MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220141

"Book Music" Representation for Temporal Data, as a Part of the Feature Extraction Process: A Novel Approach to Improve the Handling of Time-Dependent Data in Secondary Use of Healthcare Structured Data

Emmanuel CHAZARD[a], Pierre BALAYE[a], Thibaut BALCAEN[a], Michaël GENIN[a], Marc CUGGIA[b], Guillaume BOUZILLE[b], Antoine LAMER[a]

[a] Univ. Lille, CHU Lille, ULR 2694 - METRICS, CERIM, Public health dept, F-59000 Lille, France
[b] LTSI, Université de Rennes 1, CIC & CIC-IT Inserm 1414, CHU Rennes, F-35000 Rennes, France

Abstract

Book music is extensively used in street organs. It consists of thick cardboard, containing perforated holes specifying the musical notes. We propose to represent clinical time-dependent data in a tabular form inspired from this principle. The sheet represents a statistical individual, each row represents a binary time-dependent variable, and each hole denotes the "true" value. Data from electronic health records or nationwide medical-administrative databases can then be represented: demographics, patient flow, drugs, laboratory results, diagnoses, and procedures. This data representation is suitable for survival analysis (e.g., Cox model with repeated outcomes and changing covariates) and different types of temporal association rules. Quantitative continuous variables can be discretized, as in clinical studies. The "book music" approach could become an intermediary step in feature extraction from structured data. It would enable to better account for time in analyses, notably for historical cohort analyses based on healthcare data reuse.

Keywords:
Data reuse, feature extraction, survival analyses.

Introduction

Secondary use of clinical data, or clinical data reuse, consists of "non-direct care use of personal health information" [1, 2], notably for research purposes. This concept takes its origins from Fayyad's "knowledge discovery in databases" [3]. Structured data denote data that can be described in a table, excluding unstructured data such as free-text or medical images, or data with inconstant data model (e.g., JSON, XML, etc.). Data reuse of structured data usually consists of 5 steps [4]:

1. data pre-processing (from native data to standardized and cleaned data warehouse),
2. feature extraction (from data warehouse data to individual analyzable information),
3. statistical and graphical mining (from individual information to associations),
4. expert filtering and reorganization, and
5. decision making.

Data warehouses contain data that are not always directly suitable for statistical analyses: data are broken down into dozens of tables (at least one per entity), are made of multivalued qualitative data with too many categories, are deeply unbalanced, and are mostly missing (never at random, due to the indication bias related to the request for complementary examinations). The feature extraction process (formerly called "data transformation"; see step 2 above) [3, 4] enables to make those data suitable for statistical analyses: one only table (with one row per individual, one column per variable, potentially many columns), where qualitative variables have few modalities, variables are few unbalanced, and missing data are rare.

Many studies using data reuse of structured healthcare data aim at producing epidemiological knowledge, by enhancing retrospective observational cohorts [1, 5, 6]. Unlike case-control studies, retrospective cohort studies make it possible to consider the time-dependent nature of the variables in statistical models, whether they are the outcome, or the covariates. However, to our knowledge, the time-dependent nature of variables is generally under-exploited in many data reuse studies [7]. In traditional, form-based clinical research, the data are probably processed intelligently by the form filler. Data from the form have then a simplified structure, and intrinsically consider temporal aspects (e.g., a pathology is considered present when it is observed at least once in the six months preceding the patient's inclusion). We believe that the complexity of this process is insufficiently reproduced in many data reuse studies, and that this problem could be solved by defining a framework for processing time-dependent data. Computer representation and exploitation of time-dependent data have already been the subject of important work [8, 9], that has not penetrated the field of data reuse.

Our goal is to propose an approach inspired by book music to guide feature extraction for retrospective cohort studies obtained from healthcare data reuse.

Methods

Book music, analogy with time-dependent data

In the first part, we will describe book music, and conduct an analogy with clinical structured data.

