ClinMine: Optimizing the Management of Patients in Hospital

Clarisse Dhaenens, Julie Jacques, Vincent Vandewalle, Maxence Vandromme, Emmanuel Chazard, Cristian Preda, Alexandru Amarioarei, Porpimol Chaiwuttisak, Cristina Cozma, Grégoire Ficheur, et al.

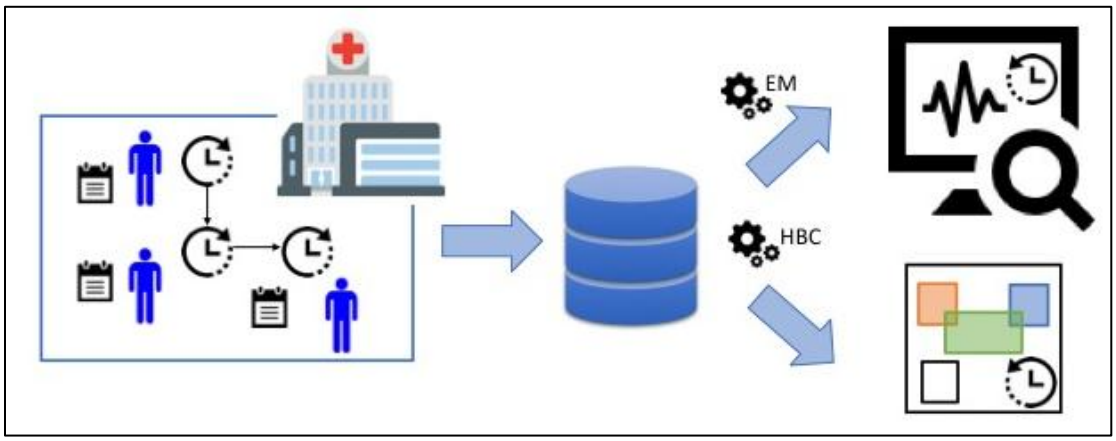
ClinMine: Optimizing the management of patients in hospital

C. Dhaenens, J. Jacques, V. Vandewalle, M. Vandromme, E. Chazard, C. Preda, A. Amarioarei, P. Chaiwuttisak, C. Cozma, G. Ficheur, M.-E. Kessaci, R. Perichon, J. Taillard, R. Bordet, A. Lansiaux, L. Jourdan, D. Delerue, A. Hansske

Abstract

A better understanding of "patient pathway" thanks to data analysis can lead to better treatments for patients. The ClinMine project, supported by the The French National Research Agency (ANR), aims at proposing, from various case studies, algorithmic and statistical models able to handle this type of pathway data, focusing primarily on hospital data. This article presents two of these case studies, focusing on the integration of temporal data within analysis. First, the hypothesis that some aspects of the patient pathway can be described, even predicted, from the management process of the hospital medical mail is studied. Therefore a specific functional data analysis is driven, and several types of patients have been detected. The second case study deals with the detection of profiles through a biclustering of the patients. The difficulty to simultaneously deal with heterogeneous data, including temporal data is exposed and a method is proposed. Results on real data show the effectiveness of the proposed method.

Graphical Abstract



Keywords

Hospital information system, patient pathway, heterogeneous data, temporal data, statistical analysis, optimization algorithms, electronic health records.

1. Introduction

1.1 - Context of the project

The concept of "patient pathway" (or patient trajectory) is now the central part of every discussion on the evolution of the health care system in France. The patient pathway is defined by the French Ministry of Health as "the global, structured and continuous care of patients, as close as possible to their homes" [MSS1]. Beyond the lexical evolution, this represents a paradigm shift that implies a need to move away from the compartmentalization of the hospital, social medicine, and liberal medicine worlds. According to that definition, these domains should no longer be studied separately: the approach should be more inclusive and patient-centered. The different successive steps of the medical care the patient receives should be studied together.

Previously, the patient pathway could also refer to the succession of steps in the handling of a patient within a hospital [Barr96]. This may refer to a sequence of procedures (e.g. clinical examination, laboratory dosage, biopsy, and then surgery), a sequence of clinical stages (e.g. inflammatory, proliferative, and maturation phases), or a sequence of medical units (e.g. emergency unit, surgery, intensive care, conventional hospitalization, rehabilitation care, and then housing unit).

Thus, the concept of patient pathway has corresponded to several different scales: inside a hospital, the succession of touchpoints between a patient and the health care system, various geographic divisions (e.g. regions and departments), and now the territorial scale introduced by the GHT reform (for *Groupement Hospitalier de Territoire*, a regulation that requires public hospitals to group together on a geographical basis) [MSS2].

Through several case studies, the ClinMine ANR project focused on several of these scales: the scale, inside a hospital, of a succession of medical units forming a patient stay; a patient's pathway among different hospitals within a same region; and finally the set of one patient's stays in one hospital. The last one is seen through the prism of the information available in the hospitals' databases.

Whatever the scale or granularity, a patient pathway through a structure or zone implies several successive *steps*. In a hospital, each step is linked with various information having different goals: administrative (identity, contact information, social security, etc.), medical (previous history, risk factors, diagnoses, procedures, drug administrations, etc.), medico-economic (expenses and billing), or related to the various logistic chains present inside hospitals in order to organize the care (patient mobility and housing, analyses, meals, laundry, staff, etc.). Each of these information types correspond to documents created in paper form, and, since the advent of hospital information systems in the last decades, numeric documents of various types:

- Text in natural language (e.g. discharge letters, operative report)
- Qualitative data (e.g. ICD10 diagnostic codes [Classif10], procedure codes, ATC drug codes [Classif12])
- Quantitative data, discrete or continuous (e.g. laboratory measures)
- Static or animated images
- Temporal series (e.g. electrocardiogram output)

Therefore, studying a patient pathway amounts to studying a composite data structure. This structure contains a set of invariant and permanent data concerning the patient; these data have a very low (or null) probability of changing over time (e.g. birth date, blood type). It also contains a sequence of timestamped steps and, for each of these steps, a non-uniform set of data of various types. Non-uniform means that all inpatient stays do not have the same information: some information may be missing. It is worth noting that the timestamping of each step is not always available. For instance, chronic diseases may be specified in discharge letters or through ICD10 coding, but the start date of the disease is usually not documented. Therefore, the model must be able to indicate if an event took place before or after another one.

