Unambiguous identification of hospital patients
Case study at the university departments of the General Hospital
Vienna

Authors: P. Sachs, W. Gall, A. Marksteiner, W. Dorda

Department of Medical Computer Sciences
University of Vienna
General Hospital
Spitalgasse 23
A-1090 Vienna
Austria

e-mail: peter.sachs@akh-wien.ac.at

Keywords: administrative data, patient identification, record linkage
Summary

This article considers the problem of identifying patients in one or more heterogeneous personal databases. The unambiguous identification of patients is an essential prerequisite for an efficient patient care system. We discuss the problems involved in this task and suggest how they can be dealt with. The solution of automatic consolidation of patient records requires programming, organisational and work psychology measures. Following a survey of conventional identification methods, the method developed at the Department of Medical Computer Sciences, which is based on the current clinical situation at the General Hospital in Vienna (AKH – Allgemeines KrankenHaus), is described in detail. The basic principle is to identify patients unambiguously by means of an ID (IZAHL) derived directly from the personal data. Thereby a deterministic technique without probability weighting is used - all compared information must correspond completely. The article closes with a critical survey of experience gathered to date.
1. Introduction

Several different specialised computer systems are used to deal with the various aspects of clinical practice today. These systems are tailored to the tasks they perform. This dedicated use ensures that the tasks are accomplished efficiently, but a large amount of technical and organisational adaptation is required for the data to be exchanged between systems. There are normally two types of systems in hospitals: administrative patient management systems, and those handle medical and scientific patient data.

The main task of an administrative system is to oversee the charging parameters. Patients are entered into the system and identified on the basis of personal details provided by the patients themselves. In the optimal case patients are identified by referring to their ID-card. However, in some cases (e.g. loss of card, accident) the identification process has to rely on conventional techniques. This is done either orally or from a registration form. Registration in this way is by nature extremely susceptible to error because of the vagaries of human communication. Where information is provided orally, the phonetic characteristics of a name can be misheard, while errors with written forms can occur because of legibility problems or spelling mistakes. The fact that patients of different nationalities and cultures are involved only exacerbates the situation. A further serious shortcoming of this type of identification is that name changes for example through marriage are inadequately allowed for, one has to ask patients specifically to give their name at birth to answer this problem. One of the most important influences on the identification process, however, is the absence of motivation in most cases to link current patient data with previous data, the main object of the system is to clearly demarcate one administrative procedure from all others. In other words, the current record is all that is required for charging of medical services. For a charging department dealing with individual medical operations, it is important to associate a particular operation with a particular patient but of little relevance if this patient happens to be entered several times in the system with different identifications.

The medical patient data management system is subject to fundamentally different rules. Here, the correct linking of all patient data can be life-saving since correct treatment can be guaranteed only if all medical information is available. The same applies if a chronological anamnesis is required. This is of central significance for the doctor who needs to know a patient’s previous history in order to make an overall assessment.
In the field of scientific research it is also necessary to have access to all data from each patient and to be able to link these data. If patient data are not consolidated, or if they are consolidated incorrectly, the results will be incorrect - as is the case where findings are wrongly attributed by combining data from two separate patients (homonym error) or where one patient has several IDs, which would give multiple records for a single patient (synonym error). This problem carries over to hospital management as well and can even affect health policy decisions in general. Incorrect figures can lead to incorrect assumptions.

Since administrative and scientific systems inside and outside a hospital often work together and since patients in the course of their life have contact with different medical departments, it would help communication between systems if really unique patient identification could be generated as the basis for an efficient patient care system and reliable scientific research.

2. Problem

Efforts to consolidate patient data coming from heterogeneous databases with different identification schemas have a long history. Automated methods for linking patient records have been described as early as the 1970s [1].

