**Step-by-step integration of multimodal data from liquid-based biopsy improves the predictive performance of lipid accumulation in the liver of mice.**

Martin David 1,2, Monbet Valérie 2, Leroyer Patricia 1, Oliviero Nolwenn 1, Turlin Bruno 1, Salim Zerrouki 3, Fautrel Alain 1, Ropert Martine 1, 3, Sire Olivier 4, Loréal Olivier 1

1 UMR NuMeCan, INSERM U1317, INRAE U1341, University of Rennes, CHU of Rennes, Rennes, France

2 University of Rennes, IRMAR UMR-CNRS 6625, Rennes, France

3 Department of Biochemistry, University Hospital, Rennes, France

4 Bretagne Sud University, IRDL, FRE CNRS 3744, Vannes, France

$ Corresponding authors.

Corresponding author:

**INTRODUCTION**

Metabolic Associated Fatty Liver Diseases (MAFLD) are one of the most common chronic liver diseases worldwide, affecting at least 10% of Europeans and 20% of the US population (1). MAFLD mainly occur in a context of obesity and/or metabolic syndrome, which associates, according to the International Diabetes Federation (2), an abdominal obesity and at least two of the following factors: high triglyceride levels, low HDL cholesterol, hypertension treated or not, and high glycemia (3). Such clinical features are facilitated by excessive food uptake, especially a Western diet, and a sedentary lifestyle. Hepatic steatosis, the first step of MAFLD, is characterized by an abnormal lipid accumulation within hepatocytes and may evolve with the appearance of hepatic lesions, such as hepatocyte ballooning, and inflammation, that characterize the development of Metabolic Associated Steato-Hepatitis (MASH) (4).

To date, the histopathologic examination of liver biopsy remains the gold standard for MAFLD staging and grading. However, performing hepatic biopsy is an invasive procedure, requiring short hospitalization for patients, and may induce abdominal pain and, in rarer cases, severe complications (4,6,7). Moreover, it is a costly procedure. Taken together, these elements explain that, despite the high prevalence of MAFLD in the general population, a hepatic biopsy cannot be carried out systematically and routinely repeated during the follow-up when evaluating the efficiency of therapeutic recommendations. Importantly, alternative tests have been proposed, including the evaluation of biological parameters determined *a priori* and imaging (8,9). In addition, the use of omics approaches, together with machine learning algorithms, have been proposed (10). However, a minimally invasive, easy-to-perform, and inexpensive method to diagnose MAFLD and assess its severity, as well as for monitoring the effectiveness of therapeutic management, is still needed (11, 12).

Noteworthy, members of our group and others (14) reported previously that mid-infrared (MIR) vibrational spectroscopy on serum could be useful in evaluating the presence of hepatic steatosis (15–17). However, improvement is still needed to optimize the performance of predictive models using liquid biopsies. We hypothesize that both the integration of multimodal data obtained from serum, together with a refinement of the strategy of their integration, could improve the performance of predictive models during MAFLD. Therefore, our objective was to evaluate whether a combination of data of different natures through adapted machine learning methods may predict hepatic steatosis. For this purpose, we used a mouse model receiving obesogenic diets and their controls, supplemented or not with iron, for which abnormal metabolism and/or a potential role in the progression of disease has been suggested in Humans. Therefore, we decided to create predictive scores through the integration of biochemical metabolic data routinely obtained from serum for diagnostic purposes with those obtained using the MIR approach and those obtained by evaluating essential metal metabolism parameters, including those that have been previously associated with MAFLD and/or the occurrence of steatotic liver, such as copper and iron. We identified that a novel strategy of data integration may greatly improve the predictive model in this condition, opening the door for a potential use in other diseases.

**RESULTS**

**The high-fat, high-carbohydrate diet induces a heterogeneous steatohepatitis.**

The experimental design was developed to generate a heterogeneous pathological landscape of the liver (Figure 1A). Two experimental variables were introduced: mice were subjected either to a control diet or to a high-fat, high-carbohydrate (HFHC) diet, and, independently, received injections of either dextran or iron-dextran. Iron-dextran administration is a well-established method to induce systemic and hepatic iron overload in mice (REF). In this study, the steatosis (i.e., hepatic lipid accumulation) has been evaluated through 3 biological scales: (i) the histological score, which is the gold standard, informs about the steatosis at the tissular scale, (ii) the hepatic triglyceride (TG) content informs about the cellular stage of steatosis, and (iii) the mRNA level of *Fsp27* gives insight into steatosis at the molecular scale (the mRNA level of *Fsp27* is widely recognized as a direct marker of steatosis (REF)).