1.2 - Aims of the project

In this context, the unprecedented availability of hospitals data gives the opportunity to improve decision-making and to discover best practices for healthcare delivery [Denton13].
The ClinMine project originated from the observation that no method (statistical, algorithmic nor hybrid) was available to handle this specific type of sequential composite heterogeneous data, with a non-fixed structure and likely to suffer from completeness issues.
ClinMine's ambition is to propose, from various case studies, algorithmic and statistical models to handle this type of pathway data, focusing primarily on hospital data.

1.3 - Article content

This article presents two of the case studies of the ClinMine project: the study of the management process of the hospital medical mail and the construction of biclusters of patients in order to identify patient's profiles.
Indeed, the development of information and communication technologies in hospitals allows the installation of new processes such as digital dictation [Hansske06]. This is what is set up in the 40 services of the GHICl. In this context, mail can become a tracer element of the various intra-hospital care pathways of patients. Because of this, we can follow the chronology of dictating the various documents. Indeed the French organization makes that there is a strong relationship between the medical secretary and the doctors through the use of digital dictation. As a result, the various documents dictated in these tools make it possible to highlight the chronology of patient care through the succession of the various document dictation steps concerning the patients in care. This includes medical notes, biology reports, radiology reports, surgical reports of surgery, internal consultation reports by other specialists for surgery account of the specialty hosting the patient and to finish the mail of exit. This list thus shows the existence of a succession of stages which make it possible to retrace the different paths of the patients inside the hospital.

As the study of the dynamic of patient's pathway needs to take into account time stamps of data, the integration of temporal data was a key point of the project. It will be particularly discussed within this article. Hence, the next section will present state of the arts regarding first the analysis of temporal data, and in particular functional data, and then the biclustering approaches for heterogeneous data. In Section 3, entitled Results, the presentation of the methods developed during the project and their applications to hospital data are exposed. Then, in s-Section 4 a discussion on the difficulties and the results obtained is given.

2. State of the art

This section focuses on two important aspects dealing with the integration of temporal data within analysis, and present state of the art on the subject. It will then, first, define and make a literature review on functional data analysis and then in a second time on bi-clustering.

2.1 – Functional data analysis

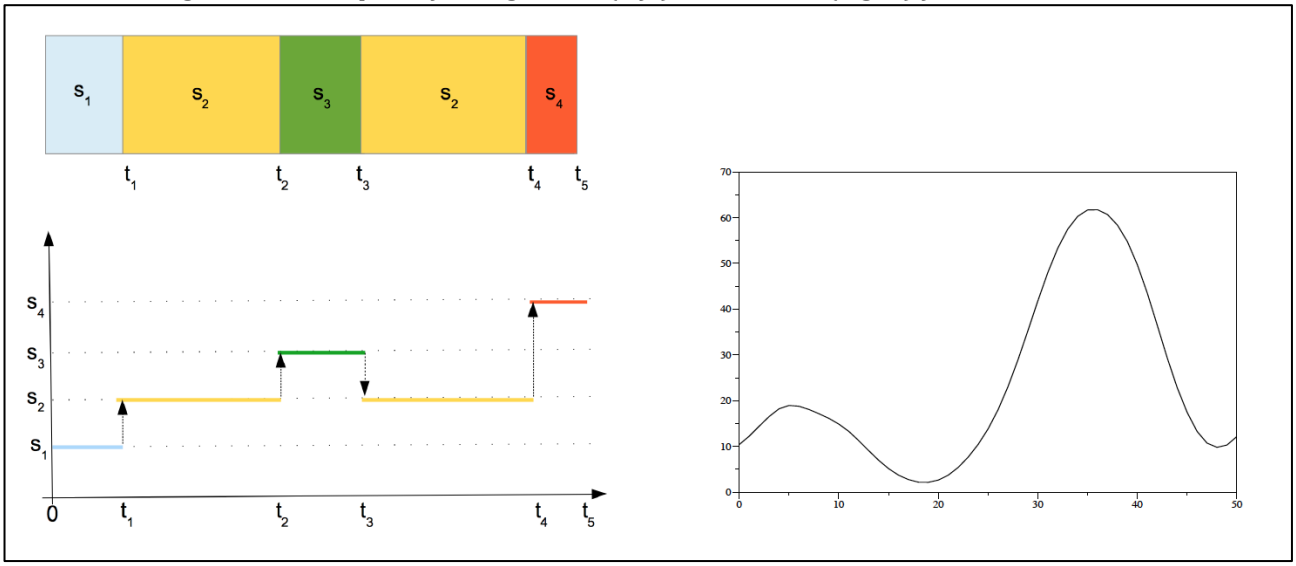
2.1.1 - Definition of functional data

Functional data analysis considers data depending on time. At a fixed time point, data can be represented as a random vector whose components are real random variables. Overtime that data can be seen as a set of curves or functions. Among a considerable record of papers on the subject, the monographs of Ramsay and Silverman [Ramsay02], [Ramsay05] and Ferraty and Vieu [Ferraty06] still remain references presenting the main methodologies for visualisation, denoising, classification and regression when dealing with functional data represented by real-valued random variables. What about the categorical case? If we think to a patient at hospital and we want to model his path (units visited during his stay), this is clearly a kind of categorical functional data.

In order to model such a path, we consider, in this project, the case where the underlying stochastic model generating the data is a continuous-time stochastic process $X = \{X_t, t \in T\}$ such that for all $t \in T$, X_t is a categorical random variable rather than a real-valued one. Let (Ω, A, P) be a probability space, $S = \{s_1, \ldots, s_m\}$, $m \geq 2$, be a set of m states and $X = \{X_t ; t \in T, X_t \in S\}$ be a family of categorical random variables indexed by T. Thus, a path of X is a sequence of states s_{ij} and times points t_i of transitions from one state to another one: $\{(s_{i1}, t_1), (s_{i2}, t_2), \ldots\}$, with $s_{ij} \in S$ and $t_i \in T$.

We call the sample paths of the process X categorical functional data. Figure 1 presents graphically scalar and categorical functional data.

Figure 1: Examples of categorical (left) and scalar (right) functional data.



2.1.2 - State of the art on categorical functional data

To the best of our knowledge, and quite surprisingly, there is no recent researches devoted to categorical functional data despite its ability to model real situations in different fields of applications: health and medicine (status of a patient over time), economy (status of the market), sociology (evolution of social status), and so on. As a start point in research on this topic we consider the works in [Boumaza80], [Deville82], [Deville83] and [Saporta81]. These works are devoted to the extension of factorial techniques (canonic and multiple correspondences analysis) towards functional data. Applications of these techniques are presented in [Heijden97] for analyzing career data and in [Preda98] for studying spectral properties of the transition probability matrix of the stationary Markovian jump process with continuous time.

