DETERMINISTIC RECORD LINKAGE OF HEALTH DATA AS PREPARATORY WORK IN MODELLING AND SIMULATION – USE CASE: HOSPITALIZATIONS IN AUSTRIA

Barbara Glock[(a)], Florian Endel[(b)], Gottfried Endel[(c)]

[(a)]dwh simulation services, Neustiftgasse 57-59, 1070 Vienna, Austria
[(b)]TU Wien, Institute for Analysis and Scientific Computing, Wiedner Hauptstrasse 8-10, 1040 Vienna, Austria
[(c)]Main Association of Austrian Social Security Institutions, Kundmanngasse 21, 1031 Vienna, Austria
[(b)]dexhelpp, Decision Support for Health Policy Planning, Neustiftgasse 57-59, 1070 Vienna, Austria

[(a)]barbara.glock@dwh.at, [(b)]florian.endel@tuwien.ac.at, [(c)]gottfried.endel@hvb.sozvers.at

ABSTRACT

Modelling and simulation as decision support in the health care sector often requires real world data. Complex models covering a variety of different areas within one model, for example outpatient and inpatient sector together, if treatment paths are examined, utilize different data sources. If those data sources are not linked, only point images are possible or the modeller has to define assumptions covering the gap of unlinked data. Therefore, a good record linkage allows more precise and reliable models and as a consequence better decision support. Within this paper a deterministic record linkage of two different data sources of the inpatient sector is proposed and tested. The results show a matching of 99.94% for initially 1.27 million data entries of one source. The linkage gives additional access to data from the outpatient sector. More information concerning a single patient is available, which can be utilized in different decision support models.

Keywords: record linkage, health data, modelling and simulation, decision support

1. INTRODUCTION

Models used for decision support in the health care sector are usually parameterized with real world data. Data sources range from aggregated data of publications, over raw data from studies to routinely collected data from insurance carriers, and others. This data usually is on patients and their specific problems within a specific context, for example a study on a specific disease, where information on prior diseases is not required. This is usually disease-centred. Models, especially within micro-based simulation methods as agent-based modelling, for example a model simulating general treatment chains of patients in the inpatient *and* outpatient sector researching the health care provision in specific regions, are, on the contrary, patient-centred according to the research question (Wurzer, Lorenz, Rößler, Hafner, Popper and Glock 2015). Here data on all of a patient's diseases is required. More information, longitudinal data like patient history or co-morbidities, concerning a patient's disease is needed. Data is needed in a individual-centred manner and not in a disease-centred manner; individuals are relevant. A linkage of different data sources opens up this necessary information. In (DuVall, Fraser, Rowe, Thomas and Mineau 2012) a case study, similar to the project presented within this paper, is described and they also argue for the necessity of a record linkage of different sources.

In this paper, routinely collected data from the inpatient sector, provided by the Main Association of Austrian Social Security Institutions and stored within the GapDRG database (see section 2 for details) is researched. Due to data privacy issues, routinely collected data of different sources is pseudonymized (e.g. MBDS minimum basic data set from the Federal Ministry of Health, lacking a personal identifier till the year 2015). This makes statistical analysis as well as modelling and simulation for decision support and health care planning very difficult. Data from insurance carriers (e.g. from the data source FoKo in Austria) is *event based*: whenever a hospital reports a new hospital admission or separation, a new data entry is generated, resulting in split *episodes*. To enable efficient, significant and quality assured data analysis (and further parameterization of models) for patient centred assertions, record linkage of these episodes is required.

A linkage for data of 2006 and 2007 hospital stays has already been implemented (Endel, Endel and Pfeffer 2012). New data for 2008 to 2011 is available from the insurance carrier of Lower Austria and the aim is to find a personal identifier for episodes provided by the Federal Ministry of Health (MBDS) based on linked episode-based events from insurance carriers (FoKo). The previously developed linkage routines cannot be applied to the new data any more (at least not a 100%), due to novel challenges that come with the new data and altered circumstances. But the basic algorithm of the previous linkage will be further developed, since it has been very successful in the past.

With this new linkage it is possible to access data for patients in the inpatient *and* outpatient sector together, so that information is available in a patient centred manner that makes models for decision support more reliable.