Healthcare data in data reuse epidemiological studies

In a second part, we will analyze healthcare data that are the most frequently encountered in data reuse studies, and evaluate whether they can fit the "book music" principle. We will analyze the most important data types described in the HL7 Reference Information Model [10], in the OMOP data model [11], and in previous works [12]. We will not interest on native data

(e.g., the patient's birthdate) but on data that usually have to be analyzed (e.g., the patient's age). Those data consist of:

- Demographic data (age, gender)
- Flow data (e.g., admission, hospital stay, ICU stay, discharge, medical appointment, etc.)
- Encoded observations, such as diagnoses
- Encoded procedures, such as therapeutic procedures, and corresponding resources
- Measurements, such as laboratory results, or results of clinical scores
- Drug administrations

Such data can more or less be found in electronic health records of inpatients and outpatients, and in nationwide medical-administrative databases [10–12].

Statistical methods for time-dependent data

In a third part, we will list the most frequent methods of time-dependent-data analysis, and evaluate whether they can fit the "book music" principle. To achieve this, we will interest in the Cox model with repeated outcomes and changing covariates [13], which to our knowledge takes into account all the temporal characteristics of the different survival analysis techniques. We will also interest in temporal association rules (TAR) [7], a broad set of more recent methods, also known under heterogeneous terms (sequential patterns, temporal patterns, temporal rules, temporal association, etc.). As this field of study is wide and heterogeneous, we will use the unified terminology proposed by Segura-Delgado et al. [7].

Results

Book music, analogy with time-dependent data

Book music is a medium for storing music. It is made from thick cardboard, containing perforated holes specifying the musical notes to be played (Figure 1). Each note of the instrument is represented by a virtual row whose location on the card is fixed (for example on Figure 1, the highest horizontal rows correspond to the highest notes). The note is played as long as the row is perforated. When the row is not perforated, the note is not played. The perforated rows are separated by non-perforated rows, to preserve the strength of the score. The length of the card is not limited, and corresponds to the length of the music piece. Book Music was extensively used in mechanical organs (or fairground or street organs; Figure 2).

From an information point of view, we can consider each statistical individual as a sheet of book music. Each virtual row represents a time-dependent binary variable, that can be on (perforated) or off (not perforated), as a function of time. As in the mechanical organ, the time corresponds to the offset from the left edge of the score. At a given position (a vertical section), each binary variable is thus in a precise state, and the set of binary variables describes the state of the statistical individual at this precise moment. Some variables change state little, like the held notes of a score, and others consist of point-events, like the *staccato* notes of a score.

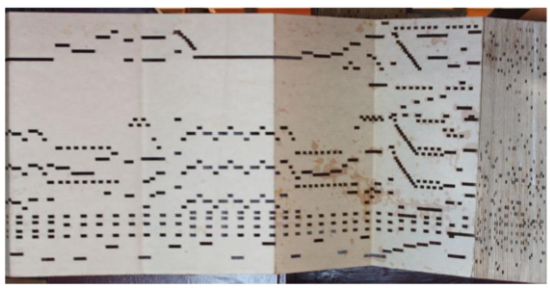

Figure 1 – Book music (resized picture - Credit: Richard Ash - Creative Commons Attribution-Share Alike 2.0)

Figure 2 – Street organ (Credit: Roman Bonnefoy - Creative Commons Attribution-Share Alike 3.0)

Healthcare data in data reuse epidemiological studies

Figure 3 show an example of transformation from data described in a data warehouse to "book music" data, and the corresponding tabular representation (right part). We will comment on some parts of this schema, to show the versatility of this approach.

The "demographic data" part of Figure 3 illustrates that a qualitative variable (gender) can be summarized as a time-dependent binary variable, which is not necessarily constant over time. The age, a continuous variable inferred from birth and admission date, is represented as 3 time-dependent binary variables in this example.

In the "patient flow" part of Figure 3, we can see that the analyst can focus on the information that interests him/her: admission, discharge, transfers, entire inpatient stay, or staying in an intensive care unit for instance.

