COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 106 (2012) 127–138

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

EPILEPSIAE – A European epilepsy database

Matthias Ihle[a,b,*], Hinnerk Feldwisch-Drentrup[a,b,c,d], César A. Teixeira[e], Adrien Witon[f], Björn Schelter[b,g], Jens Timmer[b,g,h], Andreas Schulze-Bonhage[a,c]

[a] Epilepsy Center, University Hospital Freiburg, Germany
[b] Center for Data Analysis and Modeling (FDM), University of Freiburg, Germany
[c] Bernstein Center for Computational Neuroscience, University of Freiburg, Germany
[d] Department of Neurobiology and Biophysics, Faculty of Biology, University of Freiburg, Germany
[e] Centre for Informatics and Systems (CISUC), University of Coimbra, Portugal
[f] Centre de Recherche de l'Institut du Cerveau et de la Moelle épinière (CRICM), INSERM UMRS 975 – CNRS UMR 7225-UPMC, Hôpital de la Pitié-Salpêtrière, Paris, France
[g] Department of Physics, University of Freiburg, Germany
[h] Freiburg Institute for Advanced Studies, University of Freiburg, Germany

ARTICLE INFO

Article history:
Received 1 February 2010
Received in revised form 6 July 2010
Accepted 16 August 2010

Keywords:
Database
Schema
Epilepsy
Seizure prediction
EEG

ABSTRACT

With a worldwide prevalence of about 1%, epilepsy is one of the most common serious brain diseases with profound physical, psychological and, social consequences. Characteristic symptoms are seizures caused by abnormally synchronized neuronal activity that can lead to temporary impairments of motor functions, perception, speech, memory or, consciousness.

The possibility to predict the occurrence of epileptic seizures by monitoring the electroencephalographic activity (EEG) is considered one of the most promising options to establish new therapeutic strategies for the considerable fraction of patients with currently insufficiently controlled seizures.

Here, a database is presented which is part of an EU-funded project "EPILEPSIAE" aiming at the development of seizure prediction algorithms which can monitor the EEG for seizure precursors. High-quality, long-term continuous EEG data, enriched with clinical metadata, which so far have not been available, are managed in this database as a joint effort of epilepsy centers in Portugal (Coimbra), France (Paris) and Germany (Freiburg).

The architecture and the underlying schema are here reported for this database. It was designed for an efficient organization, access and search of the data of 300 epilepsy patients, including high quality long-term EEG recordings, obtained with scalp and intracranial electrodes, as well as derived features and supplementary clinical and imaging data. The organization of this European database will allow for accessibility by a wide spectrum of research groups and may serve as a model for similar databases planned for the future.

* *Corresponding author at*: Epilepsy Center, University Hospital Freiburg, Breisacher Str. 64, 79106 Freiburg, Germany. Tel.: +49 761 270 9313.
E-mail address: matthias.ihle@uniklinik-freiburg.de (M. Ihle).
0169-2607/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2010.08.011

1. Introduction

Epilepsy is one of the most common diseases of the human brain, with a prevalence of more than 3 million patients in Europe alone. Epilepsy is characterized by sudden changes in brain dynamics that lead to abnormal synchronization of extended brain networks, the so-called "seizures". These seizures are characterized by transient impairments of sensation, thinking and motor control. In most patients, seizures are infrequent, occupying much less than 0.1% of the time. Due to their apparently unpredictable occurrence, patients are, however, suffering from restrictions in several domains, e.g., physically due to the risk of trauma, socially due to driving and occupational restrictions, psychologically due to a feeling of helplessness [1,2]. Furthermore, a continuous prophylactic medical treatment is presently being offered that renders patients liable to side effects [3].

One third of the patients presently do not respond to a continuous prophylactic treatment in maximally tolerated dosages. Particularly for this large patient group, new treatment concepts have to be developed. Such concepts could be based on prediction-based warnings to patients or on prediction-based closed-loop interventions. These would change continuous treatment to timely targeted medical or electrical interventions prior to seizures [4].

For this purpose, the EU-funded project "EPILEPSIAE" (www.epilepsiae.eu) aims at the development of EEG-based seizure prediction algorithms that capture changes in electroencephalography (EEG) dynamics and use these for a warning of the patient. Up to now, visual inspection did not reveal precursors for seizures. So far, one of the factors limiting the evaluation of prediction algorithms have been limitations in the quality and duration of long-term EEG data available for a valid evaluation of seizure prediction methods and their performances [5,6], and limited access of groups with knowledge in the field of time series analysis to such EEG data.

Accordingly, both in Europe and in the USA efforts are made to overcome this obstacle for the development of seizure prediction. The EPILEPSIAE project will gather the largest and most comprehensive epilepsy database existing worldwide. It is based on the common effort of three European epilepsy centers (Freiburg, Germany; Paris, France; and Coimbra, Portugal), which contribute EEG data from long-term monitoring of epilepsy patients as well as standardized annotations and clinical metadata. After its completion, it is planned that this database project will offer access to data for research groups throughout the world and will collaborate with a US database that is presently being designed.

We here report on the design and the resulting schema of the database. Additionally, we present client applications and the current status regarding content: the input and management of large data sets.

2. Background

Seizure prediction is based on the identification of precursors in long-term EEG time series. A major concern is access to suitable data for the systematic application and evaluation of seizure prediction methods. Particularly for research groups with no direct contact to epilepsy centers, access to clinical data often imposes severe constraints for the progress of research. This has been pinpointed already in 2001 [7] when the importance of freely available datasets was emphasized: "One recent initiative, the development of open databases that could serve as a repository of clinical data that are difficult and expensive to obtain, provides an emerging strategy that should prove indispensable for testing competing algorithms". This demand for comprehensive databases that can be used for a cross comparison of methods developed on the basis of different, publicly available data sources has led to the publication of some EEG dataset collections in the prediction community, which can be freely accessed.

The Bonn EEG database [8,9] consists of data of five subjects, including different channels per patient, recorded with a sampling rate of 173 Hz. Datasets are discontinuous. For each dataset, about 40 min of EEG are provided.

The Flint Hills Scientific, L.L.C., Public ECoG Database [10] (supported by NIH/NINDS Grant No. 3R01NS046602-03S1) consists of a total of 1419 h of continuous intracranial recordings at 249 Hz for ten patients. Additionally, the database contains meta information about the 59 contained seizures and information about the electrode locations. It provides the EEG from all recording electrodes, which range from 48 to 64 per patient.

The Freiburg EEG database [11] contains invasive long-term EEG recordings of 21 patients, acquired with a sampling rate of 256 Hz obtained during invasive pre-surgical epilepsy monitoring at the Epilepsy Center of the University Hospital of Freiburg, Germany. For each patient, the recordings of three focal and three extra-focal electrode contacts are available. In contrast to the other databases, here, a clear separation between ictal and interictal phases for each of the patients is given. For ictal events, files with epileptic seizures and at least 50 min pre-ictal data are provided; the interictal data contain at least 24 h of EEG recordings without seizure activity.