The accuracy sought through automation must naturally be reconciled with a desire for simplicity and speed. Attempts have been made to make procedures less expensive [2] and to minimise the amount of personally-identifying information in medical records (‘scrubbing’) [3]. Statistical methods can be used to identify whether two discrepant patient records refer to the same individual or not [4, 5]. Many studies have attempted to devise general strategies for linking patient data using procedures for evaluating individual patient information (before using a certain linkage-strategie the Shannon entropy as a measure of the minimum set of the available information is calculated) [6] or different deterministic [7] or probability [8] techniques or a combination of both [9]. The various linking methods have also been compared and the effectiveness of the individual methods examined [10]. The problems arising as a result of automated patient record generation have been investigated on a number of occasions in recent years [11]. Even if the quality of personal identifying data is poor [12], the linkage process can be carried out by using a probability matching method to build best pairs from two files. Linking several files with a large database is an efficient method to provide data verification within the files and the database itself [13]. Most represented linking methods rely on static files of patient data. The advantage here is that there is no time limit in processing two or more such files.
Within one hospital organisation the process of linking patients is straightforward: Most data structures are widely compatible, all patient data, especially those which are necessary for linking, are available without any restriction and detected errors can be fed back easily to the originator and corrected. More serious problems appear when patient data come from heterogeneous sources. In many European countries the linkage of patient data within the framework of medical research is controlled by law. Patient data are to be rendered anonymous in an irreversible way before they can be linked with data of other provenance. For example, anonymous linkage is necessary when the data originate from several follow-ups [14] or if the data contain sensitive information [15]. To be able to link those data, all sources of patient data should use the same encryption algorithm. In order to ensure data security on the one hand and to reduce false links on the other hand, several quality assessment methods of anonymous record linkage are proposed [10], [15].

In spite of these protracted efforts and the use of established methods, there still remains a small number of ambiguous cases in which manual intervention is required.

2.1 Patient records

Patient data in a hospital can come from various sources. Most hospitals have a central admissions and discharge office, but only for admitted patients. Out-patients are entered in the out-patient clinics. The quality of the data depends to a large extent on where the data are entered, which system consolidates the disparate data, the tightness of the hospital organisation, the use of control mechanisms and, not least, the general attitude of the staff entering the data towards data quality.

Inconsistencies in the administration can be cleared up most effectively while the patient is still available. Provided that patients do not make deliberately false statements, they are in the best position to know whether they have been treated in the department before, how their name is spelt, and which of the various patients with the same name they are. Should patients be unable to provide the data (e.g., if they are unconscious or a newly born baby), a temporary ID can be created and modified later.

Important prerequisites for the successful functioning of this input system are good training for persons involved in patient administration, on-going quality control and user-friendly, dedicated software.

Whenever an entry is made in a hospital information system, be it for an admission or for an out-patient, and before temporary identification (admission number, out-patient number, etc.) is provided, it must be established whether the person concerned is a new admission. If so, a new ID is generated by the system. This ID can be linked to
the temporary ID and to all future features relating to the patient. If the patient already exists in the database, the temporary ID is merely linked to the existing unique patient record.

Communication with an information system is almost always carried out on-screen. The interaction can be considerably reduced if the patient can present any former documentation (e.g., out-patient card, barcode, name labels, laboratory print-out). If this is not the case, the patient’s data must be entered on-screen. The following information is essential to avoid confusion:

- sex
- date of birth
- name at birth
- surname
- first name
- social security number

Nationality and address can also be used for differentiation, if necessary. Assuming that the information given is not falsified, each individual item provides identification with a certain degree of reliability. The primary object is to find out if the patient is already known by the system. The best point in time for linking the patient’s data to already known information is during his physical presence in the hospital. When patient linking has to be done without administrable information of the patient, and there are multiple assignments possible, the staff should follow clear rules in determining who is the patient. Because wrong linking often leads to calamities, the linking actions of the clinical staff should be monitored automatically by a computer-program. This program can simulate the deterministic decision process of the clinical staff and try to find out mistakes. Most published examples are using probability weighting algorithms. Weights are given to the personal variables according to their degree of discrimination. To avoid synonym errors spelling transformations can be introduced.

One of the most important factors in the identification process, however, is the motivation to link current patient data with earlier data. For administrative systems, which are designed principally with charging in mind, the linking of data from a current admission or treatment with earlier ones is of little importance.

The opposite is true of scientific research, where there is an obvious need to have access to all patient data. Multiple admissions of a single patient should not be inter-
interpreted as single admissions by a multitude of patients as otherwise even the simplest counting procedures will produce incorrect results.