The "laboratory results" part of Figure 3 illustrates that it is possible to use different cutoffs on a same functional variable, to define for instance hypokalemia, hyperkalemia, and severe hyperkalemia. An interpolation algorithm can also be defined (e.g., linear interpolation, last observation carried forward, etc.): in real life, contrary to therapeutic trials, measurements are performed on free dates and therefore not synchronized. It is also possible to trace measurements themselves, regardless of their results. If the analyst does not want to infer missing data as normal values, he/she can also explicitly identify time ranges with missing information.

The "drugs" part of Figure 3 illustrates it is possible to define binary values based on administered doses that were first aggregated on a daily base.

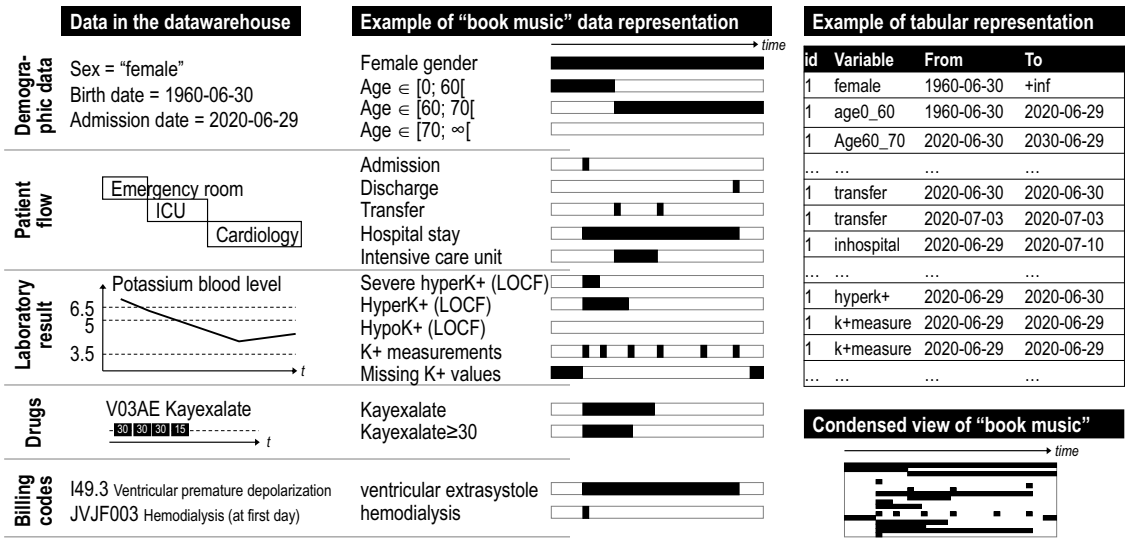

Figure 3 – Example of transformation into "book music" data. Times were truncated to dates only (K+: potassium, hyper/hypoK+: hyper/hypokalemia, LOCF: last observation carried forward)

The "billing" part of Figure 3 illustrates it is possible to describe codes as observations during all the stay, or at a precise date.

In any case, right part of Figure 3 illustrates a tabular representation of those data, based on an "entity-attribute-dates" model. It requires only 1 table made of 4 columns. Only the "true" values (the "holes") are stored (missing values can still be explicitly traced in the rare cases where they should not be interpreted as "false"). A patient identifier enables to link all his/her attributes. This table would be an intermediate form common to many studies. Then, a final transformation stage would produce a table suitable for statistical analyses, with 1 row per patient. This final step is described in the "discussion" section.

Statistical methods for time-dependent data

The "book music" data representation seems compatible with the following characteristics of the Cox model with repeated outcomes and changing covariates [13]. Regarding the covariates, it enables to describe binary, qualitative, and ordinal time-dependent variables. Regarding the outcome, it enables to describe time-dependent binary outcomes, possibly repeated, as well as censored data (censoring can be described as another binary event).

