Heterogeneous incomplete multi-view data for Neurotoxicity biomarkers Identification

Quentin Ruin[1], David Balayssac[2], Issam Falih[1], Engelbert Mephu Nguifo[1]

[1] Université Clermont-Auvergne, CNRS, Mines de Saint-Étienne, Clermont Auvergne INP, LIMOS, 63000 Clermont-Ferrand, France
[2] Faculté de Médecine, UMR INSERM 766, BP 38, F-63001, Laboratoire de Toxicologie, Clermont-Ferrand, France

quentin.ruin@uca.fr; dbalayssac@chu-clermontferrand.fr
issam.falih@uca.fr ; engelbert.mephu_nguifo@uca.fr

Résumé

Il y a un intérêt croissant pour évaluer les troubles neurologiques au stade précoce de la découverte de médicaments. Cependant, ces conditions sont difficiles à surveiller en raison d'une compréhension insuffisante de leur mécanisme sous-jacent et de leur nature plutôt asymptomatique. La découverte de biomarqueurs spécifiques et complémentaires pour la toxicité du système nerveux périphérique est donc très précieuse. Dans ce but, une nouvelle source de données est créée par nos partnenaires pharmacologue universitaires et industriels que nous analysons dans cet article. Nous proposons un flux de travail pour analyser, extraire et combiner les indicateurs prédictifs sur deux ensembles de données in vivo. Un ensemble de techniques d'apprentissage automatique et de fouilles de données ont été utilisées pour extraire des informations neuropathogènes d'une liste de biomarqueurs.

Mots-clés

Données hétérogènes, Imputation de données, Classification, Données incomplètes, Multi-vue non alignées, Identification des biomarqueurs, Neurotoxicité

Abstract

There is a growing interest in assessing neurological disorders in early stage of drug discovery. However, those conditions are difficult to monitor due to a weak understanding of their underlying mechanism and their rather asymptomatic nature. The discovery of specific and complementary biomarkers for Peripheral nervous system (PNS) toxicity is therefore highly valuable. For this purpose a novel data source is created by academic and industrial expert partners that we analyse in this paper. We propose a workflow to analyse, extract and combine the predictive indicators on two sets of in vivo data. A range of machine learning and data mining techniques were used to extract neuropathogenic information from the compiled list of biomarkers.

Keywords

Heterogeneous data sources, Data imputation, Classification, Neurotoxicity, Biomarker identification, Incomplete data, Non-aligned multi-view.

1 Introduction

Chemotherapy-induced peripheral neuropathy (CIPN) is a common adverse effect of neurotoxic anticancer drugs (e.g. platinum derivatives, taxanes, vinca alkaloids, bortezomib and thalidomide) on the peripheral nervous systems (PNS). CIPN symptoms correspond to a distal and symmetric neuropathy with paresthesia (tingling, numbness) and dysesthesia (neuropathic pain). CIPN may last several years after the end of anticancer therapy [4], and can profoundly decrease the quality of life of patients [12, 11]. No gold standard is clearly defined for screening and treatment of CIPN [3]. Peripheral nervous system toxicity (PNS Tox) is poorly predicted by the current in vitro and in vivo preclinical studies that are performed during the research and experimental procedures. The study presented in this paper is related to the NeuroDeRisk project (Innovative Medicines Initiative (IMI2)), whose aim is to provide novel, validated tools for improving the preclinical prediction of various effects of pharmaceuticals on the nervous system, and to de-risk drug candidates. In this paper we are solely focusing on the data perspective and the main contribution can be summarised as follow:

- Performing thorough analysis on two sets provided by the project of *in-vivo* data
- Performing pre-processing techniques as creating new features for temporal data, exploring several missing value techniques to generate missing values, heterogenous features (nominal, continuous, temporal data).
- The most influencing features were selected to build an in silico predictive model able to discriminate animal models.

- Established a correlation between different molecular, behavioral and histological endpoints.

The rest of the paper is organized as follows: Section 2 describes the dataset along with its challenges and issues. Section 3 presents the general workflow and outlines its modules along with the experimental results. Finally, Section 4 presents the conclusions drawn from the study and outlines directions for future work.