Although this database ranks among the most comprehensive of the currently available EEG databases and is used by more than 180 research groups worldwide, it is clear that there are still drawbacks: first, there is the general lack of long term, continuous recordings. Second, there is only limited information on clinical metadata and annotations.

Overall, all these databases consist of unstructured EEG recordings, supplemented by some clinical information. For example, the correlation between seizures, their origin and propagation and, the localization of electrodes or, other exact information about the brain topography or the epilepsy characteristics are not available when using the data of these databases.

As provider of the Freiburg database, we collect information about the purposes of the database users. The use for research on seizure prediction was most frequently named, followed by seizure detection. But it became apparent that the application domain is much wider. It ranges from detection and analysis of interictal spikes over the automatic classification of EEG signals, for instance through machine learning, to the general application of time series techniques.

Additionally, various direct user requests prove the general demand for metadata about the EEG in addition to the raw recording data, like the type and localization of the

epilepsy, the electrodes, various details about the seizures, sleep stages, etc. Not infrequently surface EEG recordings have been requested, whereas all current databases only contain recordings from intracranial electrodes.

There is a clear need for more comprehensive, generalized epilepsy databases in excess of mere EEG dataset collections.

This lessons learned as provider as well as user of EEG databases and from scientific studies performed on EEG data were integrated into the currently emerging EPILEPSIAE database. From the start, it had the goal of fulfilling the demand for such a comprehensive epilepsy database with long term, continuous EEG data from surface and intracranial recordings enriched with supplemental information about the EEG such as electrodes, seizures, their semiology and events during the non-ictal phases. To serve also as a special epilepsy database, in contrast to a pure EEG database, it was decided to include extensive information about clinical data including results of pre-surgical evaluations, imaging data, potential surgeries and outcome as well.

In order to facilitate the evaluation of a multitude of features, i.e., time series derived from the EEG for the purpose of seizure prediction, it was decided to integrate also these features into the database along with information about the underlying algorithm, the execution parameters and channel dependencies. The rationale behind this decision was that, depending on the algorithm and the size of the input, computational power needed can be high and calculation time can comprise weeks per feature. Since these calculations are often based on each other, such a caching of results saves a lot of time and resources. This is particularly important as we could show [12] that the combination of different feature algorithms can considerably improve prediction performance as compared to the usage of just a single feature.

3. Design considerations

Since large EEG data sets are important for the statistical validity of prediction results, one of the most important goals of the project was to constitute the largest pool of data, by far surpassing currently available sources. The project's plan schedules a database content of 300 datasets, thereof 50 with invasive electrodes, within a three-years period. The database should be open to additional datasets integrated after the project duration, possibly also including data from experimental epilepsy models.

Having a database with such a high quantity of data raises the question of data access. Although for relational databases SQL [13] is the standard, we additionally need to provide a user-friendly way for the systematical selection of datasets from the database without knowledge of either SQL or details of the database schema. Therefore, a graphical client interface with input masks for selected queries and the possibility for general purpose SQL queries was desirable. Additionally, the access to the raw EEG data, being stored either in database tables or in flat files, was targeted to be carried out from inside the database.

Not only high data quantity, but also high data quality was a fundamental consideration during the design of the database. Not all of the recordings conducted at the hospitals are equally suited. Therefore, criteria were defined that datasets have to fulfill in order to be included into the database: each dataset must include a continuous recording time of at least 96 h (4 days), contain at least five clinically manifest epileptic seizures, and there must be at least five seizures with interictal intervals of at least 4 h between each other. Subclinical electrographic events were not considered separately. Moreover, standardized annotations for EEG evaluation had to be defined to assure identical judgement of onset times, patterns or propagations within the consortium of epilepsy centers where data were annotated.

Additionally, difficulties arising from the distributed nature of the joint effort had do be solved. Different EEG system setups had to be integrated depending on the needs and structures found at individual sites. Partners may have special requirements for the database schema and distinct ideas about the ideal database design.

The enormous amount of data processed poses problems for data transfer. Accordingly, a distributed architecture with each partner hosting only its own data but transparently accessing the partner's data via internet would be an elegant way, whereas replicated databases offer advantages regarding data access and security and are an option as data are relatively static.

Lastly, the general acceptance of the database in the research community is another important goal. It could serve as a model for newly evolving seizure prediction databases as the US database, and it could develop into a standard by attracting new partners after the project's initial term, further enlarging the pool of available data. Sharing expertise is thereby as important as providing access to the database to other research groups.

4. Database schema

The initial step for the design of a database is the conceptual design that defines and describes the excerpt of the real world that is of particular interest for the considered database application. The most commonly used conceptual model is the Entity-Relation (ER) model [14,15], which models the real world in form of entities, usually recognizable concepts, either concrete or abstract, and defines relationships between them. Furthermore, it associates the entities with describing attributes that have a type or domain defining the possible values an attribute may have.

Subsequently, this conceptual model is translated into the relational model [16] that is the basis of all relational databases, such as the Oracle database [17] used at all participating sites of the EPILEPSIAE project.

Thereby, the entities as well as the relationships are directly translated into relational tables. Depending on their complexity, relationships obtain separate tables or are attached to tables corresponding to entities. We here present the relational schema of the EPILEPSIAE database.

Since the schema is too comprehensive to be discussed and presented here in all detail, we will highlight only the most interesting design decisions while large parts of the schema will be listed in tables.

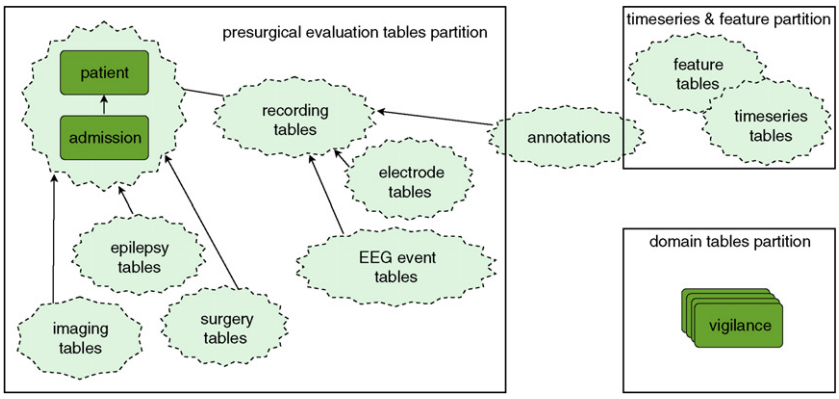

Fig. 1 – Partitions (depicted by the rectangles), table groups (clouds) and tables (green boxes) of the database schema. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

The general structure of the schema is summarized in Fig. 1. In this figure, tables are shown in form of green boxes, while the light green clouds represent groups of thematically related tables. On a higher level, the figure partitions the schema into three parts, indicated by the bigger framed boxes: (1) dataset tables holding all information about the patient's pre-surgical evaluations, (2) the feature and time series related tables and (3) the domain tables. While the former two parts, presented in the following two subsections, are internally hierarchically organized, the last generally contains independent tables.