The paper is structured as following: section 2 gives an overview of the used data bases and the challenges that come with linking the data. Section 3 gives some information on the state of the art and describes the record linkage method. Section 4 presents the results and in section 5 conclusions are drawn.

2. DATA SOURCES TO BE LINKED

The GapDRG database - *General Approach for Patient-oriented Ambulant DRGs* - of the Main Association of Austrian Social Security Institutions (HVB) stores routinely collected data from different sources. Available data in general are as following:

- prescriptions,
- inpatient sector including diagnoses, treatments, duration of hospitalization, etc.,
- outpatient sector, including diagnoses, treatments, etc.,
- sick leaves including duration of sick leaves,
- data on medications,
- and others.

In GapDRG1 data is available for 2006 and 2007 for Austria. Here a linkage already has been conducted. In GapDRG2 data from 2008 to 2011 is available from the insurance carrier of Lower Austria.

The following two databases, both stored within GapDRG2, covering the inpatient sector from 2008 to 2011 for Lower Austria are being linked:

- FoKo (FOlgeKosten): data from insurance carriers
- MBDS (Minimum Basic Data Set): data from the Federal Minsitry of Health

A hospital reports admissions and discharges separately to those institutions, as presented in figure 1.

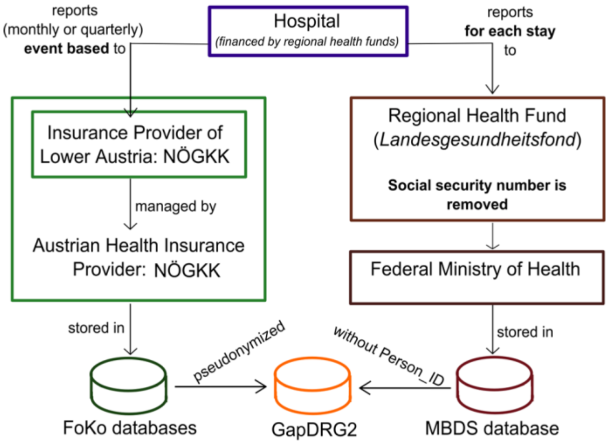

Figure 1: Databases to be Linked are FoKo (insurance carrier) and MBDS (Federal Ministry of Health).

Reports to the insurance carrier of Lower Austria are on a monthly or quarterly basis. This means that hospital stays may be split into so called *episodes*. On the other side, the hospital also reports every inpatient episode to the Regional Health Fund where the social security number is totally removed due to data privacy issues. Those data entries (for each stay) are then transferred to the Federal Ministry of Health. Here it has to be kept in mind, that it is not possible any more to determine if two hospital stays belong to one patient or not. Data from the insurance carrier of Lower Austria are then stored into the FoKo data bases. Here due to data privacy issues the data entries are pseudonymised, but it is still possible to determine if two hospital stays belong to the same patient or not.

FoKo and MBDS basically contain information on identical hospital stays with slightly different additional information. In MBDS further, more detailed information on the hospital episodes is available, like duration of the stay and additional to the main diagnoses up to four additional diagnoses, but also information on what procedures have been performed. But in MBDS there is no personal identifier. In FoKo on the other hand, there is a personal identifier, which enables further joining of the data to other data bases where the same identifier is used (sick leaves, medication, outpatient sector, etc).

Aim of this linkage is now to *find a unique person identifier for each episode in MBDS based on the information in FoKo.*

2.1. Challenges

The data linkage of course comes with some old and new challenges (Breitenecker, Urach, Miksch, Popper and Weisser 2011), that will be met within the proposed linkage:

1. **Reporting**: As mentioned before the hospital reports not only at the end of a hospital stay to the insurance carrier, but also during a stay. So in FoKo every time a new data entry is generated, as can be seen in figure 2 (green barrens).

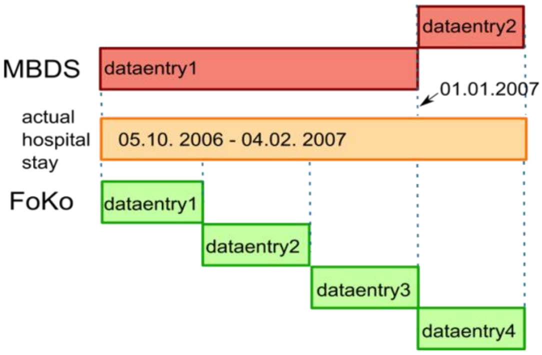

Figure 2: A Fictive example of Split Episodes in FoKo and in MBDS of One Hospital Stay.