2.1.3 - Clustering categorical functional data

Given a sample of paths (trajectories) of X, we present a model-based methodology of clustering categorical functional data. Instead of the classical setting considering a fixed length of the paths of X, i.e. the process is observed over a fixed length of time $T = [0, T]$, $T > 0$, we consider that the process X has an absorbing state and thus, we allow sample paths of different lengths. In the Markovian framework, based on the likelihood function, we derive an EM algorithm for clustering categorical functional data. An application on clustering medical discharge letters according to their status of dictating, type-writing and delivery to the end-user (patient or medicine) is presented in Section 3.

2.2 - Biclustering to identify patient pathways

2.2.1 - Biclustering definition

Biclustering can be seen as *two-way clustering*, where clustering is applied to both the rows and columns of a data matrix. Biclustering methods produce a set of *biclusters*, each corresponding to a subset of the rows and a subset of the columns of the data matrix. The goal of such methods is to group rows that are similar according to *some* of the columns, and vice-versa. In the context of the project, a row (or instance) in the data represents a patient, and a column (or attribute) represents information available on this patient. Since biclustering is an unsupervised method, it is best applied to data without any precise question in mind, in order to discover new, unexpected insight from the data. Biclustering on medical data may be used for various applications. The simplest one, although already quite useful, is the identification of groups of patients amongst the hospital's "customers": patients sharing a common subset of attributes among their personal information or their healthcare history, or both. Including temporal information allows identifying common patient pathways.

2.2.2 - Designing Biclusters

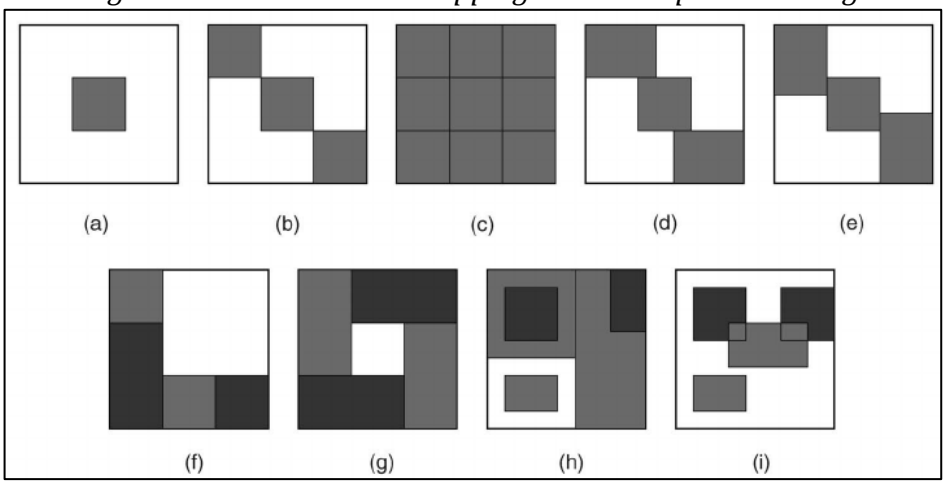
Several types of biclusters exist (similar values, similar row values, similar column values) [Pontes15], and biclustering algorithms are usually designed to detect a subset of these existing types. This design choice depends on the data, context and application. In this work, we use medical data and try to detect sets of patients or hospital stays sharing the same characteristics for some of the attributes. Therefore, we want a method able to detect biclusters with similar values on columns.

Another design choice concerns the overlapping (or lack thereof) between biclusters and the data matrix coverage. More precisely, this choice can be split into three questions:

- Can the biclusters have row-overlapping? That is, can two (or more) biclusters include the same row(s)?
- Can the biclusters have column-overlapping? That is, can two (or more) biclusters include the same column(s)?
- Does the whole data matrix need to be covered by biclusters? Some biclustering methods aim to cover the data space in what could be seen as a checkerboard structure. Others extract interesting clusters and may leave regions of the data space unused because deemed less interesting.

Figure 2 shows several types of overlapping biclusters. Types (a), (b) and (c) allow no overlapping. Type (d) has overlapping on columns, and type (e) has overlapping on rows. Type (i) corresponds to the more general case; where overlapping is allowed on rows and columns, and the biclusters do not need to cover the whole data. As previously, the adequate design choice must be motivated by the data, context and application. We consider in this study the case illustrated by type (i). This choice is made because we suppose that one patient may belong to several interesting groups, and one attribute may be used to describe several of these groups.

Figure 2: Biclusters – overlapping and data space covering



2.2.3 - Literature review on biclustering

One important characteristic of the considered medical data is its heterogeneity, that is, the fact that it includes data of various types, both qualitative and quantitative.

Although many biclustering methods have been proposed since the seminal work of Cheng and Church on the subject [Cheng00], the overwhelming majority of those focus on biological data analysis: detection of gene expression, protein interaction, microarray data analysis, etc. Therefore, most biclustering algorithms are designed to handle only one type of data, either numeric or binary (see [Pontes15] for a recent review on the subject). One algorithm, SAMBA (Statistical Algorithmic Method for Bicluster Analysis) [Tanay04], deals with heterogeneous data by integrating data from different sources. It remains, however, focused on numeric data and does not handle other data types.

As mentioned in [Busygin08], biclustering has also been applied to other data mining problems, such as text mining [Dhillon01] or recommendation systems [Yang02]. However,

application to hospital data, such as proposed in the current study, has not yet been explored in the literature.

Another characteristic of the data considered here is the inherent temporal aspect of many of its constituents: acts and diagnoses are performed at specific times, patients go through various medical units during their stay, etc. Temporal data has been explored in the context of biclustering, notably by the study on Order-Preserving SubMatrices (OPSM) [Ben-Dor03]. However, as far as we know, none of the existing applications to data mining tasks involve heterogeneous and temporal data.

Besides heterogeneity, another peculiarity of medical data is its size: hundreds of thousands of records (patients) with thousands of acts or diagnoses. This high dimensionality is coupled with high sparsity in the data matrix, as only a small fraction of all possible medical events happen during a given hospital stay. Recent studies such as [Alqadah15] are starting to explore biclustering applications on such large and sparse data, but do not consider heterogeneous data.

3. Results

3.1 Data and data warehouse

As mentioned before, since the advent of hospital information systems in the last decades, an increasing number of numeric documents of various types are available. The data used for this retrospective study issues from the activity of the 750-bed Lille Catholic Hospitals (St-Philibert and St-Vincent-de-Paul hospitals). Since January 2012 it represents 1,081,477 documents (letters, reports,…) and 486,169 stays.