A domain table represents an attribute domain and not an entity or relationship of the ER diagram like all other tables. The concerned domains cannot directly be mapped to one of the built-in data types of the database, such as the domain of seizure types. Usually these tables have two attributes: the id, referenced by other tables, and a value, containing a longer textual description of the domain value. For example, 'simple partial seizures' can be referenced through the id 'SP' in the domain table *seizure_type*.

Using separate tables instead of customized, but database-specific data types has the advantage of portability, because tables and references are available with every relational database. Since the contents of domain tables are static and not tied to a specific patient, they are not distributed alongside the dataset archives. It rather must be assured that they are available at every deployment of the EPILEPSIAE database.

4.1. *Pre-surgical evaluation datasets*

The largest partition in Fig. 1 is the one with the dataset tables containing all the data collected during the pre-surgical evaluation of patients. Such a pre-surgical evaluation has the goal to assess risks and benefits of an epileptic surgery for patients if other treatment methods failed. The comprehensive data about patient and epilepsy that is thereby collected makes such datasets very interesting for the purpose of seizure prediction research. In addition to the hitherto only relevant long-term, continuous EEG recordings with its annotated events, various examination results and imaging data may offer additional approaches to seizure prediction.

The dataset partition in Fig. 1 shows two tables directly: the patient and the admission table. Although the dataset tables and the references between them, strictly speaking, span a graph and not a tree, there is a hierarchical structure in the schema with the patient table at the root. Since the database is pseudonymized, the patient table contains an attribute for the pseudonym coding, gender and, age at which the epilepsy appeared the first time (*onsetAge*) for each patient.

The admission table is the only one referencing the patient table and acts as the actual point of reference for a pre-surgical evaluation, since all hereafter presented table groups are in direct or indirect relation to it. The admission table provides general information about the admission like the date, the patient's age, the hospital, if it was a pre-surgical evaluation (*presurgical*, boolean), if there was a subsequent surgery (*surgicalDecision*), and if surface electrodes (*sEEG*) respectively invasive electrodes (*iEEG*) were used.

Each dataset usually provides exactly one admission per patient, although, in some cases, there may be several admissions, each of them fulfilling the inclusion criteria separately. Such admissions do not count as separate datasets as they will have the same patient specific EEG characteristics. Thereby, each dataset corresponds to one record of the patient table and contains the records of all tables that are directly or indirectly related to this record in the patient table.

The remaining tables holding the data of pre-surgical evaluations are divided into the thematically related groups of epilepsy, imaging, surgery, recording, electrode, EEG event and, annotation related tables. The epilepsy related tables provide information about the patient's epilepsy characteristics and examination results, while imaging data and interpretations are recorded in the group of imaging tables. Information about potential surgeries and follow-up examinations are stored in the surgery table group. Tables 1–3 shortly survey the constituent tables of these groups.

In contrast to this tabular presentation, we describe in the following subsections the recording, electrode and, EEG event and annotation related table groups in detail and explain the fundamental design decisions/considerations.

Table 1 – This group of tables contains information resulting from the examinations of the patient conducted during the admission that has the goal to evaluate if the patient is suitable for treatment with surgery.

Table	Reference	Description
Etiology	Admission (1:1)	Holds information about the cause of the epilepsy. Several common causations, e.g., hippocampus sclerosis, an inflammation or a tumor, are explicitly listed as booleans. Additionally, there is a text field in case the etiology is not contained in the list
Cognitivefunction	Admission (1:1)	Aggregate the results of potential neuropsychological examinations into six attributes (attention, verbal respectively non verbal declarative memory, executive functions, language and visuospatial functions) attaching a result with a range of five different values from far below average to far above average. Additionally, the date of the examination and potentially a commentary are recorded
Seizuretypefrequency	Admission (1:n)	Usually, the frequency of seizures that the patient suffered from before the admission is determined by questioning the patient. This is stored in the seizureTypeFrequency table whereby for the different types (simple partial, complex partial, secondarily generalized) of seizures their number during the last six months before the admission is recorded
Complication	Admission (1:n)	Provides a free textual description of possible complications
Medication	Admission (1:n)	Associates a medicament from a given list stored in a domain table with a dosage, a date and, in case of a repetition, an end date
Eeg_focus	Admission (1:n)	Holds information about one or several possible epilepsy foci of the patient, thereby assigning a localization to each of the foci. Localizations are a reference to the domain table, aggregating a certain brain lobe (frontal, temporal, etc.), a subregion (basal, mesial, etc.), and lateralisation

Table 2 – This group of tables contains information about a surgery that is possibly performed if the prospect of success is given after the pre-surgical evaluation and the patient consents.

Table	Reference	Entity	Description
Surgery	Admission (1:n)	Epilepsy surgery	General information about the surgery like the date or the type of the surgery
Histology	Surgery (1:1)	Histology of the resected specimen	Holds detailed information about the surgery's histology
Surgerycomplication	Surgery (1:1)	Complication during surgery	Holds boolean fields for prevalent complications like infections or bleedings
Surgerylocalization	Surgery (1:*n*), localisation (*n*:1)	Localization of surgical intervention	Relates several localization to the surgery (the same as in the *eeg_focus* table)
Follow_up	Surgery (1:*n*)	Follow_up exams for surgery	Records all follow-up examinations for the surgery. It associates the date, respectively the interval between surgery and follow-up examinations with the outcome classification according to Engel [31]

Table 3 – Imaging tables.

Table	Reference	Description
MRI	Admission (1:1)	Interpretation of MRI images
MRI_files	MRI (*n*:1), files (1:1)	MRI files, depending on the used file format there may be separate files for each slice
SPECT	Admission (1:1)	Interpretation of SPECT images
PET	Admission (1:1)	Interpretation of PET images

4.1.1. *Recordings*

In theory, the EEG is recorded continuously during the whole admission of a patient. Practically, there is sometimes the need to restart such a recording. Reasons may be changes in the recording setup, like the addition of supplementary surface electrodes for an invasive recording, recording gaps due to other investigations or, technical problems of the recording system. Whether a new recording is started after a crash or the old one is continued with a gap depends on the EEG system software. There are different strategies for handling this among the EEG systems used by the project partners. The EEG system usually splits the recording into several blocks for storage on hard disk. Some of them cause system immanent gaps between two blocks, some cause these only in the case of a failure.