A hospital stay, lasting over the boarder of a quarter/month of the year may be split into single episodes. In this case it is possible that, on the one hand, due to errors when entering the data those entries don't point to the same hospital stay anymore and, on the other hand, that there are much more data entries (split episodes) to be matched to one data entry in MBDS. In MBDS hospital stays may also be split if the stay lasts over the boarder of the year or due to transfers.

2. **Amount of data**: In FoKo only data from the insurance carrier of Lower Austria is available and in MBDS data from other insurance carriers are included.
3. **Diagnoses**: Due to the fact that hospital stays may be split within FoKo (resp. MBDS), the recorded diagnoses also may differ for the split episodes (pointing at the same hospital stay). Furthermore, the recorded and reported diagnoses may also differ from FoKo to MBDS. Finally, the coding method itself for diagnoses in FoKo is slightly different to that in MBDS. For example the format of the ICD10 diagnoses: S82.1 versus S821.
4. **District**: A person may have one or more districts at a specific time. In FoKo the district recorded by the insurance carrier is used and in MBDS the district reported by the patient directly within the hospital stay is used. Later on the variable district, as it turns out, is also the most unreliable one.

3. RECORD LINKAGE

3.1. State of the Art

In the health care sector various record linkage methods exist and most of them are developed based on the existing data and their structure, as it is done with the linkage method in this paper. In (Silveira and Artmann 2009) a systematic review on the quality of probabilistic record linkage projects and methods is done and the paper shows that methods especially developed for existing data provide very good results.

In the very recent report by (Samhar 2017) challenges faced with semantic and syntactic interoperability of linking *event based data* as it is also done within this paper, are described. This shows once more that actual work and new methods are being developed and needed right now.

3.2. Linking Method

In GapDRG1 a record linkage has been developed for data from 2006 and 2007. The previous method will be slightly adapted and the new challenges, described in section 2.1 will be met and resolved. The record linkage basically consists of the following steps:

1. **Restrict both data bases to data of insurance carrier of Lower Austria:** In MBDS there is one variable *herkunft* that represents insurance carrier Lower Austria and in FoKo this variable is *leivtr*. Both will be restricted.
2. **Cleaning of data in FoKo:** a check is applied if data entries exist that are exactly the same, except for the artificial unique identifier. Those duplicates are eliminated.
3. **Define Matching Variables (MVs)**: After an analysis of the variables in FoKo and MBDS, some can be selected as so called matching variables (variables that represent the same information in both data bases). Those variables then have to be cleaned (also see section 2.1) for structural differences (Breitenecker, Urach, Miksch, Popper and Weisser 2011). See section 3.1. for details.
4. **Base Match**: Data entries in MBDS are unique due to the variable triple *hospital, year of stay* and *episode number* that are also matching variables. A base match, checking if a unique patient ID from FoKo exists for such a triple, is applied and assigns this patient ID to each unique episode in MBDS. See section 3.2. for details.
5. **Tests and Quality Checks**: After the Base Match some tests are applied and the remaining MVs are checked if they match as well. Based on those tests the iterative process starts where matching variables are varied. The order of these variations is derived from the tests here. For details see section 3.4.
6. **Iterative Matching Process with MVs**: the iterative process using all MVs with different matching conditions is applied. See section 3.4 for more details.

After steps 1 and 2 the remaining entries to be matched are as following:

- FoKo: 1,410,165 data entries (episodes)
- MBDS: 1,272,813 data entries (episodes)

All tests and matches are done with SQL. For the iterative matching process a lot of SQL queries are needed, especially for the combinations within the level matches. For this circumstance a automatically SQL script is generated with MATLAB by using string manipulation.

3.3. Matching Variables

In table 1 the ten identified MVs are described.

3.4. Base Match

The base match is very simple approach: if a unique patient ID in FoKo exists for the triple *hospital, episode number* and *year of stay*, this patient ID is assigned to the according (exact match of those three variables) episode in MBDS (here only one episode exists, since this triple is the primary key and identifies the data entry uniquely). A match in the other MVs is disregarded, since these three variables are the most trustworthy ones.