2.2 Patient identification

Patient records are in the form of text and the degree of precision with which they can be processed by automatic systems is by nature limited. One of the most widely used methods of managing personal data is to use an unambiguous ID under which the data are stored, recovered and linked together. In modern databases an ID used for internal labelling only is not important, because there is no technical necessity for a unique key and each patient record can be retrieved by its contents itself. IDs are rather used to access to the data or as a mean to refer to anonymous data.

2.2.1 Simple identification

The simplest form of identification is the allocation of a serial number. This has the advantage that little memory is required, although with the megabyte storage media available today, this argument has lost its relevance. The allocation mechanism is extremely simple. As a number gives no indication of the identity of the patient, it is essential that secondary access be provided (e.g., through the patient’s surname). When patients come to the hospital for the first time, they are given the next available free number. The only information that this form of identification provides is that it is possible to establish which of two patients in the same department was entered in the database first. In modern database systems a serial number is generated automatically and provided with a time stamp indicating the precise time at which the number was allocated.

2.2.2 Structured identification

The allocation of a structured number is a little more complicated. A structured number consists of one or more features, each identified by a number to discriminate different cases within the feature (e.g., SURG-1, SURG-2; SURG characterises the feature). Features can include:

- time stamp (e.g., year)
- organisation unit (e.g., department, ward, out-patient)
- patient-related feature (e.g., sex)
- combination of the above

The advantages of better structuring of all stored patients to have a part within the ID with information including meaningful items must be seen against the additional work involved in data management, particularly in the case of overlapping features and of alterations. If there is an error in one of the items the complete ID must be changed.
This leads to an additional effort, because all records including the wrong ID must be updated. This can be done automatically but in case of homonym problems an operator has to intervene.

### 2.2.2.1 Social security number

A special form of structured numbering is the Austrian social security number. It has 10 numeric positions (SSSCDDMMYY) and combines the date of birth of the citizen DDMMYY with a serial number SSS allocated sequentially including a check digit C, which is calculated as follows: each position of SSSDDMMYY is multiplied by a predetermined factor. The check digit then computes to the sum of these products modulo 11. The advantage of this form of identification is that the access code is ideally known by the patients themselves, making unambiguous identification that much more reliable than if they were merely to give their name. However, the current situation in Austria shows some serious disadvantages. One is that social security numbers only exist if the patient is insured, which is not always the case - when the patient is not resident in Austria, for example. A further disadvantage is the absence of the century in the birth date.

Across Europe, a variety of different types of social security numbers are used. To change this unfavorable situation, the Comité Européen de Normalisation (CEN) started an initiative to introduce a standarised social security number using 35 alphanumeric positions. It will contain extracts of name at birth, surname, first name, sex, date of birth and the present social security number.

Following the recommendations of CEN the social security number will be an identification of administrative and common clinical data structure for intermittently connected devices used in healthcare including machine readable cards [16].

### 2.2.2.2 LIC/MAC

In all hospitals of the Viennese community the KIS (Krankenhaus Informations System, the German equivalent of HIS - Hospital Information System) is used for administration. The identification code used - LIC (Local Identification Code) to identify the patient and MAC (Mother-Admission-Code) to mark the medical case (e.g. one hospital stay, one sequence of treatments belonging together) - represent an attempt to provide an unambiguous form of identification for all Vienna. The values of LICs and MACs are taken from the same set of values. When a patient is registered the first time a new MAC is generated and attached to him. Exactly this MAC is used as LIC, too. If the patient is readmitted with other complaints he gets a new MAC but keeps the old LIC. Watching the list of MACs then, one can obtain the
original medical case with which the patient was first registered by looking for the MAC that equals the patient’s LIC.

The structure is as follows:

**Table 1** Structure of LIC/MAC

The LIC/MAC code takes up 20 bytes of memory. The components province and hospital allow for a uniform identification system throughout Austria. There is no check digit component as identifying codes are usually acquired by means of barcodes.

From this ID, it is immediately possible to retrieve the date and location of treatment (9 = Vienna; 01 = AKH, SURG10400 = surgical department, station 400; 99 = 1999, 00 = 2000). This clear structure has not been maintained in the case of MACs for hospitalised patients, however. At the AKH, for example, all such patients are allocated to the cost centre 20311. This number is generated by the admission and discharge office for ward patients, regardless of the department to which they are admitted. If a patient is transferred to an other ward the MAC does not change.