These considerations have lead to the basic structure that is shown in Fig. 2. Here, an admission is related to one or sev-

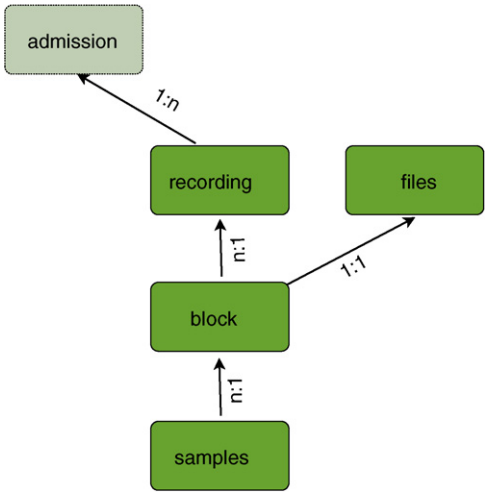

Fig. 2 – The recording related group of tables.

eral recordings, which, in turn, consist of several recording blocks.

Thereby, the recording table holds a reference to the respective admission and general information about the recording like a string identifier, timestamps for the beginning and the end, the number of contained blocks and the duration calculated from the aggregated length of the blocks minus the gaps between them. It also records the technical information that is common to all blocks like the number of channels and the sampling rate. Additionally, it contains some meta information about the EEG that does not remunerate a table of its own, e.g., the background rhythm of the recording and the localization of a slowing of activity in the EEG.

The block table contains all the block specific metadata in addition to the reference to the recording it belongs to. It has fields/attributes with information about start and end timestamps, the gap to the previous block, the block number as well as the number of channels, samples and bytes used for them (*sample_bytes*). Lastly, it holds the conversion factor that is needed to calculate the measured (micro)volt value out of the stored raw integer value for each sample.

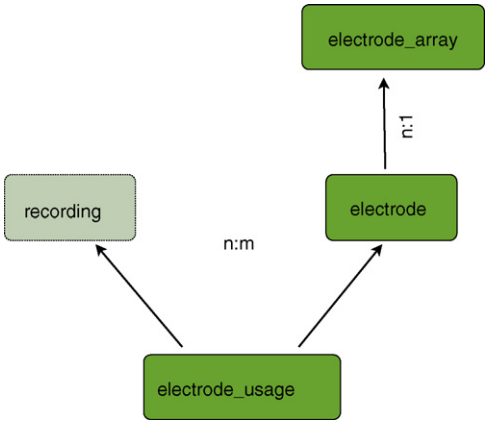

Fig. 3 – The group of electrode tables.

Additionally, if the block data is available in the local filesystem, the block table holds a reference to the files table that provides information about the storage of the block in the file system. Since records in this table refer to files on the hard disk, the tables' attributes correspond to properties of the respective file: besides fields containing the name, the path, the length and the file creation time, the md5-checksum [18] of the file is stored in the checksum field. The last field *locator* is Oracle specific and contains a BFILE blob locator [19,20]. The advantage of such a BFILE blob is that the content of the file can be accessed though the oracle database system, instead of retrieving the filename and path for accessing the file through the file system. A further remote access of the server in addition to the database server is then not necessary.

It would be impractical to use the native format of one of the individual EEG systems as format of such EEG files. We rather need a common format that can be read and written on all sites. Since none of the existing non system-specific file formats like EDF [21] and GDF [22] are accepted as a standard, and most of them are tailored to specific applications, it was decided to use a plain binary file format. This 'EPILEPSIAE binary data format' just contains the sample values and no additional header information. The information usually found in the header can be retrieved from the database. The file contains the multiplexed values of all channels for each sample. This is the most space efficient storage possibility. It simply adds up to the multiplication of the number of electrodes, samples and bytes per sample.

Since we have separated the general data about the recording block from the one about the storage of the block data, we can offer several possibilities for storing the block data. Beside the storage in files, a second storage possibility in the database is provided: the direct storage of sample values in a database table. This can happen either instead of the storage in files or in addition to it. In the latter case, this not necessarily needs to affect all samples. It may be restricted to the most interesting parts of the EEG, e.g., the seizures. It is mainly a question of storage space that depends heavily on the design of the sample table, e.g., if all channels of one sample are stored in one record or if we associate the sampling value of each channel with each new record. Another factor with great influence on the storage size is if raw integer values or the already normalized voltages in floats are stored in the binary files. These different possibilities have advantages and disadvantages concerning flexibility, storage requirements and speed, which are currently under evaluation.

4.1.2. *Electrodes*

Modeling electrodes is not a trivial task because there are several context dependent meanings for the notion of an electrode: an electrode contact measuring a voltage, an electrode as physical entity, or a channel in an EEG file representing the measured voltages. With respect to surface electrodes there is no clear distinction between the former two: each electrode has exactly one single electrode contact. But in case of invasive recordings, usually multicontact electrodes are used. Electrode names in recording files always refer to single electrode contacts. In the case that two files have a common electrode name (channel), it is not clear whether they refer to the same

physical electrode (contact). For instance, a re-implantation of invasive electrodes leads to a new electrode contact with the same name as the previous one.

In our schema, the electrode table (Fig. 3) corresponds to individual electrode contacts. It provides the fields *name* and *moniker* for the electrode name, in both the official ILAE nomenclature [23] and a potentially deviating internal name used in the EEG system. Additionally, it has a field indicating possible artifacts and a reference to a focus defined in the *EEG_focus* table.

In order to reduce the number of joins in queries, we use a common table for surface and invasive electrodes, with the Boolean attribute *invasive* indicating if the electrode is an invasive one. Hence, the table has three attributes that are only used for invasive electrodes, each of them holding one dimension of the 3D coordinates of the electrode localization in the brain. Therefore, we use the MNI coordinate system [24] and provide functions written in PL/SQL for calculating the widespread Talairach coordinates [25].

Lastly, it contains a reference to the table *electrode_array* that has the purpose to merge the contacts of invasive electrodes (the notion of a physical electrode with several contacts) as well as to cluster the more loosely coupled non-invasive electrodes into groups of related electrodes, e.g., the 10–20 system or all ECG electrodes.

Arrays have a name, a type, a configuration and, in case they are invasive, an implantation date. The name usually derives from the location of the array, e.g., 'TLA' in case of the first strip electrode in the left temporal lobe or '10–20' for the standard surface electrodes of the 10–20 system. The type is a reference to the *electrode_array_type* domain table that contains a list of the possible invasive electrodes like grids, depth electrodes or electrode groups like surface or ECG electrodes. Additionally, this table has a field indicating if the array is invasive or not. Lastly, the configuration may have additional information about the array, e.g., '8×8' in case of a 64-channel grid electrode.

Having defined the electrode contacts and grouped them into arrays, the last step is to refer them to the recordings. Thereby, the name and the position of the electrode channels in recording blocks have to be reflected. A direct reference from the recording to the electrode table is not possible since there is more than just one electrode involved in a recording. Additionally it is not possible to directly define a reference from the electrode or electrode array table to the recording table since electrodes may be used in several recordings. For instance, in case of technical problems leading to the restart of a recording, the physical electrode contact remains the same. Furthermore, there may be channels with identical names belonging to different electrodes of the same admission.