However, data available for a patient is split into several software. For example, one software will be dedicated to letters, while another deals with stay-related data or others contain lab results, demographics data or patient moves. Thus, a data warehouse was built, and fed with these different sources. The following analyses are made on subsets of this data warehouse.

3.2 - Mail workflow analysis

3.2.1 - Hypothesis

In this first case study, the evaluated hypothesis is that some aspects of the patient pathway can be described, even predicted from the management process of the hospital medical mail. Indeed, the pathway of every hospitalized patient systematically leads to the production of documents. This production follows several steps described thereafter. Each of these steps involves different actors and requires the availability of specific information and means. The studied hypothesis is that the progress of each step of the mail circuit reflects the actual availability of those information and means. A deviation from the normal course of this process would be a clue of a dysfunction in the management of the internal patient pathway, the norm being also to be defined. Except in special cases, the entire patient pathway in the hospital is covered by the mail circuit; the mail process being initiated at the patient admission. Each mail step corresponds to a key organizational, which poor progress may

indicate a dysfunction in the patient pathway (e.g. late dictation of the mail that can be related to a delay in the production of examination results or the absence of the doctor, an understaffing of the secretariat, technical difficulties, and unavailable coordinates of the corresponding doctor).

3.2.2 - Material and method

The analysis was based on a dataset corresponding to the execution traces of the internal management tool of the various states of mail used at the Lille Catholic Hospitals.
Each mail goes through several states described Table 1.

Table 1: Different states for mail

State number	**State title**	**Comment**
0	« to dictate »	A new cycle has begun. This cycle is initiated at the patient's entrance to the hospital.
1	« being dictated »	The mail is dictated by the doctor in charge of the stay.
2	« to type »	The mail has been dictated and must be entered in the text editor. It appears in the task list of the concerned department secretaries.
3	« being typed »	The mail task has been assigned to a secretary who inputs it in the text editor.
4	« to validate »	The input is done, the mail appears in the list of the dictating doctor or the persons who received his delegation.
5	« being validated »	A re-reading is done by the doctor, who possibly completes the mail (e.g. analysis results obtained in the meantime).
6	« validated / pending release »	The mail is validated and can be sent, either by paper and mail, or secure e-mail.
7	« being released »	The mail is being sent.
8	« completed »	The process is completed.

The period studied extends from January 2012 to May 2016.
Each of the following pathway results in a mail workflow:

- 345,321 consultations
- 566,065 hospitalizations and other (discharge letter, operative report, examination report)

That is a total of 911,386 pathways whose documents were traced by the internal management tool of the various states of mail.
Each “individual” corresponds to the pathway of a hospitalization document, i.e. a sequence of steps
Figure 3 shows an example of such a “trajectory” taken by a mail.

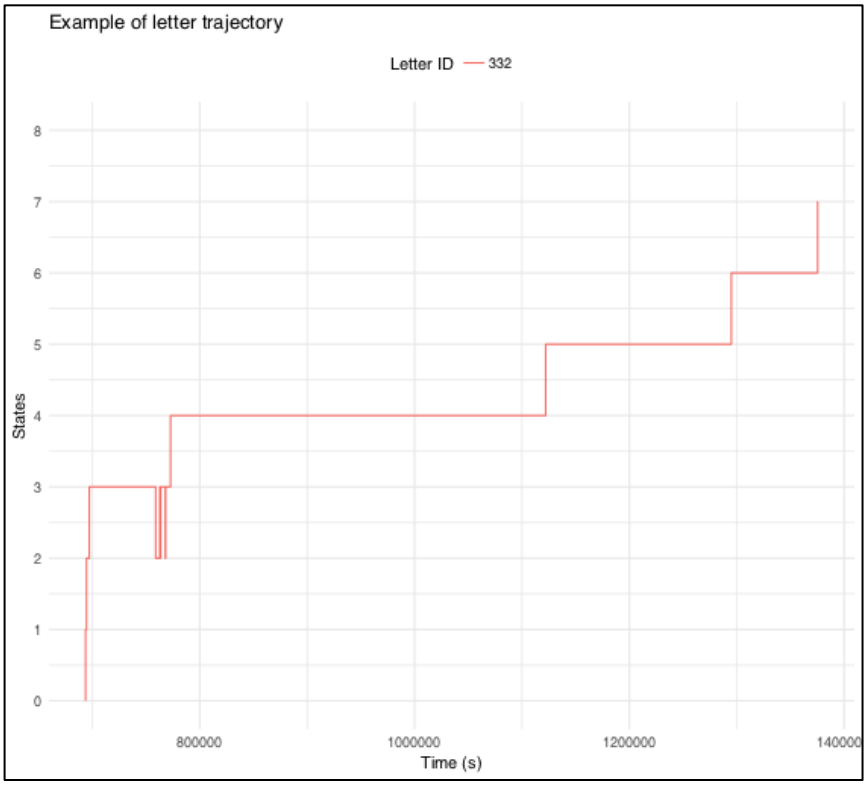


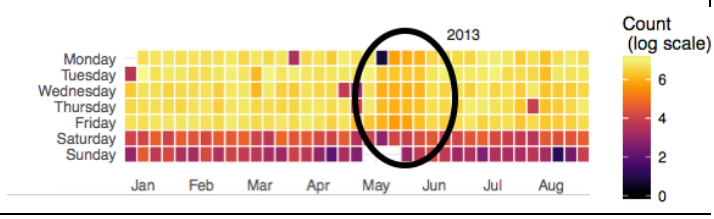
Fig. 3: Example of the sequence of steps for the path of mail 322.
The steps 0,1,2,3,2,3,2,3,4,5,6,7 are successively reached.
The total time is approximately 137 000 seconds, or about 1.58 days.

The data were enriched with co-variables related to the coding of the stay in the medico-economic classifications used in France: DRG (diagnostic-related groups) and ICD10.

3.2.3 - Analysis

A first empirical approach tried to associate the occurrence of proven, punctual or repeated, dysfunctions that are documented elsewhere (e.g. in an incident follow-up tool) with traces of the concerned stays in the management tool of the various states of mail studied here.
Lots of associations have been done, not all are detailed Table 2.

Table 2: Selection of some highlighted dysfunctions

Dysfunction found	**Trace in the analysis of mail workflows**
Migration of a computer tool over several weeks in May 2013, having severely disrupted the functioning of hospital services	An abnormally low number of mails managed, or a treatment time abnormally long, are noted as indicated by the surrounded area in the following heatmap. 