### 2.2.2.3 IZAHL/working number

An interesting form of identification that has been used for over 20 years is to be found in WAMIS (Wiener Allgemeines Medizinisches Informations System, the German equivalent of Vienna General Medical Information System) [17]. The WAMIS can be seen as a kind of predecessor system of the KIS. It was developed at the Institute for Medical Computer Science (IMC) of the University of Vienna. Till this day, on the one hand the WAMIS assumes some special tasks for some few departements of the AKH in Vienna, which the KIS cannot achieve, on the other hand the WAMIS is the central server for handling communication flows between most of the scientific systems of the IMC. The KIS and the WAMIS are completely separated systems. Both use their own patient ID.

A very important function of the WAMIS is to store and maintain patient data to guarantee unambiguous identification of hospital patients by means of the IZAHL. The IZAHL is the central patient-ID of the scientific systems of the IMC. It contains extracts from the personal data. The IZAHL structure is as follows:

**Table 2** Structure of the IZAHL

The IZAHL consists of 15 characters. As storage space was at a premium 20 years ago, however, the IZAHL was compressed so that the stored data only took up
7 bytes. The century problem caused by the reduction of the date of birth to six figures is solved by introducing a century code (0 = even century, 1 = odd century), which is hidden in the compressed form of the IZAHL. Extracting the short IZAHL gives the 15 characters of the long IZAHL plus the century code. In this way the only possible confusion would be of two patients with the same name born on the same date two centuries apart. Regarding the relationship of IZAHLs with serial number > 1 and all IZAHLs as significant random sample the probability that two patients get the same IZAHL is 0.028 (see table 5).

Most of the structural features of the IZAHL code are invariable, patient-specific data. The disadvantage of incorrect input of patient data (e.g., in the case of phonetic discrepancies) carries over to the ID itself, but this disadvantage can be remedied through wide-ranging organisational measures. The advantage of and basic idea behind the system is that all identifying features are provided by the patients themselves.

Additionally, working numbers are used to distinguish the different periods of hospitalisation of a patient. These numbers consist of a two-digit department number separated by a five-digit serial number, and a check digit. Patients who stayed at several departments have several working numbers. Some departments are using one working number to label one medical case. Thus, one patient owns several working numbers at these departments. Other departments are using working numbers to label the patient locally at the department. In this case a patient owns only one single working number, and different cases are distinguished only by their timestamp.

3. Method

The development of a general solution to the problems mentioned above was prompted by the protracted efforts at the AKH in Vienna to identify patients without ambiguity. The bodies involved in these endeavours are the IMC, which is interested in the scientific use of the data, and the ARZ (Administratives RechenZentrum - Administrative Computer Centre) of the City of Vienna, which uses the KIS for administrative purposes and passes on these data to the IMC.

When patients are entered in the KIS for the first time, they are given a unique ID in the form of an LIC. The medical case is identified by a further code, the MAC. The LIC and MAC are taken from the same set of values, in other words the patient LIC is always identical with the first MAC for that patient. This situation, which seems quite plausible at first sight, caused considerable difficulties when first implemented: if patients are entered by mistake as a new case when they visit the hospital again, a new LIC is generated. Further visits, in which new MACs are allocated, can now be asso-
associated with LIC-1 or LIC-2, i.e., the patient now has two identifiers and two sets of data linked to them.

As all patient data need to be linked for scientific research as well, the IMC is attempting to establish a relationship between KIS patient data and data that are already stored in the IMC-database (approx. 2 million patient records) and to generate a really unique patient identification, thereby laying the foundation for serious and, above all, correct scientific analysis. Till this day 7.7 millions of MACs and 1.6 millions of LICs have been stored in the KIS-database. The annual increment is approximately 600.000 MACs and 200.000 LICs.

There are several research units at the AKH, each of which - in an attempt to remain as independent as possible - has its own patient records. Since patient records have now been stored in the KIS for the last six years or so and these data are not available for scientific research, they are transferred electronically to the IMC, where they are compared with the existing patient records. As part of this automated patient integration process, records are given a unique patient ID (IZAHL) and consolidated in a central patient database.