For expressing this relationship between electrodes and recordings we designed the *electrode_usage* table. Besides the inherent references to the electrode and recording tables it has an attribute resembling the position of the electrode, i.e. the channel number, and the name of the electrode, respectively channel. This name should be the same like the internal electrode name. Again, we have here a redundant reference, in this case to the *electrode_array* table for allowing queries with fewer joins.

4.1.3. *EEG events: seizures, subclinical and interictal events*

On many occasions during the design of the database, we had to agree on compromises in order to find a trade-off between the desirable and the feasible. This applies in particular to the decision which events to take into account at all, and the level of detail of meta information about the events to record. Extracting and reviewing metadata is time-consuming, and a trade-off between manpower and usefulness for database-applications has to be made. Clinicians as well as database experts and prediction researchers took the basic decisions during initial project meetings, while some details had to be adapted later on. There was the agreement to include the following events with the listed grade of detail into the database. Additionally, the clinicians had to find a common basis on how to define and interpret the events and what constitutes them.

- Since seizures are the central aspect of the database, the greatest level of detail of all event types is provided: among them the semiology and the electrodes, at which the seizure originates and propagates. Additionally, seizures are an important part of the quality standards that decide if a dataset is suitable for inclusion in the database.
- Subclinical events, electrographic ictal patterns without clinical manifestation, are considered less important. For them, only the onset and offset times are recorded. Since the recordings include hundreds of such events for some patients, it was not feasible to annotate each subclinical event for each recording. Because these events may possibly allow further insights into the seizure generating processes, it was decided to mark at least ten subclinical events per day for each patient. While a complete analysis of all subclinical events is not possible, database users may thereby study at least some of the subclinical events that occurred.
- Interictal events, i.e., abnormal EEG activity in the phase between two seizures, cannot be annotated in full extent as well in long-term recordings. So it was decided to mark some characteristic events exemplarily: one of each interictal event type, where a type is determined by the pattern of the event and the electrode where the event has its maximum amplitude. Like with subclinical patterns, this information can be used to analyze EEG changes related to these events.

The seizure table contains a reference to both the recording and the block table. Even though redundant, this simplifies some queries that otherwise would involve timestamp arithmetic. It also records the pattern and the type of the seizure (simple, complex partial, secondarily generalized), as well as the state of vigilance, which is determined 10 s before seizure onset. Furthermore, the table has attributes for several timestamps: for both the clinical and the EEG manifestations of the seizure the onset and the offset, and, if applicable, for the first change in the EEG as well as for the first clinical sign.

Thereby, the clinical onset/offset is determined by video surveillance, which is not included in the database due to data privacy requirements, and may not be available in case of missing video. The EEG onset/offset is determined by EEG and may not be available in case of severe artifacts. Option-

ally, a seizure may be related to an EEG focus via the *focus* field referring to the eeg_*focus* table.

Directly related to a seizure is the semiology table holding detailed information about the seizure's semiology, i.e., the clinically observed signs of the seizure. Among them are ictal and post-ictal, subjective symptoms, motor, vegetative signs and aspects like language capabilities and reactivity. There is a field for every single symptom of these symptom groups. The type of these attributes is either Boolean, a reference to a lateralization or, a free textual description. Via the seizure reference it is directly connected to the corresponding seizure. Depending on the clinical standards, an additional timestamp attribute can be used to associate single symptoms or groups with a time-based order. In this case, there will be one tuple for each group with the same timestamp. Alternatively, all symptoms can be contained in one tuple without any timestamp information. This leads to an aggregated semiology without any temporal relation between the individual signs.

In the subclinicalEvent table, the onset and offset timestamps are recorded for each subclinical event. Therefore, it contains just the attributes "*onset*" and "*offset*" besides the references to the recording and block tables and to the commentary field.

Interictal events are recorded in the spike table. The naming of this table demonstrates nicely the evolution of the schema. At first, only spikes, as a special type of interictal events, were intended to be recorded in the database. By-and-by, a wider range of interictal events was included, but the original name of the table remained in use.

All types of interictal events are numbered consecutively (*type* column). Additionally, the timestamp of the event peak is recorded as well as the reference to the respective block and recording.

The propagation of seizures as well as the field extension of interictal events is recorded in the *propagation* table. Resembling the relationships between electrodes and seizures as well as between electrodes and interictal events, the table holds references to the *electrode*, *seizure* and, *spike* tables. In case of a propagation the boolean attributes *origin*, *early* and, *late* express a temporal classification of it.

4.1.4. *Annotations*

In Fig. 1, the last table group of the pre-surgical evaluation partition, the annotation subset, has an exceptional role, because it also belongs to the partition of time series related tables, being crucial for the evaluation of prediction methods.

This group consists of the tables *annotation*, *annotation_group* and, a*nnotation_channel,* and is used for general remarks, notes, annotations or, comments. Thereby, annotations may refer to recordings, seizures, algorithms or, time series and contain, besides an attribute holding the inevitable annotation text, fields for timestamps, a commentary and, a category (link to domain table).

By the group attribute (*grp*), an annotation may be associated to a group, respectively a tuple, of the table *annotation_group*. Such groups are defined for instance by a prediction algorithm, with the aim to group all its annotations together. Since this is a dynamic process during the runtime of an algorithm, it cannot be appropriately modeled by a domain. Besides a textual description this table classifies the group into categories via the reference to the domain table *annotation_category*. By the table *annotation_channel*, annotations can be related to one or more electrodes.

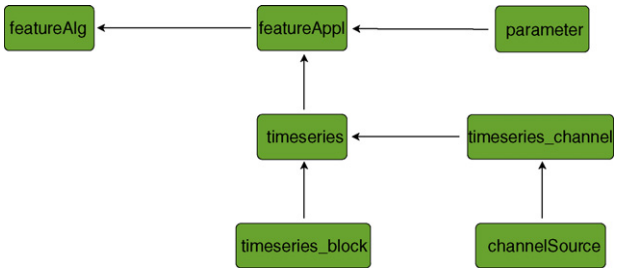

Fig. 4 – The group of time series related tables.

4.2. *Time series*

The calculation of features from EEG data is a time-consuming task. This is true especially for multivariate features for which feature time series are derived from multiple channels of the EEG time series, e.g., for each possible pair of electrode channels [26,27]. Therefore, it was decided to store these features also in the database. This "feature caching" allows a swift access to feature data. This is of importance since it was shown that the combination of different features yield significantly improved prediction results [12].

Usually, features are derived from the EEG by using a sliding-window approach: for blocks of data with a length in the order of seconds some scalar feature is calculated. Then, the window is shifted further, yielding a time series if applied consecutively, with a time distance between individual feature samples that is several orders larger than the EEG sampling distance. Yet, for multivariate features, large amounts of data are gathered, since features are derived for all pairs or even larger groups of electrode contacts. Hence, it was decided to store these features not in the relational database itself but in binary files, as also done for the raw EEG. In order to be able to access all necessary, describing metadata about the features – or, in general, time series – from the relational database, the following scheme was designed (see Fig. 4).