2 Dataset description, challenges & issues

The dataset was generated from the in-vivo experiments conducted as part of the NeuroDeRisk's project by 5 different academic and pharmaceutical partners (MSD, Novartis, SARD, Clermont-Auvergne University (UCA) and University of Florence (UNIFI)) on 7 different drugs (acrylamide, cisplatin, doxorubicin, Nova1, oxaliplatin, paclitaxel and vincristine) using 2 animal species (rat and mouse). As every partner did not necessarily sample the same parameters / features, the dataset is heterogeneous. Some partners used the same drug with different species, or different drugs with the same species, and used the same drug as another partner but with different parameters and species. At final, each partner provide two data sets one for histopathological data and the other for molecular data (behavior and biomarkers).

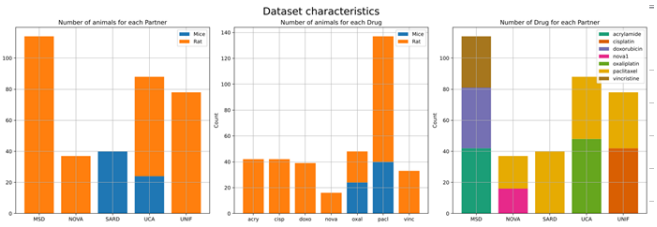

Figure 1: Animal repartition according to partners, drugs and species

Once aggregating all the molecular data and aggregating all the histopathological dataHowever, a second challenge arose due to the high dimensionality of the data. The experimental data provided by partners for the same drug did not share the same dimensions, and only a few features overlapped between the two datasets. The small number of animal studies, along with the class imbalance, is also a factor to consider, i.e., the headcount of available animals to carry out studies varies widely (from 16 to 96).

3 Heterogeneous incomplete multi-view data workflow

We present in Figure 2 the main steps to combine the histopathological and molecular data, the pre-processing and the process to predict the drug. The overall workflow can be described with three stages: data combination, data preprocessing & imputation and data analysis & building the drug classification model.

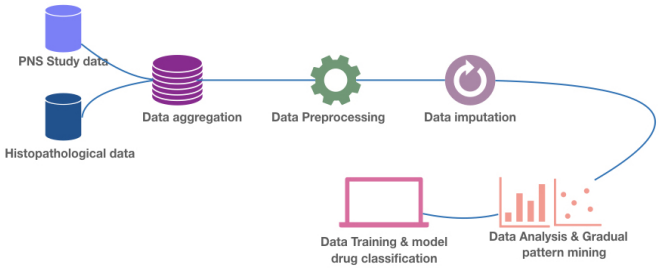

Figure 2: An overview of the overall Framework

3.1 Data aggregating

Each of the 5 partners provided two data sets, once for the histopatological data and the second describing the biomarkers along with the behavioral studies. Those sets of data can be seen as two aligned view, i.e. one to one relationship. By combining the two heterogeneous data sets of each partner, a new fused dataset is generated. The resulting data was highly dimensional (196 features).

3.2 Data Pre-processing & Data imputation

As data are heterogeneous from an experiment to another, the sparsity of the data is very high. To keep coherence and consistency, the dataset is split in subsets of animals with the same species and drug treatment (not necessarily from the same partner), then further cleaned. For each of these subsets, early sacrificed animals and features with more than half of missing values, constant values are removed. The missing values for features with less than half of missing values are filled by linear interpolation. Additionally, the weight of the animals appearing to be an important features of the impact of neurotoxic component on animal health, it is important that it is not influenced by the origin of the animals, their species or the duration of experiment they take part in. To ensure consistency two new features are created, the `weight evolution = (100*[final weight - early weight]/early weight, in %)` and the `relative weight = ([final weight - early weight]/experiment duration)`.

3.3 Data Analysis, Correlation and gradual pattern mining

In the initial exploratory phase, we are investigating the potential correlation between the levels of biomarkers and the grades of lesions observed on the nerves of the animal model. The endpoint is to have a first view of which biomarker could have an impact on the nervous degeneration, and to what extent. This study is done drug by drug, so as to identify possible drug-related behaviour. Two correlation was performed: correlation between a specific biomarker and histopathological lesions and correlation between biomarkers.

