# Classification and Similarity of Patient Records

Markus Frick

Martin Trautwein

Vivantes - Netzwerk für Gesundheit
Abteilung für Klinische Forschung
Dr. Markus Frick, Dr. Martin Trautwein

# Patient Identification?

- Data Basis: a set of clinical data records (be aware of data protection)

Question: Identify patients based on their data records containing

- – Demographic data: age, sex, …
- – Lab-Results: blood pressure, leucocytes (blood cells), HIV, …
- – Structured diagnoses (I21.2 - Akuter transmuraler Myokardinfarkt an sonstigen Lokalisationen) – for accounting purposes!!
- – And Diagnoses, Medication etc. !!!

Note that e.g. diagnoses, medication etc. are only in written document

➔ Linguistic pipeline pulls these things from the doctors letters

# Clinical Patient Records (structured)

Jane Doe: a sample feature set

| Age | Sex | Leu | Crea | MCV | J.90 | Z43.0 | R32 | KHS | cx |
|---|---|---|---|---|---|---|---|---|---|
| 65 | f | 7.51 | 0.93 | 87.6 | X | X | X | X | O |
| | | | | | | | | | |

Whatever come up here

Diagnosis (structured)

Diagnosis (un-structured)

So this feature vector is quite sparse

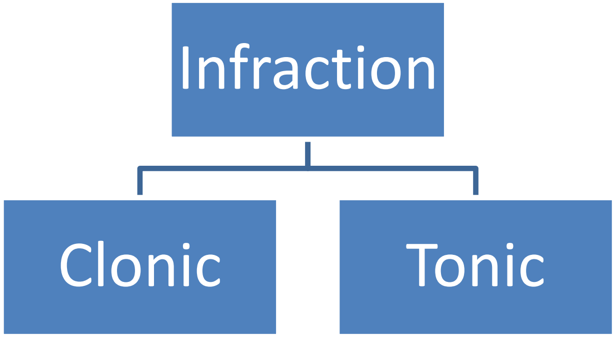
# Clinical Patient Records (structured)

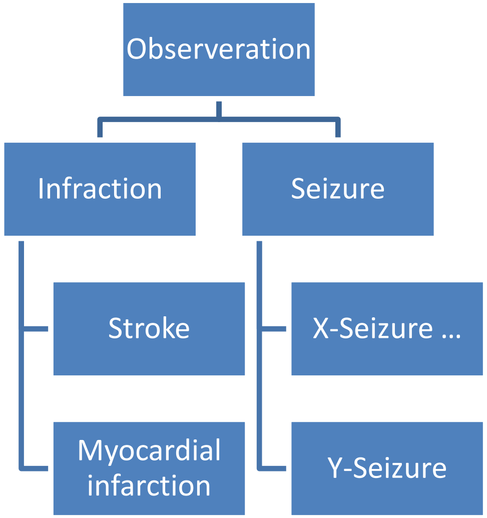
- Demographic/Visit data like age, sex and formal data like admission date, length of the clinic stay, etc.
- Laboratory results - one such result has the form: (**measurementType, value, unit, abnormalFlag**) where **measurementType** defines the kind of measurement (like blood pressure) and **abnormalFlag** is something like (very low, low, normal, high, very high).
- Structured diagnoses - these are ICD10-encoded diagnoses (e.g. **I20.8** stands for angina pectoris); these have a "prefix"-format, which means that **I25.16** is a refinement of **I25.1**

# Clinical Patient Records (unstructured)

Semantic facts - formally, a semantic fact is represented as a small labelled tree with the nodes denoting instances of medical concepts and the relations denoting relationships. The tree

- hasObservation:Seizure[hasAttribute:Tonic, hasAttribute:Clonic] i.e. a parent with two children, represents the fact that the patient has had a tonic-clonic seizure. Perhaps it's important to know that these concepts are organized hierarchically with respect to
- a "subclass" relationship (actually it's some sort of taxonomy)





# Questions – On our side

- given a patient and a selection of features, find me similar patients?? (whatever similarity means here?)
- given a condition (e.g. 30 < age < 70) on a single feature or on a set of features, can we learn that condition (find a classifier?)
- given a condition as above, can we refine the condition in such a way that the set of eligible patients remains similar (e.g can we make the 70 above a 69?)
- could we generalize the conditions such that the number of eligible patients crosses a certain threshold (assume that we want to increase the number of eligible patients without giving up too much)