The basic information about the time series is stored in the table "timeseries" that contains its name, time of calculation and more. Via "*featureAppl*", detailed information about the feature calculation is stored: on the one hand, about the algorithm the feature is calculated with ("*featureAlg*"), containing information about the user and the complete execution command with all the arguments that were used for the calculation. On the other hand, the chosen parameters are stored in the table "*parameter*". These are, for example, the window length and inter-sample-distance of the sliding window. Calculation results are saved in one or several time series blocks, resembling the related recording and block entities that are used for EEG recordings.

In order to reproduce the sources that each individual time series channel is derived from, the table "channelSource" contains for each channel of the multivariate feature time series corresponding entries. If the features are calculated from EEG data, the "channelSources" link to the electrode contacts the

feature is calculated from. Additionally, also features derived from pre-processed data or time series derived from basic features can be represented by referencing to the appropriate time series channels.

5. Status report

While the database schema was evolving during the early stages of the project, it has reached a mostly steady state where changes are rare and mostly concern domain table updates, for instance the addition of a seizure pattern. The focus of the work is currently shifting towards the preparation of the datasets and the development of client applications.

5.1. *Data input*

The number of datasets is constantly growing. Current plans aim at achieving the ambitious goal of including 300 datasets by 2011. For the input of the metadata, we have chosen not to enter all the metadata into the database by means of the EPILEPSIAE client, because this would mean a repetition of work, since the technicians and doctors already have marked such events directly in the recordings using the EEG system software.

We rather have developed a software that reads these markers directly from the original EEG files in addition to the information found in the headers of the files like sampling frequencies, starting times, etc. This program is able to fill all the data of the recording, EEG, and electrodes subsets, explained in the previous section, directly into the database. Although this procedure implied some additional work before the first dataset could be loaded into the database, it really started to pay off quickly: if there are some changes in the database schema we just have to adapt the reading software once and get the changed datasets automatically by re-reading. Furthermore, this approach is easier to integrate into the routine work at the clinics – the seizures and other EEG events need to be marked anyway. Annotations, however, have to comply with the newly developed annotation protocol [28].

This protocol specifies a syntax and semantic for the markers that are used to annotate seizures, subclinical, and interictal events. These provide additional information, e.g., about electrodes that cannot be found directly in the EEG files. In addition to the format, the protocol determines when the markers have to be set.

Of course, following this protocol requires annotating more details: it requires annotating more exhaustively and accurately than necessary for clinical purposes. This is the reason why in practice a reevaluation of the EEG recording is inevitable before inclusion into the database.

5.2. *Status of the database*

Although the Oracle database server (www.oracle.com) is the official database that is used at all partner's sites in the project, the schema, as a relational one, is quite agnostic of the underlying database system and can easily be adapted to other relational databases. Therefore, only a minor adaption of the data types is necessary.

The only exception is the locator attribute in the files table that has the type BFILE, which is a proprietary data type and only available when using Oracle. Without it, the direct file access through the database is not possible. Instead, the files must be accessed indirectly through the file system after retrieving the file name and path from the database.

The advantage is that users are not dependent of the commercial Oracle and can switch to other database systems anytime. This may be of interest either for potential partners with reduced budgets or even for the project as a whole, if further funding were insufficient to account for software costs. For instance, some project partners have successfully deployed the database on a freely available PostgreSQL database server [29,30].

Yet, when using Oracle we can take advantage of the advanced features that it offers for distributed database deployments. At the moment, three identical databases with replicated data are located at the partner's sites. This simplifies administration and minimizes network traffic. Later, a distributed database where each site hosts just it's own datasets, but provides a transparent access to all the datasets of the other sites, may be required for handling the increasing data volume, although this would be very demanding especially for the network infrastructure.

For the exchange of data between the sites two data formats were fixed. First, we have designed an archive format framing a directory structure for the EEG files, imaging files and content of the database in form of files with insert statements. These archives can then be sent by hard disks on the postal way or transferred over the Internet if the network infrastructure allows for it.

As format for the EEG files, contained in this archive, we use the binary EPILEPSIAE format, introduced in the previous section, since it is inherently supported and guarantees minimal file sizes. Additionally, a separate header file with supplemental information is provided for each of the binary files. This may be crucial for loading the data at times when the metadata cannot be read from the database.

5.3. *Client applications*

An official client for the EPILEPSIAE database is provided that allows data input, browsing through the available datasets and for querying the database. As appropriate platform for this client, we have opted for a PHP/Apache based web application that can be centrally installed and maintained. This has the advantage that access to the database, being deployed in the intranet, can stay restricted to the web server, and no direct, public access to it is needed. The web server itself may be publicly available or just from inside the intranet. In either case, a login to the application for a proper user authorization will be needed.

The client provides options for viewing the data, updating parts of it and inserting data manually. Thereby, it orientates on the hierarchical structure of the schema. On the top level it provides a list of all available datasets, respectively the list of the contained patients. After selecting one of these patients, it presents on the next level the details of the corresponding tuple of the patient table and the directly dependent tuples, in this case, of the admission table. Again, after choosing one of