Progressive modifications of the mail management system itself over two months.	The measurement system being perturbed, it is obvious that an abnormally low number of mails managed or an abnormally long treatment time are noted. However, this result is still interesting, the observation of the deviation to the norm in the traces of the mail management tool can now be used to observe finely its operation, service by service.
Selection of stays in unfavorable deviation from the average length of stay expected for this DRG, excluding holiday and weekend effect.	The unitary mail workflows corresponding to these stays effectively deviate from the norm at several levels: sequence of steps, duration of one or more steps. In this sample, the observation of the mail workflow indicates a dysfunction in the management of the patient pathway.

Through this work, a sample of out of standard or dysfunctional patient paths has allowed the GHICL to observe a corresponding set of mail workflow reports also presenting an unusual profile.

3.2.3 - Prospect

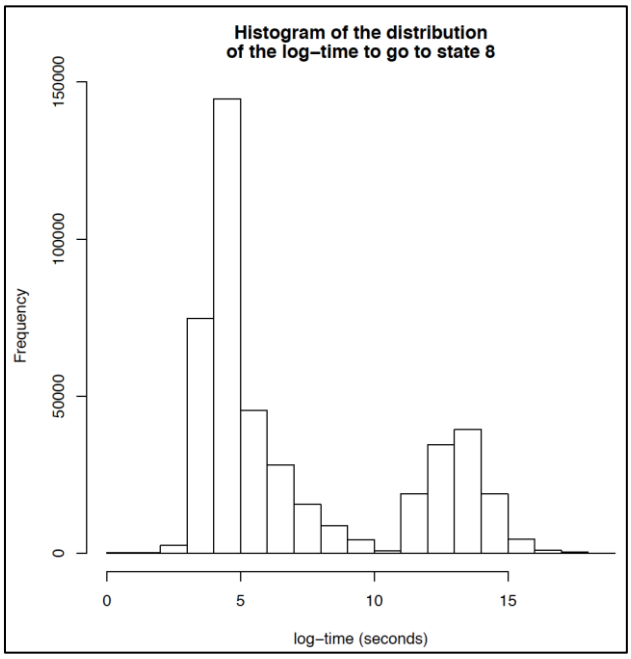
With this task, references have been constituted by CMD (1st and 2nd GHM letter (French DRG)), by the first letter of ICD10 diagnostic, by step, by service, by day for the median and mean times and for the variation of those times. This constitutes an internal standard with which each mail workflow could be confronted, either in real time (by service) or slightly delayed (just after the stay, when the coding is complete). A more detailed analysis will be carried out on all the co-variables to identify factors that can predict, as soon as possible in the mail workflow, the sign of a dysfunction in management of the patient pathway.

However, this first approach puts in light the interest of using such mail data to analyze patient pathways. Within the ClinMine project, another approach has been proposed, to deeper analyze this data and to identify patterns of pathways. This has been done using clustering.

3.2.4 - Clustering mail paths using functional data analysis

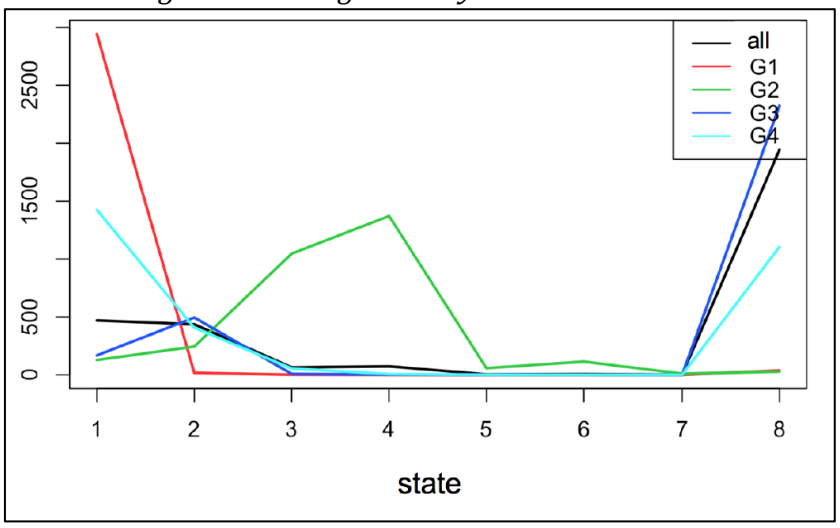
Clustering goal is to find clusters of mails having the same "pattern" of transitions between the different states (0 = "to dictate" to 8 = "complete"). Since the trajectories have different lengths, we consider that the model underlying data is a mixture of Markov continuous-time processes over the set of states {0, 1, …, 8} with initial state 0 and the absorbing state 8. We developed an EM-algorithm in order to estimate the parameter of the mixture [Preda15a].
As a sample set we use the set of 345 321 consultation letters. In order to avoid artifacts in building clusters, we considered that the time in the initial state is the same for all letters (0). Thus, clusters will describe the dynamics of the letters within services once they have been considered to be processed.
It is interested to note that the length of the trajectories (time to reach the state 8) is a bimodal distribution (see Figure 4), showing the presence of some events delaying some letters (personal change, break during the nights, etc.).

Figure 4. Distribution of log (length), cut-off at exp(10) seconds (~ 6 hours)



That behavior still exists if we consider only the length of time spent by a letter in some given state. Using the BIC criterion, we have obtained 4 clusters: Cluster 1 of size 37,783 letters, Cluster 2 of size 23,159, Cluster 3 of size 260,535 and Cluster 4 of size 23,844. For each cluster, denoted by G1, G2, G3 and G4, the Figure 5 presents the mean of times spent in each state by the letters within groups.

Figure 5: Average time by state and cluster



This figure shows that, cluster 1 (G1) is represented by letters for which the time to be dictate is larger than in the other groups. Once the dictating process is completed the letters of this group reach rapidly the state 8 (process completed). The main characteristic of cluster 2 (G2) is that the letters of this group spend more time in the states 3 (typing) and 4 (validation). The cluster 3 is the largest cluster and it has an average behavior. We note however that the average time to reach the state 8 is the largest for this cluster. Like the

Cluster 1, the Cluster 4 (G4) is also a cluster for which the dictate time is large and the time to reach the state 8 larger than that in clusters 1 and 2.

