Statistics Netherlands

BBMRI·NL

biolink-nl

2015 Scientific Paper

Record linkage for health studies: three demonstration projects

Established and edited by

G. (Gerard) van Grootheest (GGZ inGeest and VU University Medical Centre)

M.C.H. (Mark) de Groot (Utrecht University)

D. J. (Jan) van der Laan (Statistics Netherlands)

J.H. (Jan) Smit (GGZ inGeest and VU University Medical Centre)

B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)

Contents

1. Introduction

Authors
G. (Gerard) van Grootheest (GGZ inGeest and VU University Medical Centre)
M.C.H. (Mark) de Groot (Utrecht University)
D. J. (Jan) van der Laan (Statistics Netherlands)
J.H. (Jan) Smit (GGZ inGeest and VU University Medical Centre)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)

1.1 Background

Record linkage is becoming more and more common in statistical and academic research. Linkage of records makes it possible to combine data from different sources to answer research questions that are very difficult or impossible to answer using data from just one source. Although linkage can be regarded as a more efficient way of obtaining data than setting up a new collection, it is important to understand the technical, methodological and legal restrictions that may apply.

The project Biolink NL aims to report on the methodological, technical and legal aspects of record linkage of health data in the Netherlands. Biolink NL is one of the so-called rainbow projects funded by the Dutch Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL). BBMRI-NL aims to stimulate collaboration and data sharing between research institutes (mostly biobanks), building on existing infrastructures, resources and technologies. The Biolink NL project is a combined effort of researchers from a number of academic research institutes and Statistics Netherlands (CBS).

In a previous paper (Ariel et al, 2014), we reviewed the existing literature and compared the performance of different combinations of data sets, linkage variables, and linkage algorithms in a simulation study. Following this comparison of linkage approaches in a simulated setting where true links are known, the next step is to apply the same linkage methods to real datasets in which error rates and true links are unknown.

In the current paper we demonstrate the feasibility of record linkage in a number of demonstration projects that include health care data. These demonstration projects have been chosen in such a way that they differ from each other in terms of population characteristics, time span of data collection, number of records per dataset, and the way the data are collected. We describe how different approaches perform under these real-life circumstances, compare how much work needs to be invested in each approach, and in the end provide a practical guide for researchers who wish to link their data to external registrations.

1.2 Conclusions drawn from simulated record linkage

There are several advantages of studying linkage methods in a simulation rather than in real datasets. First, a simulation gives full control over the characteristics of the datasets. The researchers can add linkage variables, change the number of records, or introduce any type

of error. The second advantage is that the true matches between the created datasets are known, which means that the number of correct links, incorrect links, and missed links can be precisely reported. Based on the simulations that we performed, we described how different characteristics of the datasets influence linkage performance, and that some approaches are more susceptible for variations in these characteristics than others.

One of the determinants of linkage success is the algorithm that is used. The most basic approach is deterministic linkage, which looks for exact matches between variables in two datasets. More flexibility can be reached with a probabilistic algorithm that gives weights to any similarity between record pairs. A pair is considered a link when a certain threshold is reached. A higher sensitivity can be reached by choosing a lower threshold. However, this will also increase the chance of creating false links. Generally speaking, probabilistic methods can identify more links than deterministic linkage, but an appropriate threshold must be chosen to avoid incorrect links.

The second parameter that influences linkage results is the choice of linkage variables. Although the government and health care providers use a national identification number (NIN, in Dutch: *Burger Service Nummer, BSN*) for each citizen of the Netherlands, research cohorts are not allowed to process this number. When the use of the BSN is not possible, personal information such as sex, date of birth, name and address must be used.

Thirdly, the number of records in the two datasets and the overlap between them are important factors. A research cohort typically has fewer records than the database from which additional information is retrieved, while the latter does not cover the entire population either. In other words, the overlap is generally less than 100 percent and it is unknown which records should have a match in the other dataset. In general, both the sensitivity and precision of probabilistic linkage decrease as the overlap becomes smaller and the datasets become larger.

Fourthly, no dataset is free of error. Discrepancies between two datasets may be caused either by incorrect data entry or by the change of variables over time. Obviously, sensitivity drops as the error rate increases. The effect of error on precision however depends on the size of datasets and their overlap. Unfortunately, it can be difficult to estimate how much error the linkage variables contain. Best linkage results are achieved if both datasets have been created or updated around the same time, and if the address history is recorded. Additionally, pre-processing can help to standardise variables and remove common spelling mistakes.

1.3 Demonstration projects in Biolink NL

In this paper we describe three different linkage projects. The first two consist of an academic research cohort linked to a larger (non-academic) registry; the third entails the linkage of two datasets that both contain millions of records.

1. The Netherlands Twin Register (NTR) linked with the Achmea Health Database (AHD).
2. The KOALA cohort with a number of pharmacies in the database of the *Stichting Farmaceutische Kengetallen (SFK)*.
3. The population register (*Basisregistratie Personen, BRP*, formerly the *Gemeentelijke Basisadministratie*) and employment register (ER, in Dutch: *Werknemersbestand*).

The datasets selected for these demonstration projects differ greatly in size and coverage of the general population. Moreover, each of the linkages has certain features that impose a unique challenge. For example, the NTR consists of young twins, who share most of their personal information, such as address and date of birth. The SFK only has access to anonymised data, and linkage of the employment register with the population register is mostly challenging because of the large number of records.

Table 1.3.1 gives a short overview of the methods that were used in each linkage project. Three projects involve a TTP that performed pseudonimisation and anonymised linkage.

1.3.1 Linkage methods that were applied in the three demonstration projects

	Deterministic	Probabilistic	Probabilistic through TTP
NTR – AHD	yes	yes	yes
KOALA – SFK	–	–	yes
ER –BRP	yes	yes	–

Aims

The aim of the current study is to establish whether data linkage can be an effective and efficient way to enrich research cohorts with additional information from external sources. The quality of such enrichment is crucial when addressing research questions that would otherwise be impossible to answer or would only be addressed in smaller samples, or with lower precision.

The current paper consists of several chapters, in which we describe how the choice of the combined datasets, linkage methods, and linkage variables affect the feasibility of each linkage project and the reliability of the results. In this light, we do not only evaluate the quality of the linked datasets and try to answer the cohort's research question, but also describe the work invested in each demonstration project. In the last chapter we summarise the results and provide a number of recommendations on how to go about record linkage in diverse situations.

2. General methods

Authors
G. (Gerard) van Grootheest (GGZ inGeest and VU University Medical Centre)
D.J. (Jan) van der Laan (Statistics Netherlands)
J. A. (Jasper) Bovenberg (Legal Pathways)
M.C.H. (Mark) de Groot (Utrecht University)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)

2.1 Linkage projects

The three demonstration projects included in this study differ from each other in various ways, and each of these differences has an impact on the feasibility of record linkage. Perhaps the most important aspect of record linkage is permission to access certain information. Two linkages included in this study are based on an academic research cohort and only included subjects who gave consent for retrieval of additional information from external data sources. The third linkage was performed at Statistics Netherlands and did not require consent from any subject.

As mentioned in the previous chapter, linkage success is influenced by factors such as the availability and reliability of linkage variables, and the size and overlap of linked datasets. The availability and quality of variables may depend on the goal of data collection. For example, accurate address information is crucial for a cohort study in order to send questionnaires to its respondents, while a disease registry may not be inclined to check, keep, or update address variables. Discrepancies can also occur when different identifying variables are recorded. One database may contain given names, while another contains only initials. In some countries, such as the Netherlands, subjects can also be recorded under the name by which they are generally known (in Dutch: *roepnaam*) instead of the official given name. Just like the availability of linkage variables, the size of a dataset is often related to the goal for which a database was created. For example, a disease registry should only include a specific group of patients, while the population register (BRP) covers nearly 100% of the population. The current study includes three linkage projects so that different combinations of data sources are covered, as summarised in table 2.1.1. In this paragraph we describe how the demonstration projects differ from each other and what this implies for record linkage.

2.1.1 Summary of each demonstration project's main characteristics

	Data in external source	Dataset sizes (subjects)	Estimated overlap [1]	Expected error rate
NTR – AHD	Health insurance	30,000–1,600,000	26%	High
KOALA – SFK	Pharmaceuticals	1,700–3,000,000	20–30%	Low
ER – BRP	Population register	7,000,000[2]–14,300,000	95–100%	Low

[1] The percentage indicates how many cohort records should have a match in the external dataset.
[2] Because people can have more than one job in a certain period, the number of job records was 12.9 million.

The Netherlands Twin Register (NTR) and Achmea Health Database (AHD)

Linkage of young twins is challenging, because they share most of the linkage variables such as date of birth and address. We linked a set of more than 15,000 twin pairs with linkage consent from the NTR to health insurance data from an insurance provider (Achmea) that covers about 26 percent of the general Dutch population. We deducted healthcare use from AHD's claim records and checked whether the data corresponded with the NTR.

Besides the difficulties that result from linking twins, a major challenge in this project was the different availability of linkage variables in both datasets: the NTR distinguishes twins based on their initials and given names, while the AHD distinguishes twin siblings by their NIN and does not usually register given names.

KOALA and Foundation for Pharmaceutical Statistics (SFK)

The KOALA birth cohort consists of children born in 2000–2002 in the southern half of the Netherlands. Linkage consent was available for 1,741 children. In order to increase the linkage success, the complete address history for all KOALA participants was obtained by querying the BRP, prior to linkage with the SFK.

The SFK has dispensing records of 93 percent of the community pharmacies in the Netherlands. However, most records that the SFK received were anonymised at the pharmacy in such a way that no linkage variables were available. At the time of the current linkage project, the SFK was implementing the involvement of a TTP into the data infrastructure, in order to allow anonymous record linkage in the future.

A number of pharmacies were known to contain information about KOALA subjects. The SFK facilitated the KOALA linkage by prioritising these pharmacies in the implementation of the TTP infrastructure. Consequently, whereas approximately 20 percent of all SFK records contained a linkage key at the time of linkage, the expected overlap between both datasets was larger than 20 percent.

The employment register and the population register

The linkage between the ER and the BRP is different from the other linkages in the sense that these two datasets are very large. Approximately 7 million people had 12.9 million job records, and these records were linked with 14.3 million BRP records. As a large majority of employees should be registered in the BRP, a substantial overlap between ER and BRP is expected.

This linkage was executed at Statistics Netherlands on a consumer-level PC. In order to reduce computational load resulting from the complexity of probabilistic linkage, we carried out special processing steps such as segmentation of the data into smaller blocks.

OMEGA and the Netherlands Cancer Registry

A fourth demonstration project was planned, but its linkage was not accomplished during the project period. Our ambition was to link subjects from the OMEGA cohort study to the Netherlands cancer registry, using two approaches that were also covered by our other demonstration projects. In addition to those linkages, we proposed to link a subpopulation based on their national identification number, the BSN. Linkage based on the BSN is considered a gold standard that may be used to validate other linkage methods.

The Comprehensive Cancer Centre the Netherlands (*Integraal Kankercentrum Nederland, IKNL*) aims to record all primary diagnoses of cancer in the Netherlands Cancer Registry and receives this information from hospitals. The OMEGA study follows women who received fertility treatment in any of twelve Dutch IVF clinics since the early 1980's. Since these women were usually between the age of 25 and 40 at the time of inclusion, the cohort currently includes subjects between 25 and 75 years old. A set of 21,000 participants gave consent and was selected for linkage.

Whereas researchers from OMEGA and from Biolink NL have no access to the BSN, IKNL and the IVF clinics collected this identifier in recent years. With cooperation of the clinic where they were recruited, it would thus be possible to link OMEGA subjects to the cancer registry. As linkage based on unencrypted BSN is prohibited, this variable should be hashed and pseudonymised in the IVF clinics and in the cancer registry. Linkage based on this newly created pseudonym, conducted by a Trusted Third Party, would have served as a reference set for linkages that are based on common linkage variables such as name and address.

Since the proposed linkage involved many parties and record linkage using a (pseudonymised) BSN is controversial, much time was invested in the preparation of this project. Two of the three contacted IVF clinics were willing to cooperate; the privacy officer of the third clinic decided not to cooperate based on a more strict interpretation of BSN legislation.

IKNL's Institutional Review Board responded positively to the idea of comparing three linkage methods, of which one is considered a gold standard. Despite the project being approved at the highest level within IKNL, it did take more time than anticipated to reach agreement on methods for securing all linkage variables, especially since the BSN is involved. Although datasets, contracts, and the infrastructure have been prepared, the contracts have not yet been signed, so that linkage could not be conducted before the end of Biolink NL (mid 2015). Based on recent communication however, we expect that OMEGA can link to records from the cancer registry shortly after publication of the current paper.

2.2 Linkage procedures

Ethical, legal and social issues (ELSI)

A first step in each linkage is to gain permission to use the data of interest. Subjects who participated in research cohorts can only be linked if they provided informed consent for retrieval of their records from external data sources. On the flipside, the external data source must evaluate whether their cooperation is justified. To the extent these data sources contain medical data, they can only provide access to their data on the basis of either consent or under either of the two statutory exemptions (box). The purpose and methods of using their data must be precisely described in order to get approval from sources such as the AHD, SFK or any disease registry. Under the Code of Conduct for the Use of Health Data, a researcher processing personal health data has to set up a research protocol, in the event the researcher suspects that the privacy of the research participant might be compromised or in the event he will be processing personal data. The protocol must be submitted for approval to a privacy review board.

Statutory exemptions for record linkage of health data

Research cohort
Requires informed consent for linkage to specific registries, under the Act on Data Protection (WBP).

Clinical Care Registry
Requires informed consent for linkage to external data, under the Act on Clinical Care (WGBO).

Both require a protocol and approval for linkage from the Review Board.

The linkage of personal data collected in a non-clinical setting (in casu the first two research cohorts: NTR and KOALA) is a form of processing of personal data and hence governed by the Dutch Act on Data Protection. Under the general requirements of the Act, data which has been collected and processed for a specific purpose may be used for 'secondary processing' (i.e. processing for another purpose), provided this secondary processing is not incompatible with the primary goals of the collecting and processing of the data. The Act provides that compatibility is deemed to exist, if the secondary processing is the purpose of scientific research or statistics. This raises the question of whether linking cohort data to an external source qualifies as 'secondary processing'. Notably, the Act on Data Protection does not contain specific provisions on record linkage as such. Indeed, according to the Parliamentary Explanatory Notes to the Act, record linkage is too multi-faceted to be subject to specific regulation. It is also stated in the Notes that the record linkage issue is to be dealt with in the context of compatible use. It follows from this Note that linkage of cohort data to external resources, for scientific purposes or statistics meets the statutory requirement of compatible use.

In addition to satisfying the general requirements of the Act, however, the record linkage must also meet the Act's specific requirements. As a general rule, the Act explicitly prohibits the use of personal health data. This prohibition does not apply, if, among other grounds, the data subject has explicitly consented to this use. In addition, the Act provides that personal health data may be processed for purposes of scientific research and statistics *without explicit consent* of the data subject, if asking such consent has turned out to be impossible or would require a disproportionate effort. This 'research exception' to the requirement of (explicit) consent, is conditional on (i) the research serving a general interest (as opposed to private), (ii) the processing being necessary for the research concerned and (iii) safeguards being in place to prevent disproportionate damage to the data subject's privacy. Whether or not asking consent to record linkage from their cohort participants, is impossible or would require a disproportionate effort by the research cohort has to be determined on a case by case basis. Arguably however, for research cohorts who regularly communicate with their participants by sending new questionnaires, inviting them for repeat measurement and informing them about the cohort, it seems neither impossible nor a disproportionate effort to ask their participants for their consent to link their records to external health registries. As a rule then, research cohorts should obtain their participants' explicit consent for linkage of their data to an external registry and they should put in place safeguards to protect their privacy.

The Act has been implemented in a Code of Conduct for the Use of Health Data. Pursuant to this Code, a researcher processing personal health data has to set up a research protocol, in the event the researcher suspects that the privacy of the research participant might be compromised or in the event he or she will be processing personal data. The protocol must be submitted for approval to an ethics or privacy review board.

For each cohort included in this study, consent documents were reviewed. Research participants were explicitly asked to consent to having their cohort records linked to records kept by their healthcare providers. Technically this raised an issue, as linkage was sought with the records kept by the registry rather than by the healthcare provider. To the extent, however, the records in the registry match the records held by the healthcare provider, the use of the registry records can be seen as covered by the consent.

Security

Record linkage requires that two files with identifying information be brought together. Research participants gave consent for retrieval of their data from an external source, but the opposite is usually not true. It is therefore strictly prohibited to exchange identifying information between the cohort and the registry that concerns subjects who are not in the cohort. Therefore, any linkages carried out by Biolink NL researchers must take place at the offices of the respective registries. All files with identifying data were stored on an encrypted hard drive that did not leave the registry's office.

Linkage through a TTP however requires that the linkage take place on the TTP servers. Therefore, several steps must be carried out in order to anonymise the data. Before being transferred to the TTP, identifiers are standardised and hashed at the source; this process is irreversible. The hashed data are transformed into pseudonyms; this step is also carried out using a one-way algorithm.

Record linkage at the registries' office

Several algorithms were used to perform the linkages, the most straightforward being deterministic. Simply put, records from two datasets are regarded as a link if several identifying variables match. This implies that each variable is considered equally important and that minor spelling variations can result in missed links.

In reality, a variable such as date of birth or postal code is a stronger identifier than sex or hometown. The strength of a variable further depends on the distribution of its values in the respective datasets. Therefore, probabilistic linkage methods were used that incorporate such information. The linkage algorithm calculates weights for each possible combination of record pairs, where the assignment of these weights depends on the distribution of values in both datasets.
In both deterministic and probabilistic linkage, small differences such as spelling errors can lead to missed links. There are however methods to quantify the similarity of variables in two distinct datasets, but note that this is only possible when unencrypted identifiers are available. In the linkage between the NTR and AHD we addressed the issue of minor differences by incorporating the Jaro-Winkler distance calculation into the probabilistic linkage procedure.

In contrast with deterministic linkage, probabilistic methods give a score to each possible combination of records instead of a binary output. The researcher needs to decide which threshold score must be reached before record pairs are considered a link.

Record linkage through the Mondriaan infrastructure

The Mondriaan foundation offers an infrastructure that separates identifying variables from biomedical data, while maintaining the possibility to link datasets based on personal identifiers through a Trusted Third Party. The Mondriaan client is a piece of software that runs at the office of health care providers, such as pharmacies, hospitals, and GP practices. Because it runs locally, the software can be fed with patients' personal identifiers. Every patient record is normalized in order to reduce the effect of data entry variations. All variables within each patient record are hashed and sent to the TTP, which returns a so- called source pseudonym (SP) for each patient record. If the TTP has encountered the exact record of hashes belonging to this source before, it will return the existing SP; otherwise it creates a new SP. Based on the hashed linkage variables, a linkage weight is calculated for each possible combination of records from multiple sources. If this weight reaches a certain threshold, it is concluded that both SP's identify the same person and both records are assigned the same Link Pseudonym (LP). Linked record pairs may originate from a single or from different data sources.

The linkage algorithm mixes deterministic and probabilistic methods in a stepwise manner. If the national identification number (in Dutch: *Burgerservicenummer, BSN*) matches, the algorithm stops and concludes both SP's identify the same person. If the BSN is not available as a linkage variable, the probabilistic algorithm calculates linkage weights, but only if the date of birth and sex match.

Just like the probabilistic linkage without a TTP, the strength of a variable depends on the distribution of its values, which is incorporated in the algorithm. However, because small variations in the raw value result in a completely different hash, similarity algorithms like Jaro-Winkler cannot be used by Mondriaan.

Obtaining research data

After linkage, data were requested from the external registry for only those individuals who had been successfully linked by any of the algorithms. These medical datasets did not contain any personal identifiers and were encrypted before transfer to the cohort researchers.

2.3 Linkage quality

Validation procedures

Linkage projects in this study were validated based on different types of information. Firstly, we reviewed whether personal identifiers of both datasets were correctly and uniquely linked. The personal identifiers used for validation may include the linkage variables, but can also consist of extra information such as a previous address, phone numbers, name of spouse, mother or child, etcetera. Secondly, linkage results can be evaluated by reviewing whether the content of the newly linked dataset is in agreement with the existing cohort data. Variables used for such validation are diagnosis or prescribed medication.

In the current project, both types of validation were performed by separate researchers who had no access to each other's data files. The results from both validation steps were combined and used to estimate the sensitivity and precision of each linkage. The specific validation procedures for each linkage project are described in their respective chapters.

Indicator for the representativeness of linked datasets

The percentage of records that were linked provides information on the quality of the linkage. If the number of false links and false non-links are small, this percentage corresponds with the sensitivity. But in order to accurately calculate the sensitivity and specificity, the numbers of true links and false links must be known (Ariel et al., 2014). This information can be obtained in several ways, such as manual or clerical inspection (Karmel et al., 2010; Meray et al., 2007; Victor & Mera, 2001; Zhu et al., 2009), cross-validation using a unique identifier like a personal identification number (Weber et al., 2012), or variations within the linkage keys (Lyons et al., 2009; DuVall et al., 2010; Hser and Evans, 2008) or the linkage methods (Adams et al. 1997). These procedures are however labour intensive and give little evidence on an important aspect of the quality of the linkage, namely the representativeness. Linkages with a high sensitivity can still result in biased datasets, and linkages with a low sensitivity can still be representative of a larger population.

Overall representativeness of linkage results

We propose an indicator for the similarity of the linked records to the population under investigation. This indicator measures the effect of missed links and assumes there are little to no false links. We make use of the notion that missed links lead to similar errors as selective non-response in surveys. To measure representativeness of the response of a survey, an indicator has been developed by Schouten, Cobben and Bethlehem (2009) and Shlomo, Skinner and Schouten (2012). This indicator is based on the idea that the response of a survey is representative of a target population if the response probabilities are the same for all units in the population. Because the true response probabilities are often unknown, the response of a survey is considered representative when the average response probabilities are equal for each of the subpopulations defined by a given combination of variables, X. Based on the work by Schouten et al. (2009), we introduce a representativeness indicator for linkage results, ℓ.