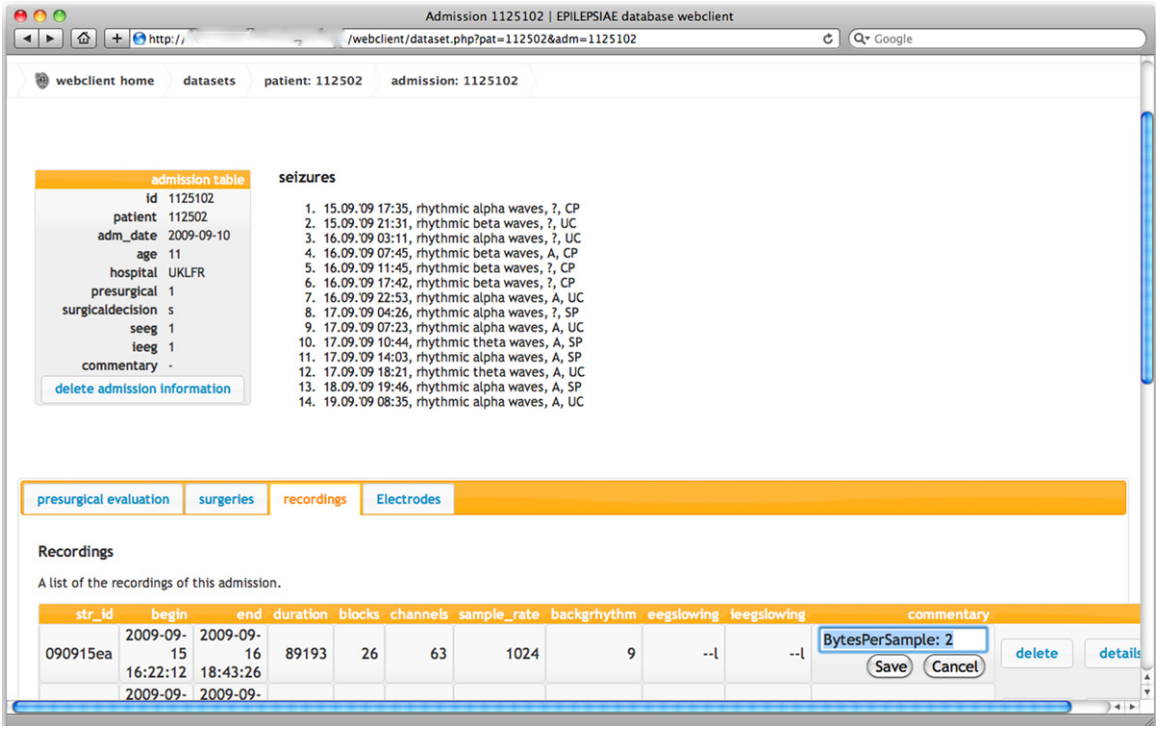

Fig. 5 – A screenshot of the EPILEPSIAE web client showing the details of a certain admission. On the top half, the content of the respective tuple of the admission table is given, while the bottom page shows a tab bar where the details of the subsets can be accessed that are directly dependent of admissions: the pre-surgical evaluation, surgery and recordings. The electrode tab on the right is an exception. It gives an overview over all electrodes and their arrays that are used in one of the recordings of the admission. The opened recordings tab shows the tuples of the recordings table that have a reference to the selected admission, that is, all recordings that belong to the admission. There are buttons for deleting and adding recordings; the content of the current tuples can directly be manipulated by clicking the respective field as shown for the commentary field of the third tuple.

the admissions for detailed view, on the next level the respective admission tuple and directly dependent ones are shown. This is shown in Fig. 5. The sequence of these levels is thereby translated into the breadcrumb navigation structure that can be seen at the top of the figure.

In addition to the data input, the systematic querying of the database is the main task of the client. Therefore, it provides a simple interface for common types of queries, for instance about seizures, which is depicted in Fig. 6. The options of the query can be gathered by form elements like checkboxes. Additionally, the client offers the possibility to view the seizure's EEG in the browser, without the possibility to change montages or other options known from EEG systems. Of course, the client offers such user-friendly interfaces for other predefined queries as well as the possibility to directly enter generic SQL queries for the expert user.

In addition to this client, there may be tailor made local clients that are used for data input at the individual project sites. For instance, in Freiburg the pre-surgical data part is entered into a pre-existing local patient's database with information about all recorded patients and subsequently transferred to the EPILEPSIAE database. Here, the web client is only used for viewing and querying the data.

6. Discussion and future plans

Albeit still in the process of completion, the EPILEPSIAE database is already by far the most comprehensive and complete epilepsy database currently existing. The arousing interest for access to the database presently offered by the Freiburg Epilepsy Center [11] shows the general need for publicly available, high quality databases of long-term, continuous EEG recordings, not only for seizure prediction but also for numerous related research communities. Apart from that, numerous participation requests from all over the world give evidence of the emerging acceptance of such a database as the de facto standard for databases in the field of epilepsy.

This acceptance of our database schema, content and, methods like the annotation protocol as emerging standards is probably the most important impact of our database.

On the other hand, the time and effort for compiling such a database is so enormous that single institutions sooner or later reach their limits. For example, out of the more than 1000 long-term EEG recordings conducted at the epilepsy Center of the University Hospital of Freiburg during the monitoring of epilepsy patients, not more that 200 meet the quality criteria of the EPILEPSIAE project.

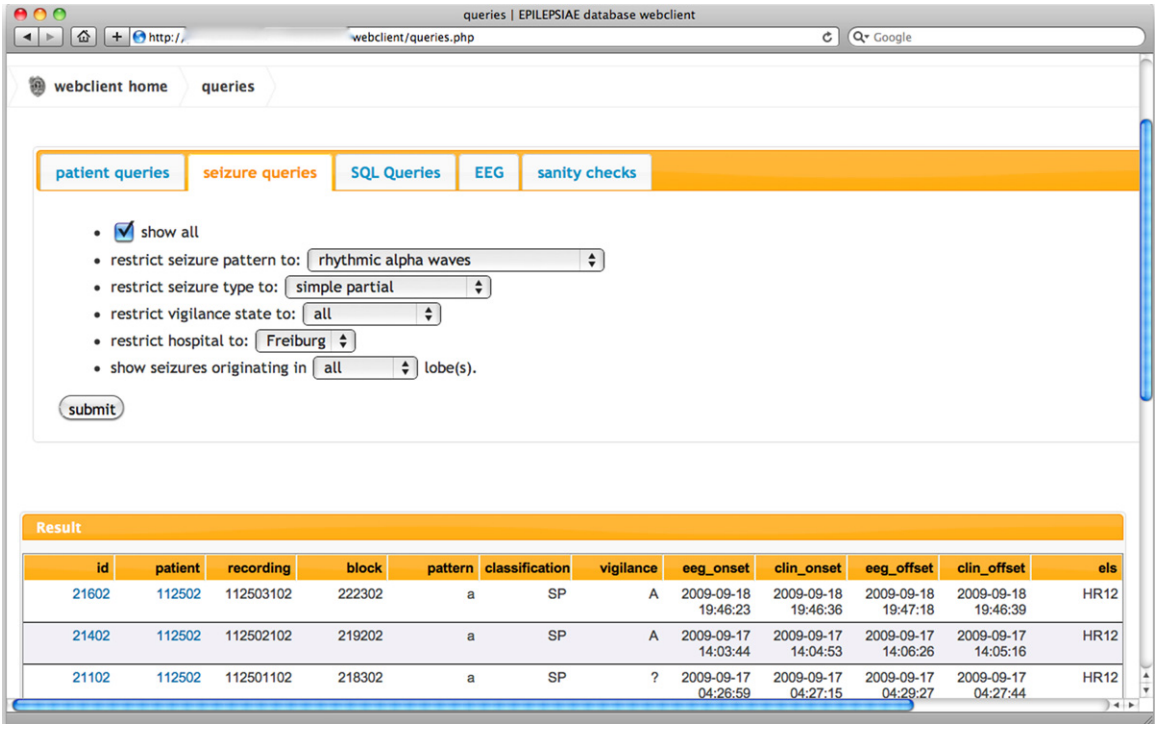

Fig. 6 – Graphical mask for seizure queries.

So, organizing a huge database like the one presented here is only possible as a joint effort, where the local datasets of several hospitals are collected and compiled into a single database. But building a database from such distributed sources is a bigger endeavor than just the mere sum of the local activities, as our experience within the project has been showing us.

