. For example, only 896 out of 2.7 million citizens are identifiable to a group of $\leq$100 by $\{town + gender\}$, so by themselves, those variables do not pose a significant privacy risk for most citizens. For readability, we replaced numbers by '-' when the total population is reached at some k.

From the numbers for quasi-identifier $\{gender + DoB + PC6\}$ it follows that approximately 99.4% of the Dutch citizens in our data set (2,758,578 out of 2,774,476) can be unambiguously identified by $\{gender + DoB + PC6\}$; and it turns out that 67.0% (1,861,081 out of 2,774,476) can still be unambiguously identified by $\{PC4 + DoB\}$.

3.2 Case: National Medical Registration

The LMR contains a large amount of information about hospital admissions and day care treatment: amongst others, it contains fields describing the hospital, the patient's insurance type, diagnosis codes, the treatment that was provided and the medical specialisms and disciplines involved [2, 4]. This information could be privacy-sensitive and it is generally treated as such, even when de-identified. The LMR data set also contains demographic data about the patient. In particular, the LMR contains the following quasi-identifier:

Definition 2. $QID_A = \{\ PC4 + gender + YoB + MoB\ \}$

Our data contains 635,679 different anonymity sets for QID_A. We use k_A to denote the anonymity set sizes for this quasi-identifier. 137,035 people, ∼4.8%, are unambiguously identifiable by QID_A, that is, they are the only person in the anonymity set, which thus has k_A=1. Furthermore, we found 212,536 citizens to have $k_A = 2$; 260,244 to have $k_A = 3$ and 282,644 to have $k_A = 4$ (most common size). Table 4 lists the statistical properties of the size of the anonymity sets established by this quasi-identifier. The municipality size is included for quick reference.

The numbers show that there is no large difference in anonymity between citizens of different-sized municipalities: the range of the medians is 1–5. The highest median anonymity set size is found in Amsterdam, the lowest is found in Terschelling. The latter means that half of the QID_A values found in Terschelling unambiguously identify a citizen.

The municipality size (column *'# of citizens'*) and median anonymity set size (column *'Median'*) have a Pearson correlation coefficient of .60. The single largest anonymity set is found in Amsterdam and is of size 40. Based on the numbers shown in table 3, the total percentage of citizens identifiable to a group of 10 or less by this quasi-identifier is ∼79.1% (2,196,950 out of 2,774,476).

Figure 1 visually represents the numbers in table 4. Some large anonymity sets exist as outliers, especially for larger municipalities, but overall anonymity is approximately the same (poor) over all municipalities.

Note that there is a difference in constraints between registry office data and the hospital admission data set: whereas the year of birth is allowed to be zero by the Dutch registry offices — e.g. for immigrants about whom the date of birth is not fully known —, the LMR requires it to be non-zero and be estimated if unknown [1]. This means that LMR-records about a person who is officially registered with zero year of birth (in our data set we only found 3) will *not* be re-identified by quasi-identifiers involving the year of birth. On the other hand, the quality of data from the LMR and BFS depends on their sources (hospitals and municipalities); it is not asserted whether each record accurately represents reality [2–4] – note that any mismatch (error) prevents linkability, and thus improves privacy for the involved individual.