The "book music" data representation seems compatible with the following characteristics of temporal association rules. According to the terminology defined by Segura-Delgado et al. [7], it can be used when time is an implied component (i.e. to filter relevant combinations), either for sequential methods or intertransaction methods. It can also be used when time is an integrated component (i.e. is part of the learning process), when the methods interest on periodicity, time-intervals (not only point-like events), or lifespans. It is also suitable for changes or incremental mining. However, while the Cox model and some temporal association rules methods are able to incorporate continuous quantitative variables, the "book music" approach does not allow so (apart from time itself), and requires that the variables be discretized. This will be discussed in the discussion section.

Discussion

The "book music" approach could facilitate and abstract the feature extraction phase, as shown in Figure 4. The first step of feature extraction would consist in generating the "book music" using a procedure independent of the study to be conducted. Then, in a project-specific way, predefined and relatively abstract functions could be used to generate the individual information table, which consists of one row per statistical individual and one column per variable. These functions would describe generic operations such as "summarize the state of the patient at such a date", or "aggregate the state over such a period with such a function".

Feature extraction consists of computations that enable to transform data into ready-to-use information. Those transformations can be performed during the first step, that enables to generate "book music" data, but Figure 5 also illustrates that some transformations can be performed afterwards.

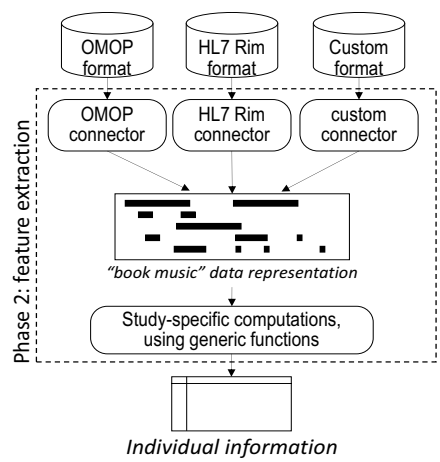

Figure 4 – The "book music" data representation, an intermediary step in the feature extraction process

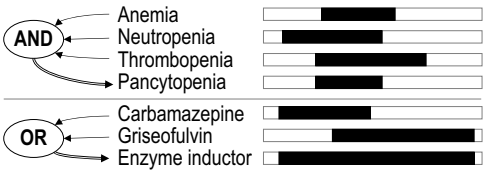

Figure 5 – Time-dependent Boolean operations

The data transformation to the "book music" representation participates in fulfilling the 5 objectives of feature extraction defined in [4]:

- It enables to reduce data complexity to one single data table, with several rows per statistical individual. The next step will be easier to implement, to get one row per individual.
- It uses domain specific cutoffs for quantitative variables, and enables to get precise start and stop dates.
- If can be used to reduce data imbalance, by mean of mappings. This can be achieved during this step, or easily at next step (see Figure 5).
- It enables to handle heterogeneous data in the form of generic time-dependent events. The nature of the data remains important for interpretation, but no longer makes a difference for the analysis methods to be implemented.
- It participates in making future results more acceptable for experts.

The main drawback of this approach is its inability to fully represent continuous variables: cutoffs must be applied. To our knowledge, this weakness has little impact in practice. Quantitative variables are most often interpreted with respect to domain-dependent thresholds. This is because medical reasoning is generally based on trees whose nodes are binary conditions. Moreover, threshold-based discretization allows one to clearly identify at which dates a condition is true. In addition, the introduction of continuous quantitative variables in regression models generally relies on the assumption of a linear or log-linear effect, which is nearly always false. It is obvious, for example, that when potassium plasma value increases from 3 to 4 it is a good thing, when it increases from 3.5 to 4.5 it is indifferent, and when it increases from 5 to 6 it is a bad thing. For this reason, to the best of our knowledge, in epidemiological studies, quantitative variables are nearly always discretized before being introduced into regression or survival models.

The "book music" approach is compatible with programs that enable to visualize time-dependent healthcare data, whether they are programs intended for care (a single patient) [14–20], or programs intended for decision analysis (several patients at once) [14, 21–23].