Table 1: The 10 Matching Variables identified in both data bases. Bold and underlined variables represent the unique identifier of MBDS data entries.

Variable	MBDS	FoKo
year of stay	is given	year of date-variable *end of stay*
begin of stay	is given	is given, but episode may be split
end of stay	is given	is given, but episode may be split
hospital	is given	is given, but 56,330 entries are missing
episode number	is given	is given, but 26,453 entries are missing
diagnosis	main and additional diagnoses are given, but only main diagnoses are used	is given, but 3 data entries are missing, others may differ from MBDS (see section 2.1, challenge 3)
age	age at discharge is given and may be inaccurate, therefore additional conditions of consistency +/- 1 year are allowed.	*person_id* is given and age can be retrieved from another data base by using this ID and variable on *birth year*. Calculation of *age* by using *birth year* and *year of hospital stay* is possible.
gender	gender is given, but may be inaccurate. "M" for male and "W" for female.	*person_id* is given and gender can be retrieved from another data base by using this ID and variable on *gender*
death	year of *end of stay* together with being 'S' for death is used.	year of death is given
district	is given	*person_id* is given and district can be retrieved from another data base by using this ID in accordance of insurance carrier and time of insurance.

In the base match 611.591 episodes in MBDS can be matched (48.05%).

3.5. Tests and Quality Checks of Base Match

Since in the base match the other matching variables have been disregarded it is of course interesting to know the degree of consistency anyway. Consistency checks are applied directly with SQL by checking if those data entries are equal (using "="). Results can be seen in table 2.

Table 2: Degree of Consistency in other MVs after the Base Match.

Matching Variables	Degree of Consistency	
	Total	In %
Begin of stay	611,489	99,98%
End of stay	611,075	99,92%
Diagnosis	579,180	94,70%
Age	315,758	51,63%
Age +/-1 OR exact match	611,124	99,92%
Gender	610,714	99,86%
District	5,657	0,92%
death	10,614	--

Consistency in *begin of stay* and *end of stay* (exact date) is very good, as well as *gender* and also *diagnoses*. The check of consistency for *age* (age at discharge in MBDS compared to calculated age in FoKo based on *end of stay* and *birth year*) is not so good. The check of consistency with +/- 1 year due to calculation inaccuracies on the other hand gives a very good degree of consistency.

In the iterative matching process (see section 3.4) matching variables are varied (all variables have to match except 1, 2, 3...) and the remaining variables are checked for consistency. This results in a huge amount of combinations of MVs that are left disregarded within the different levels. The order in which the MVs are varied within the matching process is retrieved from exactly these tests (reverse order of matching qualities), meaning that for *example begin of stay* is left disregarded in the variations at the end due to its good matching "qualities" in these tests.

3.6. Iterative Matching Process

After the base match the iterative record linkage starts with the remaining data entries in FoKo and MBDS. Here all matching variables are used together: First there are the so called *level matches* up to level 6, each of them consisting of two so called *steps*.

In each step/level whenever matches are found the next step/level is conducted with the remaining data entries of both data bases. The found matches are removed from the further matching process. When level 6 is completed, the process starts at level 1 again with the remaining data entries. This is called iteration. A schematic representation of the level matches and iterations can be seen in figure 3.

The red numbers show the time line of the linkage process starting with 1 at the base match (orange rectangle), 2 for storing the found matches of the base match and 3 transferring the remaining - still to be matched - data entries of FoKo and MBDS. Then the level matches start with level 0, meaning consistency – being equal - in all MVs. Here again the match is conducted (red 4), found matches – if person identifier is unique – stored into “Matches” (red 5) and remaining data entries used for Level 1 (red 6).

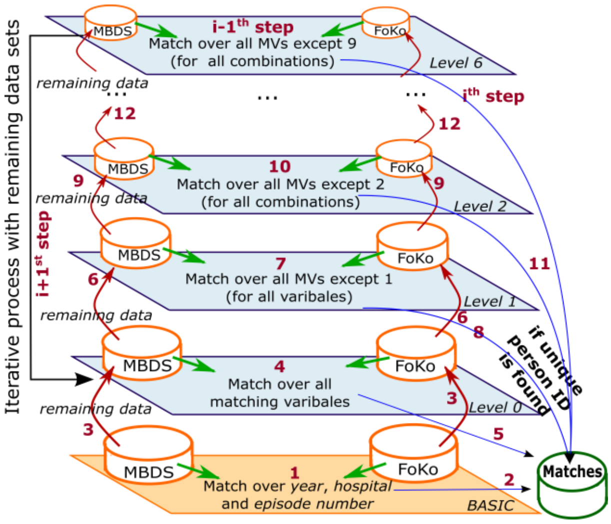

Figure 3: Schematic Representation of the Record Linkage with its Level Matches and Iterations after the Base Match.