In parallel with surveys, the representativeness indicator for linkage is based on the probability of each record to be linked. If records from two sources are linked, the resulting links are representative of a target population if all units from the population have the same probability of being included in the linked data set, or the linkage probability (this probability is not related to the m- and u-probabilities used in probabilistic linkage). In a hypothetical situation where the linkage probabilities are known, it is very easy to evaluate the linkage result by measuring the amount of variation in the linkage probabilities. A larger variation corresponds with a less representative linked dataset. However, the linkage probability is a theoretical concept that cannot be observed. What can be observed is the value of R_i, which has the value 1 if element i links (with probability ρ_i) and otherwise has the value 0 (with probability 1-ρ_i). The idea is to estimate the linkage probabilities using auxiliary variables, chosen in such a way that the linkage probabilities are optimally explained. If a set of explanatory variables X can be found and their values X_i are observed in the linked sources, the linkage probabilities ρ_i can be replaced by the linkage propensity.

$$\rho_i(\mathbf{X}) = \Pr(R_i = 1 | \mathbf{X} = \mathbf{X}_i).$$

To estimate the linkage propensities, one could use a logistic regression model like:

$$\operatorname{logit} \rho_i(\mathbf{X}) = \log\left(\frac{\rho_i(\mathbf{X})}{1-\rho_i(\mathbf{X})}\right) = \mathbf{X}_i\boldsymbol{\beta}' .$$

Using the response probabilities or linkage probabilities, Schouten et al. (2009) introduce two measures for the representativeness. They are both based on the standard deviation

of the linkage probabilities. The only difference is in the way the indicator is scaled. First, the R-indicator, which has a value between zero and one, with a value of one corresponding to a representative dataset; and second, the coefficient of variation. The last measure has the advantage that it does not depend on the size of the linked data set. This is important when evaluating linkage results, as the size of the linked dataset changes when additional records are linked in subsequent linkage steps. Therefore, we base the representativeness indicator for linkage, ℓ, on the coefficient of variation. This coefficient has a value of zero for a representative dataset, and values larger than zero for non-representative datasets. It measures the maximum relative bias in the estimate of the population mean of a variable when this variable is maximally correlated with the non-response. The representativeness indicator for linkage, ℓ, is defined as:

$$\ell(X) = \frac{S(\rho_X)}{\bar{\rho}_X},$$

where $S(\rho_X)$ is the standard deviation of the estimated linkage probabilities and $\bar{\rho}_X$ the average linkage probability.

Representativeness of subpopulations

Besides providing an overall indicator of representativeness, this method can also be used to identify which subpopulations are under- or overrepresented. Such information may be used to improve the linkage process (e.g. including specific linkage variables for these subpopulations) and identifies possible sources of bias in further analyses. Here we describe the unconditional partial coefficients of variation (De Heij et al. 2014) which can be used for this purpose. Let Z be a categorical variable with categories $k = 1,2,\ldots,K$. Z is a component of X. Then the unconditional partial coefficient of variation of Z is defined as (Schouten et al. 2011):

$$\ell_u(Z, \rho_X) = \sqrt{\sum_{k=1}^{K} \ell_u(Z = k, \rho_X)^2},$$

with $\ell_u(Z = k, \rho_X)$ the partial unconditional coefficient of variation of category k of Z given by:

$$\ell_u(Z = k, \rho_X) = \frac{1}{\bar{\rho}_X}\frac{N_k}{N}(\bar{\rho}_{X,k} - \bar{\rho}_X),$$

where $\bar{\rho}_{X,k}$ is the average estimated linkage probability for category k of Z; N and N_k are the total number of records and the number of records in category k, respectively.

The value of $\ell_u(Z, \rho_X)$ is bounded above by $1/2\bar{\rho}_X$ and below by zero. The larger the value, the larger the contribution of Z to the lower representativeness. $\ell_u(Z=k, \rho_X)_u$ is the unconditional partial coefficient of variation of category k of Z. A positive value indicates an overrepresentation and a negative value an underrepresentation. The values are between $-1/2\bar{\rho}_X$ and $1/2\bar{\rho}_X$..

Interpretation

The interpretation of the representativeness indicator for linkage is not straightforward. It is impossible to give an absolute interpretation of the indicator in a way that there is a limit below which analyses based on the linked dataset are valid and above which they are not. This is mainly caused by the fact that the result depends largely on the chosen model (the vector of X-variables), the target variables, the sources and methods that are chosen each

specific research. Only by experimenting with different models and different sources, rules of thumb can be developed for these limits. However, even without an absolute interpretation the indicators have their uses:

- The coefficient of variation gives an upper bound on the relative standard deviation of estimates of (uncorrected) population means, thereby pointing towards the possible introduction of selection bias in the linked dataset.
- The partial unconditional coefficient of variation can be used to identify subpopulations that are under- or overrepresented.
- The coefficient of variation can be used to compare different linkage algorithms with respect to representativeness.
- The indicators can be used to determine whether additional linkage steps have increase the representativeness, or whether linkage quality remains constant when periodically linking the same registries.

3. Linkage of health insurance data to the Netherlands Twin Register

Authors
A. (Adelaide) Ariel (GGZ inGeest)
T.J. (Tina) Glasner (VU University Amsterdam)
E.C.M. (Bep) Verkerk (GGZ inGeest)
C.E.M. (Toos) van Beijsterveldt (VU University Amsterdam)
G. (Gerard) van Grootheest (GZ inGeest and VU University Medical Centre)
D.J. (Jan) van der Laan (Statistics Netherlands)
M.C.H. (Mark) de Groot (Utrecht University)
S.T. (Sipke) Visser (Mondriaan Foundation)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)
D.I. (Dorret) Boomsma (VU University Amsterdam)

3.1 Introduction

In this research project we enrich data in an academic research cohort from the Netherlands Twin Register (NTR) by linking their records to health insurance data from the Achmea Health Database (AHD).

Twin research is based on the premise that genetic and environmental effects on any trait can be studied by comparing mono- and dizygotic twin pairs who are concordant or discordant for that trait. It is of greatest importance that the two siblings of a twin pair are correctly distinguished. This also poses the main challenge of this project: young twins share most of their potential linkage variables and are therefore difficult to discriminate.

The AHD contains claims for health care consumption including visits, drug prescriptions and therapies, which potentially give valuable and detailed insight into the actual health status of a person. Record linkage between datasets can be performed using identifiers such as name, address, and date of birth.

In the present chapter we describe datasets, procedures to process the datasets, and how they were linked. The research question that we aim to answer is whether existing or customized linkage algorithms are sensitive enough to locate matching records from both datasets, but at the same time sufficiently precise to correctly distinguish siblings within a twin pair. After linkage of the datasets and validation of the results, the key question is addressed whether record linkage has led to an increase of quality and a higher coverage of the NTR dataset. Whereas the record linkage produced in the current project could be used to obtain information about a multitude of diseases, we focus on attention deficit hyperactivity disorder (ADHD) to demonstrate its validity.

3.2 Description of the datasets

The Netherlands Twin Register

The NTR recruits twins and their families and investigates individual differences in mental and physical health. While the NTR collects data from twins of all ages, an important focus of the NTR is the development of psychopathology in children. With the current linkage we aim to enrich the NTR with data that relate to attention deficit hyperactivity disorder (ADHD). Our focus lies on linkage of *young* twins, who were born after 1 January 1986. Most twins were registered at birth by their parents. The NTR recruits approximately 40 percent of all twins and multiples (Van Beijsterveldt et al. 2013) and recruitment is from all strata of society (Hoekstra et al. 2010).

The NTR mainly collects data through surveys. Mothers fill in the questionnaires when children are 0, 2, 3, 5, 7, 9/10 and 12 years old. Twins receive self-report questionnaires when they reach the age of 14. Teachers provide additional information at the age of 7, 9/10, and 12, provided that the parents give permission to approach the teachers. To facilitate longitudinal comparisons, the content of the surveys is kept fairly similar. An overview of the data collection is given in Van Beijsterveldt et al. (2013).

Permission for record linkage has been requested in questionnaires since 2005. Subjects under 16 years old were only considered for linkage if their parents gave permission. When 16 or older, subjects were only linked if they gave consent by themselves. Around 90 percent of the mothers gave permission for linkage. Mothers who gave no permission for record linkage had a lower educational attainment levels, and were more often born in a non-western country. Zygosity of twins (mono- or dizygotic), age of the mother and religion did not differ between mothers who gave permission and mothers who did not (Van Beijsterveldt et al., 2013).

The Achmea Health Database

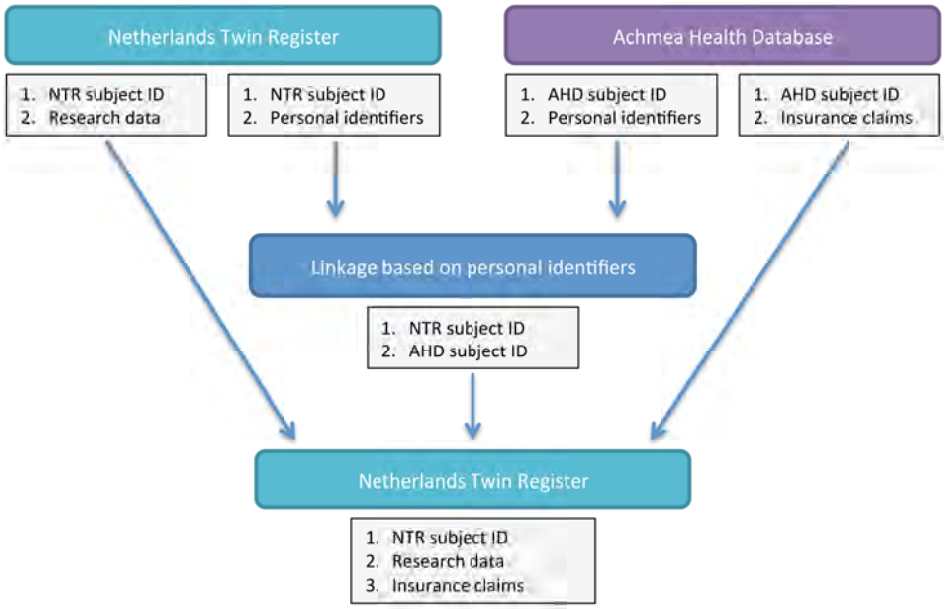
Achmea is currently the largest provider of health care insurances in the Netherlands. It maintains a database that is also accessible for scientific research purposes, the Achmea Health Database (AHD). The AHD contains information about health care consumption at general practitioners, pharmacists, dentists, hospitals, specialised psychiatric care and other health care providers. As of December 2012, the database held records of approximately 4.5 million people, or about 26.5 percent of the total Dutch population. Although Achmea has clients throughout the country, a higher coverage exists in the west, middle and east regions than in the north and south of the Netherlands. The dataset that was linked to the NTR was limited to insurances in the period from 2006 to 2013.

3.3 Methods

General approach

In order to obtain data from the AHD, a research proposal is required in which the purpose and methods of the study are explained. Variables can only be requested if they are supported by the research proposal and after approval from the AHD data access commission. The AHD usually provides only anonymous data that cannot be linked to another dataset. However, an exception was possible for the current project, which specifically focuses on the methods of record linkage. Throughout the entire process, the identifying variables were strictly separated from research and insurance data (figure 3.3.1).

3.3.1 Simplified chart depicting the separate flow of identifying variables and research and insurance data in the NTR-AHD linkage. Newly created subject ID's were used for this record linkage project.



Privacy protection

The authorisation to obtain NTR participants' health data from external sources was provided by permission from subjects themselves or from their parents. However, the NTR should never have access to AHD records of individuals who are not in the NTR or who did not provide consent. Moreover, NTR research data may never be transferred to Achmea. This means that record linkage was to be performed within the walls of Achmea's office, using files that do not contain health data.

In order to comply with these requirements, both the NTR and the AHD provided the necessary linkage variables to the Biolink researchers who performed the record linkage at Achmea's office. All files were stored on an encrypted hard drive, which never left the office. Biolink researchers had no access to any research data from the NTR or to any insurance claims from the AHD. An alternative linkage was performed using the Mondriaan infrastructure, which includes pseudonymisation and probabilistic linkage by a Trusted Third Party (TTP) as explained in chapter 2.

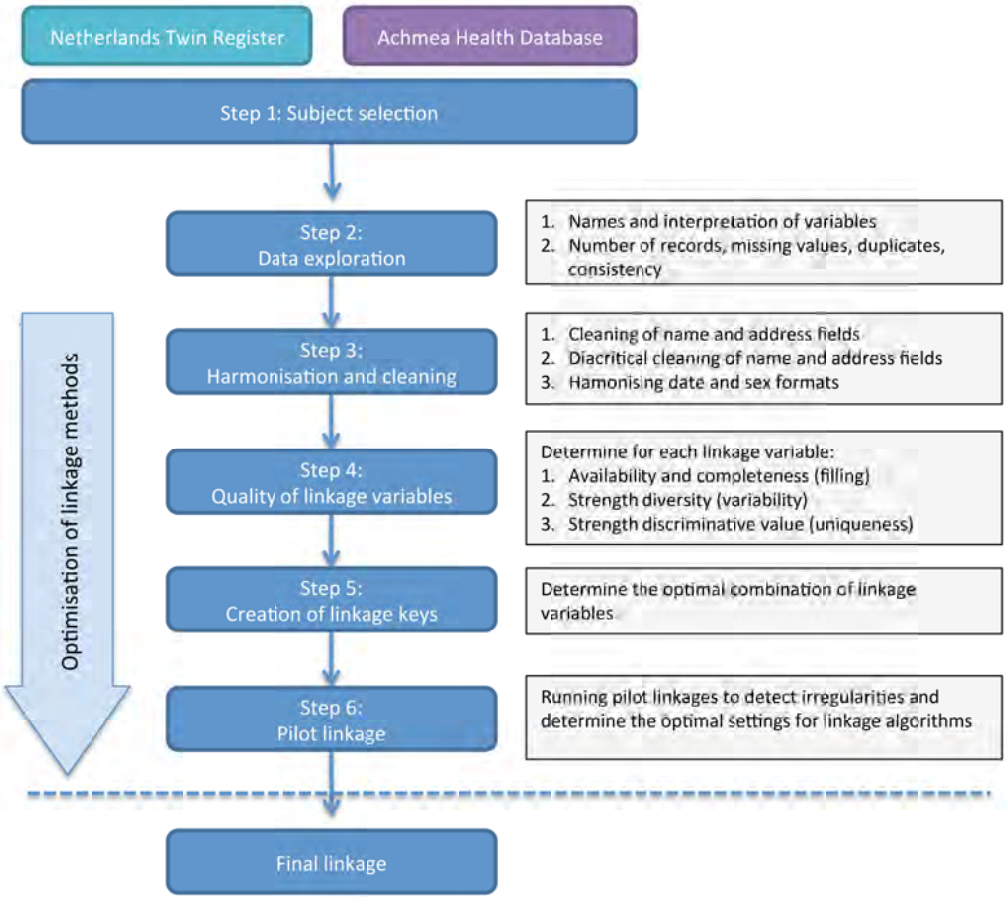
A single link file was created that contained nothing more than the person IDs of both the NTR and AHD datasets. This file contained all AHD subjects that were linked to any NTR subject in any of the linkage procedures. Achmea's data manager used this link file to extract data on ADHD medication and relevant DBC codes of all AHD records that were linked NTR subjects. These data were cleared of identifying variables and transferred to the NTR researcher through a password-protected service from Wetransfer (www.wetransfer.com).

After the record linkage was complete, the personnel who carried out the linkage validated the linkage procedures, using identifying variables that were not included in the linkage key. Validation of the consistency of the AHD data with existing information within the NTR was performed by NTR researchers.

Preparation for record linkage

We carefully selected relevant populations from both datasets and explored the availability and quality of potential linkage variables. In order to achieve optimal linkage results, several steps were carried out that improved linkage quality (figure 3.3.2).

3.3.2 Overview of the data preparation steps and retrieval of settings for the linkage

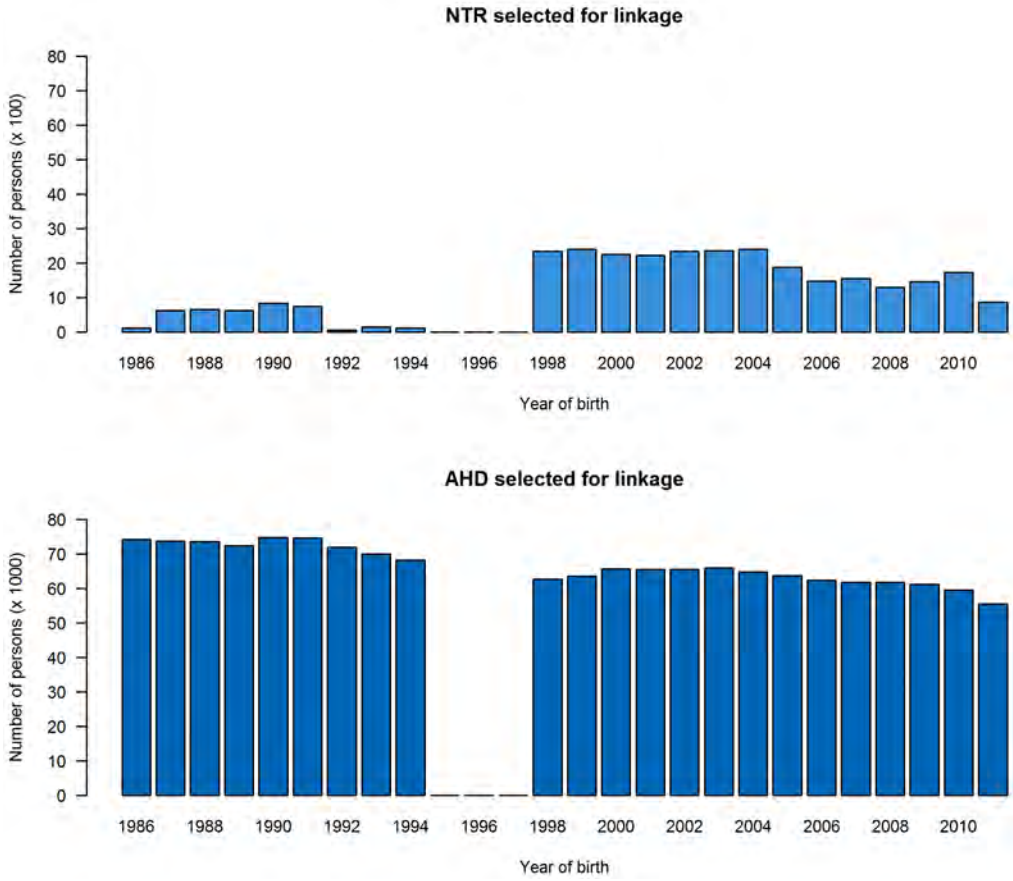


Step 1: Subject selection

All NTR subjects who were born between 1986 and 2011 and gave consent were selected for record linkage. No consent was available from subjects who had just turned 16 but had not yet been approached for a new NTR questionnaire. For this reason, no subjects born in the years 1995–1997 were linked (figure 3.3.3).

The NTR selection included 30,383 subjects from 15,555 families. Only subjects in the AHD with corresponding years of birth were selected, thereby strongly reducing the number of records to be considered for linkage. The AHD contained 1,532,675 subjects with the birth years of interest. In both databases multiple records were created for subjects with multiple addresses recorded over time.

3.3.3 Number of subjects available for linkage in both datasets by year of birth



Step 2: Exploration of datasets and linkage variables

The NTR dataset contained 38,237 different address records for 30,383 unique persons. Within the NTR dataset, 76 twin pairs did not have the exact same date of birth. The difference, generally no more than two weeks, was not caused by data entry errors but reflected the rare situation where twins were born on different days. In addition, there were nine twin pairs who had small spelling differences in their surname, indicating data entry errors. Two adolescent twin pairs had completely different surnames, possibly resulting from marriage.

The AHD dataset contained 1,602,201 address records for 1,532,675 persons. In most cases, the presence of multiple records for an individual indicated a change of address. Also, for 3,627 individuals, multiple records existed because they were registered under different surnames over time. Because 85 percent of these subjects were adult women and their surnames were entirely different, the name inconsistencies were probably due to marriage and did not indicate data entry errors. 820 individuals occurred in different records with different initials, and 22 persons had a different sex in two records. As it was unclear which values were correct, all records were included for linkage.

Both datasets contained given name, surname, initials, address, and date of birth. The NTR also provided maiden names, if applicable. The given names and initials did not necessarily correspond between the two datasets, as illustrated in table 3.3.4 (fictional data). It is important to realise that in the Netherlands, the field for the given name (*'voornaam'*) may contain either the official given names, or the more informal name by which one is generally known, *'roepnaam'*. The variable for given name in the AHD dataset could either refer to the one or the other, but was only available in 17 percent of the records. While the NTR provided

two separate fields for the official given names and for *roepnaam*, only the latter was very complete. Thus, when given names were included as a linkage variable, the *roepnaam* from the NTR was linked with the given name from the AHD.

3.3.4 Name variables in the NTR and AHD datasets (fictional data)

NTR						Achmea			
person ID		surname	given name (official)	given name (roepnaam)	initials	person ID	surname	given name	initials
N1		Postma	Anton Pieter	Twan	A.P.	A1	Postma	Anton Pieter	A.P.
N2		Postma	Gerardus Jan	Jan	G.J.	A2	Postma	Jan	G.J.
N3		Stoop		Julia	J.F.	A3	Stoop	Juliana	J.F.
N4		Stoop		Bjørn	B.	A4	Stoop	Bjorn	B.J.
N5		De Vries	Hetty	Hetty	H.	A5	Kelder	Hetty	H.
						A5	De Vries	Hetty	H.
N6		De Vries	Helma	Helma	H.	A6	De Vries	Helma	H.
N7		Huntelaar		Emma	E.	A7	Huntelaar		E.G.
N8	0	Huntelaar		Daan	D.		–		

Step 3: Data harmonisation and cleaning

Harmonisation and cleaning of both datasets is especially important for correct linkage of names that may be spelled slightly differently in the two datasets. Characters such as spaces, dashes, digits and diacritical marks were removed from all text fields; sex and date fields were stored in uniform formats.