3.1 Automatic consolidation

Table 3 shows all combinations which are possible on the basis of the ID (LIC) and patient records themselves, from which the IZAHL is formed:

Table 3  Actions depending on LIC-IZAHL constellations

Table 3 reflects the IMC’s view of the data. The ‘LIC new’ constellation, for example, means that a patient whose data are tagged to an LIC that is unknown to the IMC is entered as new irrespective of whether this is really the case. A transfer algorithm has been designed to detect errors of this sort (known LIC – new IZAHL) so that appropriate action can be taken. All these constellations are processed automatically only if all record items have the expected form.

3.2 Linking algorithm

The patient records which come from the KIS are processed by an IMC-program as follows:

Figure 1  Flowchart of the linking algorithm.
1. Records that cannot yet be processed and that do not contain sufficient information (e.g. if records are transferred partly or split) for identification are buffered and feedback requested from the KIS.

2. The incoming records are checked to see whether an earlier queried version with the same ID (MAC) is buffered. If so, the corresponding buffered data are updated and the normal data uptake routine initiated.

3. The incoming records are compared with existing records as follows:
   - The MAC is checked to see whether it is already known. If so, the associated data are checked to see whether they comply with earlier dispatches. A known MAC associated with a LIC that does not correspond to earlier received records indicates one of the two following scenarios: If the LIC is known the KIS has linked two patient records, whereas the LIC is unknown the KIS has split one patient record in two.
   - For patients who cannot be identified through an ID (LIC or MAC) an attempt is made with the help of the main identifying features (family name at birth, date of birth, sex) to form an IZAHL and to link it to existing records.
   - An algorithm working with the phonetic roots of the first name helps to identify spelling mistakes. Compared with other efforts described in the literature the transformation process of the first names intends to be rather unambitious. Therefore it is highly discriminating, because the superior aim is to avoid homonym errors.
   - If ambiguous situations occur, the data are not stored and a query is sent back for further clarification.

4. Records that overcame the first three steps (the identity of the patient is clear now) are checked to establish whether they are additional data, changes or repeat information and the appropriate action taken.

5. For new patients the IZAHL is generated. For new admissions the admission transaction is initiated and the corresponding case-ID (sequence number) generated.

6. All transactions (including discharges), important information (e.g., name change) and error messages are registered and can be viewed selectively by users (e.g., return orthopaedic out-patient visits, all admissions, deaths).
3.3 Accompanying measures

As the correct consolidation of patient records is of great significance in all systems, a qualified person was entrusted at both the KIS and the IMC with processing problems with the transfer algorithm and resolving them as rapidly as possible by means of cross-platform measures.

Every night, scans are performed in the KIS to check new data against the record database, and the results of these scans are transferred to the IMC.

4. Experience and results

The solution of the problems discussed to date through the automatic consolidation of patient records involved programming, organisational and work psychology measures. The implementation of these measures took approx. six man/years and may now be regarded as complete. A lot of IMC systems, many of which already had historical significance, were affected.

This form of identification was chosen in 1975 because at that time self-explanatory codes were preferred. The IZAHL is a structured identification. It is available on two forms – encrypted and readable. The encrypted ID can be decoded by means of a simple program to provide useful information. As each ID consisted of 7 bytes, billions of identifying codes can be generated, with the result that every genuine new admission need only to be one number higher than the previous one for it to have its own unique ID.

The WAMIS system currently contains more than 1.96 million IZAHLs. Until 1993 most patients were entered in WAMIS itself. Since then patient records have been entered in KIS and transferred electronically from there to WAMIS. Some records established by individual departments in the 1970s with the aid of punched cards have also been included in WAMIS and added to the general database.

An analysis of the IZAHLs incorporated in the system in such different ways provides information on the possibilities for error in the data used for identification and on the limits of the support algorithms. Demographic developments can also be inferred - although this is not of central interest to our study.

To decide how a patient has to be positioned against the background of all stored patient data in WAMIS several steps are to be carried out. Patient’s data (name at birth, date of birth and sex) produce a fragment of the IZAHL (without serial number). The different situations are shown in table 4.
• If the fragment does not exist the patient is clearly seen as new. In this case the IZAHL-fragment is concatenated with the serial number 1.