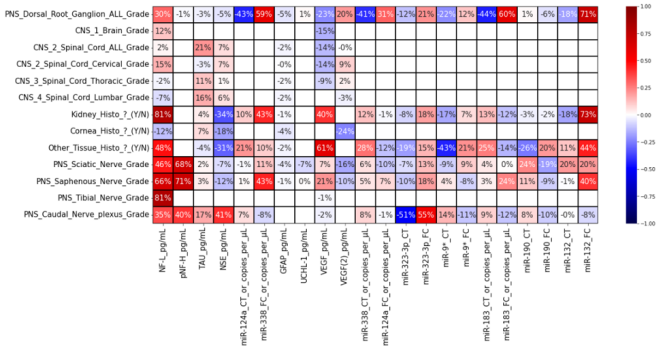

Figure 3: *Correlations between biomarkers and histopathology data*

This study is divided into two parts: calculating the absolute value of the correlations then plotting the features with the highest correlations to check their link, as a high correlation level does not necessarily mean a strong link. The blanks represent the pairs for which no data are available. To plot the highest correlations. To plot the highest correlations, we need to set a threshold of what we consider to be "high". A correlation of 0.8 is generally considered to be the minimum to assess that the correlation is strong, and 0.5 to asses that the correlation is medium.

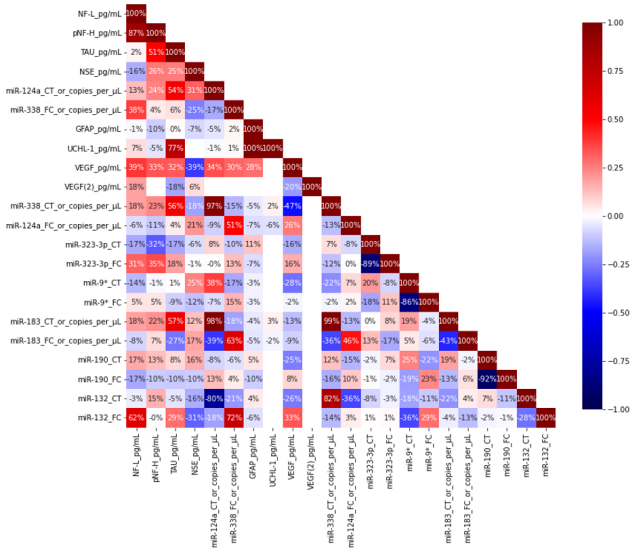

Figure 4: *Biomarker correlation matrix*

We can note that the biomarkers: pNF-H and NFL, miR-338-CT and miR-124a-CT, miR-183-CT and miR-124a-CT, UCHL-1 and GFAP, miR-183-CT and miR-338-CT, miR-323-3p-FC and miR-323-3p-CT, miR-9*-CT and miR-9*-CT, miR-190-CT and miR-190-FC are strongly linked.

Gradual patterns mining Pattern mining is a major area of data mining [5, 6, 7]. It consists in simple algorithms aiming to discover frequent co-variations of features, by counting for every set of features the direction of variation of each one and the frequency at which this co-variation appears. This process can be very time and resource consuming, as according to the number of features and individuals the number of sets can be very important. Hence, the number of outputs is also very important.

To avoid such computational and representation problem, we will here stick to the case where the sets of features are a couple biomarker/histopathology, biomarker/behaviour, dose/biomarker, dose/histopathology or dose/behaviour. Then, we will gather information about the impact of biomarkers on histopathology or the effect of the dose on the other features, which is our endpoint.

Here is an example with the individuals treated with Cisplatin drug.

Cisplatin (rats UNIFI), indivuals = 40			
Features: Dose_Level, VEGF_pg, mechanicalstimuli, Coldstimuli, weight_evolution, relative_weight, VEGF2_pg, BB_failure_last, Irwin_last			
Only dual biomarkers/histo links			
Biomarker	Histopathology	Occurence	Prop
VEGF2_pg+	BB_failure_last+	32	80%
VEGF2_pg+	mechanicalstimuli+	26	65%
VEGF2_pg+	Coldstimuli+	26	65%
VEGF_pg+	BB_failure_last+	25	63%
VEGF_pg+	mechanicalstimuli+	22	55%
VEGF_pg+	Irwin_last+	21	53%
VEGF2_pg+	Irwin_last+	20	50%
VEGF_pg+	Coldstimuli+	20	50%
VEGF_pg+	relative_weight-	16	40%
VEGF_pg+	weight_evolution-	15	38%
VEGF2_pg-	relative_weight+	15	38%
VEGF2_pg-	weight_evolution+	14	35%
Dose impact			
Dose_Level+	Coldstimuli+	28	70%
Dose_Level+	BB_failure_last+	27	68%
Dose_Level+	mechanicalstimuli+	25	63%
Dose_Level+	VEGF2_pg+	23	58%
Dose_Level+	VEGF_pg+	21	53%
Dose_Level+	Irwin_last+	18	45%
Dose_Level+	weight_evolution-	15	38%
Dose_Level+	relative_weight-	13	33%