Box 3.1 Procedures that were applied in order to harmonise both datasets prior to linkage

Procedure	Description
Basic cleaning of name and address fields	Removal of punctuation, blanks and digits from text fields and changing all cleaned identifiers in uppercase
Diacritical cleaning of name and address fields	Replacing letters having diacritical signs with only the letters. Removal of irregular characters.
Reformatting date of birth and sex	Formatting dates as 'yyyymmdd', and sex as '1' for male and '2' for female.

Step 4: Determining the availability and strength of linkage variables

The strength of a linkage variable depends on how complete it is, and how many different values occur. It is preferable to use variables with few missing values and many different values, such as postal code and surname (see table 3.3.5). Variables that have a high number of missing values may be excluded from the linkage key. The strength of a linkage variable depends not only on the number of different values, but also on the distribution of its values: it is harder to reliably link subjects who have a very common name, than those who have an uncommon name. Probabilistic linkage algorithms take this distribution of values into account.

Table 3.3.5 shows the numbers of missing values and the number of different values for each identifier in both datasets. Given names (official names and *roepnaam*) are highly discriminative but unfortunately they are usually not registered in the AHD. House number and sex both have few missing values but are less discriminative than the other variables.

3.3.5 The numbers of missing values and different values in both datasets

	NTR (N = 30,383)			AHD (N = 1,532,675)		
	number of missing values	(%)	number of different values	number of missing values	(%)	number of different values
Surname	0	0%	8,589	0	0%	150,597
Given name (official)	19,114	(62.1%)	8,760	1,279,626	(83.5%)	34,391
Given name (roepnaam)	26	(<0.1%)	5,124			
Initials	3	(<0.1%)	2,529	13	(<0.1%)	35,004
Date of birth	0	0%	5,888	0	0%	8,400
Sex	1	(<0.1%)	2	0	0%	2
Postal code	0	0%	18,950	0	0%	303,177
House number	1	(<0.1%)	686	4,029	(0.3%)	4,921

Step 5: Choosing linkage keys from linkage variables

A linkage key is the combination of a chosen set of linkage variables. When more variables are included into this key, a higher discriminative power is achieved. However, with an increasing number of linkage variables, the chance of mismatching information in any of the variables also increases. Variables that are susceptible to spelling errors, such as names, can be trimmed to the first few characters to circumvent this unwanted effect.

The ideal linkage key achieves a high unique key rate with a minimum set of variables. The strength of various linkage keys can be expressed as a unique key rate, which indicates how many records can be uniquely identified based on this key. Several linkage keys produced a unique key rate close to 100% (table 3.3.6), meaning that nearly every record could be uniquely identified with these linkage keys. These rates can however be very different for both datasets. For example, when surnames were not included in the linkage key, the unique key rate dropped considerably in the AHD, but not in the smaller NTR. When initials and given names were not included, the linkage key became weak in the NTR, but not in the AHD, where twins constitute only a small proportion of the records.

3.3.6 The unique key rates for each linkage key and for both datasets

Linkage key	Unique key rate NTR (%)	Unique key rate AHD (%)
Surname, given name, initials, DOB, sex, postal code	100.00	99.98
Surname, given name, initials, DOB, sex	100.00	99.96
Surname, given name-4[1)], initials, DOB, sex	99.99	99.96
Surname, initials, DOB, sex, postal code	99.03	99.97
Surname, initials, DOB, sex	99.02	99.94
Surname, given name, DOB, sex	99.98	98.46
Surname, given name-4[1)], DOB, sex	99.97	98.45
Surname-4[1)], given name-4[1)], DOB, sex	99.97	96.82
Given name, initials, DOB, sex	99.63	77.27
Given name-4[1)], initials, DOB, sex	99.57	77.23
DOB, sex, postcode, house number	67.38	99.20

[1)] Only the first four characters of this variable were used in the linkage key.

Step 6: Pilot record linkage
A pilot linkage was performed in order to ascertain whether matching records in both datasets show complete or partial agreement. Agreement can be expressed as the number of variables that show an exact match, but it is also possible to calculate partial agreement within linkage variables using the Jaro-Winkler (J-W) distance measurement. Whereas the previous step provided information about the strength of each linkage variable, the current step helps to identify which variables show most overlap between the two datasets and thus have a low error rate.

As the number of records in datasets becomes larger, the number of possible links increases quadratically. The resulting high computational demand can be greatly reduced if both datasets are stratified based on one or more variables that are considered very complete and reliable and links can only be made within each stratum or block. We used sex and the year of birth as blocking variables. Based upon the results from this step and the previous, we chose a linkage key that consists of variables that are available, identifying, and of sufficient quality (box 3.3).

Box 3.3 Linkage key that was used in the pilot linkage of the AHD to the NTR

Linkage key for the pilot linkage:
Surname-4, full initials, first initial, date of birth, sex, postal code.

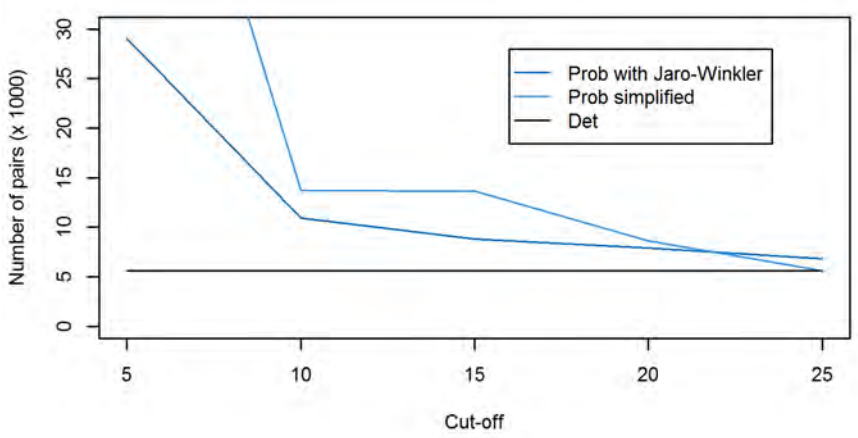
Deterministic linkage where all linkage variables were required to match resulted in 5,613 links for 5,552 NTR individuals, which was expected to be an underestimation of the true overlap between both datasets. A higher number of links could be achieved using probabilistic methods. The number of links retrieved by a probabilistic algorithm depends on how the cut-off value is set. Figure 3.3.7 shows the number of pairs that are obtained using three different linkage algorithms, with an increasing cut-off score. The choice of a reasonable cut-off score is rather subjective, but nevertheless based on a number of considerations. In general, a high value leads to results that are very similar to a deterministic linkage. If the value is set too low, the number of total links becomes unrealistically high and subjects may be linked to more than one record – especially when linking twins. Since the AHD does not contain a given name for most subjects, no information is available to validate such false links, other than the variables that are already included in the linkage key. Taking these considerations into account, a cut-off value of 20 was chosen for the ultimate linkage.

The linkage pilot showed that a number of NTR subjects were each linked to more than one AHD individual, usually a sibling. Apparently, several AHD records were similar enough to result in false positive links. Most of these false positive links could be resolved based by selecting the link pair with the highest score or by manual inspection of given names. However, there were still some multiple links with very similar scores. We detail these in the result section.

Linked record pairs from both datasets that agreed on many identifiers often disagreed on the initials: about 35 percent of the linked NTR records had a one-letter initial while

the corresponding AHD records had initials consisting of multiple letters. This discrepancy was explained by the fact that the NTR had previously deducted a missing initial from the *roepnaam*, while the AHD always recorded full initials. Because it was unclear for which records this deduction of initials was done, we did not only include the full initials, but also the first initial in the linkage keys for probabilistic linkage. Consequently, the initials had a relatively high impact in the linkage.

3.3.7 The number of link pairs obtained at different cut-off values, using different linkage algorithms.



The probabilistic pilot linkage further showed that the postal codes in 2.4% of the AHD records were invalid (the 4 digits were '0000') and should be considered a missing value. Based on the pilot results, we decided to use the linkage keys given in box 3.4.

Box 3.4 Linkage keys that were chosen for the linkage of the AHD to the NTR

Linkage key for deterministic linkage:
Surname-4, full initials, date of birth, sex, postal code.

Linkage key for simple probabilistic linkage:
Surname-4, full initials, first initial, date of birth, sex, postal code.

Linkage key for probabilistic linkage with Jaro-Winkler distance calculation:
Surname, given name, full initials, first initial, date of birth, sex, postal code.

Linkage key for linkage by Mondriaan:
Deterministic step: Date of birth, sex.
Probabilistic step: Surname, given name, full initials, postal code, address.
In addition, information such as policy number and NTR family ID were used to improve linkage.

Final record linkage

As a final step, records were linked with four different methods. Like in step 5, sex and the year of birth were used as blocking variables for the linkages that were performed at Achmea's office.

In principle, a linkage key should include as few variables as possible, while maintaining a unique combination of identifiers. Even though the linkage keys that do not include postal codes can have a very high unique key rates, the fact that the initials are not so reliable in the NTR and given names are usually missing in the AHD may still result in false links. For this reason, the postal code was included into the linkage keys. Even when records with multiple initials did not agree completely between the two datasets, the first initial was often correct. For this reason, the first initial was also used as a linkage variable beside the full initials.

Deterministic and simple probabilistic linkages were applied using a self-written procedure in SAS® 9.3 (SAS Institute, Cary, North Carolina, USA). For these linkages, surnames were reduced to their first four characters, and no given names were included in the linkage key. These limitations reduced the effect of minor disagreement between both datasets.

Probabilistic Jaro-Winkler linkage was conducted using Registry Plus™ Link Plus, a publicly available record linkage program (Atlanta (GA): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion; 2010). Unlike the deterministic and simple probabilistic linkages, this linkage included the complete surname and given name. Agreement of names was calculated using the Jaro-Winkler distance metric. The frequency distribution within variables was also taken into account. For example, agreement on a common name was assigned a lower weight than agreement on a rare name.

No blocking was necessary prior to the Mondriaan linkage, as their algorithm only links records if sex and the date of birth match completely. This algorithm made use of a few additional linkage variables, being street, house number and town.

Validation

After completion of the linkage, personal identifiers that were not included in the linkage key were used to validate whether records from both datasets were correctly and uniquely linked. In parallel, it was evaluated whether the content data retrieved from the AHD corresponded with the NTR dataset.

Validation of linkage based on identifying variables

All linked record pairs that resulted from probabilistic linkage were classified into three distinct categories of certainty. Pairs were classified as links if all identifiers matched and given names or multiple initials were available. Pairs were classified as possible links if most identifiers matched, but given names were missing. If an NTR subject was linked to more than one AHD subject, the link pairs with disagreeing given names were classified as false links and discarded.

Whereas the NTR keeps information about family relationships, this information is only implicitly available in the AHD. Family members do not necessarily have the same insurance provider, so that they do not always share the same policy number. However, if different individuals have the same insurance number, this is a strong indicator that they are from the same family. We used the insurance number to find the missing sibling if not all children from a family could be linked.

Another method by which false negatives were identified was the use of an additional dataset from the NTR that contained records of a number of mothers of the selected twins. Linking the mothers of non-linked twin pairs to the AHD allowed us to see whether they did have children on the same insurance policy.

Information on false positives was derived from the notion that a person in one dataset should be linked to no more than one person in the other dataset. In box 3.5 below we summarize the complete conditions based on family and linkage relationship, and their implications.

Box 3.5 Possible scenarios that provide information on the linkage validity

Scenario	Validity
NTR subjects from one family are paired with AHD individuals with the same insurance number.	This indicates correct links, at least at the family level.
NTR subjects from one family are paired with AHD individuals who have different insurance numbers.	Links cannot be validated based on family relations.
NTR subjects from different families are paired with different AHD individuals who have the same insurance number.	This indicates false positives, as it is unlikely that twins from different families have the same insurance number.
One NTR subject has been paired to different AHD individuals	This indicates false positives.
An NTR subject has not been linked, but the mother can be found in the AHD, together with a child with a matching date of birth.	This indicates false negatives.

Validation of content

The quality of the linked dataset was validated based on the combined content of the two datasets. The following three questions were addressed in order to provide insight into the agreement or disagreement between both sources:

a) For how many of the linked NTR participants do both data sources (NTR and AHD) indicate that the participant was treated for having ADHD?
b) For how many of the linked NTR participants does the NTR data point towards ADHD medication use or medical specialist treatment, that is not confirmed by the AHD?
c) For how many of those participants does the AHD show ADHD treatment that has not been reported to the NTR?

Data that originates from two sources may have been collected at different moments in time. Not all subjects filled in the NTR questionnaires in the same year, and may have changed their health insurance provider during the period of interest. In order to interpret the results correctly, data were aggregated into calendar years.

Two files provided by the AHD contained information on health care consumption of the linked persons. One file reported all medication that was claimed; each drug was coded according to the anatomical therapeutic coding (ATC) system and was accompanied with the date of dispense. The drug that indicates ADHD is methylphenidate (ATC: N06BA04), which is marketed in the Netherlands in the following formulations and brand names: Ritalin, Concerta, Equasym, Medikinet, Strattera, and generic methylphenidate tablets. For the present validation, only ATC codes starting with N06B (psychostimulants, agents used for ADHD and nootropics) were taken into consideration.

The second file contained records for all therapeutic contacts with a health care professional. These were coded according to the Dutch billing system, which is based on DBC codes (*diagnosebehandelingcombinatie*). DBC codes combine information about diagnosis and the type and duration of treatment. Only the DBC codes that indicated 'attention deficit or behavioural disorders' were used for the present analyses.

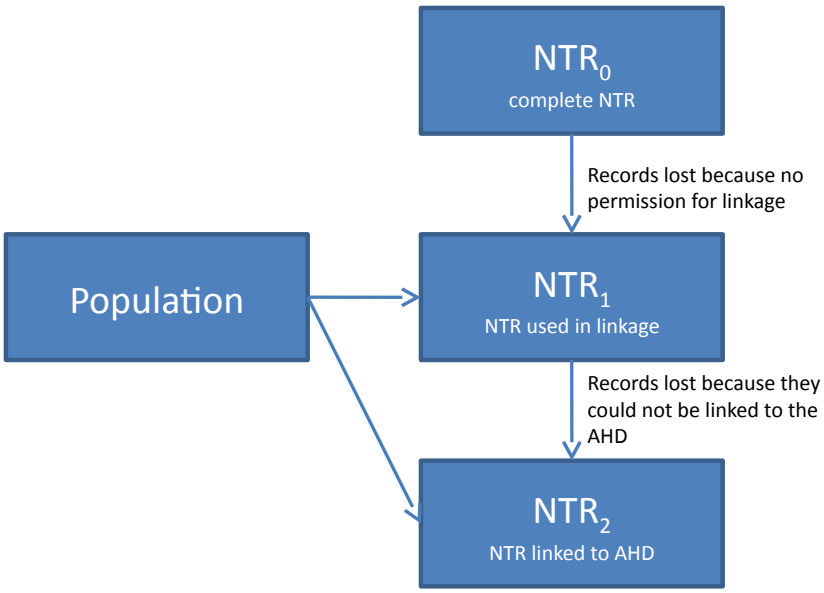
For each year from 2006 to 2013 we indicated whether NTR survey data about medication or scores on the Child Behaviour Checklist (CBCL) pointed towards ADHD. We subsequently analysed whether these data were confirmed by health insurance records. If NTR data were missing, the AHD data were used to enrich the NTR database.

As an additional measure of content validity, we compared the use of methylphenidate with measures of attention problems from the Child Behavior Checklist (CBCL) (Achenbach, 1991), completed by the twins' mother between 2006 and 2013 when the children were 7, 10, and 12 years old. Attention Problems were scored (Verhulst et al., 1996) and converted to T-scores (standardized score (separately for boys and girls) with a mean of 50 and SD =10). A possible ADHD case was defined as follows: when 2 or 3 time-points were available (age 7, 10, and 12) a case must have a T-score above 60 at all available time-points and a T-score equal or above 65 at least once. When 1 time-point was available (for 68% of the cases) a case must have a T-score equal or above 65.

Representativeness of the NTR-AHD linkage

Besides the validity of the linkage procedures, it is also important to know whether the linked dataset is representative of the target population – in this case the entire population of Dutch twins that are born since 1986. A number of steps in the record linkage process could cause the linked dataset to be less representative (figure 3.3.85). First, the NTR contains a subset of Dutch twins and participation in the NTR might be selective (NTR0). Second, only those NTR subjects who gave permission for record linkage were included (NTR1) and the question asking for permission was not included during the first years of data collection. Third, subjects could only be successfully linked (NTR2), if they had an Achmea health insurance and if the linkage variables were of good quality.

3.3.8 The (sub) populations used in the analysis of the representativeness of the NTR-AHD linkage



In order to determine the representativeness, we calculated the representativeness indicator for linkage introduced in chapter 2. This coefficient of variation was used to show the effect of each of the aforementioned steps on the representativeness. We compared NTR1 and NTR2 to the Dutch population, and compared NTR2 to NTR1.

The reference population was determined at Statistics Netherlands by selecting all children born in the period 1986–2009, who have a sibling born in the same month and to the same mother, from the population register of 31 December 2013.

The variables used in the calculation of the representativeness were selected on availability, expected correlation with the subject selection, with linkage probabilities, and with the research topic ADHD. Age, sex, and ethnicity are important variables in many research questions. Furthermore, coverage of the AHD differs in different regions of the Netherlands, and participation in surveys may correlate with urbanisation. For each dataset (the 3 NTR datasets and the population) a cross-table was created of the number of subjects by year of birth, sex, ethnicity, urbanisation, and region of the Netherlands.

3.4 Results

Linkage results

The deterministic linkage resulted in 5,552 linked NTR subjects (18% of the NTR dataset). The Mondriaan procedure linked 5,974 NTR subjects (20%), the simple probabilistic method 7,600 (25%), and the probabilistic linkage that uses the Jaro-Winkler distance algorithm linked 7,944 NTR subjects (26%) to the AHD. These 7,944 subjects belonged to 4,219 families, or 27 percent of all families in the NTR dataset. Adolescents were less likely to be linked (22%) than children (27%).

The Jaro-Winkler linkage identified an additional 467 links that were not in the deterministic and simple probabilistic linkages, but rejected 124 previously linked pairs. Verification of these 124 link pairs revealed that they were incorrectly linked siblings. The J-W linkage included nearly all links that were produced by the other three linkages. Therefore, we focus on the results of the J-W linkage in this section.

Of all the records linked by Mondriaan 98.5% were also identified by the J-W linkage. Verification of the 89 Mondriaan links that were not previously linked showed that 45 were really ‘new’ links. Another 38 cases involved subjects that were linked differently in the previous linkage, usually to the other twin. Only six linked record pairs disagreed to such an extent that we considered them as false links.

Linkage evaluation

Based on the agreement of linkage variables and the availability of initials or given names, all link pairs produced by the J-W linkage were classified into *links*, *possible links*, and *false links* (table 3.4.1). Out of 8,080 linked record pairs, 6,628 pairs (82%) showed a high agreement on all linkage variables, while siblings could be distinguished based on their names or initials. These pairs were classified as *links*.

A group of 1,404 linked pairs (17%) were categorised as *possible links*. These pairs showed a good agreement on most linkage variables, but siblings may have been incorrectly linked because given names or initials were of insufficient quality.

The fact that a number of 7,944 NTR subjects were linked to 8,080 AHD subjects implies that at least 136 links were false positives. 48 of these pairs were confirmed as *false links*, since given names did not match. The remaining 88 did not contain any given names and could not be validated; these were retained in the dataset as *possible links*.

Mondriaan seemed to employ a stricter threshold than was used in the J-W linkage and this majorly had an effect on the links that we classified as *possible links*. Whereas 87 percent of the *links* were also found by Mondriaan, only 9 percent of the *possible links* were also linked by Mondriaan. The mean weights that we calculated were 29.7 (±4.40) and 31.6 (±3.21) for the links produced in the J-W and Mondriaan linkages, respectively.

3.4.1 Detailed specification of the number of links, possible links, and false links. The number between brackets indicates the number of unique NTR subjects that was linked to more than one AHD individual.

		Number of linked NTR subjects			
		children	adolescents	total	also linked by Mondriaan
Links					
Given names are available, or the initials contain multiple letters; disagreement on at most one variable.	Twin siblings can be distinguished based on name, initials, or sex.	2,468 (0)	221 (0)	2,689 (0)	2,576 (0)
Given names are missing; disagreement on at most one variable.	Twin siblings can be distinguished based on initials or sex.	3,646 (0)	293 (0)	3,939 (0)	3,192 (0)
Possible links					
Given names are missing; disagreement on at most one variable.	Twin siblings have the same initials and sex.	28 (9)	12 (3)	40 (12)	18 (5)
	Unknown: only one sibling was included.	3 (0)	76 (0)	79 (0)	71 (0)
Given names are missing; disagreement on more than one variable.	Twin siblings can be distinguished based on initials or sex.	937 (28)	177 (2)	1,114 (30)	35 (0)
	Twin siblings have the same initials and sex.	53 (26)	7 (2)	60 (28)	1 (0)
	Unknown: only one sibling was included.	1 (0)	86 (18)	87 (18)	2 (0)
Agreement on given names and other identifiers; disagreement on the initials.	The initials of siblings appear to be swapped in the AHD.	21 (0)	3 (0)	24 (0)	6 (0)
False links					
Duplicate links that were discarded because of disagreement on given names.		25 (0)	23 (1)	48 (1)	2 (0)
Total		7,182 (63)	898 (26)	8,080 (89)	5,903 (7)

Detailed evaluation of linkage variables

Two linked record pairs disagreed on the date of birth: the day and month appeared to be swapped in one of the datasets. One record pair was linked despite the fact that the surname disagreed completely. As all other identifiers matched, these three linked pairs were classified as *links*.

The NTR dataset contained 300 twin pairs in which the siblings had the same sex and initials. When such twins were linked but no given name was recorded in the AHD, the siblings could not be discriminated. This situation applied to 50 linked twin pairs (i.e. 100 twin siblings). From 789 NTR families, only one child was available for linkage and no information about the other sibling's initials or sex was provided. In the AHD, the given name or the initials appeared to be swapped between the two siblings of 24 twin pairs. For example, Carine had the initial J., while her twin sister Janet had the initial C. In these situations it was not possible

to validate whether the correct sibling was linked; these were classified as *possible links* but may be excluded from further analyses.

Family relations
Approximately 26 percent of the Dutch have a health insurance with Achmea, and around 26 percent of the NTR records were linked with the AHD, indicating a fairly low number of missed links. Nevertheless, a number of such false negatives were identified using information from family members.