• If the fragment is unique and there is full concordance in the subordinate fields (family name, first name, social security number) the patient is identified and the data is linked to this IZAHL.

• If there is no concordance the patient is new and is associated with sequence number 2.

• If the fragment exists more than once some further steps are necessary. The subordinate fields are compared.
  • If the comparison does not show any concordance the patient is clearly seen as new. In this case to the highest existing serial number of the fragment is increased by 1 and concatenated with the fragment.
  • If there is a unique concordance the patient’s data is linked to that IZAHL which shows the unique concordance.
  • If the there is multiple concordance the linking process can only be carried out manually.

Table 4  IZAHL-fragments matching

Some 57,100 IZAHLs have a serial number > 1. These patients have been declared as new automatically (see table 4) or manually by datatypists. In almost 54,000 of those cases the serial number was generated correctly due to some lack of concordance in subordinate fields.

The personal data of the remaining 3000 IZAHLs are in fact identical in all respects with existing IZAHLs but have been declared as new patients wrongly, instead of being linked to existing IZAHLs. Those errors exclusively result from human failure. Specifically, it was due to inadequate training of the operators in one department. In good faith and despite the test routine indications, the operators insisted on entering the data as a new record. The operating error would have been discovered much earlier if the system had not simply accepted the rejection without comment but had asked for confirmation. The 3000 wrong links represent the state of the data at the end of 1998. They are currently corrected.

In all data of the 1,96 million IZAHLs there can be an unknown number of errors. The detection of these errors can lead to different actions for example to the linkage of two patients. The following sections examine the possible errors in the various data fields.
4.1 Sex

Except for rare cases of sex change, sex is a constant patient feature. In the event of a sex change it is permissible to have two records, since two different personalities are effectively involved.

As with all input fields, errors are possible here that cannot be rectified by the checking algorithms. The idea of comparing the patient’s first name with a sex-specific set of first names would not work in practice. Although common names like Hans and Fritz are sex-specific, even a name like Hansi would not be detected by the algorithm. In addition, various combinations of input fields are possible, thereby giving a “false positive” result (two wrongs can accidentally make a right).

4.2 Date of birth

Although a checking algorithm can identify formal errors (e.g., 31 June or 29 February in a non-leap year), it can do nothing about numbers being reversed. No program can verify whether someone was born on 12 or 21 December. With 1919 and 1991 the operator can check whether the patient was admitted to the geriatric or paediatric clinic, but that is all.

4.3 Surname and name at birth

The most frequent reason for a link to an existing IZAHL being rejected was an incorrectly written surname or name at birth. Automatic checking routines are almost impossible here because names change with time and there are many possible ways in which they can be spelt (Vaclavek and Vaclavik, for example, are equally common and Waclawek and Waclawik are not necessarily spelling mistakes but different names). An automatic comparison in a checking algorithm provides no assistance here. When data are being entered on screen, the operator can make decisions by displaying patients with almost identical names and comparing the rest of the data. Such parallel displays are only useful, however, when the numbers involved are manageable; the operator cannot be expected at busy times to search through a long list of names. During training sessions the importance of the name at birth should be emphasised. Registration forms and screen templates should be designed so the name has to be written only once if surname and name at birth are identical. Provided that it is written correctly, the name at birth, like the sex, is an unchanging parameter. The correct interpretation of a surname can be complicated if a title or abbreviation of a title is used intermittently. This problem can be resolved by having a separate field for title.

4.4 First name
Demographic changes have resulted in the frequent occurrence of names from many different cultures. A checking algorithm, which should logically have a multi-level structure, should take this into account. Names that appear to have been spelt wrongly (Gerte instead of Grete, for example) can be corrected only if there is no doubt as to the origins of the name. Any attempt to apply German spelling rules to African or Asian names would only cause confusion. Nor is it possible to identify the sex or nationality of the person on the basis of the first name. The checking algorithm for first names had therefore to be adapted on numerous occasions to changing circumstances. Fashions - double names, for example - also reduce the reliability of checking algorithms if the names are not indicated precisely. There was a case of quadruplets born at the AKH, for example, who were all given double names in which one of the components was identical (Marie-Anna, Marie-Louise, Marie-Lotte and Marie-Therese).