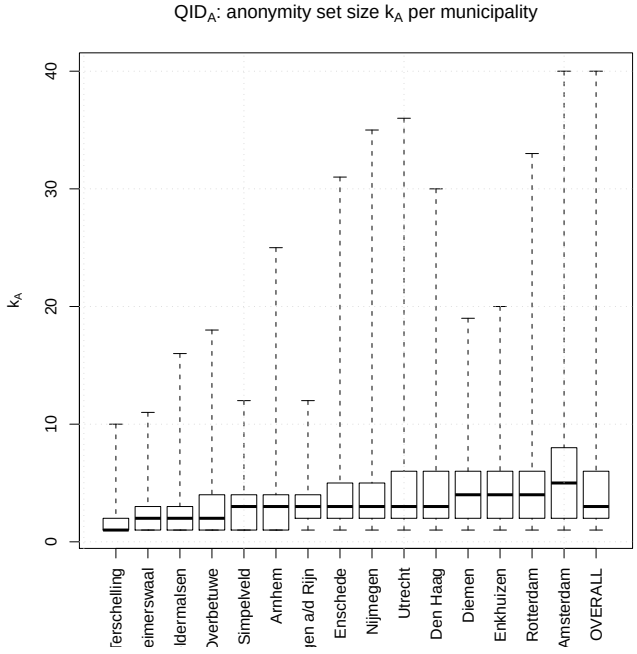

Fig. 1. Box-and-whisker plot showing anonymity set sizes k_A, per municipality. Whiskers denote the minimum and maximum values; the boxes are defined by lower and upper quartiles and the median value is shown.

3.3 Case: Welfare Fraud Statistics

In the BFS data set, we recognised the following as a potential quasi-identifier:

Table 4. Statistical summary of k_A, divided by municipality (ordered by median)

Municipality:	# of citizens	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Amsterdam	766,656	1	2	5	6	8	40
Rotterdam	591,046	1	2	4	5	6	33
Enkhuizen	18,158	1	2	4	4	6	20
Diemen	24,679	1	2	4	4	6	19
Den Haag	487,582	1	2	3	4	6	30
Utrecht	305,845	1	2	3	4	6	36
Enschede	156,761	1	2	3	4	5	31
Nijmegen	161,882	1	2	3	4	5	35
Arnhem	147,091	1	1	3	3	4	25
Millingen a/d Rijn	5,915	1	2	3	3	4	12
Simpelveld	11,019	1	1	3	3	4	12
Geldermalsen	26,097	1	1	2	2	3	16
Overbetuwe	45,548	1	1	2	3	4	18
Reimerswaal	21,457	1	1	2	2	3	11
Terschelling	4,751	1	1	1	1	2	10
OVERALL	2,774,476	1	2	3	4	6	40

Definition 3. $QID_B = \{$ *municipality + gender + YoB + MoB* $\}$

Our data contains 34,790 different anonymity sets for QID_B. 2,186 people, ∼0.07%, are unambiguously identifiable by QID_B. Furthermore, we found 3,552 citizens to have $k_B = 2$; 5,064 to have $k_B = 3$ and 5,508 to have $k_B = 4$. The total percentage of citizens identifiable to a group of 10 or less is ∼2.14% (59,597 out of 2,774,476). The single largest anonymity set is found in Amsterdam and is of size 733.

Table 5 lists the statistical properties of k_B per municipality. The numbers show that regarding the BFS, large differences in anonymity exist between citizens of different-sized municipalities: the range is 1–733. The highest median anonymity set size is 310, found in Amsterdam, the lowest is 2, found in Terschelling. Municipality size and median anonymity set size have a Pearson correlation coefficient of .99; the median anonymity set size is rather constant at ∼0.04% (1/2,500) of the population size.

Figure 2 visually represents the numbers in table 5. Note that the range on the vertical axis is much larger than in figure 1. It is clear that citizens from large municipalities tend to have much stronger anonymity than citizens from small municipalities, which is something to remember when dealing with de-identified data about citizens from small municipalities.