We first validate that the livers of mice exhibit varying degrees of steatosis. As expected, the mice that undergo a high-fat, high-carbohydrate (HFHC) diet develop steatohepatitis at all biological scales. The HFHC diet induced an increase in hepatic steatosis (tissular scale), compared with the CTL and IRON groups (+40% cells, p<0.001; +36% cells, p<0.001, respectively; Table 1). Combined HFHC diet and iron supplementation also led to an increase of hepatic steatosis compared with CTL and IRON groups (+14% of cells, p<0.01; +9% of cells, p<0.05, respectively; Table 1), but a lower hepatic steatosis level was observed compared with mice receiving HFHC diet alone (-27% of cells, p<0.05, Figure 1A). The hepatic TG concentration was greatly increased in HFHC diet animals compared with CTL and IRON groups (+204%, p<0.001 for both groups, Figure 1C), whereas iron alone did not significantly modulate the HTG concentration compared with CTL (Figure 1C). The results are similar at the molecular scale (Figure 1D). The other

While histological quantification remains the gold standard, a notable discrepancy was observed between three biological scale assessments of hepatic lipid accumulation. As shown in Figure 1E, within the range of 20 to 40 mg of triglycerides per gram of liver, which can be considered as an early stage of lipid accumulation, histological scoring fails to detect a clear increase in hepatic lipid accumulation. In this context, the question remains about the right target variable that we need to predict. This question is not addressed in this study; thus, we create a synthetic variable that sums up the three biological scales.

**A synthetic variable summarizing the steatohepatitis.**

In this context, we construct a synthetic variable that integrates the three indicators of lipid accumulation in the liver: (i) at the molecular scale (mRNA level of Fsp27), (ii) at the cellular scale (based on the triglyceride content), and (iii) at the tissular scale. To generate the synthetic variable summarizing hepatic lipid accumulation, a principal component analysis (PCA) is performed.

We first validate the interpretation of the principal components by calculating confidence intervals for Pearson’s correlation coefficients between the PCA coordinates and the original variables measured in each mouse (Figure 1D). The first principal component (PC1) is consistently associated with all markers of lipid accumulation in the liver, indicating that higher PC1 scores correspond to more severe hepatic steatosis (Figure 1D.(2)). This interpretation is supported by positive association with alanine aminotransferase (ALAT) levels and the inverse relationship with the ASAT/ALAT ratio (Figure 1D.(1)), consistent with diet-induced liver damage (REF). On this basis, we select the coordinates in Principal Component 1 as the main outcome variable that we have to predict.

**A multi-modal approach appears essential to encompass all the information in the dataset.**

In this study, we constructed several datasets (Figure 1A). The first dataset consists of biochemical measurements from serum. All biochemical data, stratified by experimental condition, are presented in Table S2. The second dataset is derived from mid-infrared (MIR) spectroscopy performed on serum, providing molecular fingerprints specific to each mouse. An example of a spectrum is shown in Figure S2. The third dataset comprises quantitative measurements of metals and trace elements in the blood, presented in Table S3.

We first explore the interest in integrating these diverse data types (Biochemical data, MIR data, and metallomic data). The method is presented in the Materials and methods section. Briefly, we compute (i) the pairwise sample distance matrix from the three variables summarizing hepatic lipid accumulation, (ii) the pairwise sample distance matrix from one type of data, and finally, we perform (iii) a Mantel test comparing both pairwise sample distance matrices. The outcomes of the Mantel test give the distribution of Pearson’s r between both pairwise sample distance matrices. It means that data sharing information should exhibit a high level of Pearson’s r. The results are presented in F. It

Then, we perform the same demonstration as previously. However, we compare the degree of sharing information between types of data. Our goal is to assess the redundancy or non-redundancy between types of data. In this context, Figure 1F reveals that the MIR dataset shares limited information with the serum biochemistry data. However, the metallomic dataset appears to share more information with serum markers, suggesting potential redundancy.

From these observations, we conclude that (i) each dataset contains predictive information relevant to hepatic steatosis, and (ii) each dataset provides distinct and potentially complementary insights. In this context, we propose an innovative boosting approach composed of three sequential models, where each model is trained on the residuals of the preceding one (Figure 3). We hypothesize that this strategy improves predictive performance by effectively leveraging complementary, non-redundant information distributed across datasets, surpassing even other multimodal integration approaches.

**A strategy for integrating multimodal data drives the predictive performance of steatohepatitis.**

In the following section, we develop predictive models for the synthetic variable that characterizes hepatic steatosis. The primary objective is to assess the potential added value of incorporating diverse data types. A secondary objective is to identify the most effective strategy for integrating multimodal data to enhance predictive performance.

A detailed description of the predictive modelling approaches is provided in the Materials and Methods section. Briefly, we evaluate one unimodal approach and two distinct strategies for multimodal data integration. To contextualize model performance, we first estimate a baseline random prediction effect, representing the expected performance range of models with no true predictive capacity (illustrated by the grey box in Figure 4A). Models that exceed this baseline are considered to possess informative value.

The first strategy consists of building separate Random Forest models for each dataset (i.e., unimodal modelling). This allows the assessment of predictive value for each data type independently (Figure 4A). Among the unimodal models, those based on serum biomarkers and metallomic profiles demonstrate the highest predictive accuracy (Figure 4A).