4.5 Nationality and address

When the linking process must be carried out manually nationality and address of the patients are included into the decision. Both nationality and address can change. Complete concordance provides indication of identity, but incomplete concordance does not mean that a new patient is involved. Changes in nationality as a result of nationalisation are relatively frequent (YU ⇒ A). The new states that have come into existence as a result of changes in the political situation also contribute to this frequency (YU ⇒ SLO/HR, CS ⇒ CZ/SK).

4.6 Effect on automatic linking

To prevent the risk of incorrect linking of two non-identical patients, errors should not normally be corrected automatically. Patients whose identity is unclear should be stored on the waiting list or be entered only as new patients. The waiting list should be processed only by responsible and experienced staff in an interdisciplinary database management section. These persons will also process the many log messages that are generated in the case of unclear new entries and will link records where necessary. In spite of these extensive security arrangements, there is still the possibility that non-identical patients will be linked. The probability is estimated at 1 in 500,000. Over the 23 years in which the WAMIS system has been routinely used, three such cases have been discovered. Even if there are one or two other cases that have not been discovered, the frequency of this type of error is still very low. Nevertheless, the systems should be equipped to rectify even these rare instances.

To enable WAMIS-users to find patients by LICs or MACs datasets with the linkages of LIC⇔IZAHL and MAC⇔working number are maintained. Based on these
datasets a study performed during the first year of the ARZ-IMC data transfer concerning the ratio of LICs to IZAHL produced the results shown in table 5:

**Table 5** LICs per IZAHL

Thus, 189 patients were registered four times in the KIS. Thanks to the good contacts and concentration of administrative data operators and through comparison with KIS data in the IMC records as well, the data quality in the KIS has improved markedly.

The conflicts occurring during consolidation have the disadvantage that they make extra work, but they have the advantage of improving the data quality. If a new patient can be linked with several known patients, for example, these records can often be consolidated into a single record. The documents for these patients can then be combined to form a single history as well.

The percentages of synonym and homonym errors in these data can be estimated when analysing the data which are transferred to the IMC. Regarding table 5 the percentage of synonym errors (all IZAHLs with more than one LIC) is approximately 5 %.

## 5. Discussion

In most of the described methods in the literature the linking process is rather unproblematic, because the data to be linked come from two or more static files. These files can be adapted as needed, the linking components can be analysed, sorted, transformed and adjusted. Ambiguous situations can be solved easily because there is no pressure of time. In some cases, one file including unique patients has to be linked to others including patients more than once, so that the structure of linkage is one-to-many [9]. The purposes of the linking exercises reach from local solutions, e.g. to identify any duplicate record within one file, to global solutions, e.g. to find a national unambiguous ID of one individual to achieve a common basis for the exchange of patient’s data. Depending on these purposes, linking errors can take effect immediately, they can have a very lasting effect, or they are never detected.

The Vienna situation is characterised by several peculiarities. A very difficult problem to deal with lies in the permanent flow of data. Personal data come online round-the-clock and must be identified immediately as described. Successful identification is the prerequisite for the correct storage of subsequently obtained findings. Findings of buffered patients must be buffered, too, and are missing in scientific studies. To improve this unfavorable situation it is intended to store these findings with a preliminary ID and to correct this ID later when the linkage is clear.
Due to the rather inscrutable status of the KIS database and a comparatively small number of supposed synonym and homonym errors in the WAMIS database the structure of the linking process is many-to-many. Therefore, rejected linkage causes a very dynamic process between both concerned systems. Before the data can be linked it is necessary to find out which database has caused the error. If the error originates from the KIS database, the data have to be corrected and sent again. If the error originates from the WAMIS database, the data have to be fetched from the buffer and the information about the error has to be fed back to the KIS as a precaution.

Personal data are entered into the KIS when the patient is present at the counter and they are transferred a few seconds after to the IMC. Therefore, in most cases the patient can help when problems occur.

With the increasing significance of the European Union, one of the desired additions to the registration procedure is an improvement in the phoneticization algorithm to take account of the various patient nationalities [16].

Although the importance of unique patient identification is uncontested, political decisions and ethical considerations (e.g. people cannot be tattooed with an ID immediately after birth, by force) also have to be taken into account in practice.