Although siblings are not necessarily insured under the same policy number, this variable could be used to identify false negatives. When both siblings of a twin pair were linked, the siblings did usually share the same policy number (96%). There were 320 twin pairs where only one of the siblings was linked. When the AHD was searched for the insurance number, sex, and date of birth, 92 of the missing siblings were found. These false negatives mostly disagreed with the NTR on postal codes and initials.

In order to estimate the total number of false negatives in this linkage project, we used information provided by the twins' mothers. Linkage variables were available for approximately 9,800 mothers of not-linked children. Using these linkage variables, 118 mothers were found in the AHD, leading to the identification of 80 twin pairs that did match on both sex and date of birth. These false negatives mostly disagreed with the NTR on postal codes, initials, and last names.

Content evaluation

Validation of content was carried out on 7,813 NTR subjects that were linked to 7,901 AHD-subjects. This number is smaller than the number of subjects that were actually linked, because children who turned 16 in the period between linkage and actual transfer of the insurance data were not linked: permission provided by their parents lost validity at that moment. The 88 NTR subjects who were linked to two different AHD records and could not be resolved in the previous validation steps were retained in these results.

Insurance period
We verified that all linked NTR individuals had a basic insurance ('*basisverzekering*') with Achmea at any moment from 2006 to 2013; supplemental insurance packages were irrelevant for the present topic. Table 3.4.2 shows how many of the linked subjects were present in the AHD per calendar year.

Health insurances are typically bought for entire calendar years. The start and end date of each insurance policy might be used to check whether subjects had been insured the entire year if NTR data are not confirmed by the AHD data. However, we found that most subjects who were insured for a shorter period than nine months were young children born during that year. The number of insured subjects is highest in 2013 for the simple reason that many linked children were not yet born in earlier years.

Medication, diagnosis and treatment in the AHD
The AHD file with the reimbursed drugs covered the period 2006 to 2012; data from 2013 were not yet available at the time of retrieval. For each dispense, the date of prescription, ATC code, and cost of purchase were provided. The total number of drug dispenses within ATC category N06B was 4,713; these drugs were received by 211 AHD subjects (linked to 209 NTR subjects).

The AHD file with diagnosis and treatment codes covered the period 2007 to 2013. Each record contained the date of diagnosis, DBC code, and cost of treatment. In total, 1,392 DBC codes indicated a psychiatric diagnosis or treatment for 630 AHD subjects; 121 of these subjects had a DBC code for attention problems or conduct disorders. The number of registered DBC's was much smaller in the years 2007–2009 than in more recent years, possibly because this new declaration system was not fully implemented in mental health care. In addition, the data suggest that a large number of DBC's from the year 2013 were not yet registered in the AHD.

The number of subjects with drug treatment and psychiatric treatment are specified in table 3.4.2. In each year, more subjects received ADHD drugs than psychiatric treatment: drug use may continue after the initial diagnosis has been made. Even so, 45 subjects with an ADHD-related DBC did not use any drug within category N06B. Thus in total, 256 subjects had an ATC or DBC code related to ADHD.

A few minor discrepancies were found: 22 records in the ATC file were claimed in a year that no valid insurance was found for the concerning subjects. In these situations, subjects were considered as insured in that year.

3.4.2 The number of linked subjects and the prevalence of ADHD, per calendar year and at any moment between 2006 and 2013

	Insured subjects per year		Subjects with ADHD medication		Subjects with ADHD diagnosis and treatment	
Year	N	percentage of NTR1	N	percentage of insured subjects	N	percentage of insured subjects
2006	3,382	14.4	22	0.7	–	–
2007	3,642	14.5	34	0.9	5	0.1
2008	3,789	14.4	46	1.2	16	0.4
2009	4,893	17.6	69	1.4	30	0.6
2010	5,110	17.3	102	2.0	63	1.2
2011	5,143	16.9	144	2.8	65	1.3
2012	5,276	17.4	163	3.1	68	1.3
2013	6,391	21.0	–	–	39	0.6
2006–2013	7,901	26.0	211	2.7	121	1.5

Medication and behaviour scores in the NTR

For the period between 2005 and 2013, the NTR had data on medication from surveys 2 to 12 for 17,238 children, which included the ATC code, date of prescription, and date of the survey. In this period, 432 NTR subjects (2.5%) reported methylphenidate use. This information on general medication use was available for 3,127 of the 7,813 linked subjects; for 86 of the linked subjects (2.8%) the use of methylphenidate was reported at any moment. CBCL scores (collected at age 7, 9/10, or 12) were available for 2,292 linked NTR children born between 1998 and 2005. Comparison of ADHD medication use and treatment from AHD to NTR data

The AHD confirmed the use of methylphenidate for 64 of the 86 linked subjects (74%) who reported this drug in the NTR surveys between 2005 and 2013. DBC codes for ADHD were found for 20 of the 86 subjects (23.3%). The twelve cases that were not confirmed by the AHD had no insurance with Achmea in the years they filled in their NTR survey.

Table 3.4.3 shows the agreement between AHD and NTR data per year. Linked record pairs were consistent if both the NTR and the AHD data indicated that the participant had taken ADHD medication or if both sources indicated no use of ADHD medication. If only one of the sources indicated ADHD medication use, the case was classified as inconsistent. Most linked record pairs were not included either because NTR data were missing or the participant was not insured with Achmea for the year in question.

3.4.3 Consistency and agreement of methylphenidate use in the AHD and NTR datasets per year

	2006	2007	2008	2009	2010	2011	2012
Consistent							
ADHD medication in both sources	5	2	4	7	15	10	5
No ADHD medication in either source	490	539	319	536	754	501	550
Inconsistent							
ADHD drugs in NTR but not AHD	0	1	1	2	0	3	1
ADHD drugs in AHD but not NTR	1	0	0	0	0	1	2

Further examination of the twelve inconsistent record pairs showed that they were not necessarily caused by incorrect linkage. In ten cases, both datasets indicated the use of ADHD medication, but in slightly different time periods. For instance, medication use started in December according to the NTR survey, but in January according the AHD database. Two inconsistencies occurred in a single twin pair that was disconcordant on ADHD. It is uncertain whether the error occurred in the record linkage process, in the AHD dataset, or in the NTR dataset.

Comparison of CBCL scores and ADHD treatment

Children who had received ADHD treatment (methylphenidate and/or a related DBC) had significantly higher mean T-scores for attention problems (65.1 ±14.1) than those who had not received treatment (48.7 ±9.0; p<0.001). 29.3% of the children defined as an ADHD case based on CBCL scores by the NTR did receive ADHD treatment, versus 2.9% of the children who were not considered an ADHD case (p<0.001; table 3.4.4). A total of 44 percent of the children who were treated for ADHD were recognised as having attention problems by the NTR. This indicates that only looking at attention problems may be too limited and that information on other symptoms needs to be included (see Derks et al. 2007).

3.4.4 Professional treatment of subjects who are considered ADHD cases based on CBCL scores

		ATC or DBC for ADHD				
		no		yes		total
High CBCL score for attention problems	No	2,064	(97.1%)	61	(2.9%)	2,125
	Yes	118	(70.7%)	49	(29.3%)	167

Representativeness of the NTR-AHD linkage

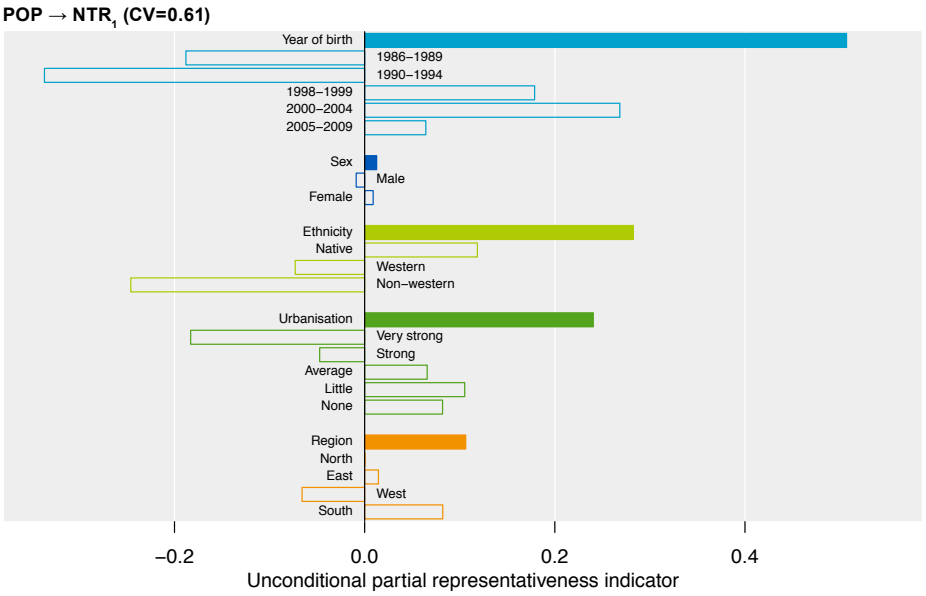
Figure 3.4.5 shows the coefficient of variation and the partial coefficients of variation for two NTR subsets (NTR1 and NTR2), compared to each other and to the population. The variable that has the largest effect on the non-representativeness of the NTR datasets is the year of birth. Subjects who were 16 or older (born before 1998) were strongly underrepresented.

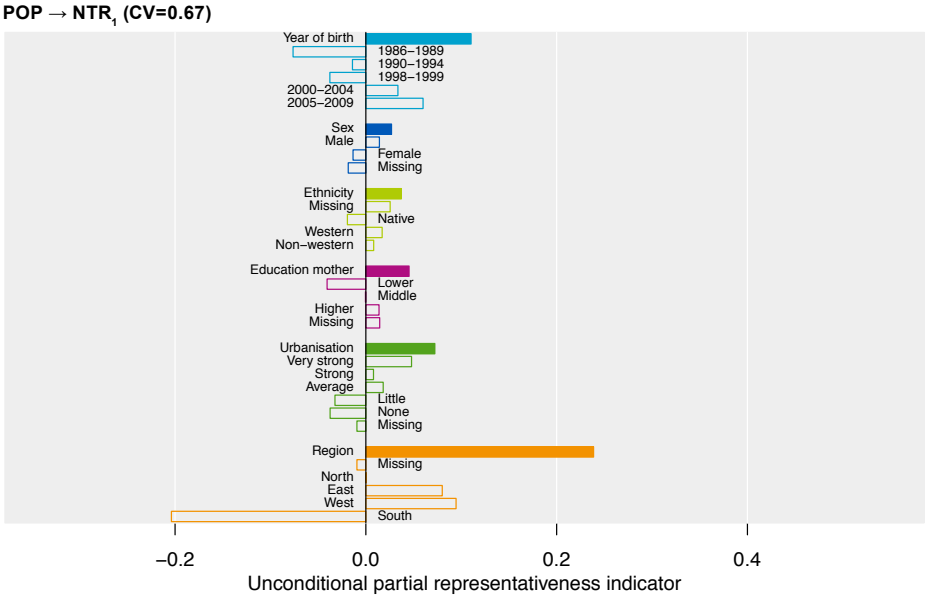
This is mainly caused by the absence of permission for linkage in this group, but the younger subjects also have a higher chance of being linked. Furthermore, twins living in urbanised municipalities and twins with a non-native background are underrepresented in the NTR.

The linkage slightly decreased the representativeness of the datasets (the coefficient of variation increased from 0.61 to 0.67). This was mainly caused by regional differences in coverage of the NTR and the AHD. NTR1 contained relatively many subjects from the south and few subjects from the west. After linkage however, the south was underrepresented and the east and west were overrepresented.

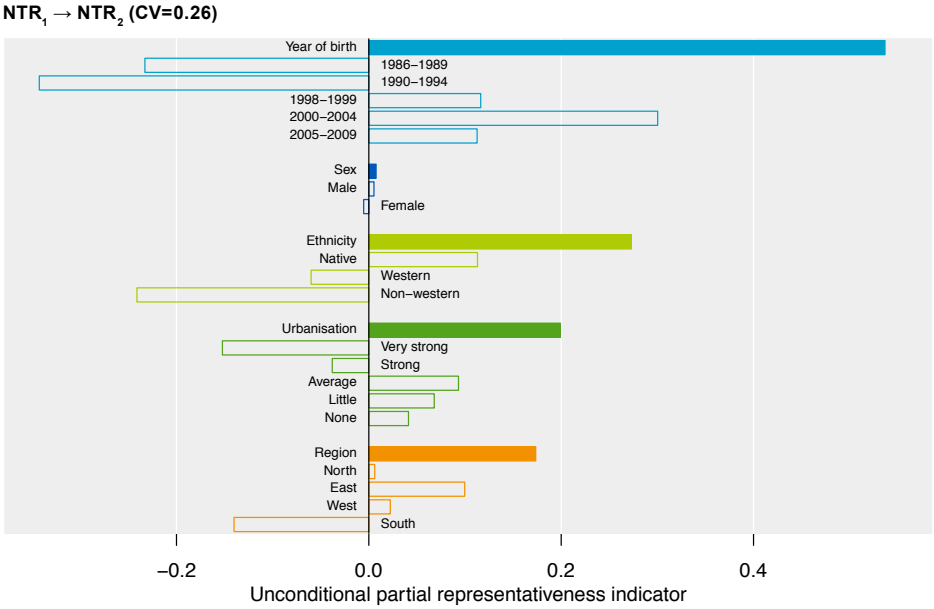
The datasets NTR1 and NTR2 were representative of the population regarding sex. Females only had a slightly smaller chance of being linked than males. We suggest this is the result of adolescent women changing their name when marrying.

3.4.5 Representativeness of the NTR subjects selected for linkage (NTR1) and the successfully linked dataset (NTR2), compared to the population and to each other





3.4.5 Representativeness of the NTR subjects selected for linkage (NTR1) and the successfully linked dataset (NTR2), compared to the population and to each other (end)



3.5 Discussion

The current project demonstrates that a dataset consisting entirely of twins can be successfully linked to an existing data source. The results show that detailed health data can be obtained for almost 8,000 subjects through record linkage, indicating that it is a serious alternative for sending out questionnaires.

Detailed health insurance records were retrieved for 7,813 subjects, which is 26 percent of the selected NTR population. However, the number of linked subjects who were insured with Achmea within a year was not higher than 21 percent (in 2013). Given that Achmea insures approximately 26 percent of the Dutch population, it is possible that the current linkage was not sufficiently sensitive to identify 100 percent of the true matches. The difference may however also be explained by the differential coverage of the Netherlands, which caused a small decrease in the representativeness of the linked population. While the selected NTR subjects were more often from the south and less often from the west, the AHD has relatively more customers in the east and west of the Netherlands and fewer in the southern provinces.

The variables that should be included into the linkage key depend strongly on the sizes and characteristics of both datasets. As the NTR dataset consisted of twin pairs, other linkage keys were needed than for linkage of records of singletons. When the surname and address variables were removed from the linkage key, the unique key rate dropped in the AHD, but not in the smaller NTR. In contrast, 99 percent of the AHD subjects could be uniquely identified by the combination of address, sex and date of birth, but since young twins usually live on the same address, this combination was not very powerful in the NTR.

To correctly distinguish twin siblings, it is necessary to assign extra value to the initials, given names and sex in the calculation of linkage scores. One issue we encountered however was that given names in both datasets could either refer to the official given names, or to the

name by which someone is commonly known, the *roepnaam*. Related to this issue was the finding that the initials were often incorrectly recorded in the NTR. Linkage results could be improved if both datasets used the same definitions for variables.

Linkage quality
The number of linked subjects was 43 percent higher when applying a probabilistic method with distance calculation than when linkage was based on a simple deterministic approach. Based on the linkage variables, 82 percent of the linked records were classified as *links* and 17 percent as *possible links*, and 1 percent as *false links*.

Our results show that linkage by a Trusted Third Party based on pseudonymised variables is a good alternative if linkage based on unencrypted identifiers is not possible. Whereas the number of links obtained by Mondriaan was only 75 percent of the Jaro-Winkler linkage, 97 percent of these links were evaluated as reliable *links*.

The retrieved content was validated by looking at information related to ADHD. Whereas a small number of inconsistencies were found between the NTR and AHD datasets, they were mostly explained by different periods of sampling rather than incorrect linkage. Our finding that the agreement between the NTR and AHD datasets was high supports the conclusion that the majority of subjects were linked correctly.

Enrichment of the NTR through record linkage
Information about medication use was missing in the NTR dataset for 4,686 or the 7,813 linked subjects (60%). The records of 4,686 NTR subjects were enriched through linkage with the AHD. The current record linkage has confirmed a number of known ADHD cases and identified 146 new subjects using methylphenidate and 101 receiving psychiatric care for ADHD, who were not previously recognized as ADHD-cases within the NTR.
While we focused on ADHD diagnoses and medication in order to interpret the quality and added value of the data retrieved from the AHD, the produced linkage tables can also be used to enrich the NTR with information about other disorders than ADHD.

Conclusion
Record linkage with sources of data such as a health insurance database can be an efficient way of data collection in cohort research, and our study shows that record linkage is also possible when collecting data of twin pairs. Although a minority of all NTR subjects were covered by the AHD, the linkage resulted in an enrichment of the NTR dataset. Exact information about the dosage and frequency of drugs was obtained without contacting subjects with detailed questionnaires. With a total of almost eight thousand retrieved records, the size of the linked dataset is sufficiently large for epidemiological research of non-rare conditions.

4. Linkage of community pharmacy records to the KOALA birth cohort study

Authors
D. (Dianne) de Korte (Maastricht University)
A. (Adelaide) Ariel (GGZ inGeest)
E.C.M. (Bep) Verkerk (GGZ inGeest)
G. (Gerard) van Grootheest (GGZ inGeest and VU University Medical Centre)
W. (Willem) de Bruijn (Utrecht University)
S.T. (Sipke) Visser (Mondriaan Foundation)
M. (Monique) Mommers (Maastricht University)
M.C.H. (Mark) de Groot (Utrecht University)
J.D.L. (Jan Dirk) Kroon (Dutch Foundation for Pharmaceutical Statistics)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)
C. (Carel) Thijs (Maastricht University)

4.1 Introduction

Common childhood diseases such as asthma are often studied in longitudinal birth cohort studies. Cohort studies are studies that follow a group of subjects over a long period of time and gather information about for example the development of children or changes in health status. Questionnaires are commonly used to collect information because these methods are much cheaper than the use of interviewers. Moreover, one can reach large groups in a relatively short period of time. However, the quality of the information obtained in this way may be incomplete or inaccurate in content and over time (Bethlehem, 2011). For example, parental reports of physician's diagnoses or medication use may be inaccurate or missing. A concern in longitudinal studies is that inevitably, participants will be lost to follow-up due to non-response, leading to attrition bias.

To improve the quality of the data, cohort studies may benefit from linkage to existing medical registries. This may be beneficial for two reasons. First, information obtained through such a linkage may complement data that was collected through questionnaires. Second, it may enrich the cohort database with new information on missing data or provide information on children who were otherwise lost to follow-up. Despite these benefits, linkage between cohorts and medical registries is not yet common practice in the Netherlands. It is currently largely unspecified how such a linkage must be established and how large the added value can be. In this chapter we demonstrate the record linkage between the KOALA Birth Cohort Study and a national pharmaceutical database.

The KOALA Birth Cohort Study (KOALA is a Dutch acronym for: Child, Parent and Health: Lifestyle and Genetic constitution) investigates the influence of the intrauterine and early childhood environment on child development. KOALA is an example of a birth cohort study

that has informed parental consent for obtaining medical data from pharmacies and general practitioners (GPs) for the majority of its participants. This permission makes KOALA very suitable for evaluating record linkage, because KOALA is allowed to evaluate in a non-anonymous manner whether medical registry data is correctly linked to KOALA participants. In this demonstration project, we use record linkage based on anonymised personal variables to obtain detailed information from pharmacies. The quality of the retrieved pharmacy records is evaluated by comparing them with existing information in the KOALA dataset.

Within this demonstration project, the focus is on medication for asthma and ADHD. Prevalence of diagnosis and medication use for both conditions is high among children. Childhood asthma is an outcome of interest in KOALA and many other birth cohort studies, and some birth cohort studies already harmonised their definitions to increase comparability and collaboration. Moreover, most prescribed medication for asthma and ADHD is very specific and therefore straightforward to evaluate.

The aim of this Biolink NL demonstration project was to link KOALA research data to a large pharmaceutical drug-dispensing database (the Dutch Foundation for Pharmaceutical Statistics *Stichting Farmaceutische Kengetallen, SFK*). Linkage was performed through a Trusted Third Party (TTP) linkage infrastructure. In this chapter we describe the datasets used, the linkage procedure, and the evaluation of the linked dataset. The linkage quality was validated by determining whether SFK patients were correctly linked to KOALA participants. Subsequently, it was determined whether retrieved SFK records on asthma and ADHD medication corresponded to KOALA research data. Thirdly, we evaluated the added value of this linkage in terms of enrichment and efficiency.

4.2 Description of the datasets

KOALA
KOALA has followed approximately 2,800 mothers and their children since the year 2000. Women were included during pregnancy and have repeatedly responded to questionnaires and on-site measurements. Its first scientific focus lies on allergies, asthma, and inflammatory and infectious diseases. The second focus is growth and development, including overweight and cardiovascular and metabolic risk factors for adult disease. Special attention is paid to the relation between children's development and lifestyle.

As of late 2014, children in KOALA were between 12 and 14 years of age. They have been followed with questionnaires since birth at ages 3, 7, 12 and 24 months, and then yearly between 5 and 10 years of age. When the children were 5–7 years old, parents of 1,754 KOALA children gave their permission to obtain medical information from the child's general practitioner and pharmacist.

SFK
The Dutch Foundation for Pharmaceutical Statistics (SFK) has been collecting drug dispense data for monitoring and analysing the use of drugs in the Netherlands since 1990. As of 2014, the SFK received data from 93 percent of a total of 1,974 community pharmacies in the country on voluntary basis. Pharmacies submit information about the dispensed drugs and materials to the SFK, together with the area code, birth year and sex of the patient. A number

of statistical procedures have been conducted to ensure the representativeness and the quality of SFK data (Griens et al. 2011).