While the database was designed for the use of research on seizure prediction, its general structure may also be employed for other medical research databases. Especially for time series databases, significant parts of the scheme may be reused, which also would allow to apply evaluation software written for the EPILEPSIAE database to similar databases.

6.1. *Future plans*

Since the funding of the project ends after the initial grant period, the future financing of the database has to be addressed. There are high maintenance costs particularly of the hardware and its hosting in the computing centers that require an exploitation plan to ensure some financial revenue, e.g., by charging for the use of the data. A possible extension of the data pool could be accomplished by adding further partners. This may involve adaptations of the schema.

Furthermore, a planned collaboration with the American epilepsy database that will start in 2010 may have implications for the database scheme to allow for cross-linking or special import/export functions.

Acknowledgements

We would like to sincerely thank the clinical teams at the Epilepsy Centers in Freiburg, Paris, and Coimbra. This work was supported by the European Union (Grant 211713), the German Federal Ministry of Education and Research (BMBF grant 01GQ0420), and the Excellence Initiative of the German Federal and State Governments.

REFERENCES

[1] J. Murray, Coping with the uncertainty of uncontrolled epilepsy, Seizure 2 (1993) 167–178.
[2] A. Schulze-Bonhage, F. Sales, K. Wagner, R. Teotonio, A. Carius, A. Schelle, M. Ihle, Views of patients with epilepsy on seizure prediction devices, Epilepsy & behavior 18 (2010) 388–396.
[3] A. Schulze-Bonhage, A. Buller, Unpredictability of seizures and the burden of epilepsy, in: B. Schelter, J. Timmer, A. Schulze-Bonhage (Eds.), Seizure Prediction in Epilepsy: From Basic Mechanisms to Clinical Applications, Wiley-VCH, Berlin, 2008, pp. 1–10.
[4] B. Schelter, J. Timmer, A. Schulze-Bonhage (Eds.), Seizure Prediction in Epilepsy: From Basic Mechanisms to Clinical Applications, Wiley-VCH, Berlin, 2008.
[5] M. Winterhalder, T. Maiwald, H.U. Voss, R. Aschenbrenner-Scheibe, J. Timmer, A. Schulze-Bonhage, The seizure prediction characteristic: a general framework to assess and compare seizure prediction methods, Epilepsy Behav. 4 (2003) 318–325.
[6] B. Schelter, M. Winterhalder, T. Maiwald, A. Brandt, A. Schad, A. Schulze-Bonhage, J. Timmer, Testing statistical significance of multivariate time series analysis techniques for epileptic seizure prediction, Chaos 16 (2006) 013108.

[7] L. Glass, Synchronization and rhythmic processes in physiology, Nature 410 (2001) 277–284.
[8] The Bonn EEG database http://epileptologie-bonn.de/cms/front_content.php?idcat=193.
[9] R. Andrzejak, K. Lehnertz, C. Rieke, F. Mormann, P. David, C. Elger, Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: dependence on recording region and brain state, Phys. Rev. E 64 (2001) 061907.
[10] The Flint Hills Scientific ECoG database http://www.fhs.lawrence.ks.us/PublicECoG.htm.
[11] The Freiburg EEG database http://epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project/eeg-database.
[12] H. Feldwisch-Drentrup, B. Schelter, M. Jachan, J. Nawrath, J. Timmer, A. Schulze-Bonhage, Joining the benefits: Combining epileptic seizures prediction methods, Epilepsia 51 (2010) 1598–1606.
[13] D. Chamberlin, R.F. Boyce, SEQUEL: A Structured English Query Language. SIGMOD Workshop, vol. 1, 1974, pp. 249–264.
[14] P. Chen, The entity-relationship model-toward a unified view of data, in: ACM Transactions on Database Systems, ACM Press, New York, 1976, pp. 9–36.
[15] P. Chen, Entity-relationship modeling-historical events, future trends, and lessons learned, in: M. Broy, E. Denert (Eds.), Software Pioneers: Contributions to Software Engineering, Springer, Berlin, 2002, pp. 296–310.
[16] E.F. Codd, A relational model of data for large shared data banks, in: Communications of the ACM, ACM Press, New York, 1970, pp. 377–387.
[17] Oracle http://www.oracle.com.
[18] R. Rivest, The MD5 Message-Digest Algorithm, RFC 1321, 1992.
[19] N. Mukherjee, B. Aleti, A. Ganesh, K. Kunchithapadam, S. Lynn, S. Muthulingam, K. Shergill, S. Wang, W. Zhang, Oracle SecureFiles System, in: Proc. VLDB Endow, 2008, pp. 1301–1312.
[20] P. Mazur, J. Murlewski, M. Kaminski, B. Sakowicz, D. Makowski, Comparison of large object storage methods in Oracle database version 11g, CAD Systems in Microelectronics, 2009, pp. 233–236.
[21] B. Kemp, A. Värri, A.C. Rosa, K.D. Nielsen, J. Gade, A simple format for exchange of digitized polygraphic recordings, Electroencephalogr. Clin. Neurophysiol. 82 (1992) 391–393.
[22] General data format: http://arxiv.org/abs/cs.DB/0608052.
[23] H. Jasper, Report of the committee on methods of clinical examination in EEG, Electroencephalogr. Clin. Neurophysiol. 10 (1958) 370–375.
[24] A.C. Evans, D.L. Collins, S.R. Mills, E.D. Brown, R.L. Kelly, T.M. Peters, 3D statistical neuroanatomical models from 305 MRI volumes, in: Proc. IEEE-Nuclear Science Symposium and Medical Imaging Conference, 1993, pp. 1813–1817.
[25] J. Talairach, P. Tournoux, Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System – an Approach to Cerebral Imaging, Thieme Medical Publishers, New York, 1988.
[26] R. Sowa, A. Chernihovskyi, F. Mormann, K. Lehnertz, Estimating phase synchronization in dynamical systems using cellular nonlinear networks, Phys. Rev. E 71 (2005) 061926.
[27] A. Müller, H. Osterhage, R. Sowa, R.G. Andrzejak, F. Mormann, K. Lehnertz, A distributed computing system for multivariate time series analyses of multichannel neurophysiological data, J. Neurosci. Methods 152 (2006) 190–201.
[28] M. Ihle, C. Gierschner, V. Navarro, M. LeVan Quyen, F. Sales, N. Silva, A. Schulze-Bonhage, Standardization of EEG Annotations for the European Epilepsy Database EPILEPSIAE, in: 4th International Workshop on Seizure Prediction, Kansas City, 2009.
[29] M. Stonebraker, L.A. Rowe, The design of POSTGRES, SIGMOD Rec. 15 (1986) 340–355.
[30] M. Stonebraker, G. Kemnitz, The POSTGRES next generation database management system, Commun. ACM 34 (1991) 78–92.
[31] J. Engel, Outcome with respect to epileptic seizures, in: J. Engel (Ed.), Surgical Treatment of the Epilepsies, Raven Press, New York, 1993, pp. 609–622.