This unimodal approach serves as a reference to evaluate the benefits of multimodal integration. In the first integration strategy, we concatenate all datasets into a single input matrix and apply two machine learning algorithms (Random Forest and Gradient Boosting) (Figure 4B). However, this pooled-data strategy does not improve predictive performance relative to the best unimodal models (Figure 4A and 4B).

To overcome the limitations of naïve integration, we propose a sequential integration strategy based on a residual learning framework, conceptually inspired by the Gradient Boosting algorithm. In this approach, models are trained in sequence, with each model learning to predict the residuals (i.e., the unexplained variance) of the previous one. This allows for incremental improvements in prediction by leveraging the non-redundant information distributed across datasets. The sequence of data modalities is selected to maximize the reduction in residual variance at each step, and the order of data integration does not influence the predictive performance (data not shown).

Our results show that this residual-based modelling approach outperforms models trained on any single data modality, including serum markers alone (Figure 4C). Additionally, our algorithm enables the extraction of variable importance scores from the aggregated model, offering interpretable insights into the lipid accumulation severity. These results are summarized in Table 2.

**DISCUSSION**

1. **Résumé les principaux résultats :**

* **L’intégration de données multimodal améliorer la prediction**
* **La strategies d’intégration est importante.**

1. **Proposez une hypothèse de pourquoi notre strategies améliore la prediction.**
2. **Petit paragraphe sur l’intéprétation des variables importantes.**

**CONCLUSION AND APPLICATIONS**

**FIGURE LEGENDS**

**Figure 1. A multimodal approach is essential to fully capture the information contained in the dataset.** **(A)** Experimental design of the study. One hundred mice are assigned to either a control diet or a high-fat, high-carbohydrate (HFHC) diet. In addition, they receive either dextran or iron-dextran injections, the latter inducing hepatic iron overload. Multiple analyses are conducted on liver tissue, alongside datasets derived from blood-based analyses. The goal of this study is to predict steatohepatitis. **(B)** Relationship between the molecular and cellular biological indicators of steatohepatitis. **(C)** Relationship between the cellular and molecular biological indicators of steatohepatitis. **(D)** Principal component analysis (PCA) of three steatosis-related variables: histological steatosis score, hepatic triglyceride concentration, and **Fsp27** mRNA expression. **(E)** Pearson's r coefficients between PCA coordinates (from panel C) and five categories of biological variables. The figure shows the 95% confidence intervals of the correlations for: (1) blood markers of hepatic injury, (2) variables describing hepatic lipid accumulation, (3) fibrosis-associated variables, (4) inflammation-related markers, and (5) iron metabolism-related variables.

**Figure 2. A multi-modal approach appears essential to encompass all the information in the dataset.** **(A)** . In other words, it means that datasets share information with steatosis-related variables.**B** In this case, it means that datasets share information with serum biochemical variables.

**Figure 3. Overview of the integration strategies of the multimodal data dedicated to predict steatosis.** **(A)** Biochemical, MIR, and metallomic data obtained from serum were used to predict steatosis following three different strategies. **(B)** Unimodal strategy means that each type of data is considered independently. Multimodal strategies mean that types of data are considered dependently in the analysis. The first one is a classic multimodal strategy consisting of pooling datasets, while the multistep and incremental integration of the data categories is considered here as an innovative strategy. **(C)** The Root Mean Square Error (RMSE) will be used to compare predictive performance.

**Figure 4. Comparative predictive performance of single and aggregated models for hepatic steatosis prediction.** **(A)** Overview of the three modelling strategies used in the study. Strategy 1: Single model. A Random Forest model is trained independently on each dataset (blood markers, mid-infrared (MIR) spectroscopy, and serum metallome). Strategy 2: Aggregated models. Sequential models are trained on residuals from the previous model to integrate complementary information from multiple datasets. Strategy 3: Pooled data. An Extreme Gradient Boosting (XGBoost) model is trained on all datasets combined into a single input. **(B)** Root Mean Square Error (RMSE) of models trained on individual datasets. Performance is compared to a null distribution (grey) representing the expected error under random association. Models based on blood markers, selected MIR variables (SelVar), and metallome all show significant predictive power, with blood markers yielding the lowest RMSE. **(C)** RMSE of aggregated models integrating two or more datasets. The multi-model aggregation strategy (blood markers → MIR → metallome) provides the best predictive performance. Asterisks indicate levels of statistical significance compared to the null distribution (\* p < 0.05, \*\*p < 0.01, \*\*\* ***p*** < 0.001).

**TABLES**

**Table 1. titre**

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Association | Interpretation (for MIR only) | References |
| Blood markers | | | |
| Cholesterol | ↑ |  |  |
| MIR | | | |
| 1730 cm-1 | ↑ |  |  |
| 1522 cm-1 | ↑ |  |  |
| 1344 cm-1 | ↑ |  |  |
| 3033 cm-1 | ↑ |  |  |
| Metallome | | | |
| Zinc | ↑ |  |  |
| Copper | ↑ |  |  |

**Table 2. Key variables contributing to the prediction of the synthetic outcome summarizing steatohepatitis in mice.**