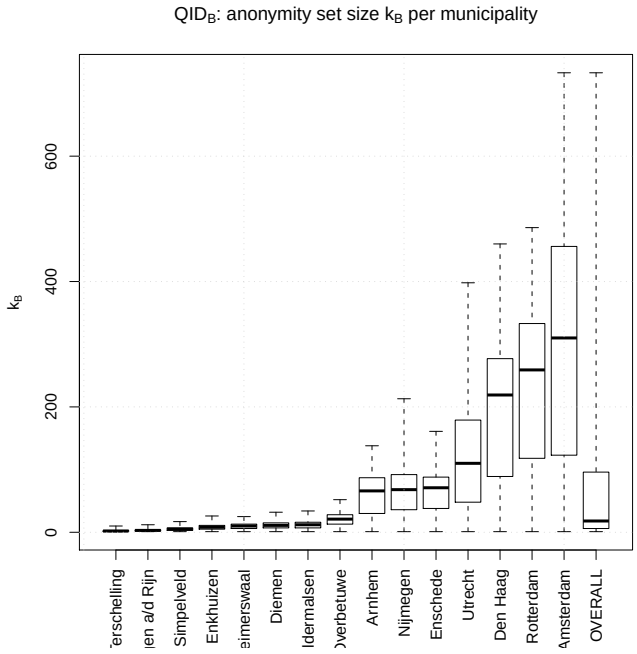

Fig. 2. Box-and-whisker plot showing anonymity set sizes k_B, per municipality. Whiskers denote min-max values.

Table 5. Statistical summary of k_B, divided by municipality (ordered by median)

Municipality:	# of citizens	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Amsterdam	766,656	1	123	310	296	456	733
Rotterdam	591,046	1	118	259	228	333	486
Den Haag	487,582	1	89	219	188	277	460
Utrecht	305,845	1	48	110	121	179	398
Enschede	156,761	1	38	71	64	88	161
Nijmegen	161,882	1	36	68	66	92	213
Arnhem	147,091	1	30	66	60	87	138
Overbetuwe	45,548	1	13	21	20	28	52
Geldermalsen	26,097	1	7	12	12	16	34
Diemen	24,679	1	7	11	11	15	32
Reimerswaal	21,457	1	6	10	10	13	25
Enkhuizen	18,158	1	5	8	8	11	26
Simpelveld	11,019	1	3	5	5	7	17
Millingen a/d Rijn	5,915	1	2	3	3	4	12
Terschelling	4,751	1	1	2	3	3	10
OVERALL	2,774,476	1	6	18	80	96	733

4 Related Work

Various extensions and enhancements on k-anonymity have been devised, such as l-diversity [8] and t-closeness [7]. k-anonymity attempts to make it hard for an adversary to link records to individuals, i.e., it protects against identity disclosure, but still allows adversaries focussing on some subset of k-anonymous records to make educated guesses about specific variables by looking at the distribution of those variables. l-diversity and t-closeness, for example, attempt to also make it hard for an adversary to do this, and are applied as a complement to k-anonymity.

In 2006, Arvind Narayanan and Vitaly Shmatikov demonstrated new statistical de-anonymisation attacks against the publicly released Netflix Prize data set containing de-identified movie ratings of about 500,000 subscribers of Netflix [9]. The authors showed that, given a little prior knowledge of a certain subscriber, it is possible to identify, with high certainty, records related to that subscriber in the anonymised data set. The authors show that their findings apply in general to multi-dimensional microdata.

In his short paper revisiting Sweeney's work, Philippe Golle mentions a lack of available details about the data collection and analysis involved Sweeney's work as a reason for being unable to explain the big difference between the outcome between both studies: in Golle's study of the 2000 U.S. Census data, only ∼63% of U.S. citizens turned out to be uniquely identifiable, as opposed to ∼87% that Sweeney determined by studying the 1990 U.S. Census data. This may be attributed to inaccuracies in the source data. By using registry office data we are confident that our results (for the Dutch population) are likely to be highly accurate.

5 Discussion

We determined the identifiability of Dutch citizens using information about postal code, date of birth and gender. We studied real registry office data of approximately 2.7 million citizens, ∼16% of the total population, obtained from 15 of 441 Dutch municipalities of varying size. We assessed the re-identifiability of records about these individuals in known data sets about hospital admissions and welfare fraud.

It turns out that approximately 99.4% of the sampled population is unambiguously identifiable using PC6 postal code, gender and date of birth, and 67.0% by PC4 and date of birth alone. Regarding the quasi-identifier found in the LMR data set, approximately 4.8% of the sampled population is unambiguously identifiable and 79.1% is identifiable to a group of 10 or less. Regarding the quasi-identifier found in the BFS data set, approximately 0.07% of the sampled population is unambiguously identifiable and 2.14% is identifiable to a group

of 10 or less; for small municipalities, however, the anonymity set sizes become much smaller and re-identifiability higher.