Although the SFK has been collecting this large amount of pharmaceutical data, it has only been possible to study its data on the level of individual patients within a pharmacy. Because the pharmacies provide a very limited set of personal identifiers to the SFK, it is not possible for the SFK to link records from pharmacies to each other or to any other dataset without including content variables into the linkage key (Florentinus et. al 2006). Consequently, it is also impossible to tell whether an individual has visited more than one pharmacy.

Currently, the SFK is implementing the extraction of pharmacy patient data through a Trusted Third Party (TTP) using the Mondriaan client software as described in chapter 2. In this setup, personal identifiers are irreversibly hashed at the pharmacy, and sent to the TTP. The TTP has algorithms in place to look up subjects in a dictionary using the hashed identifiers. For subjects not yet present in the dictionary for a specific source, a so-called source pseudonym is generated. If a combination of linkage variables has been previously encountered before, the same existing pseudonym is assigned to that record. This source pseudonym is returned to the Mondriaan client software and exported to the SFK as a unique patient identifier. Based on the source pseudonyms it is possible for the TTP to link records from different pharmacies belonging to the same individual. This recognition of individuals between different pharmacies and sources avoids duplicate patient entries in the SFK, but also enables direct linkage to external datasets such as KOALA, provided that they also generate source pseudonyms by the same TTP. Dispense data are not sent from the pharmacy to the TTP, but are sent from the pharmacy to the SFK separately.

4.3 Methods

General approach

Whereas the information on drug dispenses originates from many different pharmacies, record linkage took place between the KOALA and SFK datasets. By delivering data to the SFK, pharmacies agree that their data may be used for scientific research, provided that the identity of pharmacies is not revealed. It was therefore sufficient to sign a contract with the SFK and not with each separate pharmacy.

As of 2014, most pharmacies provide data to the SFK through the existing system, which does not include a combination of personal variables that can uniquely identify individual subjects. However, as the SFK is currently implementing the new infrastructure through which the SFK receives pseudonymised personal identifiers, data from a limited number of pharmacies were available for linkage to the KOALA cohort. Whereas the proportion of pharmacies with this upgraded information system was still quite modest at the time of linkage (20%), the implementation was executed with priority for pharmacies that KOALA respondents claimed to visit. In this way, the chances of linkage success were increased. Throughout the entire process, any variables that could identify subjects or pharmacies were kept strictly separated from research and medication data (figure 4.3.1). A detailed description of the Mondriaan linkage method is given in chapter 2.

After linkage of the SFK to the KOALA birth cohort, a number of pharmacies were visited in order to validate the linkage results. Personal identifiers as registered in the two datasets were compared during these visits.

Privacy protection

Data was only retrieved from those children whose parents gave permission for retrieval of data from pharmacies and general practitioners. Any variables that could identify a person were kept within the environment of the pharmacies; any identifying information about pharmacies was kept by the SFK unless a pharmacy agreed to be visited by researchers from Biolink NL for validation purposes.

The SFK data manager attached the requested pharmaceutical data to the Link Pseudonyms and e-mailed this information to the corresponding KOALA researcher as a zipped and password-protected Excel spreadsheet. The password was sent separately in a text message to a KOALA-researcher's personal cell phone.

For the linkage validation, the KOALA researcher sent identifying information of linked subjects to one of the Biolink NL researchers. The files containing this information were first secured using AES-256 encryption and subsequently transferred using a Dutch academic file sharing service called SURFfilesender (http://www.surf.nl). Passwords required to download and to open the files were sent to the Biolink NL researcher's cell phone. Files with identifying information were only opened for linkage validation and were then deleted locally and from the server.

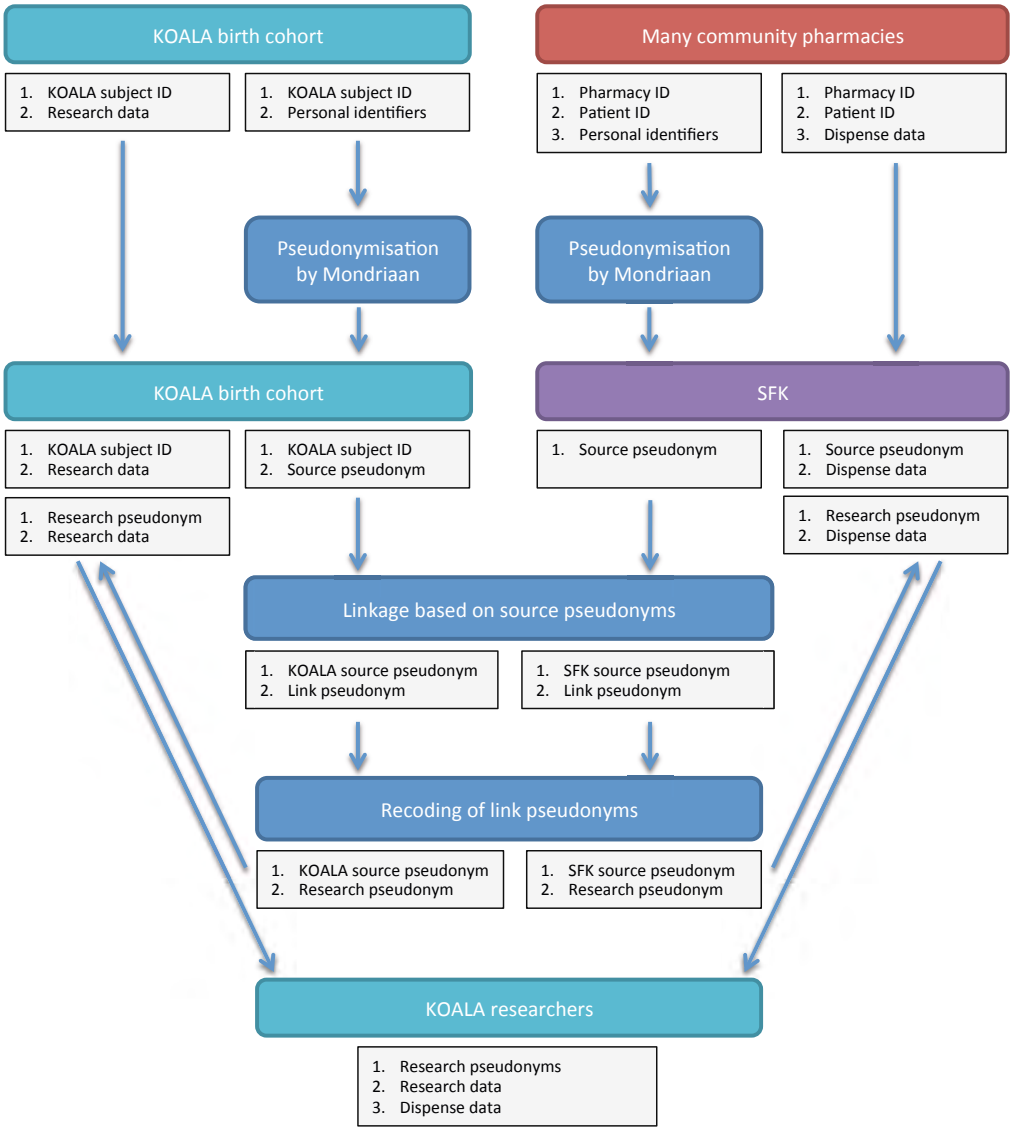
Preparation for record linkage

Step 1: Subject selection

The KOALA Birth Cohort Study originally consisted of 2,834 children. Candidates for record linkage were defined as children whose parents had given informed consent to obtain medical information from the child's general practitioner and pharmacist. This selection included 1,754 subjects.

As no medical or identifying information of any patient could be revealed at any moment, except if they provided explicit consent to KOALA to retrieve such data, no pre-selection was necessary on the side of the SFK. The factor that limited the potential number of links on the side of the SFK was whether a subject's pharmacy provided data through the new infrastructure or not. Based on the overlap with the pharmacies that parents mentioned when they were asked where they usually collected their prescriptions, we estimated that approximately 130 subjects should be linked.

4.3.1 Chart of the KOALA-SFK linkage. depicting the separation of identifying variables and research and pharmaceutical dispense data in the KOALA-SFK linkage



Step 2: Ensuring KOALA data quality by updating variables based on the BRP
Before the actual linkage, personal identifiers including the full address history of the selected KOALA subjects were verified and updated by querying the Municipal Personal Records Database (*Basisregistratie Personen, BRP*; formerly *Gemeentelijke Basisadministratie, GBA*). This BRP check was successfully completed for all 1,754 subjects; 13 children had however emigrated and were excluded from further linkage efforts. Any errors in name and address fields were corrected in the KOALA registration and a separate record was created for each historical address. After this step, the KOALA dataset contained 1,741 subjects.

We assumed that the reliability of linkage variables in the patient administration of the pharmacies was reasonably high. Already in 1992, Herings et al. showed that 98.6 percent of the patients who visited more than one pharmacy were recorded with the exact same personal details. We expect that nowadays especially the BSN, date of birth and sex are accurately registered by health care providers, because these variables are required for

reimbursement of medical costs from a patient's health insurance. As the current project focuses on relatively recent dispenses, little error was expected in address fields.

As mentioned in chapter 3, it is important to realise that given names as recorded in administrative databases (*'voornaam'*) may refer to the official given names, but can sometimes refer to the more informal name by which one is generally known, *'roepnaam'*. KOALA provided the complete official given names as registered in the BRP, rather than the *roepnaam* that was sometimes recorded as well. Surnames were not used in the linkage key, because this variable was recorded under different variable names in different pharmacies. As the TTP only received encrypted data, it could not be established remotely which variable contained the surname. Sex and date fields were delivered in uniform formats by both data sources. It was not necessary to further harmonize the datasets, because the Mondriaan client removes characters such as spaces, dashes, digits and diacritical marks from all text fields before pseudonymising and uploading the data.

Step 3: Retrieving medical data of KOALA participants from general practitioners
In 2014, short questionnaires were sent to the GPs of 1,598 KOALA participants for whom the required consent was given, in order to obtain additional information on the prescription of medication. Detailed information of 1,006 subjects (63%) was received, coded into categories by a research assistant and reviewed by one of the KOALA researchers, and added to the KOALA dataset. The KOALA dataset thus contained information about the use of medication as reported by the parents, as well as prescription information provided by the GPs.

Record linkage procedures

Only the 1,741 subjects for whom a complete address history was obtained from the BRP were offered for record linkage. As is described in chapter 2, the linkage algorithm used by Mondriaan is probabilistic, but has some deterministic features as well. The variables that were included into the linkage key are given in box 4.1. Because surnames were not sent from the Mondriaan client to the TTP, this variable could not be used in the linkage key.

The computational load of probabilistic linkage can sometimes be greatly reduced by blocking, i.e. restricting the comparisons only to record pairs that match on certain variables. The Mondriaan linkage procedure uses the date of birth and sex as the blocking variables.

Once record linkage was completed, Mondriaan sent the linkage results to KOALA and the SFK separately. Both parties received only a table with their respective Source Pseudonyms and newly created Link Pseudonyms (LP). The SFK collected all medication data (ATC-code, brand name and generic name, the date dispensed, dosages and prescribers) for each LP and sent these to KOALA.

Box 4.1 Linkage key that was used in the linkage of the SFK to KOALA

Linkage key for linkage by Mondriaan:

Deterministic step: Date of birth, sex.
Probabilistic step: Given name, full initials, postal code, street, house number, town.

4.4 Evaluation

Validation of linkage based on identifying variables
Since the SFK has no access to personal identifiers, the evaluation of this linkage could only be performed by looking up personal records at individual pharmacies. The SFK sent an e-mail to each pharmacy where any KOALA subjects were found, asking whether they were willing to cooperate in the validation procedure. This request was followed up by a telephone call one week later. Of the 53 pharmacies where links were identified, 17 agreed to facilitate the validation. Most pharmacies were visited by our researchers; a few remote pharmacists with only one or two linked patients helped us out over the telephone.

After a pharmacy agreed, a KOALA researcher secured the required personal identifiers and the scanned consent form with passwords and sent them to the Biolink NL researchers, who opened the files only at the pharmacy. Under supervision of the pharmacist, these researchers checked whether the personal identifiers of linked records matched between KOALA and the pharmacy database. This evaluation was based on agreement on four of the linkage variables and on the surname, which was not included in the linkage key. In total, 125 linked KOALA subjects were evaluated (47%).

In some cases, a KOALA subject was expected at the pharmacy but was not linked. These cases were looked up in the pharmacy database by entering the surname, date of birth and/or postal code. If a match was found and most other variables agreed, this indicated a false negative.

Biolink NL researchers who visited the pharmacies only looked at personal identifiers and had no access to medical information of KOALA subjects or of any other person registered in the pharmacies' databases.

Validation of content
The pharmaceutical information obtained after linkage was also validated for asthma and ADHD-related medication. The use of these drugs was assessed differently for each of the three data sources (box 4.2). Using the linked dataset, the validity of the SFK data was evaluated by comparing it with the combined data from parents and GP's, which we regarded as the reference: if both the parents and the GP's reported the respective medication, it was considered a *Yes*; if both reported no such medicine, it became a *No*. Thus, only the children whose parents and GP provided consistent information were used to validate the retrieved SFK records.

For both asthma and ADHD medication, we calculated 2x2 tables with sensitivity and specificity with 95 percent confidence intervals (CI). Sensitivity in this case indicates how many of the subjects who were categorised as *Yes* in the reference set also had a record of the drug of interest in the retrieved SFK dataset. Specificity indicates how many of the subjects who were categorised as *No* in the reference set indeed had no drug record of interest in the SFK dataset.

Box 4.2 Definitions of asthma and ADHD medication for different data sources

Asthma medication

KOALA questionnaires:

If 'yes' was answered to the question: 'Did your child use medication for wheezing or asthma that was prescribed by a doctor?' or if any type of inhaled asthma medication was reported after the question: 'If yes; what medication and how often was it used?'

General practitioner:

If any of the following medication was prescribed: inhaled beta-agonists (e.g. Salbutamol, Fenoterol, Terbutaline, Salmeterol, Formoterol, Indacaterol, *Ventolin, Serevent*), inhaled corticosteroids (e.g. Beclometason, Budesonide, Fluticason, *Pulmicort, Flixotide*), and combination inhalers (e.g. *Combivent, Symbicort*),

SFK:

If any of the following medications were dispensed: inhaled beta-agonists (ATC code R03AC), inhaled corticosteroids (ATC code R03BA), and combination inhalers (ATC code R03AK).

ADHD medication

KOALA questionnaires:

If 'yes' was answered to the following question: 'Does your child use one of the following medications: Ritalin, Rilatine, Concerta, Equasym, Medikinet, Strattera (Atomoxetine)?'

General practitioner:

If any of the following medication was prescribed: methylphenidate (e.g. *Ritalin, Concerta, Medikinet, Equasym*), dexamphetamine, or atomoxetine (e.g. *Strattera*).

SFK:

If one or more of the following medications were dispensed: methylphenidate (ATC code N06BA04), dexamphetamine (ATC code N06BA02), or atomoxetine (ATC code N06BA09).

4.5 Results

Linkage results

Record linkage between the SFK and KOALA resulted in 264 linked KOALA subjects (15% of the subjects who were selected for linkage). These subjects were linked to a total of 444 patient records, originating from 53 pharmacies. Drug dispense records were retrieved for 248 subjects; 16 subjects were linked but pharmaceutical data was missing. In addition, one subject had too many missing data in the KOALA dataset. These 17 subjects were included in the linkage validation but not in the content validation.

The TTP assigned linkage scores to each record pair; the required minimum score for the current linkage was set at 50. The linked records had a mean score of 106 points and no linked record pair had a score lower than 90 points. The linkage scores of record pairs that did not reach the threshold and were thereby not linked, were not saved and thus not reported.

Linkage validation

Of the 125 KOALA subjects included in the linkage validation, nine had multiple records within one pharmacy; three subjects were each identified in three different pharmacies. The records of 69 subjects (55%) showed an exact agreement on all five identifiers used for linkage validation, in any of the pharmacies (table 4.5.1). Issues with the other 56 links were mostly related to the initials (N = 53). These discrepancies were often caused by the difference between the official names and the name by which one is commonly known, the *roepnaam*. For example, someone named *Charlotte* may be registered as *Lotte* and thus be assigned a different initial. Furthermore, nine link pairs showed a small disagreement in the surnames, and for one case information on sex was missing in the pharmacy. Date of birth and postal code agreed for all linked subjects.

4.5.1 Agreement patterns of validated links

	Agreement pattern*	N
Subjects with one patient ID in one pharmacy	11111	60
	12111	27
	13111	3
	13191	1
	19111	13
	21111	3
	22111	5
	29111	1
Subjects with two different patient IDs in one pharmacy	11111, 11111	5
	12111, 12111	2
	13111, 19111	1
Subjects with three different patient IDs in one pharmacy	11111, 11111, 19111	1
Subjects with three different patient IDs in three different pharmacies	11111, 11111, 11111	1
	11111, 12111, 12111	1
	11111, 11111, 21111	1

As the only disagreements that were found were issues with initials and minor variations in surnames, we considered all links as correct. The 125 links that were validated and the 139 links that were not both had a mean linkage score of 106 points. Thus, we assume that the links that were validated are of the same quality as the ones that were not verified.

False negatives

Based on the answers that parents gave when asked where they usually pick up their prescribed drugs, we identified 18 KOALA subjects who should have a match in any of the pharmacies that were visited for the linkage validation of this project, but were not linked.

Three of these negatives appeared to be part of a twin pair. As the linkage algorithm could not distinguish twins, they were not linked. Ten potential false negatives were found with a manual search in the pharmacies' systems. These ten negatives were evaluated using the same criteria as the positive links (table 4.5.2). Four subjects showed agreement on all five identifiers and the other six showed only minor disagreement, confirming that they

were indeed false negatives. Some missed links were explained by the finding that some pharmacies coded an unknown sex as 'O'. Based on these results, the Mondriaan Foundation decided to adjust their procedure to handle this situation in future linkages.

The remaining five potential false negatives were not found in their reported pharmacy, but we did find records of their parents. Apparently, no medication was ever dispensed to these subjects in these pharmacies; they were considered as true negatives.

4.5.2 Agreement patterns of missed links (false negatives)

	Agreement pattern[1]	N
Subjects found by a manual search	11111	4
	11191	2
	12111	2
	19111	2

[1] The pattern indicates the agreement of variables between the pharmacy records and KOALA. The variables taken into consideration were: surname, initials, date of birth, sex, and postal code. 1 = exact agreement; 2 = minor disagreement (small deviation); 3 = disagreement; 9 = missing.

Content validation

A number of basic characteristics of the linked subjects and the subjects who were selected for linkage are given in table 4.5.3. Unsurprisingly, subjects with symptoms or a diagnosis of asthma or ADHD were more likely to be found in the SFK than subjects who did not have these conditions for which medication is commonly prescribed ($p<0.01$). Furthermore, the region of residence differed significantly between subjects who were successfully linked and those who were not ($p<0.01$): subjects living in the east of the country were underrepresented and those from the south were overrepresented after linkage.

4.5.3 Characteristics of KOALA subjects

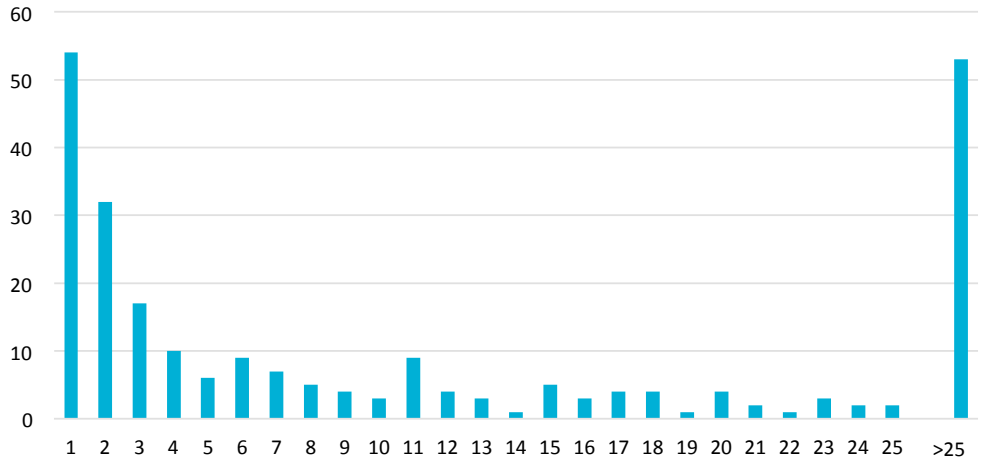
	KOALA subjects selected for linkage	KOALA subjects linked to the SFK
Total number	N = 1741	N=248
Sex		
Male	866 (49.7%)	133 (53.6%)
Female	875 (50.3%)	115 (46.4%)
Year of Birth		
2001	783 (45.0%)	111 (44.8%)
2002	654 (37.6%)	90 (36.3%)
2003	304 (17.5%)	47 (19.0%)
Region of residence[1] North		
East	17 (1.0%)	1 (0.4%)
West	260 (14.9%)	14 (5.6%)
South	141 (8.1%)	24 (9.7%)
	1,323 (76.0%)	209 (84.3%)
Asthma symptoms[2]	230 (13.2%)	51 (20.6%)
Asthma diagnosis[3]	187 (10.7%)	48 (19.4%)
ADHD diagnosis[4]	83 (4.8%)	17 (6.9%)

[1] Based on province in the year 2007. North: Friesland, Groningen, Drenthe. East: Overijssel, Gelderland, Flevoland. West: Noord-Holland, Zuid-Holland, Zeeland. South: Noord-Brabant, Limburg.
[2] Wheeze between the ages of 5 and 10 years, reported by the parent.
[3] Doctor's diagnosis of asthma between the ages of 5 and 10 years, reported by the parent.
[4] Doctor's diagnosis of ADHD between the ages of 9 and 10 years, reported by the parent.