Figure 5: *Patterns identified for the Cisplatin individuals. The symbol represents the direction of evolution, '+' being an increase and '-' being a decrease*

Here, we can see for example that for 80% of the animals, when the biomarker VEGF2 increase, there is a corresponding increase of the histopathological BB_failure_last, and for 40% of the animals, an increase of VEGF2 goes along with a decrease of the relative weight. Moreover, for the dose effect, 70% of the animals

experience have an increase of the behavioral markers cold stimuli when the dose increase.

3.4 Drug Classification model

Classification aims to discriminate the data according to a label and using a set of features for decision making. Here the criterium to discriminate is the dose level, and the features are described by behavioural observations, histopathological grades and biomarkers alike. We build a model based on a consensus of five classical Machine Learning algorithms (Naïve Bayes [10], Decision Tree [9], Random Forest[1], k Nearest Neighbour and Support Vector Machine [8]), and perform a majority vote (i.e. choosing the most represented class given by the 5 outputs) to assign a class (dose) to every animal. Thus, we ensure that a given algorithm does not favour a certain subset of animals (a certain species or a certain drug).

Experiment	Precision	f1-score	Accuracy
Acrylamide	0.95	0.95	0.89
Cisplatin	0.67	0.67	0.62
Doxorubicin	1	1	0.95
Nova1	1	1	1
Oxaliplatin	0.57	0.57	0.57
Paclitaxel	0.74	0.74	0.7
Vincristine	0.85	0.85	0.82
Mean	0.83	0.83	0.79

Table 1: Result of dose and drug classification model

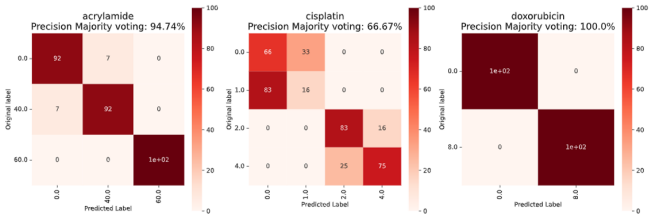

Figure 6: *Dose classification confusion for 3 examples of datasets. The labels are the dose levels in mg/kg/dose*

For each of this algorithms, we use the Leave-One-Out (LOO) cross-validation method, that consists in using all the individuals except one to train the model (training sample) and use it to predict the class of the last one (test sample). By doing so for every individual, we ensure that the efficiency is not dependent on the characteristics of a randomly chosen partition for training and testing, and then, the results will not vary too wildly from one run of the method to another. Efficiency is then calculated by comparing the true classes of the individuals and those predicted by the model. Table 1 presents the result of drug with dose level classification.

4 Discussion

The following experimental features correlate the most with the animal dosing of PNS toxicants, regardless of the species (rat or mice) or the drug (Acrylamide, Cisplatin, Doxorubicin, Nova1, Oxaliplatin, Paclitaxel and Vincristine): (1) Biomarkers: pnF-H, NF-L. (2) Histopathology: Other_Tissue_Histo, PNS_Dorsal_Root_Ganglion, Cornea_Histo. (3) Behavioral: paw_p_last, cold_plate_last, mechanical stimuli, relative_weight, weight_evolution. Furthermore, it was seen that biomarkers correlate with histopathology findings in a dose dependent maner: the higher the dose of the drugs, the better the correlation between biomarkers and histopath findings.