3.2.5 - Discussion

This analysis allows identifying several typologies of letters that help to understand the internal organization of the hospital and may lead to some ameliorations.
G1 consists of several different behaviors of physicians using their digital dictation equipment. Working hypotheses are: lengthy letters typical of certain medical specialties; dictation is carried out in several stages; digital dictation workflow management software manages session without closing them automatically. G4 follows the same pattern, but on a smaller scale.
G2 seems to regroup letters that need several roundtrips between physicians and secretaries, to accurately validate the content.
G3 regroups the normal and expected running of the digital dictation process.

3.3 - Bi-clustering to identify patient pathways

The second case study of the project, presented here, deals with the analysis of patients pathways through BiClustering.

3.3.1 - The ClinMine methodology

In the ClinMine project, the bi-clustering problem is dealt as a bi-objective combinatorial optimization problem [Dhaenens16]. Indeed, selecting a subset of patients that share a subset of common characteristics is by definition a combinatorial problem. Moreover, it is possible to evaluate the quality of the biclusters regarding several criteria. We chose in our approach to consider two objectives; the quality (that measure the homogeneity of the bi-cluster) and the size (larger bi-clusters more interesting than smaller ones).
Regarding the size of the problem (number of patients and number of possible characteristics), it is unlikely to generate all the possible bi-clusters.
Hence, in this project we study two different approaches: a metaheuristic and a constructive heuristic.

3.3.2 - A bi-objective local-search for bi-clustering – LS-HBC

LS-HBC (Local Search-based Heterogeneous BiClustering) has been designed to extract biclusters from heterogeneous, large-scale, sparse data matrices [Vandromme17]. Experiments, conducted on synthetic data, in which we manage different parameters, such that the size of the bi-clusters to find, the noise within bi-clusters..., show that the proposed algorithm, LS-HBC, is able to generate large biclusters of high quality, outperforming three other standard biclustering methods from the literature in this regard. However, LS-HBC fails to achieve diversity in the generated solutions, and tends to fill the archive with biclusters very close to one another. The method has been tested on a real data set, extracted from a hospital's information system. On this data, LS-HBC is able to detect small but relevant biclusters, describing real groups of patients.

3.3.3 - A constructive heuristic - HBC

HBC (Heterogeneous BiClustering) has also been designed to extract biclusters from heterogeneous, large-scale, sparse data matrices. HBC takes advantage of the data sparsity and uses a constructive greedy heuristic to build a large number of possibly overlapping

biclusters. Moreover, a particular attention is given to the underlying temporal aspect present in a large part of medical data [Vandromme16a]. Results on synthetic data show that the proposed method outperforms several other biclustering methods, and is the best suited to generate solutions of good quality and to recover biclusters hidden in the data. The model is extended to handle temporal events, with slight modifications to the algorithm. Comparison with OPSM [Ben-Dor03], the standard method for temporal data, shows that HBC performs consistently better. Experiments on heterogeneous real-life data show that the method scales well on large sparse sets, exhibiting good performance and reasonable runtime.

3.3.4 - Examples of biclusters found on medical data

We run the biclustering process on real-life medical data sets as extracted from a hospital's information system. These data sets form a good sample of the data we wish to be able to deal with, exhibiting heterogeneity, high sparsity and large size [Jacques15]. Dataset MRB comes from a case study for methicillin-resistant bacteria, while PMSI and PMSI 2013 are more generic and gather information on hospital stays (acts, diagnoses, prescriptions, medical units, etc.). Note that datasets PMSI and PMSI 2013 both contain a large majority of temporal attributes (medical events), and only *age* and *gender* as non-temporal. On the other hand, MRB includes a few dozens numeric and unordered symbolic attributes, describing additional data on antibiotics. Table 3 presents the main characteristics of the three studied data sets.

Table 3: Characteristics of real datasets used

	Number of rows	Number of columns	Number of non-missing	% non-missing
MRB	194,715	11,450	2,584,252	0,12%
PMSI	102,896	11,413	1,413,057	0,12%
PMSI_2013	49,231	7,941	746,566	0,19%

We present here some biclusters found on the real data sets, and show that these make sense from a medical point of view.

[Bicluster1] Extracted from the MRB data set,
contains 55 patients and 4 attributes:

- {Medical unit 1} Infectious and parasitic diseases
- {Medical unit 1} Prophylactic (preventive) isolation
- {Medical unit 1} Staphylococcus aureus
- {Medical unit 1} Methicillin-resistant agents

This bicluster is fairly straightforward to understand; it gathers the patients having methicillin-resistant staphylococcus aureus, who have been placed in isolation.

[Bicluster2] Extracted from the PMSI data set,
contains 1102 lines and 4 attributes:

- {Medical unit 1} Diabetes mellitus
- {Medical unit 1} Resting 12-lead electrocardiography
- {Medical unit 1} Retinography in colored or monochromatic light
- {Medical unit 1} Check-up after treatment for other affections

This bicluster includes diabetic patients. The retinography is performed since diabetes often causes eye problems; the electrocardiography is performed because coronary diseases are the main comorbidities associated with diabetes.

In the two examples shown so far, all the events happen at the same time (i.e. in the same medical unit). One last example [Bicluster3] is presented to exhibit the algorithm's ability to build biclusters with a true sequential aspect.

[Bicluster3] Extracted from PMSI data set,
contains 598 lines and 3 attributes:

- {Medical unit 1} Thorax radiography
- {Medical unit 2} Hormonal, nutritional and metabolic diseases
- {Medical unit 2} Circulatory system diseases

In this bicluster, the radiography happens **first** in medical unit 1, then the two diagnoses **after** in another medical unit : medical unit 2. This makes sense, since a radiography is often performed as a first act, before directing a patient to the appropriate unit. This example shows how biclustering may identify patient's flows within hospitals, as soon as temporal data are managed.

Regarding the results exposed, the combinatorial approach proposed to deal with biclustering of patients flows data has been validated and manage to extract useful information to identify common characteristics of patients flows.
It is therefore envisaged to test this model on datasets coming from one other similar hospital: the public hospital of Valenciennes, that has a similar size.

4. Discussion

The presentation of these two case studies leads to several conclusions and discussions. They are reported here regarding several point of views, in order to point out several types of challenges.

From a statistical point of view, data heterogeneity and the high dimension are the main issues on the statistical analysis. The dimension of data is closely related to the temporal feature, thus the stochastic process based models such the Markov processes are of big interest. However, it is not sure that the Markov behavior (given the past, the future depends exclusively on the present) is a realistic hypothesis in the framework of medical data. Our guess is that is not the case. More appropriate models (eg, long memory processes) must be investigated. Merging data analysis methods and stochastic models are, in our opinion, the challenge for this kind of investigations.