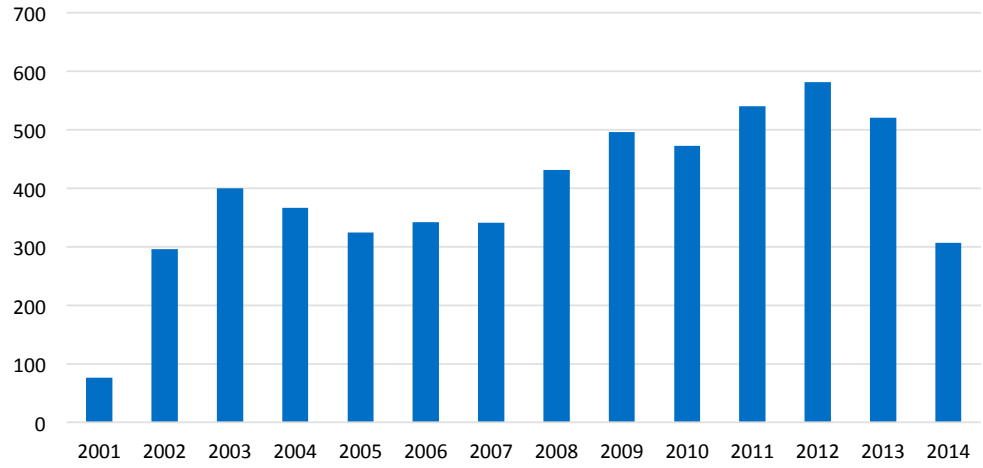
Drug dispense records from the SFK

In total, 5,495 drug dispense records were received for 248 subjects. Each record mentioned the type and brand of medication, ATC code, date of dispense, dosage, and the prescriber (such as a specialist, general practitioner, or psychologist). The number of dispenses per subject varied from 1 to 873, with a median of 6 (figure 4.5.4). The number of dispenses per year are shown in figure 4.5.5.

4.5.4 The number of retrieved drug dispenses per linked subject (n=248 subjects)



4.5.5 The number of retrieved drug dispenses per year (n=248 subjects)



Validity of SFK asthma and ADHD medication records

Table 4.5.6 shows the number of subjects for whom any SFK records indicated the use of asthma or ADHD medication, and whether this information agreed with the use reported in the reference set that was based on information from parents and GPs. Of the 26 subjects whose parent and GP both reported asthma medication, 18 had a record of asthma medication in the SFK dataset. For eight subjects, no asthma medication records were found, in spite of the fact that they were successfully linked. The sensitivity for SFK asthma medication was thus 69 percent (95% CI: 48–86%).

Six cases of asthma medication were found in the SFK dataset while no such use was reported in the reference set (specificity 89%, 95% CI: 78–96%). In all of these six cases, only one or two dispenses were found in the SFK dataset, each before the age of four years old. Although none of these children had any asthma symptoms or diagnosis between the ages 4 and 10 as recorded in the KOALA dataset, we found that five of them were reported by parents to wheeze before the age of two years old.

For ADHD medication, no disagreement was found between the KOALA and SFK datasets for those subjects whose parents and GP's gave consistent information. As all five cases of ADHD drugs in the reference were also indicated in the SFK dataset, sensitivity was 100% (95% CI: 48–100%). Since no records of ADHD drugs were found for subjects categorised as No in the reference, specificity was also 100 percent (95% CI: 93–100%).

4.5.6 Comparison of asthma and ADHD medication as registered in the KOALA and SFK datasets as used for validation

	Reported by parent and GP	Indication in SFK dataset	No indication in SFK dataset	Total
Asthma medication in KOALA reference set				
Yes	Yes for both	18	8	26
No	No for both	6	49	55
ADHD medication in KOALA reference set				
Yes	Yes for both	5	0	5
No	No for both	0	49	49

4.6 Discussion

While 130 links were expected, the actual number of linked subjects was higher. Apparently it was not unusual that KOALA subjects collected medication in another pharmacy than was reported on the consent form. We found that the specificity of the current linkage was very high (100%); no false links were identified. These good results may be partly caused by the enhanced chance of linkage through enriching the database with a full address history and complete official given names by querying the BRP prior to the actual linkage.

It was not possible to calculate the sensitivity of the linkage between KOALA and the SFK, as the number of true matches between both sources is not known. Moreover, the number of pharmacies that has implemented the Mondriaan pseudonymisation infrastructure was small, and is currently increasing. Linkage to the SFK resulted in the identification of pharmacy records for 15 percent of the selected KOALA subjects.

When compared with the best standard (parental report and general practitioner agreeing on medication), the sensitivity of the linked SFK information was 69 percent for asthma medication and 100 percent for ADHD medication for children with at least one drug dispensation record. Specificity was 89 percent for asthma medication and 100 percent for ADHD medication. If the number of linked records had been higher, it would have been possible to also calculate the sensitivity and specificity for specific types of medication and for timing of its use. In the current results however, the number of retrieved records was too small to reliably calculate such figures.

At the moment of linkage, SFK records from only 20 percent of all Dutch pharmacies were available for linkage within the Mondriaan infrastructure. Therefore, it was unsure whether subjects who were not linked were not registered in the SFK dataset at all (i.e. individuals with no history of medication dispenses), or that they have received drugs in a pharmacy that is not yet covered for linkage. The latter has probably also limited the sensitivity of linked SFK information in the content validation. With a higher (preferably a 100 percent) availability of pharmacies in the Mondriaan infrastructure, we expect the sensitivity of the content to improve.

It is important to realise that subjects who never picked up any medication from a community pharmacy may never be represented in the SFK dataset, even if all of these pharmacies have implemented the new infrastructure in the future. These subjects cannot be distinguished from those who received medication from the hospital pharmacy during a hospital stay, or from a pharmacy abroad. SFK data can thus not be used with 100% certainty to conclude that a subject did not use a medicinal product at all.

With respect to the specificity of the SFK records, even when we used the most stringent criterion (no parent report nor GP report of medication) we cannot exclude the possibility that false positives in the SFK are in fact truly dispensed medications. An indication for this is that five out of six false positive asthma medications reported by SFK concerned one or two prescriptions of inhalant medication before the age of 4 years (when an asthma diagnosis cannot be established) in children with parentally reported wheezing, which may have been prescribed by a specialist for short term use for a wheezing episode and not reported to the GP. An advantage of SFK in such cases is that the exact timing, dosage, dispensing history and prescriber are known from SFK, which may help to explain discrepancies with other data sources.

Furthermore it is important to realise that the three data sources that are used in this study (parental questionnaires, general practitioners, and pharmacies) in fact do not represent the exact same thing. Parents were asked whether their child ever used medication for asthma and/or ADHD. General practitioners were asked whether they ever prescribed medication for asthma and/or ADHD to their patient, and the SFK database provided information on whether medication for asthma and/or ADHD was ever dispensed.

Strengths and limitations
Strengths of this demonstration project include the opportunity to obtain high quality personal identifiers, the use of a TTP infrastructure, and the availability of parental informed consent, which enabled technical validation of the linkage within participating pharmacies. Limitations include the current incomplete implementation of the Mondriaan infrastructure in the pharmacies, which limits the number of cases with linked SFK data, and results in incomplete medication histories for cases that were successfully linked to SFK. Therefore a demonstration of the full potential of linkage to the SFK was not possible.

Enrichment
Retrieved data from SFK are well coded and structured and therefore easy to interpret. SFK medication records are very precise and include the type of medication and generic name according to the international ATC coding system, as well as dosage and date of dispense. Moreover, linking to SFK is less time consuming and less prone to data entry error as, for instance, data that are obtained from GPs through questionnaires and manually entered into a database.

In order for the SFK to become useful for enrichment, it is important that the number of pharmacies that deliver pseudonymised records reaches a figure as close to 100 percent as possible (including hospital pharmacies). In that case, linkage to SFK can enrich cohort data, first, to complement missing data from questionnaires, from non-response to questionnaires, or due to loss-to-follow-up. Second, linkage to SFK may even substitute follow-up by questionnaire for these specific data. For collaborating cohorts it can offer a harmonised way of collecting medication data.

Medication data can sometimes be used to complement or confirm medical diagnoses. We chose asthma and ADHD as examples for this demonstration study because medication for these conditions is quite specific. For many other conditions there is no specific medication to substitute diagnostic information, for instance: antibiotics are used for many different conditions although ATC codes sometimes include indications (e.g. acne) or site of treatment (e.g. oral, dermatological, eye or ear). In addition, medication data may be interesting not as an outcome but as an exposure, and for that purpose, dose and duration in SFK is an invaluable source of exposure data. Retrospective reporting of medication use by subjects in a questionnaire may be flawed if the subject already knows the outcome, e.g. the course of the disease after the medication was initiated or discontinued. Historical data that are recorded before the disease outcome has occurred, as in the SFK database, is therefore more valid.

Conclusion

For cohorts or biobanks that have sufficient personal identifiers of a high quality, linkage to a pharmaceutical database such as the SFK is a feasible method to obtain detailed patient information. As a growing number of pharmacies is providing pseudonymised dispense data to the SFK, cohorts/biobanks can benefit from linkage with the SFK to enrich their data. This method of data collection adds more detail to existing records without the need to contact subjects or their parents. Standardised variables such as drug type, brand, dosage, and dispense date can be obtained through record linkage and are useful for enrichment of single cohorts and biobanks, and for harmonisation between cohorts and biobanks on medication as exposure or outcome.

5. Linking the Dutch Population Register and the Employment Register

Authors
D.J. (Jan) van der Laan (Statistics Netherlands)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)
M.C.H. (Mark) de Groot (Utrecht University)
G. (Gerard) van Grootheest (GGZ inGeest and VU University Medical Centre)

5.1 Introduction

All European countries are required to hold a census once every ten years. In the Netherlands the census is carried out by Statistics Netherlands. The Dutch censuses of 2001 and 2011 were based on administrative sources (Statistics Netherlands, 2014) rather than on a traditional door-to-door survey. Using administrative sources has numerous advantages (Wallgren and Wallgren, 2014) the most important being cost efficiency. The Population Register (in Dutch: *Gemeentelijke Basisadministratie, GBA*) plays a central role, as this register ought to contain information on all regular inhabitants of the Netherlands. However, some inhabitants are not registered even though they should be. Such undercoverage leads to underestimation of the population size.

Capture-recapture methods can be used to estimate population sizes, and therefore the undercoverage of registers (Fienberg, 1972 and IWGDMF, 1995). This is usually done by randomly sampling elements from the population. A second sample is captured after a period that is long enough for the second sample to be independent of the first, but short enough for the population to remain constant. The population size can be estimated from the number of elements that were present in both samples. When the overlap is large, the population size is close to the sample size. When the overlap is small, the population is much larger than the sample size. A similar approach is used for the census. However, various registers are used instead of random samples (Gerritse et al. 2015, Van der Heijden et al. 2012). A prerequisite of this method is that the linkage between the registers is very accurate as false links lead to an underestimation of the population size, and missing links to an overestimation. In this project the Population Register (PR) was linked to the Employment register (ER) containing information on persons receiving income from an employer.

Unlike research cohorts, Statistics Netherlands has access to the BSN – a social security number serving as the national identification number used by all government agencies – and can use this identifier to link registers deterministically. However, as the linkage between the registers needs to be perfect for the capture-recapture method to work, and as there may be quality issues with the BSN of foreign employees, probabilistic linkage methods were used to link the remaining records that could not be linked deterministically.

We faced two main challenges in linking the two resources. First, the datasets are large: the PR contains information on approximately 17 million persons and the ER contains 13 million

records on 7 million persons. Size could be a problem for probabilistic record linkage where all possible pairs are investigated in principle.

Second, there will be little overlap between the sources because only those records from the ER that do not link to the PR on the BSN will be linked to the (complete) PR.

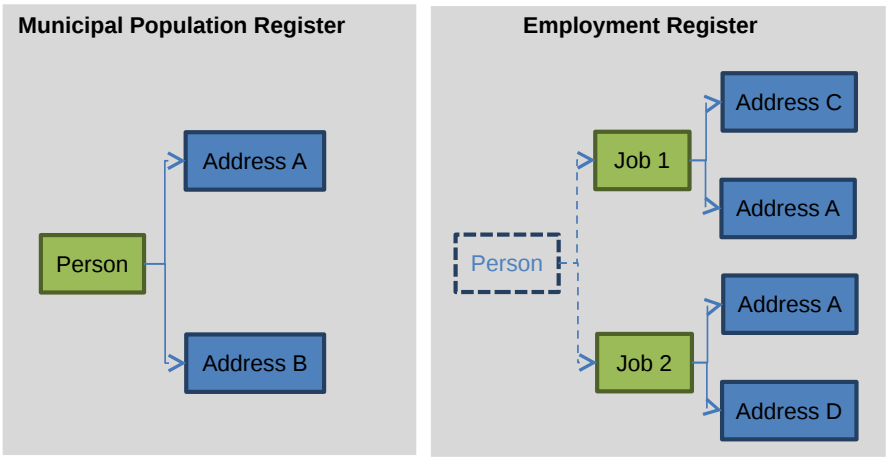
As was discussed by Ariel et al. (2014) this may lead to problems in estimating the m and u-probabilities required for probabilistic record linkage. This leads to the following research questions for this chapter: Can probabilistic linkage be used to improve the linkage of administrative sources? Can the algorithms cope with such large datasets? How is the probabilistic method affected by selecting only data that could not be linked deterministically?

5.2 Description of data sources

The PR contains personal details such as address information, marital status, nationality and sex of all registered current end former Dutch residents. The file that was available for linkage concerned the entire year 2010. When someone's information changes, an additional record is produced. Therefore someone can have multiple records and each record is valid for a certain time (figure 5.2.1). Since changes in marital status and nationality are irrelevant for the record linkage, these have been filtered out. Furthermore, since we were linking against employed persons, persons born before 1933 or after 1997 were removed from the file, resulting in a dataset of 14.3 million records.

The ER contains information on income from jobs. This register is maintained by the Employee Insurance Agency (*Uitvoeringsinstituut Werknemersverzekeringen; UWV*) and is used to check someone's rights to unemployment benefits. Statistics Netherlands uses the register for some of its income and labour statistics. Someone can have several jobs at the same time and within the same period. As is the case in the PR, changes of address information can lead to multiple records per job (figure 5.1). This resulted in a total of 12.9 million records.

5.2.1 Structure of the two data sources. The primary units in each of the files are marked green



5.3 Methods

General approach

One assumptions of the capture-recapture model is that persons included in both registers are identified as such: false positive links lead to an underestimation of the population size, and false negatives lead to an overestimation. Therefore, the linkage needs to be of a high quality and a two-step approach has to be used for the linkage. First the records are linked deterministically using the BSN. For foreigners working in the Netherlands there may be quality issues with the BSN (missing or incorrect values). Incorrect values can be caused by typing errors (these can be detected using the check digit), using a different number than registered in the PR, or using someone else's number. Therefore, the records from the ER that could not be linked in the deterministic step were linked using probabilistic methods to the full PR.

Subject selection

The PR contains information on all individuals registered at some time during 2010. Since we were linking against employed persons, only individuals born between 1933 and 1997 were included in the dataset used in the linkage.

We selected all jobs that were (partially) within 2010 from the ER. And for each job we selected all corresponding addresses with a starting date before 1 June 2011 and an end date after 1 June 2009. The larger period for addresses was used to allow for delays in registering addresses. These selections led to 14.3 million records for the PR and 12.9 million records for the ER.

Linkage procedures

The deterministic linkage was performed using the BSN. The variables used in the probabilistic linkage included the date of birth and address variables 1, 2 (box 5.1). Since Statistics Netherlands does not use names, these are not available for linkage. The steps below give an overview of the complete procedure followed to link the two data sets together.

Box 5.1 Linkage keys used to link the employment register to the population register

Deterministic step: BSN.
Probabilistic step: Day of birth, month of birth, year of birth, sex, postal code (4 digits), postal code (2 letters), house number and suffix.

Step 1: Deterministic linkage based on BSN

The PR was linked deterministically to the employees register based on BSN. Records from the ER that did not link were linked probabilistically to the complete PR in the following steps. All records from the ER that were not linked by BSN were selected for probabilistic linkage. As people can have more than one job, the number of records in the PR was not reduced for the probabilistic step.

Step 2: Data standardisation

Suffixes of house numbers can be written in various ways in the Netherlands. For example, '1', '\1', '–1', '–01', 'I' are actually all the same suffix. Therefore, it was decided to use a similarity

score for the suffixes. This score was 1 when the field matched exactly, and 0 when the fields were completely different. The similarity score was incorporated into the weight calculation of the pair as described in our previous report (Ariel et al., 2014). First, all text was converted to capital case, symbols such as '–', '\' and '.' were removed, as well as white space at the beginning and end of a field. Fields that matched exactly after these procedures received a similarity score of 1. Variations were subsequently corrected according to a number of rules. Depending on the rule these received a similarity score between 0.90 and 0.95 (based on intuition). Remaining scores received a weight as calculated by the Jaro-Winkler algorithm (Winkler, 1990).
When calculating the m and u-probabilities, suffix fields with a similarity score of 0.8 and higher were considered a match. In practice, values of 0.8 and higher indicate very similar suffixes using the Jaro-Winkler similarity metric.

Step 3: Estimation of m and u-probabilities
The m and u-probabilities are usually estimated using the EM algorithm. However, the EM algorithm starts performing worse when the number of true matches in the generated pairs becomes smaller than approximately 4 percent. In this step we only want to link those records from the ER that were not linked deterministically, so the number of actual matches is a factor 100 below this 4 percent. Alternatively, the m-probabilities can be estimated using a reference dataset (Herzog, Scheuren and Winkler, 2007). This was available in the form of the deterministically linked dataset. It was relatively easy to determine the probability of errors in the linkage variables (the m-probabilities), given that two records match on BSN. As this is a huge reference set, these probabilities can be estimated accurately and stably.

It is possible to estimate separate m-probabilities, depending on the values of a variable. This was done for the suffix of the house number. Few errors are made in most addresses without a suffix. However, we see many variations in the addresses with a suffix such as 'I' instead of '1', or 'A01' instead of 'A-1'. Therefore, we estimated separate m-probabilities for addresses with and without suffixes. When estimating the m-probabilities the corrections as discussed in step 2 were applied.

The u-probabilities were estimated using the generated pairs. The u-probabilities give the matching probability of variables from the two data sources, while the record pair is actually not a match. As the number of true matches in the pairs is very small (less than 1%), we can consider all pairs as non-matches.

The u-probabilities can likewise be calculated by computing the fraction of pairs for which the variables match between the two data sets. Again because the dataset is large, it is possible to estimate different u-probabilities depending on the value of the variable. For example, for a common house number such as '1' the probability that two records will agree is much higher than for an uncommon one such as '5,432'. Therefore, a higher weight is assigned to a pair that agrees on an uncommon value, because it is more likely to be a match. This was done for house numbers and their suffixes. They were divided into 12 groups for which 12 different u-probabilities were calculated.

Step 4. Probabilistic linkage
Approximately 196 thousand WNB records that were not linked in step 1 were selected for probabilistic linkage with the PR. As it is necessary to generate all possible linking pairs between both dataset for probabilistic linkage, a total of 14.3 million times 196 thousand = $2.8 \cdot 10^{12}$ pairs should be created. Because this number is too large for practical processing,

we blocked on postal code and date of birth separately. The advantage of blocking on two separate blocking keys is that a record with an error in say a postal code can still be found when blocking on date of birth and vice versa. Blocking on postal code resulted in 8.8 million pairs. Blocking on date of birth resulted in 138 million pairs.

Using the m and u-probabilities generated in the previous step, weights were calculated for each pair, then a threshold was determined for the weight. Record pairs above the threshold were classified as link; record pairs with a weight below the threshold were not linked. The threshold was determined 'by eye' (Herzog, Scheuren and Winkler, 2007): twenty records just above and below the threshold were reviewed for different values of the threshold. We determined the plausibility for each record that it corresponded to the same person. The threshold was moved until a value was found above which most records could be considered true matches. Because both the PR and the ER can contain multiple records per person and because the same record could match with multiple records, an optimisation was performed forcing each record from the ER to be linked to no more than one person in the PR. Pairs were selected in such a way that the total weight of selected pairs was optimised (Christen, 2012).

The steps in the previous paragraph were applied for each of the two blocking variables. This generated two sets of linked pairs. These two sets were combined and a deduplication was performed as just described. This resulted in one set of linked records.

5.4 Results

In the next sections we first show the results of the linkage procedure, i.e. the number of links made. Then we will discuss two specific aspects of the linkage procedure, namely the determination of the m- and u-probabilities and the determination of the weight threshold. This is followed by an evaluation of the linkage quality.

Overview of linkage results

Table 5.4.1 gives an overview of the results of each of the steps in the linkage procedure. Most of the records from the ER (95.7%) could be deterministically linked to the PR using the BSN. Of the remaining records, 361 thousand contained foreign addresses. A large part of these records will contain information on border workers living in Belgium or Germany and do not belong to the population. Moreover, the only additional information available for linkage were date of birth and sex, and since these are not identifying enough, records with a foreign address could not be linked to the PR. These records were therefore excluded from the probabilistic linkage. In the end 196 thousand records were linked to the 14.3 million records from the PR.

After determining the weight threshold (see below) and removing duplicate pairs, blocking on postal code resulted in 3,795 pairs and blocking on date of birth in 3,341 pairs. When these are combined and duplicate pairs are removed, we end up with 3,813 linked records. Therefore blocking on date of birth besides blocking on postal code resulted in only 18 additional linked records. Of the 196 thousand pairs that could be linked, eventually only 1.9 percent was linked.

5.4.1 Records in each of the steps of the linkage procedure including the final number of pairs selected

Data sources and deterministic linkage	
Population register	
Number of records	14,336,108
Employment register	
Number of records	12,859,325
Deterministically linked	12,301,990
Foreign address	361,356
To probabilistic linkage	195,797

Probabilistic linkage	
Blocking on date of birth	
Number of pairs	137,829,400
Selected pairs	3,438
After deduplication	3,341
Blocking on postal code	
Number of pairs	8,835,502
Selected pairs	3,956
After deduplication	3,795
Final number of pairs	3,813

Calculation of m- and u-probabilities

Table 5.4.2 shows the estimated m- and u-probabilities for the probabilistic linkage for each of the two blocking variables: postal code and date of birth. Since the m-probabilities are estimated from the deterministically linked data sets they are the same in both tables. The high m-probabilities for date of birth and sex indicate a high quality of these variables in both data sets. The m-probabilities for address variables are much smaller. This is partly due to the fact that both registers can contain multiple records for everyone who had a change of address. For example, if someone changed from address A to B during the year and this is registered correctly in both registers, then the probability of the address matching for this person is still only 50 percent. What can be noted, however, is that the m-probability for the suffix is lower than for the other address variables, indicating a higher error rate in this variable.