The proposed solution has been successful on the whole, although there have been a certain number of implementation problems:

- As with all similar systems, the operator was given the responsibility for patient identity/non-identity. This was justified by pointing out that the operator, by asking the patient directly, is in the best position to make the correct decision. Unfortunately, not all operators are sufficiently motivated to carry out this decisive step properly. Poor results have been achieved in cases where operators are administrative rather than medical personnel, since the consolidation of patient admission records is only of secondary importance for the administration. Training and background information could help to improve the situation with the operators involved.

- The components of the KIS make it too easy to decide that patients are non-identical, as the main purpose of the KIS is to assist the administration. For this reason an appreciable percentage of patients have several different IDs.

- Random sampling showed that patients with one IZAHL frequently had several (up to eight) different LICs (see Table 5).
One of the long-term health policy aims should be to ensure that all medical data for a patient, irrespective of when and where they were collected, can be consolidated to form a complete medical history that will accompany the patient throughout his/her life.

References


[16] Medical Informatics: Person name data definitions to be utilised within electronic data interchange. CEN/TC251/WG3 N302. 1996.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Length</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Province</td>
<td>1</td>
<td>'9'</td>
</tr>
<tr>
<td>Hospital</td>
<td>2</td>
<td>'01'</td>
</tr>
<tr>
<td>Cost centre</td>
<td>9</td>
<td>'SURG10400'</td>
</tr>
<tr>
<td>Year</td>
<td>2</td>
<td>'99'</td>
</tr>
<tr>
<td>Number</td>
<td>6</td>
<td>Sequential</td>
</tr>
</tbody>
</table>
**Table 2** Structure of the IZAHL

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Length</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name at birth</td>
<td>6</td>
<td>First six letters</td>
</tr>
<tr>
<td>Date of birth</td>
<td>6</td>
<td>DDMMYY</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>M/F/U</td>
</tr>
<tr>
<td>Number</td>
<td>2</td>
<td>Sequential</td>
</tr>
</tbody>
</table>
**Table 3** Actions depending on LIC-IZAHL constellations

<table>
<thead>
<tr>
<th>LIC</th>
<th>IZAHL</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>New</td>
<td>New</td>
<td>New admission</td>
</tr>
<tr>
<td>New</td>
<td>Known</td>
<td>Consolidation after verification</td>
</tr>
<tr>
<td>Known</td>
<td>Known</td>
<td>Update after verification</td>
</tr>
<tr>
<td>Known</td>
<td>New</td>
<td>Error</td>
</tr>
</tbody>
</table>
Table 4 IZAHIL-fragments matching

<table>
<thead>
<tr>
<th>Fragments matching</th>
<th>Concordance in subordinate fields</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>New patient (sequence = 1)</td>
</tr>
<tr>
<td>1</td>
<td>Yes</td>
<td>Existing patient found</td>
</tr>
<tr>
<td>1</td>
<td>No</td>
<td>New patient (sequence + 1)</td>
</tr>
<tr>
<td>&gt;1</td>
<td>Yes – Unique</td>
<td>Existing patient found</td>
</tr>
<tr>
<td>&gt;1</td>
<td>Yes - Ambiguous</td>
<td>Assigning manually</td>
</tr>
<tr>
<td>&gt;1</td>
<td>No</td>
<td>New patient (sequence + 1)</td>
</tr>
</tbody>
</table>
Table 5 LICs per IZAHL

<table>
<thead>
<tr>
<th>LICs</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>413.267</td>
</tr>
<tr>
<td>2</td>
<td>18.107</td>
</tr>
<tr>
<td>3</td>
<td>1.533</td>
</tr>
<tr>
<td>4</td>
<td>189</td>
</tr>
<tr>
<td>5</td>
<td>33</td>
</tr>
<tr>
<td>&gt;5</td>
<td>13</td>
</tr>
</tbody>
</table>
Data from KIS

sufficient

yes

information

new

check

MAC
	new

no

admission

no

information

on buffer

known

comply

yes

with earlier dis-
patches

fetch from

update

buffer

store to

buffer

no

new

check

LIC

known
Figure 1  Flowchart of the linking algorithm