Level 1 means consistency in all MVs except 1. Here “except 1” means that the MV(s) are varied, so all of the MVs are disregarded once. The order, which variable is left outside is in reverse order as the degree of consistency after the base match (see table 2).

Level 2 means that all except 2 MVs have to match, and so on. There are 10 MVs and as a consequence the number of variations in level 1 is 10 (each of the MVs is left disregarded once). In level 2 at each step 2 MVs are left disregarded, resulting in 45 variations (combination of the 10 MVs). In level 3 there are 120 variations. As a result – programming those variations by hand would be very time intensive – the whole record linkage is semi-automatic (SQL Code that is partially constructed with MATLAB).

Furthermore, within the level matches there are two steps (not shown in figure 3) concerning the excluded MV(s) – of course these two steps are applied in levels 1 to 6 and not in level 0:

- Step 1: allows missing data, meaning that at least one (either in FoKo or in MBDS) MV within the MV(s) that are excluded is NULL (missing).
- Step 2: allows missing data OR inconsistency, meaning either one or both entries is NULL or contradicting (missing or contradicting).

Example: Level 1/Step 1: a match is included if gender in FoKo is “M”, gender in MBDS is NULL and all other MVs are the same. Level 1/Step 2: a match is included if gender in FoKo is “M”, gender in MBDS is “W” and all other MVs are the same.

Step 2 guarantees that errors in reporting as mentioned within the challenges in section 2.1 are eliminated. In the first step only checks with missing data are permitted, and only then in the second step contradicting values are permitted, because in step 1 no one knows if it would be a match if the entry is available and in step 2 it is for sure that it is contradicting (or an error in reporting). Here again the order of the single parts within the matching process is relevant to guarantee the best possible outcome of the matches.

This is done until level 6. Further levels, like level 7, would be very unreliable, meaning that it is a match except in 7 (out of 10) MVs. After level 6 the next iteration starts at level 1 again with the remaining data entries in FoKo and MBDS until no entries are left or until no matching entries can be found any more. This procedure guarantees the best possible matches due to the chronology of the different matches.

4. RESULTS

The results of the record linkage can be seen in table 3. Most matches (after the base match) have been found whenever the *episode number* was excluded. This is a rather long number where errors may be more likely.

Table 3: Results of the Record Linkage Together with Information in which Level and Iteration it was found.

Iteration / Level	In Foko remaining	In MBDS remaining	Match	Match total in %
Start	1,410,165	1,272,813	--	--
Base Match	794,294	661,222	611,591	48.05%
1 /1	794,256	661,184	38	48.05%
1 / 2	780,741	647,656	13,528	49.11%
1 / 3	232,976	99,767	547,889	92.16%
1 / 4	189,629	18,262	81,505	98.56%
1 / 5	198,599	18,235	27	98.56%
1 / 6	184,342	1,418	16,817	99.88%
2 / 3	184,310	1,384	34	99.89%
2 / 4	184,285	1,359	25	99.89%
2 / 5	183,764	830	529	99.93%
2 / 6	183,654	713	117	99.94%
3 / 4	183,653	712	1	99.94%
3 / 5	**183,650**	**709**	**3**	**99.94%**

After level 3 (3 MVs have been excluded) most of the remaining matches have been found (92.16%) and that is due to the fact that the previous mentioned triple of *hospital*, *episode number* and *year of stay* have been excluded (Step 1 or Step2, NULL or contradicting). After 3 *iterations* no significant number of matches can be found (in iteration 3 only 4 additional matches occur) and the record linkage stopped with a total of 99.94% assigned patients to hospital episodes in MBDS, leaving 709 data entries in MBDS without personal identifier. This *deterministic* record linkage, if being applied to the same data again, produces the same results as shown in table 3. Within the proposed record linkage method the sequence of the single steps of the matching is important, because it guarantees that in each step the best possible outcome is produced.