The u-probabilities decrease as the house number becomes higher. Lower numbers are much more common than higher numbers with number 1 being very prevalent. This probably indicates that the number is also used when the true number is unknown. The high u-probability for addresses without a suffix indicates that addresses with a suffix are very common. Any finding that two records are both lacking a suffix is therefore not a strong indication that these two records belong to the same person (hence the small weight of 0.18). However, if two records both have one of the more rare suffixes (such as E), this is a strong indication that both belong to the same person (hence the high weight of 5.53).

5.4.2 The m- and u-probabilities and the corresponding weights in variable or non-matches of the variables used in the probabilistic linkage

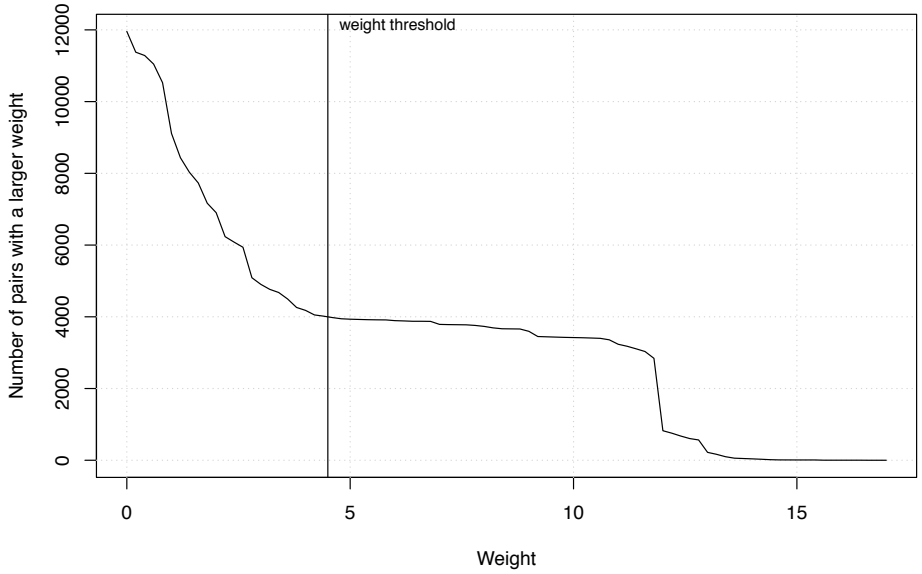
			Blocking on postal code			Blocking on date of birth		
Variable	**Value**	**m-prob.**	u-prob.	w_{match}	$w_{no\text{-}match}$	u-prob.	w_{match}	$w_{no\text{-}match}$
Day of birth	1	0.997	0.041	3.19	−5.86	–[1]	–[1]	–[1]
	Other	0.997	0.033	3.42	−5.87	–[1]	–[1]	–[1]
Year of birth	–	0.998	0.024	3.72	−6.09	–[1]	–[1]	–[1]
Month of birth	–	0.997	0.084	2.48	−5.71	–[1]	–[1]	–[1]
Sex	–	0.994	0.512	0.66	−4.36	0.502	0.68	−4.38
Postal code first 4 digits	–	0.814	–[1]	–[1]	–[1]	0.001	7.28	−1.68
Postal code last 2 letters	–	0.775	–[1]	–[1]	–[1]	0.003	5.54	−1.49
House number	1	0.776	0.467	0.51	−0.87	0.502	0.68	−4.38
	2	0.776	0.273	1.05	−1.18	0.043	2.89	−1.45
	4	0.776	0.307	0.93	−1.13	0.029	3.30	−1.47
	3	0.776	0.132	1.77	−1.36	0.024	3.48	−1.47
	5	0.776	0.122	1.85	−1.37	0.021	3.62	−1.48
	6	0.776	0.173	1.50	−1.31	0.018	3.78	−1.48
	7	0.776	0.210	1.31	−1.26	0.026	3.40	−1.47
	8	0.776	0.198	1.37	−1.28	0.022	3.57	−1.48
	9	0.776	0.443	0.56	−0.91	0.019	3.69	−1.48
	10	0.776	0.174	1.49	−1.31	0.028	3.31	−1.47
	11	0.776	0.164	1.55	−1.32	0.018	3.78	−1.48
	Other	0.776	0.062	2.53	−1.43	0.007	4.66	−1.49
House number suffix	*No suffix*	0.963	0.893	0.08	−1.06	0.802	0.18	−1.68
	A	0.604	0.218	1.02	−0.68	0.067	2.20	−0.86
	B	0.604	0.284	0.75	−0.59	0.043	2.64	−0.88
	C	0.604	0.145	1.43	−0.77	0.019	3.46	−0.91
	1	0.604	0.146	1.42	−0.77	0.011	4.02	−0.91
	2	0.604	0.137	1.48	−0.78	0.010	4.09	−0.92
	H	0.604	0.110	1.70	−0.81	0.004	5.01	−0.92
	3	0.604	0.141	1.45	−0.77	0.007	4.50	−0.92
	D	0.604	0.089	1.91	−0.83	0.006	4.64	−0.92
	E	0.604	0.055	2.39	−0.87	0.002	5.53	−0.92
	'BS', ..., '58'	0.604	0.106	1.74	−0.81	0.024	3.22	−0.90
	Other[2]	0.604	0.016	3.64	−0.91	0.019	3.44	−0.91

[1] Variables used as blocking variables are not used as linkage variables and no weights and u-probabilities are calculated for these variables.
[2] Because use was made of a similarity score the weight used is between w_{match} and $w_{no\text{-}match}$.

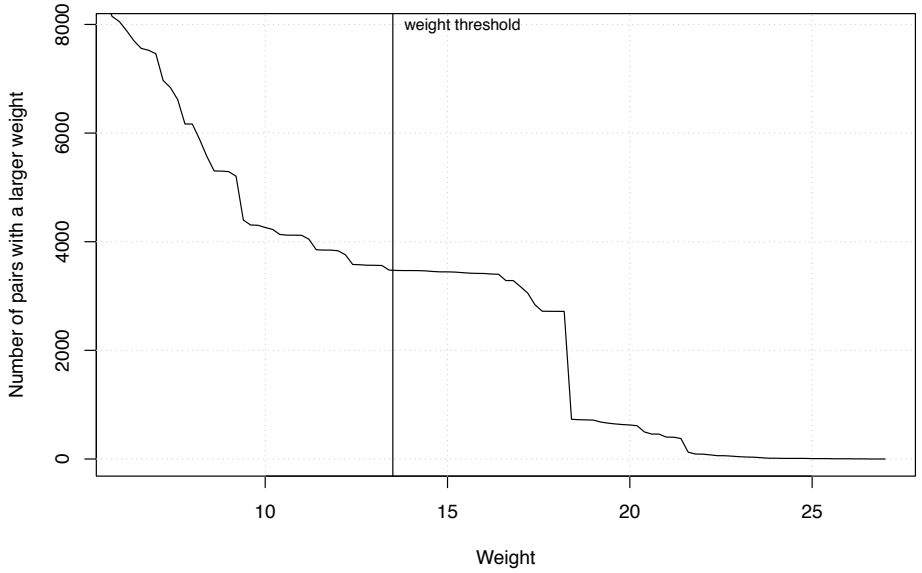
Selection of the weight threshold

Figures 5.4.3 and 5.4.4 show the number of pairs designated as links as a function of the threshold. As the threshold increases the number of links decreases. Both figures show the same pattern. At very high weights (above 12 and 16 respectively for the two blocking variables) both have a region with a relative large number of pairs. Here pairs match on all of the important variables. The region below that hardly contains any pairs. For blocking on postal code this region runs from 5 to 12 and for the blocking on date of birth from 10 to 16. Below this the number of pairs increases exponentially and most will not be true matches. By manual inspection a threshold of 4.5 was eventually determined for blocking on postal code and a threshold of 13.5 was set for blocking on date of birth. With a weight above this threshold pairs were likely to be true matches, whereas pairs with a weight below were not. In both cases the threshold is chosen approximately in the flat region in the figures. Since the number of pairs in this region is small, the exact position of the threshold in this region has little influence.

5.4.3 Pairs designated as links as a function of the threshold when blocking on postal code. The selected weight threshold is indicated using a vertical line.



5.4.4 Pairs designated as links as a function of the threshold when blocking on date of birth. The selected weight threshold is indicated using a vertical line.



In order to gain more insight in the selected pairs, we show the distribution of weights for each pattern in table 5.4.5. The patterns are ordered by decreasing median weight (the 50th percentile column in the table). A pattern is a certain combination of matching and non-matching variables. For example, all variables match in the first pattern (111111). This pattern had obviously received the largest weight. In the sixth pattern (111010) the address suffix and sex of the two records in the pair did not match. Using the maximum and minimum weights for each pattern, it is possible to determine from which patterns pairs are selected as matches. With a threshold of 4.5, all pairs in patterns 1–6 are selected. These are pairs with non-matching sex and/or house numbers and/or suffixes A few pairs from patterns 7 and 8 were selected because the house number suffix was rare and received a high weight.

5.4.5 The distribution of weights for each observed pattern for postal code as a blocking variable. The table also shows the number of pairs for each pattern and the uniqueness of records in de PR for the given set of keys

	Pattern										Percentiles				
No.	yr	mon	day	gndr	num	sffx	N. pairs	unique key	min	max	10	25	50	75	90
1	1	1	1	1	1	1	3,248	100.00%	10.6	16.5	11.7	11.9	11.9	11.9	12.9
2	1	1	1	1	1	0	175	99.89%	9.7	13.6	10.5	10.8	11.0	11.2	12.2
3	1	1	1	1	0	1	232	99.82%	8.7	10.6	8.9	8.9	9.0	9.0	9.2
4	1	1	1	1	0	0	107	99.68%	7.8	9.6	7.8	7.9	8.1	8.3	8.4
5	1	1	1	0	1	1	120	99.91%	5.8	9.2	6.1	6.9	6.9	6.9	7.6
6	1	1	1	0	1	0	36	99.79%	4.9	7.6	5.2	5.7	6.0	6.3	6.9
7	1	1	1	0	0	1	172	99.66%	3.7	4.7	3.9	3.9	4.0	4.0	4.1
8	1	0	1	1	1	1	355	99.88%	2.4	7.3	2.7	3.5	3.7	3.7	4.7
9	1	1	1	0	0	0	52	99.50%	2.8	4.5	2.9	2.9	3.1	3.3	3.3
10	1	0	1	1	1	0	359	99.72%	1.5	5.3	2.1	2.6	2.7	3.4	3.9
11	1	1	0	1	1	1	982	99.74%	1.6	6.2	1.6	2.6	2.6	2.6	3.6
12	0	1	1	1	1	1	1,337	99.71%	0.8	5.4	1.1	1.4	2.1	2.1	3.0
13	1	1	0	1	1	0	918	99.51%	0.4	4.2	1.0	1.5	1.6	2.0	2.8
14	0	1	1	1	1	0	1,096	99.49%	-0.3	3.7	0.8	1.1	1.1	1.4	2.3
15	1	0	1	1	0	1	1,792	98.99%	0.5	4.4	0.7	0.8	0.9	0.9	1.0
16	1	0	1	1	0	0	748	98.86%	-0.6	1.7	-0.3	-0.3	0.0	0.1	0.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
63	0	0	0	0	0	1	2,335,333	5.17%	-23.4	-19.7	-23.4	-23.3	-23.3	-23.3	-23.2
64	0	0	0	0	0	0	809,416	2.99%	-24.5	-22.0	-24.4	-24.4	-24.2	-24.0	-23.9

Table 5.4.6 shows the same information for the probabilistic linkage where date of birth is used as a blocking variable. With the weight of 13.5 that was selected, only errors in either sex or suffix were allowed. Some records with errors in the house number were selected if they matched on a rare suffix.

5.4.6 The distribution of weights for each observed pattern for date of birth as a blocking variable. The table also shows the number of pairs for each pattern and the uniqueness of records in de PR for the given set of keys

	Pattern									Percentiles				
No.	postc4	postc2	gndr	num	sffx	N.pairs	unique key	min	max	10	25	50	75	90
1	1	1	1	1	1	3,248	100.00%	16.6	24.0	17.4	18.3	18.3	18.3	21.6
2	1	1	1	1	0	175	99.89%	15.5	22.2	16.0	16.4	17.2	17.7	20.3
3	1	1	0	1	1	120	99.91%	12.2	17.2	12.2	13.3	13.3	15.7	16.6
4	1	1	1	0	1	232	99.82%	12.2	15.5	12.2	12.2	12.2	12.2	13.4
5	1	1	0	1	0	36	99.79%	10.4	13.2	10.8	11.3	12.2	12.7	13.0
6	1	0	1	1	1	348	99.90%	9.5	13.2	10.0	10.2	11.3	11.3	11.3
7	1	1	1	0	0	107	99.68%	10.3	11.2	10.3	10.3	11.1	11.1	11.1
8	1	0	1	1	0	125	99.76%	8.9	10.2	8.9	8.9	9.8	10.2	10.2
9	0	1	1	1	1	1,601	99.51%	7.6	12.6	7.6	8.1	9.4	9.4	9.4
10	0	1	1	1	0	686	99.29%	5.7	8.3	6.3	6.6	7.5	7.5	8.3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
32	0	0	0	0	0	20,072,203	0.16%	−10.7	−8.1	−10.7	−10.7	−10.7	−9.9	−9.9

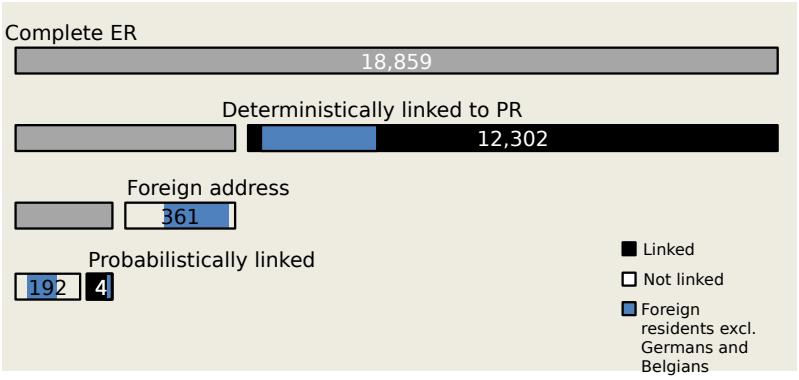
Linkage quality

The BSN does not change over time and is usually obtained directly from the identification cards of the employees. It is in the interest of most employers and employees that this number is correctly registered (furthermore, the BSN contains a check digit which allows for the detection of small typing errors). Therefore, the BSN will be correctly registered for most records and few problems with linkage quality are expected in the part of the register that could be linked deterministically. This was also visible in the high m-probabilities for sex and date of birth (see table 5.4.2). The quality for the probabilistic linkage is difficult to evaluate. There are no content variables that can be used to evaluate the likelihood of a pair being a true or false match; this can only be done with linkage variables. However, as there were very few linkage variables (date of birth, sex, and address) and very large datasets, one or two errors in the linkage variables will cause the remaining linkage variables to no longer be unique. Tables 5.4.5 and 5.4.6 also show the percentage of unique records in the PR for each of the sets of keys. For example, when the number or suffix is ignored there are 99.68 percent unique records and therefore the chance of accidentally linking two unrelated records is approximately 0.32 percent. It is therefore difficult to judge on the linkage variable alone whether or not two records belong to the same person. In this linkage there were no content variables available for evaluation so no content valuation was performed. However, as tables 5.4.5 and 5.4.6 show, with the selected thresholds for the weights the fraction of false links will be small, e.g. in the group of 1,601 pairs with an error in the 4 digits of the postal code, there will be at most 8 (0.49% of 1,601) accidentally linked pairs.

Representativeness

It is important for analyses on the linked data set that the linked data are representative of the target population. The fraction of successfully linked records gives only a partial indication of representativeness. The coefficient of variation, ℓ, as a measure for the representativeness was introduced in chapter 2. The target population for the calculation of this indicator was defined as the employed foreign residents in the Netherlands, except residents with a Belgian or German address. They were excluded because the majority of them are crossborder workers and are not of interest for the research question. As figure 5.4.7 shows, these employees are found in the deterministically linked, the probabilistically linked, and in the non-linked parts of the ER.

5.4.7 The different parts of the ER (sizes are not to scale, numbers x 1,000)

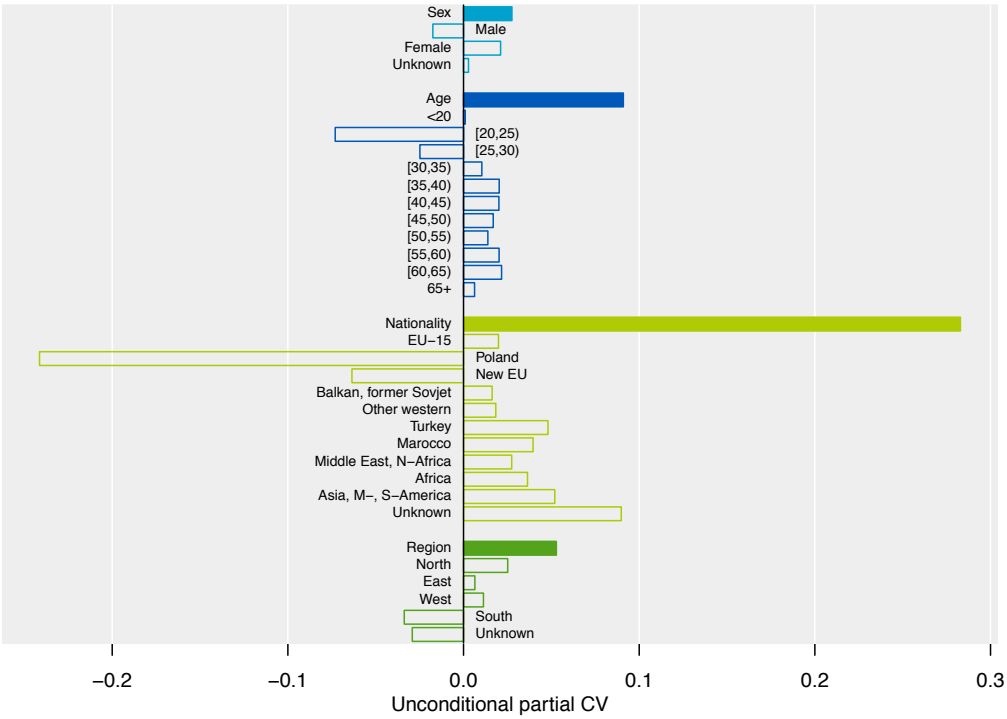


The registers are linked with a combination of deterministic and probabilistic methods. In the first step, the records are linked deterministically on a BSN that is widely available in administrative data sources in the Netherlands. The remaining records are linked

probabilistically as described above. In total 84.6 percent of the records of the ER of the target population could be linked. The probabilistic linkage has led to an increase of 0.3 percent of the total number of linked records in the target population.

The covariates sex, age, nationality and region were used in the calculation of the linkage propensities. Using these propensities, we arrived at a coefficient of variation of 0.29. So a relative bias of almost 30 percent is possible in the estimate of the mean of a variable that is strongly correlated with the linkage probability when the estimate does not correct for the selective linkage probabilities. The probabilistic linkage decreased ℓ from 0.295 to 0.294; an almost negligible decrease.

5.4.8 Partial coefficients of variation showing the contributions of subgroups to the overall representativeness



It is also possible to calculate partial coefficients of variation, ℓ_u, which express the contributions of the various subgroups to the overall ℓ; these are shown in figure 5.4.8. Nationality makes the largest contribution to the lack of representativeness of the linked dataset. People from Poland and the other countries that joined the EU after 2004 are underrepresented. Any analysis of this data set should therefore take nationality into account. Furthermore, men, people aged 20–30 and people living in the south of the Netherlands are underrepresented. However, these categories will correlate with the underrepresented nationality groups.

Specification of work and resources

The linkage had to be performed within the secure environment at Statistics Netherlands. Within this environment it is almost impossible to install new software or to introduce new hardware. Therefore the linkage had to be performed using R on machines with very limited memory. Much effort has gone into adapting the existing R-code for record linkage for

supporting large data sets that do not fit into memory, and speeding up the record linkage process. Eventually, the code was completely rewritten.

Because the data sets were so very large, another bottleneck was the analysis and preparation of the linkage files. For example, the ER consisted not of one file but of four separate files: one containing the jobs, one containing the BSN of job holders, one containing their address information, and one containing their personal information such as date of birth and sex. These had to be combined into one file. Furthermore, all variables had to be investigated and the most obvious coding errors had to be corrected. For example, missing postal codes were sometimes coded using non-existent postal codes such as '0000XX'. There were many errors in the field coding for Dutch and foreign addresses (many addresses coded as foreign were actually Dutch). The size of the files made this type of inspection and preparation quite cumbersome. For example, calculation of a simple frequency table could take 20 minutes, making this process quite time-consuming.

Nevertheless, the entire linkage process as described in section 5.3 was performed on standard office hardware and the entire process runs in a few hours.

5.5 Discussion

Statistics Netherlands has little experience in using probabilistic record linkage since deterministic records linkage is usually done with the BSN present in both files. However, as Statistics Netherlands is switching to the use of more existing external data sources which do not always contain the BSN, probabilistic record linkage is becoming more important.

In the deterministic linkage step 95.67 percent of the records could be linked. The probabilistic linkage resulted in an additional 3,813 records, increasing the percentage to 95,70. There are a number of reasons for the relatively small number of additional probabilistic links. First, the quality of the BSN used in the deterministic record linkage was very high as most employers and employees benefit from a correctly registered BSN. The BSN also contains check so that typing errors can easily be detected. Also the number of linkage variables and the power of the linkage variables were very limited. Although probabilistic record linkage allows for errors in the linkage variables, there was very little room for allowing errors in the linkage variables as that would lead to large number of incorrectly linked records. Furthermore, although address consists of four variables (first 4 digits postal code, next 2 digits postal code, house number and suffix), the entire address is regularly registered[1)] incorrectly in one of the registers. In such cases the records will not be linked. A strong linkage variable such as name would therefore have helped much.