Regarding the optimization point of view, the heterogeneity and in particular the temporal feature is also a great challenge. It requires the development of specific methods and may also require pre-treatment of data. The impact of these pre-treatments has to be evaluated. In the ClinMine project, such an approach is proposed (efficient pre-treatments are exposed) and has shown the ability of optimization methods to deal with such data. Now, the next challenges are to be able to always add some new types of data and to develop methods able to deal with such a diversity.

From a medical point of view, the two case studies give to the Lille Catholic Hospitals a better view of their mail workflow and patients' profiles. Hypothesis generated by both methods can now be checked through the classic way. Some results already allowed to improve ICD10 encoding or detection and profiling of groups of patients. In this work, the

“patient pathway” was in a stay scope. Data about previous stays was not analyzed because the patient could have stay in others hospitals: a previous stay is found for about 10% of stays at the Lille Catholic Hospitals. One perspective could be to conduct this work on a national data warehouse when it will be available. This would allow a greater scope than stay-scope.
Regarding the biclustering, HBC finds a lot of bi-clusters (more than 1000), making harder to identity the useful bi-clusters and profiles. Methods to select or filter the most appropriate bi-clusters to show to the decision maker still need to be investigated. Another challenge resides in proposing new user interfaces dedicated to display the obtained bi-clusters and navigate between them.

Beyond the two case studies presented here, the results of the ClinMine project are the support of new challenges for medical practice.
Firstly, the actual disease classification is often insufficient to show the complexity of clinical situations. The ClinMine methodological approach should offer opportunity to improve the possibility of patient stratification. In future, patient stratification will be related to demographic data, to clinical symptoms and evolution, to biological or genomic informations, to physiological and imaging parameters. This stratification is able to identify patients cluster with common characteristics eligible to a personalized treatment. This strategy is already available for some diseases with a predominant factor of susceptibility to a specific treatment such some subtypes of cancer. Nevertheless, this stratified approach should be generalized in more complex situation with multiple pathways. For instance, the cognitive disorders (i.e. memory impairment, attention trouble, executive dysfunction) are related to several pathological conditions: neurodegenerative diseases (in particular Alzheimer’s disease), inflammatory diseases (such multiple sclerosis), neurovascular diseases like stroke, vascular risk factors (hypertension, diabetes, hypercholesterolemia), drugs… These cognitive disorders are developing during several years, with consultation or hospitalization in different medical departments. The clustering tools developed in ClinMine would be able to propose a new stratification of cognitive disorders with the possibility to develop specific therapeutic approach in each cluster [Bordet17].
Secondly, the so-called big data is postulated to offer new opportunity to the resourcing of pathophysiological and therapeutic hypothesis in medicine, by identification of new associations between symptoms, paraclinical parameters, drugs, environmental factors, even though these new hypothesis should have to be confirmed or developed through classical translational and clinical research. This big data approach could be particularly relevant to detect weak signal, as observed in management of Adverse Drug Events. Indeed, in this case, the association between a symptom or a syndrome and the intake of drug can be difficult to detect when it is a long term effect, such observed in amphetamine-induced valvulopathy or valproic acid-induced brain malformation. Hospital database, in addition to others, could be accurate to detect such signal by application of ClinMine methodology.
Thirdly, the inclusion in clinical trials remains an important goal for medical progress. The ClinMine tool should propose a solution to optimization of patient identification fitting with more precise inclusion criteria related to patient stratification. Coupled to automatic detection systems, a screening of patients using pathways identified by ClinMine should provide rapidly and precisely the population necessary for a high-quality clinical research, improving the capacity to conclude and sparing required budgets [Bordet15].

5. Conclusion

ClinMine's ambition was to propose, from various case studies, algorithmic and statistical models to handle patient pathway data. Two of these case studies were exposed in this article. Both of them give special attention to temporal data, as it is necessary to understand patient pathway.
The first case study examined if some aspects of the patient pathway can be described, even predicted from the management process of the hospital medical mail. Hence execution traces of the internal management tool of the various states of mail have been studied and a clustering realized. Clusters obtained indicate some dysfunctions that could help to better organize services in order to optimize patients in hospital.
The second case study aimed at grouping patients sharing some similar characteristics. A biclustering approach is proposed and methods developed managed to produce interesting biclusters.
These two case studies, and other from the project, validate the approaches proposed in the project ClinMine.
Along the project, one of the difficulties was to be able to discuss between partners and in particular to understand what data and results proposed could represent. Thus the visualization of data appeared to be an important issue for such a project, but is also a difficult part because of the heterogeneity of data and the consideration of temporal data.

Acknowledgments

This work was supported by the French National research Agency [ClinMine ANR-13-TECS-0009].

6. References

[Alqadah15] Faris Alqadah, Chandan K. Reddy, Junling Hu, and Hatim F. Alqadah. ***Biclustering neighborhood-based collaborative filtering method for top-n recommender systems***. Knowledge and Information Systems, 44(2): 475-491, 2015.

[Barr96] Barr JE, Cuzzell J. ***Wound care clinical pathway: a conceptual model***. Ostomy Wound Manage. 1996 Aug;42(7):18-24, 26.

[Ben-Dor03] Amir Ben-Dor, Benny Chor, Richard Karp, and Zohar Yakhini. ***Discovering local structure in gene expression data: the order-preserving submatrix problem***. Journal of computational biology, 10(3-4):373–384, 2003.

[Bordet15] Bordet, R., Lang, M., Dieu, C., Billon, N., & Duffet, J. P. (2015). ***Early results from a multi-component French public-private partnership initiative to improve participation in clinical research–CeNGEPS: a prospective before-after study***. BMC medical research methodology, 15(1), 67.

[Bordet17] Bordet, R., Ihl, R., Korczyn, A. D., Lanza, G., Jansa, J., Hoerr, R., & Guekht, A. ***Towards the concept of disease-modifier in post-stroke or vascular cognitive impairment: a consensus report.*** BMC medicine, 15(1), 107, 2017.

[Boumaza80] Boumaza ***R. Contribution à l'étude descriptive d'une fonction aléatoire qualitative***, Thèse de 3ème cycle, Université Paul Sabatier, Toulouse, France, 1980.

[Busygin08] Stanislav Busygin, Oleg Prokopyev, and Panos M Pardalos. ***Biclustering in data mining.*** Computers & Operations Research, 35(9):2964–2987, 2008.

[Cheng00] Yizong Cheng and George M Church. ***Biclustering of expression data***. In Intelligent Systems for Molecular Biology, volume 8, pages 93–103, 2000.

