**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 frequency 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). 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 performed on serum could be useful in evaluating the presence of hepatic steatosis (15–17). MIR spectroscopy mainly explores the chemical bonds existing between the CHNOPS atoms that constitute proteins, lipids, and carbohydrates found in the studied biofluids. Such MIR analysis is minimally invasive, fast, costless, and could be iterated frequently during the follow-up of patients. However, improvement is still needed to optimize the predictive models. In this study, we propose a third dataset as a predictor of the lipid accumulation in the liver: the metallomic, which gives information about the trace elements in the blood.

We hypothesize that a combination of biochemical, MIR, and metallome data that can be all obtained from serum could improve the characterization of hepatic lipid accumulation during MAFLD. Therefore, our objective was to evaluate whether a combination of multimodal data and machine learning methods may predict hepatic steatosis. For this purpose, we used a mouse model receiving control (CTL), or a high-fat and high-carbohydrate diet (HFHC). In addition, animals were submitted to or not to parenteral iron supplementation (IRON and IRON/HFHC) to potentially increase the impact of the HFHC diet.

**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.

**Multi-step data integration improves the predictive performance of steatohepatitis.**

In the following section, we aim to develop predictive models for the synthetic variable that characterizes hepatic steatosis in the mouse cohort. The third goal is to evaluate the potential added value of incorporating alternative data types in future human studies. A secondary objective is to determine the best strategy for multimodal data integration into the model.

The complete predictive methods are presented in the Materials and Methods section. Briefly, we consider one unimodal data approach and two integration approaches of multimodal data. The first strategy yields in a construction of unique model (RandomForest) with a unique dataset. For each data types is associated a prediction. We also estimate the random effect of prediction, representing the expected performance range of a model with no true predictive power (represented by the grey box, Figure 4A). The further a model performs above this threshold, the more informative it is considered. Among all single-modality models, those based on serum markers and metallome profiles yield the best predictive performance.

Then this strategy is used as a control strategy and will help answer the question: Does multimodal data improve the predictive performance? In this context, we first compare the integration strategy that uses the pooled data as an input (Figure 3B). We test two types of algorithms for this data integration strategy: the Random Forest and the Gradient Boosting. The pooled data integration strategy does not outperform the single-modality algorithm (Figure 4B).

In this context, we construct a machine learning algorithm integrating the non-redundant information distributed across the different datasets. The strategy we adopt is conceptually inspired by the gradient boosting algorithm. In such an algorithm, the model is trained sequentially, with each subsequent model aiming to predict the residuals of the previous one. Residuals here refer to the portion of variance in the target variable not explained by the prior model. Thus, if each dataset contributes unique information, the sequential models can incrementally improve prediction performance by capturing previously unexplained variance.

We demonstrate residual-based modelling strategy presents higher predictive performance for hepatic steatosis than the model fitted only on serum-derived markers (Figure 4C). Parler de l’ordre sequential des types de données.

The importance of the variables from the aggregated models also gives an overview of the severity of the lipid accumulation (Table 1).

**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 correlation 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)** **B**

**Figure 3. 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).

|  |  |  |  |
| --- | --- | --- | --- |
| 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 1. Key variables contributing to the prediction of the synthetic outcome summarizing steatohepatitis in mice.**