As far as we know, we are the first to study re-identifiability using authoratitive registry office data. Comparing to Sweeney and Goll (who used census data), our study uses registry office data, which is the authoritative data source during passport issuance. Our data was not prone to the intricacies of survey-based data collection. We only cover a portion of the Dutch citizens, ∼16%, but are confident that the results for that portion are accurate. We also provide the minimum and maximum anonymity set sizes that can be expected to be found anywhere in the Netherlands.

The results suggest that, considering the quasi-identifier in the National Medical Registration data set, someone who is able to access registry office data can re-identify a large portion of records with relatively high certainty. Considering the quasi-identifier in the Welfare Fraud Statistics data set, the re-identification risk is generally lower, but strongly depends on municipality size.

One could argue about the plausibility of the threat scenario underlying the two cases we picked: we assume an adversary who is able to access non-public records from both registry offices and Statistics Netherlands. Access to the data sets at Statistics Netherlands is only granted to qualified applicants, for specific purposes, under specific conditions of confidentiality [15]. Thus, obtaining data may require an investment that is disproportional to the expected gain of re-identifying records from these particular data sets to begin with. We note, however, that our results apply to *any* de-identified data set containing the assessed quasi-identifiers. Also, registry offices are not the only source for identified data, and *any* identified database containing these quasi-identifiers with sufficiently large coverage of the total population may be used; suitable data sets may also exist at, e.g., information brokers, marketing agencies and public transport companies. Besides, preventing registry office data itself from being used for re-identification may be difficult: the 441 municipalities are autonomous gatekeepers to their citizen's data and that citizen data is already necessarily exchanged on a regular basis for a variety of legitimate purposes [11]. It is hard to protect data that has many legitimate users and uses.

We believe that our results are useful as input for privacy impact assessments involving data about Dutch citizens. It remains a matter of policy what value of k can be considered *sufficiently strong* anonymity for particular personal information.

References

1. Tieto Netherlands Healthcare BV. *LMR Gebruikershandleiding*, 2009.
2. CBS. *Documentatierapport Landelijke Medische Registratie (LMR) 2005V1*, March 2007.
3. CBS. *Documentatierapport Bijstandsfraudestatistiek (BFS) 200901-06V1*, November 2009.
4. CBS. *Documentatierapport Landelijke Medische Registratie (LMR) 2007V1*, July 2009.

5. CBS. Website: Cbs - gemeentelijke indeling op 1 januari 2009, 2009. http://www.cbs.nl/.
6. CBS. Website: Cbs - ziekenhuisopnamen - dataverzameling, 2009. http://www.cbs.nl/.
7. Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *23rd International Conference on Data Engineering*, pages 106–115, 2007.
8. Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3, 2007.
9. Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pages 111–125, Washington, DC, USA, 2008. IEEE Computer Society.
10. Atzo Nicolaï. Kst99754: Modernisering gemeentelijke basisadministratie persoonsgegevens, 2006.
11. NVVB. *Schema voor schriftelijke verzoeken om gegevensverstrekking uit de GBA*, January 2010.
12. Andreas Pfitzmann and Marit Hansen. A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management. http://dud.inf.tu-dresden.de/Anon_Terminology.shtml, December 2009. v0.32.
13. Latanya Sweeney. Uniqueness of simple demographics in the u.s. population, 2000.
14. Latanya Sweeney. *Computational disclosure control: a primer on data privacy protection*. PhD thesis, Massachusetts Institute of Technology, 2001. Supervisor: Abelson, Hal.
15. Leon Willenborg and Ton de Waal. *Statistical Disclosure Control in Practice*, volume 111 of *Lecture Notes in Statistics*. Springer, 1996. ISBN: 978-0-387-94722-8.