The results of this linkage are in line with the results obtained from a previous simulation study (Ariel et al., 2014). There it was found that in case of large files and little overlap, the probabilistic methods do not perform much better than the deterministic method. In the current demonstration project, a deterministic linkage using all variables except the suffix, would have resulted in almost the same number of additional links. The main reason for this is that the number of linkage keys in this linkage and the strength of the linkage keys was limited.

1) A different address in the population register and employment register does not have to be incorrect. In the population register one had to register the address at which one lives most of the time. At the employer one will register the address where the employee wants the employer to contact him (e.g. to send the pay slip). These do not necessarily have to be the same.

Therefore, records quickly become no longer unique when errors are allowed in the linkage keys. Probabilistic linkage is still preferred when the number of linkage variables is larger, if one makes use of a distance metric between variables (such as the Jaro-Winkler distance for names), or of the uniqueness of variables (such as rare last names). Because of the small amount of overlap between the two files that were probabilistically linked, we had to estimate the m-probabilities using the deterministically linked dataset. The disadvantage of this is that it is uncertain that the estimated m-probabilities also apply to the remainder of the ER that needs to be probabilistically linked. Although error numbers in that part of the register are likely to be higher, the effect will be small when the relative order of the m-probabilities remains the same.

5.6 Conclusion

The linkage of administrative sources can be improved by using probabilistic linkage methods. The algorithms are able to cope with the size of the datasets and, after some modifications to the standard linkage procedure, probabilistic linkage was possible on datasets with very little overlap. Because a large number of records were removed in the deterministic linkage, the overlap between the remainder and the PR was very small. Because of this, the standard method for estimating the m- and u-probabilities, the EM-algorithm, could not be used. On the other hand, the deterministically linked dataset is a very powerful training dataset and we were able to estimate the m-probabilities using this dataset.

Very few records were additionally linked using probabilistic linkage because of the high quality of the BSN in these files and the relative weak power of the variables used in the probabilistic linkage. Therefore, it is questionable in such cases as this whether probabilistic linkage is worth the effort. Subsequent deterministic linkage steps where absolute similarity of the linkage variables is no longer required would probably give similar results with less (computational) effort. However, as we mentioned in the introduction, the linked dataset will be used together with other data sources to estimate the under-coverage of the PR using capture-recapture methods. In order to apply these methods, all records in both registers belonging to the same person should be linked, otherwise the under-coverage will be under- or overestimated. Therefore, the exact number of links is not that important, as long as we can be sure we have perfect linkage. Although the number of additional links gained with the probabilistic linkage is limited, the effect this can have on the estimates of the under-coverage can be considerable, especially when looking at specific populations. Therefore, it is important to put as much effort as possible into the linkage process.

6. Summary

Authors
G. (Gerard) van Grootheest (GGZ inGeest and VU University Medical Centre)
M.C.H. (Mark) de Groot (Utrecht University)
D.J. (Jan) van der Laan (Statistics Netherlands)
J.H. (Jan) Smit (GGZ inGeest and VU University Medical Centre)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)

6.1 Background

Record linkage makes it possible to combine data from different sources at the level of individual subjects. This approach to data collection can help answer research questions that are difficult or impossible to answer using data from just one source, or may be used to improve the quality of existing datasets. In the present paper, we demonstrated the feasibility of record linkage in three demonstration projects. Two of these linkages included health care data. We investigated how different approaches perform under real linkage circumstances where error rates and the overlap between two datasets are unknown.

We described how the choice of the datasets to be combined, linkage methods, and linkage variables affect the feasibility of a linkage project and the reliability of the results. In each demonstration project, specific requirements applied for privacy protection, both technically (data security) and procedurally (consent, permission, contracts). Topics that were covered in the different demonstration projects included the linkage algorithms and privacy-protecting measures such as pseudonymisation and the involvement of a Trusted Third Party (TTP).

The aim of the current study was to establish whether data linkage can be an effective and efficient way to enrich research cohorts with additional information from external sources. The quality of such enrichment is a crucial element when studying research questions that would otherwise be impossible to answer.

6.2 Conclusion

The demonstration projects within Biolink NL have shown that record linkage is a powerful alternative to obtaining research data through interviews, mail or web surveys. Detailed information can be retrieved without directly or repeatedly contacting large groups of respondents. The number of successfully linked subjects, however, depends not only on the quality and availability of linkage variables, but also on the algorithms that are used. Consequently, these aspects can have an influence on the representativeness of the linked dataset and should be taken into account when analysing the data.

Linkage variables

The current record linkages were performed as a proof of concept; the actual content of the linked datasets was used to describe the quality and added value of record linkage. Relatively much time was spent on characterising the quality of linkage variables and the agreement of

linkage variables in the linked datasets. If the quality of these variables is expected to be high, it may not be necessary to assess these extensively.

With the linkage of the Netherlands Twin Register (NTR) to insurance data from the Achmea Health Database (AHD), we showed that the strength of potential linkage keys depends strongly on the characteristics of the respective datasets. While given names and initials are important to distinguish between young twins, these variables are much less important for the correct identification of the majority of subjects in a large database such as the AHD, of which twins constitute only a small fraction. Surnames, however, were shown to be more important for the unique identification of subjects in the AHD than in the smaller NTR database. The choice of linkage variables to be used for record linkage is thus related to the characteristics of the research population, but also depends on the size of the databases.

The second demonstration project involved the linkage of the KOALA birth cohort study to pharmacy records from the Foundation for Pharmaceutical Statistics, SFK. A strong feature of this linkage was the availability of the complete address history of KOALA subjects. Furthermore, these data were checked through querying the population register, thereby ensuring a high quality of linkage variables in the KOALA dataset. Although it was not possible to accurately assess the number of missed links (false negatives), validation of the linkage results showed that the number of incorrect links (false positives) was very small.

The third demonstration project encompassed record linkage between the Employment Register (ER) and the Population Register (PR) and did not involve medical research data. This linkage was executed at Statistics Netherlands, and made use of different linkage variables than the first two projects. Firstly, the two registers were linked based on the BSN. Only those records that were not linked in this first (deterministic) step were selected for probabilistic linkage based on other personal variables. Although these linkage variables were expected to be of good quality, the linkage results were probably significantly influenced by the fact that only address variables and no names could be used. It was concluded that the probabilistic step added a small but specific group of subjects who could not be linked on BSN.

Linkage methods

Deterministic linkage on name and address fields is relatively straightforward and linkage of the NTR to the AHD showed that a few thousand subjects were successfully linked with this strategy. The results were however greatly improved by the application of probabilistic methods, especially when the Jaro-Winkler similarity metric was included in the calculation of linkage weights.

Similarity calculations were not possible in the linkages that were performed through a TTP, as such a strategy requires the linkage variables to be hashed prior to linkage. We found that the TTP produced a smaller number of linked subjects than the Jaro-Winkler linkage, indicating a reduced sensitivity. However, as the number of false positives in the anonymised linkage was small, it may be concluded that such a privacy-preserving linkage can still result in high quality datasets. The KOALA-SFK linkage confirmed this picture: whereas the number of false negatives and thus the sensitivity of the linkage could not be determined, the fact that no false links were identified indicated a high specificity.

Unlike the research cohorts, Statistics Netherlands was allowed use the BSN as a linkage variable to link the ER and the PR. Only those subjects who were not linked based on their BSN were selected for a probabilistic linkage based on sex, date of birth and address. This

additional step identified a relatively small number of extra links, which demonstrates the high quality of the BSN as a linkage variable in these datasets.

Representativeness of linked datasets

In the methods section we introduced the idea that indicators of representativeness that are used to describe the effects of survey non-response can also be used to describe the representativeness of a linked population, compared to the population which this group should represent. This representativeness indicator for linkage results, ℓ, is of practical use for assessing the representativeness of linked records. Moreover, it can be used as a powerful tool to make decisions during the linkage process, e.g. for determining the thresholds for probabilistic linkage. We calculated ℓ in two of the demonstration projects and used it to show how subjects with certain characteristics were under- and overrepresented after linkage.

In the NTR-AHD linkage, loss of representativeness was mainly caused by different regional coverage of the two data sources. Although the representativeness indicator was not calculated in the KOALA-SFK linkage, region was also a significant factor in this situation. The variable that had most influence on the representativeness of the ER-PR linkage was nationality.

Quality of retrieved data

In two linkage projects (NTR-AHD and KOALA-SFK) information was retrieved from an external database. The quality of this information obtained through linkage was evaluated by comparison with previously collected information in the cohorts' databases. In general, the newly obtained information complemented existing data and few discrepancies were found. The strongest feature of the linked datasets was the fact that the level of detail was much higher than could usually be achieved by questionnaires. So even when only a subpopulation of the original research cohort was successfully linked, the high precision of retrieved records resulted in very valuable datasets.

Ethical, legal and social issues

Research involving human subjects can only take place if informed consent is given. The enrichment of existing research cohorts with data from other registries is only allowed if subjects gave permission for retrieval of their data from external sources. If such permission is available and the data request is well motivated, the two parties can conduct record linkage under the premise of confidentiality.

As described in chapter 2, one of our intended demonstration projects could not be accomplished within the time available. Informed consent was available for linkage of the OMEGA cohort to the Dutch cancer registry, which has been linked to research cohorts in the past. Nevertheless, it turned out to be very difficult to reach agreement between all parties involved. A positive response was initially received from the cancer registry's Institutional Review Board. However, our rather unconventional proposal to validate linkage procedures using the pseudonymised BSN required contracts and data protection that was not straightforward. We learned that approval of an Institutional Review Board does not guarantee immediate access to the requested data; record linkage can only take place with cooperation of the information security officer and legal counsel. We recommend that these officers be involved in an early stage of data application.

The fact that one of the three approached IVF clinics decided not to cooperate in this project confirmed that linkage based on the pseudonymised BSN trough a TTP is considered controversial. In this sense Statistics Netherlands holds a special position: unlike in research cohorts the use of BSN is allowed here.

Conclusion

If consent and linkage variables are available and of good quality, record linkage can be an effective way of combining information from multiple sources. We found in our previous study on simulated datasets (Ariel et al., 1014) that the most suitable approach for a specific linkage project depends on several aspects, such as the size and expected overlap of the datasets. This finding is confirmed by the current results. Further important aspects in record linkage include the need for anonymised linkage and the desired balance between sensitivity and specificity.

When a fair amount of error is expected in the linkage key, a probabilistic approach should result in a larger number of links than a simple deterministic linkage, especially if a distance calculation is included in the algorithm. If access to unencrypted personal details is not possible, linkage by a TTP is a viable alternative, albeit at the cost of sensitivity.

Compared with survey research based on questionnaires or interviewing, record linkage can be considered as an efficient method of data collection. A large advantage is that subjects do not need to be contacted if the data of interest are already recorded elsewhere. However, the efficiency of record linkage may differ per situation. Firstly, gaining approval to link data from external data sources to an existing cohort can be a very time consuming process, without the guarantee that linkage will be successful. Secondly, the number of linked records and the representativeness of the linked population must be large enough to apply solid statistical analyses. This notion can be compared with survey response. Nevertheless, the objective information recorded by health care providers often has a level of detail that can hardly be achieved by surveys. Even when only a subset of the population can be linked with an external data source, the combined data can thus help to address research questions that would otherwise be unanswered.

References

Adams MM, Wilson HG, Casto DL, Berg CJ, McDermott JM, Gaudino JA, McCarthy BJ. 1997. Constructing reproductive histories by linking vital records, *American Journal of Epidemiology*, 145:339–348.

Ariel A, Bakker B, de Groot M, van Grootheest G, van der Laan J, Smit J, Verkerk B. 2014. Record Linkage in Health Data: a simulation study. Statistics Netherlands.

Bakker BFM, Daas P. 2012, Some Methodological Issues of Register Based Research, *Statistica Neerlandica*, vol. 66, nr. 1: 2–7

Bergman L, Beelen M, Gallee M, Hollema H, Benraadt J, van Leeuwen F. 2000. Risk and prognosis of endometrial cancer after tamoxifen for breast cancer. The Lancet; 356(9233):881–887.

Bethlehem JG, Cobben F, Schouten B. 2011. Handbook of Nonresponse in Household Surveys. John Wiley & Sons, Hoboken, NJ.

Bozkurt O, de Boer A, Grobbee DE, de Leeuw PW, Kroon AA, Schiffers P, Klungel OH. 2009. Variation in Renin-Angiotensin system and salt-sensitivity genes and the risk of diabetes mellitus associated with the use of thiazide diuretics. Am J Hypertens; 22(5):545–51.

Christen P. 2012. Data Matching: concepts and techniques for record linkage, entity resolution and duplicate detection, New York, Springer.

De Heij V, Schouten B, Shlomo N. (2014), RISQ 2.1 manual. *Tools in SAS and R for the computation of R-indicators and partial R-indicators*, available at www.risq-project.eu.

Derks EM, Hudziak JJ, Boomsma DI. 2007. Why more boys than girls with ADHD receive treatment: a study of Dutch twins. Twin Res Hum Genet; 10(5):765–70

DuVall, SL, Kerber RA, Thomas A. 2010. Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators, *Journal of Biomedical Informatics*, vol. 43, pp. 24–30.

Eussen SR, de Jong N, Rompelberg CJ, Garssen J, Verschuren WM, Klungel OH. 2010. Effects of the use of phytosterol/-stanolenriched margarines on adherence to statin therapy. Pharmacoepidemiol Drug Saf.; 19(12):1225–32.

Fellegi IP, Sunter AB. 1969. A Theory for Record Linkage. Journal of the American Statistical Association 64:1183–1210.

Fienberg S. 1972. The multiple recapture census for closed populations and incomplete 2k contingency tables, *Biometrika*, 59, pp. 409–439.

Fienberg SE, Manrique-Vallier D. 2009. Integrated methodology for multiple systems estimation and record linkage using a missing data formulation. Asta-Advances in Statistical Analysis 93:49–60.

Florentinus SR, Souverein PC, Griens AMG, Groenewegen PP, Leufkens HGM, Heerdink ER. 2006. Linking pharmacy dispensing data to prescribing data of general practitioners. BMC Medical Informatics and Decision Making. 6:18

Gerritse SC, van der Heijden PGM, Bakker BFM. 2015. Sensitivity of population size estimation for violating assumptions in loglinear models, accepted by Journal of Official Statistics.

Griens, AMGF, Lukaart, JS, van der Vaart, RJ. 2011. Facts and Figures 2010. Foundation for Pharmaceutical Statistics, The Hague, The Netherlands. www.sfk.nl Accessed in June 2014.

Herings, RMC, Bakker A, Stricker, BH, Nap G. 1992. Pharmaco-morbidity linkage: a feasibility study comparing morbidity in two pharmacy based exposure cohorts. J. Epidemiol Community Health, 46:136–140.

Herzog TN, Scheuren FJ, Winkler WE. 2007. *Data Quality and Record Linkage Techniques*, New York, Springer.

Hoekstra C, Willemsen G, van Beijsterveldt CEM, Lambalk CB, Montgomery GW, Boomsma DI. 2010. Body composition, smoking, and spontaneous dizygotic twinning. *Fertility & Sterility*, 93, 885–893.

Hser, YI, Evans E. 2008. Cross-system data linkage for treatment outcome evaluation: Lessons learned from the California Treatment Outcome Project, *Evaluation and Program Planning*, vol. 31, pp. 125–135.

IWGDMF (International Working Group for Disease Monitoring and Forecasting). 1995. 'Capture-, recapture- and multiple record systems estimation. Part 1. History and theoretical development', *American Journal of Epidemiology*, 142, pp. 1059–1068.

Karmel R, Anderson P, Gibson D, Peut A, Duckett S, Wells Y. 2010. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. *Bmc Health Services Research* vol. 10, nr. 4.

Lyons, RA, Jones KH, John G, Brooks CJ, Verplancke JP, Ford DV, Brown G, Leake K. 2009. The SAIL databank: linking multiple health and social care datasets. *Bmc Medical Informatics and Decision Making* vol. 9; nr. 3.

Meray N, Reitsma JB, Ravelli AC, Bonsel GJ. 2007. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number, *Journal of Clinical Epidemiology*, vol. 60, pp. 883–891.

Schelleman H, Stricker BH, Verschuren WM, de Boer A, Kroon AA, de Leeuw PW, Kromhout D, Klungel OH. 2006. Interactions between five candidate genes and antihypertensive drug therapy on blood pressure. Pharmacogenomics J. Jan-Feb;6(1):22–6.

Schouten B, Cobben F, Bethlehem J. 2009. Indicators for the representativeness of survey response, Survey Methodology, Vol. 35, No. 1, pp. 101–113.

Schouten B, Shlomo N, Skinner C. 2011. Indicators for Monitoring and Improving Representativeness of Response, Journal of Official Statistics, vol. 27, pp. 231–253.

Shlomo N, Skinner CJ, Schouten B. 2012. Estimation of an Indicator of the Representativeness of Survey Response, *Journal of Statistical Planning and Inference* Vol. 142, No. 1, pp. 201–211

Shlomo N, Schouten B. 2013. *Theoretical properties of partial indicators for representative response*, technical report, available at www.risq-project.eu.

Statistics Netherlands (2014). *Dutch Census 2011: Analysis and methodology*, The Hague/ Heerlen, Statistics Netherlands.

Van der Heijden PGM, Whittaker J, Cruyff M, Bakker B, van der Vliet R. 2012. People born in the Middle East but residing in the Netherlands: invariant population size estimates and the role of active and passive covariates, *Annals of Applied Statistics*, 6, nr. 3, pp. 831–852.

Victor TW, Mera RM. 2001. Record linkage of health care insurance claims, *Journal of the American Medical Informatics Association*, vol. 8, pp.281–288.

Vink JM, van Kemenade FJ, Meijer CJ, Casparie MK, Meijer GA, Boomsma DI. 2010. Cervix smear abnormalities: linking pathology data in female twins, their mothers and sisters. Eur J Hum Genet. 19(1):108–11.

Wallgren A and Wallgren B. 2014. Register-based Statistics: Administrative Data for Statistical Purposes. *Wiley Series in Survey Methodology*, New York, Wiley.

Winkler WE. 1990. String comparator metrics and enhanced decision rules in the Felligi-Sunter model of record linkage. *Proceedings of the section on survey research methods*, American Statistical Association, pp. 354–359.

Winkler WE. 1994. Advanced Methods for Record Linkage. Proceedings of the Section on Survey Research Methods, American Statistical Association, 467–472.

Yancey WE. 2002. Improving EM Parameter Estimates for Record Linkage Parameters. Proceedings of the Section on Survey Research Methods, American Statistical Association.

Weber, SC, Lowe HJ, Das A, Ferris TA. 2012. A simple heuristic for blindfolded record linkage, *Journal of the American Medical Informatics Association*, vol. 19, pp. e157–e161.

Zhang, LC. 2012. Topics of statistical theory for register-based statistics and data integration, *Statistica Neerlandica*, vol. 66, nr. 1, pp. 41–63.

Zhu, VJ, Overhage MJ, Egg J, Downs SM, Grannis SJ. 2009. An Empiric Modification to the Probabilistic Record Linkage Algorithm Using Frequency-Based Weight Scaling, *Journal of the American Medical Informatics Association*, vol. 16, pp. 738–745.

Authors

Established and edited by

G. (Gerard) van Grootheest (GGZ inGeest and VU University Medical Centre)
M.C.H. (Mark) de Groot (Utrecht University)
D. J. (Jan) van der Laan (Statistics Netherlands)
J.H. (Jan) Smit (GGZ inGeest and VU University Medical Centre)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)

Chapter 1:

G. (Gerard) van Grootheest (GGZ inGeest and VU University Medical Centre)
M.C.H. (Mark) de Groot (Utrecht University)
D. J. (Jan) van der Laan (Statistics Netherlands)
J.H. (Jan) Smit (GGZ inGeest and VU University Medical Centre)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)

Chapter 2:

G. (Gerard) van Grootheest (GGZ inGeest and VU University Medical Centre)
D.J. (Jan) van der Laan (Statistics Netherlands)
J. A. (Jasper) Bovenberg (Legal Pathways)
M.C.H. (Mark) de Groot (Utrecht University)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)

Chapter 3:

A. (Adelaide) Ariel (GGZ inGeest)
T.J. (Tina) Glasner (VU University Amsterdam)
E.C.M. (Bep) Verkerk (GGZ inGeest)
C.E.M. (Toos) van Beijsterveldt (VU University Amsterdam)
G. (Gerard) van Grootheest (GZ inGeest and VU University Medical Centre)
D.J. (Jan) van der Laan (Statistics Netherlands)
M.C.H. (Mark) de Groot (Utrecht University)
S.T. (Sipke) Visser (Mondriaan Foundation)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)
D.I. (Dorret) Boomsma (VU University Amsterdam)

Chapter 4:

D. (Dianne) de Korte (Maastricht University)
A. (Adelaide) Ariel (GGZ inGeest)
E.C.M. (Bep) Verkerk (GGZ inGeest)
G. (Gerard) van Grootheest (GGZ inGeest and VU University Medical Centre)
W. (Willem) de Bruijn (Utrecht University)
S.T. (Sipke) Visser (Mondriaan Foundation)
M. (Monique) Mommers (Maastricht University)
M.C.H. (Mark) de Groot (Utrecht University)
J.D.L. (Jan Dirk) Kroon (Dutch Foundation for Pharmaceutical Statistics)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)
C. (Carel) Thijs (Maastricht University)

Chapter 5:

D.J. (Jan) van der Laan (Statistics Netherlands)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)
M.C.H. (Mark) de Groot (Utrecht University)
G. (Gerard) van Grootheest (GGZ inGeest and VU University Medical Centre)

Chapter 6:

G. (Gerard) van Grootheest (GGZ inGeest and VU University Medical Centre)
M.C.H. (Mark) de Groot (Utrecht University)
D.J. (Jan) van der Laan (Statistics Netherlands)
J.H. (Jan) Smit (GGZ inGeest and VU University Medical Centre)
B.F.M. (Bart) Bakker (Statistics Netherlands and VU University Amsterdam)

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress: Statistics Netherlands, Grafimedia
Printed by: Statistics Netherlands, Grafimedia
Design: Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
verkoop@cbs.nl
Fax +31 45 570 62 68
ISBN 978 90 357 2025-1