[Classif10] World Health Organization. ***International Statistical Classification of Diseases and Related Health Problems*** [Internet]. 2010 [cited Sept 6, 2017]. Available on: http://www.who.int/classifications/icd/en/

[Classif12] World Health Organization. ***Anatomical Therapeutic Chemical classification*** [Internet]. 2012 [cited Sept 6, 2017]. Available on: http://www.whocc.no

[Denton13] Brian T. Denton. ***Handbook of healthcare operations management***. New York: Springer, 2013

[Deville82] Deville J.C. ***Analyse de données chronologiques qualitatives : comment analyser des calendriers ?,*** Annales del'INSEE, No. 45, 45—104, 1982.

[Deville83] Deville J. C., Saporta G. ***Correspondence analysis with an extension towards nominal time series,*** Journal of Econometrics, 22, 169—189 , 1983.

[Dhaenens16] Clarisse Dhaenens, Laetitia Jourdan, ***Metaheuristics for Big Data*****,** 212 pages Wiley-ISTE, ISBN-13: 978-1848218062, ISBN-10: 1848218060, 2016

[Dhillon01] Inderjit S Dhillon. ***Co-clustering documents and words using bipartite spectral graph partitioning.*** In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 269–274. ACM, 2001.

[Ferraty06] Ferraty F., Vieu P. ***Nonparameric Functional Data Analysis. Theory and Practice***, Second Edition, Springer Series in Statistics, 2006.

[Fisset14] Benjamin Fisset, Laetitia Jourdan, Clarisse Dhaenens. ***A framework for multi-objective Clustering***, Conference on Metaheuristics and Nature Inspired Computing, META 2014.

[Hansske06] Hansske Arnaud. ***Impact des technologies de santé sur l'organisation***, Revue Hospitalière de France, 512, 18—22, 2006.

[Heijden97] Heijden P.G.M., Teunissen J., van Orle C. ***Multiple correspondence analysis as a tool for quantification or classification of career data***, Journal of Educational and Behavioral Statistics, 22, 447—477, 1997.

[Jacques15] Julie Jacques, Julien Taillard, David Delerue, Clarisse Dhaenens, and Laetitia Jourdan. ***Conception of a dominance-based multi-objective local search in the context of classification rule mining in large and imbalanced data sets***. Applied Soft Computing, 34:705–720, 2015.

[MSS1] http://solidarites-sante.gouv.fr/systeme-de-sante-et-medico-social/parcours-des-patients-et-des-usagers/

[MSS2] http://solidarites-sante.gouv.fr/professionnels/gerer-un-etablissement-de-sante-medico-social/groupements-hospitaliers-de-territoire/

[Pontes15] Beatriz Pontes, Raúl Giráldez and Jesús S Aguilar Ruiz. ***Biclustering on expression data: A review***. Journal of biomedical informatics, 57:163–180, 2015.

[Preda98] Preda C., ***Analyse harmonique qualitative des processus markoviens de sauts stationnaires***, Scientific Annals of Alexandru Ioan Cuza University of Iasi (Romania), Computer Science Section, Tome VII, 5—18, 1998.

[Preda14a] Cristian Preda (2014), ***Categorical Functional Data and Space-time clusters of lymphoma in northern France –a scan statistics approach***, Symposium on Modern Biotechnological Advances for Human Health – BAHH May 20-22, 2014, Bucharest, Romania.

[Preda14b] Cristian Preda, Gilbert Saporta (2014). ***PLS regression for multivariate***

functional data, The 7th International Conference of the ERCIM (European Research Consortium for Informatics and Mathematics) Computational and Methodological Statistics (ERCIM 2014)

[Preda15a] Cristian Preda, Cristina Elena Preda, Vincent Vandewalle, ***Clustering categorical functional data***, The 8-th International Conference ERCIM on Computational and Methodological Statistics (CMStatistics 2015), 12-14 December 2015, London, UK.

[Preda15b] Cristian Preda, ***Regression models with categorical functional data***, The 8th International Conference ERCIM on Computational and Methodological Statistics (CMStatistics 2015), 12-14 December 2015, London, UK.

[Preda16] Cristian Preda, Vincent Vandewalle, ***Clustering categorical functional data***, 147th ICB Seminar, Tenth International Seminar on Statistics and Clinical *Practice, May 15 - 18, 2016, Warsaw, Poland*

[Ramsay02] Ramsay J.O., Silverman B.W. ***Applied functional data analysis. Methods and studies***. Springer Series in Statistics, Springer-Verlag, New York, 2002

[Ramsay05] Ramsay J.O., Silverman B.W. ***Functional Data Analysis***, Springer Series in Statistics, Springer-Verlag, New York, 2005.

[Saporta81] Saporta G. ***Méthodes exploratoires d'analyse de données temporelles***, Cahiers du B.U.R.O., No. 37-38, Université Pierre et Marie Curie, Paris, 1981.

[Tanay04] Amos Tanay, Roded Sharan, Martin Kupiec, and Ron Shamir. ***Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data***. Proceedings of the National Academy of Sciences of the United States of America, 101(9):2981–2986, 2004.

[Vandromme15] Maxence Vandromme, Julie Jacques, Julien Taillard, Laetitia Jourdan, Clarisse Dhaenens, ***Handling numerical data to evolve classification rules using a Multi-Objective Local Search***, Metaheuristics International Conference, MIC 2015

[Vandromme16a] Maxence Vandromme, Julie Jacques, Julien Taillard, Laetitia Jourdan and Clarisse Dhaenens, ***A scalable biclustering method for heterogeneous medical data***, Second International Workshop on Machine Learning, Optimization and Big Data - MOD 2016.

[Vandromme16b] Maxence Vandromme, Julie Jacques, Julien Taillard, Arnaud Hansske, Laetitia Jourdan and Clarisse Dhaenens, ***Extraction and optimization of classification rules for temporal sequences: application to hospital data***, Knowledge Based System, Vol. 122, pp. 148-158, 2017

[Vandromme17] Maxence Vandromme, Julie Jacques, Julien Taillard, Laetitia Jourdan and Clarisse Dhaenens, ***A local search-based multi-objective metaheuristic for biclustering on heterogeneous data***, Metaheuristic Internationale Conference, MIC 2017.

[Yang02] Jiong Yang, Wei Wang, Haixun Wang, and Philip Yu. ***δ-clusters: Capturing subspace correlation in a large data set***. In Data Engineering, 2002. Proceedings. 18th International Conference on, pages 517–528. IEEE, 2002.