Despite a common focus, the approach we propose differs from the illustrious previous work. In 1983, Allen et al. proposed an approach to describe imprecisely dated events [8]. The fact is that, in structured reusable databases, clinical data are all precisely dated: we mean that, for example, each diagnostic code has a precise date of coding, but this date does not necessarily represent the reality of the disease itself. It is now commonly accepted that this inconsistency is implicitly taken into account by the analyst, and does not need to be explicitly described in the data. In 1992, Shahar et al. proposed a temporal abstraction system for patient monitoring [9]. This approach was oriented towards the individual analysis of a patient's data, and not the constitution of a learning base, as the time was not yet ripe for machine learning [24]. The development of a complex temporal query language was based on the belief that tasks could be separated and performed by different profiles (computer engineers, physicians, etc.). The fact is that today's trend is to train data scientists mastering the algorithmic, statistical and domain-specific aspects at the same time. They do not necessarily need a new framework, but rather a set of packaged functions, all available in the same environment, such as the R software.

The perspectives are summarized in Figure 4. First of all, we will have to develop generic mechanisms to facilitate the transformation of relational data into "book music" form. These mechanics will be specified, and their use simplified when the native data follow a known data model such as OMOP or HL7 RIM [10, 11], and could be developed quite easily for non-standard data models (top of Figure 4). We will then need to abstract and propose standardized functions to complete the feature extraction step, by transforming the "book music" data into a ready-to-analyze table, including one row per individual (bottom of Figure 4). The complete feature extraction process will be made of both steps.

Conclusions

The "book music" approach could become an important step in feature extraction, simplify and secure this phase. This approach could help to better account for time in analyses, especially for historical cohort analyses based on the secondary use of structured healthcare data.

Acknowledgements

This research did not receive any funding. Authors have no conflict of interest to declare.

References

[1] S.M. Meystre, C. Lovis, T. Bürkle, G. Tognola, A. Budrionis, and C.U. Lehmann, Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress, *Yearb. Med. Inform.* **26** (2017). doi:10.15265/IY-2017-007.

[2] C. Safran, M. Bloomrosen, W.E. Hammond, S. Labkoff, S. Markel-Fox, P.C. Tang, D.E. Detmer, and null Expert Panel, Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper, *J. Am. Med. Inform. Assoc. JAMIA*. **14** (2007) 1–9. doi:10.1197/jamia.M2273.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, From Data Mining to Knowledge Discovery in Databases, *AI Mag.* **17** (1996) 37.

[4] E. Chazard, G. Ficheur, A. Caron, A. Lamer, J. Labreuche, M. Cuggia, M. Genin, G. Bouzille, and A. Duhamel, Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features, *Stud. Health Technol. Inform.* **255** (2018) 15–19.

[5] C. Safran, Reuse of clinical data, *Yearb. Med. Inform.* **9** (2014) 52–54. doi:10.15265/IY-2014-0013.

[6] C. Safran, Update on Data Reuse in Health Care, *Yearb. Med. Inform.* **26** (2017) 24–27. doi:10.15265/IY-2017-013.