Pulling together biomarker/histopathology correlations, dose classification and gradual pattern mining can bring valuable information on the relevance of behavioral, biomarker and histopathology assays. Regarding biomarkers, NF-L and pnF-H appear to be the most valuable parameters to take into account consider, as they are strongly correlated to most of the histopathology assays they were measured with, discriminatory in some drug dose classifications and show frequent gradient patterns with most of the data they were measured with. Moreover, they are present in most of the subsets. The TAU biomarker seems interesting to a lesser extent. Similarly, from the histopathology point of view Other_Tissue_Histo and PNS_Dorsal_Root_Ganglion shows the most correlation with biomarkers and the most discriminatory power in relation to the dose. Sciatic and Saphenous Nerves degradation are also strongly dependent of the dose according to gradient patterns. Eventually, Decision Trees show a significant importance of the behavioral assays on the dose classification, especially the cold and mechanical stimuli that are strongly present on gradient patterns with the dose and biomarkers as well as weight features that are often found instrumental in the very present in decision trees.

5 Conclusion

To help identifying peripheral ,nervous system (PNS) toxicity and build in silico predictive models, the most influencing biomarkers, histopathological observations and behavioral experiments were mined from a range of experimental conditions gathered by the consortium partners. The data set tackles different and challenging issues that we tackle by a range of data mining and machine learning techniques. However, different directions still to carry on. Future works include incomplete multi-view classification based non-negative matrix factorisation [13] and domain adaptation [2].

Acknowledgments

This research is funded by the NeuroDeRisk Project H2020 Innovative Medicines Initiative (IMI2).

References

[1] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[2] Mourad El Hamri, Younès Bennani, and Issam Falih. Hierarchical optimal transport for unsupervised domain adaptation. *Machine Learning*, 111(11):4159–4182, 2022.

[3] B Jordan, A Margulies, F Cardoso, G Cavaletti, HS Haugnes, P Jahn, E Le Rhun, M Preusser, F Scotté, MJB Taphoorn, et al. Systemic anticancer therapy-induced peripheral and central neurotoxicity: Esmo–eons–eano clinical practice guidelines for diagnosis, prevention, treatment and follow-up. *Annals of Oncology*, 31(10):1306–1319, 2020.

[4] Nicolas Kerckhove, Aurore Collin, Sakahlé Condé, Carine Chaleteix, Denis Pezet, and David Balayssac. Long-term effects, pathophysiological mechanisms, and risk factors of chemotherapy-induced peripheral neuropathies: a comprehensive literature review. *Frontiers in pharmacology*, 8:86, 2017.

[5] Jerry Lonlac, Yannick Miras, Vincent Mazenod, and Engelbert Mephu Nguifo. An Interactive Platform for Extracting Gradual Patterns from Multivariate Temporal Data. In *HAL-03986063*, 2023.

[6] Jerry Lonlac and Engelbert Mephu Nguifo. A novel algorithm for searching frequent gradual patterns from an ordered data set. *Intelligent Data Analysis*, 24(5):1029–1042, 2020.

[7] Benjamin Negrevergne, Alexandre Termier, Marie-Christine Rousset, and Jean-François Méhaut. Para miner: a generic pattern mining algorithm for multi-core architectures. *Data Mining and Knowledge Discovery*, 28:593–633, 2014.

[8] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

[9] J. Ross Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.

[10] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[11] Marie Selvy, Nicolas Kerckhove, Bruno Pereira, Fantine Barreau, Daniel Nguyen, Jérôme Busserolles, Fabrice Giraudet, Aurélie Cabrespine, Carine Chaleteix, Martin Soubrier, et al. Prevalence of chemotherapy-induced peripheral neuropathy in multiple myeloma patients and its impact on quality of life: a single center cross-sectional study. *Frontiers in Pharmacology*, 12:637593, 2021.

[12] Marie Selvy, Bruno Pereira, Nicolas Kerckhove, Coralie Gonneau, Gabrielle Feydel, Caroline Pétorin, Agnès Vimal-Baguet, Sergey Melnikov, Sharif Kullab, Mohamed Hebbar, et al. Long-term prevalence of sensory chemotherapy-induced peripheral neuropathy for 5 years after adjuvant folfox chemotherapy to treat colorectal cancer: a multicenter cross-sectional study. *Journal of clinical medicine*, 9(8):2400, 2020.

[13] Zhe Xue, Junping Du, Dawei Du, Wenqi Ren, and Siwei Lyu. Deep correlated predictive subspace learning for incomplete multi-view semi-supervised classification. In *IJCAI*, pages 4026–4032, 2019.