A huge benefit and further development of this new linkage process, compared to the linkage in GapDRG1, is the fact that more information concerning the level and iteration of the found match is available: an additional data entry to each match provided as a string shows the number of iteration and number of level it was found in. This means that the modeller or data analyst can decide (on its own or based on the research question) which match will be accepted in further analysis or for parameterization of simulation models. More information on which specific matching variable was excluded in the level/iteration while the match was found can – hypothetically – be obtained as well, if necessary.

Another result of this linkage is the fact, that, if more data of the same structure, for example for the whole of Austria is provided, the linkage method can be applied quickly. Finally, with this matching where for (nearly) each episode in the inpatient sector a personal identifier was found, other information for this person is available, like prescription data, sick leaves, services in the outpatient sector, etc (see section 2 on data in the GapDRG database). Now, for example, the prevalence of various diseases can be calculated by using inpatient and outpatient data together reducing uncertainties by using only one data source. This is important in modelling and simulation, because less assumptions have to be made.

5. CONCLUSION

Using good data for parameterization in modelling and simulation is essential to designing a model, especially when it is used in such a sensitive area as decision support in the health care sector. Different data sources give different point images of the current or past situation and therefore gaining more information (linked information) is nearly impossible. Usually a lot of assumptions are made, which makes the model unreliable. In this paper, a deterministic record linkage of two different data sources concerning hospitalizations is further developed and tested. With these linked data sources it is now possible to retrieve much more information than just on hospitalizations, as it is now possible to add information from outpatient care, sick leaves or prescriptions. As a result, for example, modelling of whole treatment chains is possible or modelling the health care provision in different regions of Austria.

A record linkage of historic data within the database GapDRG from the Main Association of Social Security Institutions has been performed before. The proposed improved procedure met new challenges like different sizes of data sources (Lower Austria vs. Austria) or record errors due to split episodes or simple syntactic errors (diagnoses) which have been addressed. The main innovations of this procedure include a significant improvement of previously developed methods, mainly concerning reproducibility, stability and adaptability to new data and a documentation on every single step of the linkage procedure, allowing researchers to comprehend the origin of a link and adapt their data analysis strategies.

The procedure achieves the best possible outcome for the new data sets and is highly suitable to be used within new data in a semi-automatic way, which also enables new simulations faster. In (Bohensky 2010) a systematic review is done and they conclude that an incomplete record linkage is very problematic for further analysis. The record linkage proposed within this paper is nearly complete and therefore suitable for further use for parameterization of simulation models.

ACKNOWLEDGMENTS

The K-Project Dexhelpp as part of COMET – Competence Centres for Excellent Technologies - is sponsored by BMVIT, BMFW and the city of Vienna. The COMET project is organized by FFG.

REFERENCES

Bohensky M. A., 2010. Data linkage: a powerful research tool with potential problems. BMC Health Serv Res 10, 346.

Breitenecker F., Miksch F., Popper N., Urach C., Weisser A., 2011. Dealing with Health Care Data of the Austrian Social Security System. Medical Decision Making, 31 (1), E51.

DuVall S. L., Fraser A. M., Rowe K., Thomas A., Mineau G. P., 2012. Evaluation of record linkage between a large healthcare provider and the Utah Population Database. J Am Med Inform Assoc 19, e54-59.

Endel F., Endel G., Pfeffer N., 2012. Routine Data in HTA: Record Linkage in Austria’s GAP-DRG Database. Poster at ISPOR 15th Annual European Congress. November 3-7. Berlin, Germany.

Samhar M., 2017. The ‘PEARL’ Data Warehouse: Initial Challenges Faced with Semantic and Syntactic Interoperability. Studies in Health Technology and Informatics pp. 156–160. doi:10.3233/978-1-61499-753-5-156

Silveira D. P., Artmann E., 2009. Accuracy of probabilistic record linkage applied to health databases: systematic review. Rev Saude Publica 43, pp. 875–882.

Wurzer G., Lorenz W., Rössler M., Hafner I., Popper N., Glock B., 2015. ((MODYPLAN)) – Early-Stage Hospital Simulation with Emphasis on Cross-Clinical Treatment Chains. Proceedings of the Symposium on Simulation for Architecture and Urban Design, pp. 97-100. April 12 – 15, Washington D.C. (USA).