[7] A. Segura-Delgado, M.J. Gacto, R. Alcalá, and J. Alcalá-Fdez, Temporal association rule mining: An overview considering the time variable as an integral or implied component, *WIREs Data Min. Knowl. Discov.* **10** (2020) e1367. doi:https://doi.org/10.1002/widm.1367.
[8] J.F. Allen, Maintaining Knowledge about Temporal Intervals, in: D.S. Weld, and J. de Kleer (Eds.), Read. Qual. Reason. Phys. Syst., Morgan Kaufmann, 1990: pp. 361–372. doi:10.1016/B978-1-4832-1447-4.50033-X.
[9] Y. Shahar, and M.A. Musen, A temporal-abstraction system for patient monitoring, *Proc. Symp. Comput. Appl. Med. Care*. (1992) 121–127.
[10] Reference Information Model (RIM) Downloads | HL7 International, (n.d.). http://www.hl7.org/implement/standards/rim.cfm (accessed May 3, 2021).
[11] M.J. Schuemie, Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD, *Pharmacoepidemiol. Drug Saf.* **20** (2011) 292–299. doi:10.1002/pds.2051.
[12] E. Chazard, B. Merlin, G. Ficheur, J.-C. Sarfati, PSIP Consortium, and R. Beuscart, Detection of adverse drug events: proposal of a data model, *Stud. Health Technol. Inform.* **148** (2009) 63–74.
[13] M. Zhou, Understanding the Cox Regression Models With Time-Change Covariates, *Am. Stat.* **55** (2001) 153–155. doi:10.1198/000313001750358491.
[14] N. Martignene, T. Balcaen, G. Bouzille, M. Calafiore, J.-B. Beuscart, A. Lamer, B. Legrand, G. Ficheur, and E. Chazard, Heimdall, a Computer Program for Electronic Health Records Data Visualization, *Stud. Health Technol. Inform.* **270** (2020) 247–251. doi:10.3233/SHTI200160.
[15] R. Bade, S. Schlechtweg, and S. Miksch, Connecting Time-oriented Data and Information to a Coherent Interactive Visualization, in: Proc. SIGCHI Conf. Hum. Factors Comput. Syst., ACM, New York, NY, USA, 2004: pp. 105–112. doi:10.1145/985692.985706.
[16] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman, LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records, in: B.B. Bederson, and B. Shneiderman (Eds.), Craft Inf. Vis., Morgan Kaufmann, San Francisco, 2003: pp. 308–312. doi:10.1016/B978-155860915-0/50038-X.
[17] D.A. Aoyama, J.-T.T. Hsiao, A.F. Cárdenas, and R.K. Pon, TimeLine and visualization of multiple-data sets and the visualization querying challenge, *J. Vis. Lang. Comput.* **18** (2007) 1–21. doi:10.1016/j.jvlc.2005.11.002.
[18] D. Goren-Bar, Y. Shahar, M. Galperin-Aizenberg, D. Boaz, and G. Tahan, KNAVE II: The Definition and Implementation of an Intelligent Tool for Visualization and Exploration of Time-oriented Clinical Data, in: Proc. Work. Conf. Adv. Vis. Interfaces, ACM, New York, NY, USA, 2004: pp. 171–174. doi:10.1145/989863.989889.
[19] A.A.T. Bui, D.R. Aberle, and H. Kangarloo, TimeLine: Visualizing Integrated Patient Records, *IEEE Trans. Inf. Technol. Biomed.* **11** (2007) 462–473. doi:10.1109/TITB.2006.884365.
[20] J.S. Hirsch, J.S. Tanenbaum, S. Lipsky Gorman, C. Liu, E. Schmitz, D. Hashorva, A. Ervits, D. Vawdrey, M. Sturm, and N. Elhadad, HARVEST, a longitudinal patient record summarizer, *J. Am. Med. Inform. Assoc. JAMIA*. **22** (2015) 263–274. doi:10.1136/amiajnl-2014-002945.
[21] K. Wongsuphasawat, and D. Gotz, Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization, *IEEE Trans. Vis. Comput. Graph.* **18** (2012) 2659–2668. doi:10.1109/TVCG.2012.225.
[22] A. Perer, F. Wang, and J. Hu, Mining and exploring care pathways from electronic medical records with visual analytics, *J. Biomed. Inform.* **56** (2015) 369–378. doi:10.1016/j.jbi.2015.06.020.
[23] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman, Temporal Event Sequence Simplification, *IEEE Trans. Vis. Comput. Graph.* **19** (2013) 2227–2236. doi:10.1109/TVCG.2013.200.
[24] C.A. Kulikowski, Beginnings of Artificial Intelligence in Medicine (AIM): Computational Artifice Assisting Scientific Inquiry and Clinical Art – with Reflections on Present AIM Challenges, *Yearb. Med. Inform.* **28** (2019) 249–256. doi:10.1055/s-0039-1677895.

Address for correspondence

Pr Emmanuel Chazard ; CERIM, faculté de Médecine, F-59045 Lille Cedex, France ;
emmanuel.chazard@univ-lille.fr ;
Phone: +33 3 20 62 69 69.