Towards Precision Medicine with Infrared Molecular Profiles: Identifying and Explaining Subgroups

Lea Gigou[1,2,3(✉)], Kosmas V. Kepesidis[1,2,3], and Ferenc Krausz[1,2,3]

[1] Chair of Experimental Physics - Laser Physics, Ludwig-Maximilians-Universität München, Garching, Germany
lea.gigou@campus.lmu.de
[2] Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics, Garching, Germany
[3] Center for Molecular Fingerprinting, Budapest, Hungary

Abstract. A common approach to precision medicine is stratifying individuals into homogeneous subgroups based on defined criteria. In previous studies, infrared molecular fingerprinting (IMF), a blood-based profiling method that captures broad molecular information, has been proposed for personalized health monitoring due to its low intra-individual variability relative to the population-level variability. In a personalized setting, deviations from the healthy state may thus be more sensitively captured. To enable this in practice, subgroups with reduced interpersonal variability must be identified. This study explores the existence, prediction, and explainability of such subgroups within IMFs. Using a cohort of 4032 healthy individuals with up to 5 visits each, we show that subgroups of reduced interpersonal variability exist. The first three principal components (PCs) of the high-dimensional IMFs are sufficient to define subgroups where within-subgroup variability approaches the levels of intra-individual variability. Machine learning models are trained to predict these PCs from routine clinical chemistry, IMF measurement parameters, and participants' characteristics (demographics, lifestyle, and health-related variables). The PCs are successfully predicted with inaccuracies close to or below intra-individual variability. Using Shapley Additive Explanations, we identify key factors behind subgroup formation, ensuring interpretability.

Keywords: Subgroup identification · High-dimensional data · Precision medicine · Infrared molecular fingerprinting

1 Introduction

Precision medicine aims to individualize clinical decision-making to improve patient outcomes. This can be achieved by leveraging subgroups of individuals similar to a patient of interest. This narrows the reference population to

R. Bellazzi et al. (Eds.): AIME 2025, LNAI 15735, pp. 176–180, 2025.
https://doi.org/10.1007/978-3-031-95841-0_33

which the patient is compared and on which decisions are based, thus enabling more individualized conclusions [1]. Infrared molecular fingerprinting (IMF), a profiling method capturing a wide range of molecular information via vibrational spectroscopy of blood plasma, has been suggested as a candidate for personalized medicine. Owing to its generality, IMF may capture various health aberrations, as it has, e.g., been demonstrated for several cancer types [2,3]. IMFs have been shown to be stable over clinically relevant timescales with significantly lower intra- than inter-individual variability. This makes the method suitable for a precision medicine setting [3].

For diagnostics or health monitoring, where deviations from a healthy state are to be sensed, we define subgroups as groups of reduced IMF variability compared to the overall population. In these subgroups, the healthy state is more narrowly defined and may thus improve the sensitivity to health deviations. To apply this in clinical practice, subgroups must be determined using patient characteristics, such as demographic, clinical, or lifestyle-related variables (hereafter referred to as markers). It is standard practice to define subgroups for a target outcome, here reduced IMF variability, using machine learning [5]. As a first step toward precision medicine with IMFs, this work investigates the existence and identification of subgroups of reduced IMF variability among healthy individuals while tackling the challenge of high-dimensional data by finding an appropriate low-dimensional representation.

2 Methods

Data. We use a longitudinal cohort of 4032 healthy individuals with four or five visits each [4]. For each individual, IMFs (493 features, see [3] for details on measurement and preprocessing, L1 instead of L2 normalization is used here) and standard clinical chemistry panels (27 parameters) are available. Additionally, information on demographics (3 markers), lifestyle (4 markers), and medical characteristics (14 markers) is at hand, alongside information on medications (70 types) and measurement parameters (8 markers).

The Existence of Subgroups. First, principal component (PC) analysis is performed to yield a low-dimensional representation of the IMFs. In this representation, the existence of subgroups is investigated by matching fingerprints. For this, a confidence interval per individual is constructed in the first few principal components, assuming Gaussian-distributed fingerprints. An IMF is defined as matching an individual if it lies within that person's confidence interval. Variability between measurements and among or between individuals is reported via standard deviation.

Predicting Subgroups. Machine-learning (ML) models are trained to predict the first five PCs from markers (one model per PC). The tested models include: linear regression (basic, lasso, ridge, elastic net), support vector regression, random forests, gradient boosting (XGB, LightGBM, sci-kit learn), and MLP regressor.

Initially, all models are trained using a loose hyperparameter tuning and cross-validation at different numbers of features (features ranked by mutual information). The best-performing models (Ridge, LightGBM, XGB) are then selected for further refinement. Since all selected models are regularized, feature selection is skipped. For these models, hyperparameters are tuned using Bayesian search cross-validation. The final prediction models are then trained in a train-test split (3226/15204 individuals/measurements for training and 806/3808 for testing for LightGBM and XGB, 1607/6654 and 419/1753 for Ridge regression due to missing marker values). To avoid information leakage, all measurements of one individual are assigned to either the train or the test/one fold of the cross-validation and the PC transformation is fitted on the train set only for the final models.

Explaining Subgroups. Model predictions are explained on the test set using Shapley Additive Explanations (SHAP values) with an independent masker and reported aggregated over the models. To this end, the mean absolute SHAP values per marker are standardized by their sum per model, then weighted by the explained variance ratio of the corresponding PC and summed across PCs. The resulting values can be interpreted as the percentage of IMF variability attributed to each marker by the models.

3 Results

The Existence of Subgroups The first five PCs of the IMFs account for 94.5 % of the total variability (50.0 %, 25.1 %, 10.3 %, 5.7 %, and 3.4 %). Figure 1 (a) shows the number of matching measurements per individual when determining matches based on two, three, or four PCs. The number of matches rapidly declines with more PCs, with close to zero matches for four PCs. When investigating the between-measurement variability of matching measurements determined using three PCs at a confidence level of 0.95, it reaches levels of the within-person variability (see Fig. 1 b). It can be concluded that groups of matching fingerprints with strongly reduced variability exist and can be identified based on the first three PCs.

Predicting Subgroups. Using the above findings, predicting the first three PCs is sufficient to define subgroups. To explore subgroup formation and the lack of matches when including PC four, the first five PCs are investigated in the following. Figure 2 (a) shows the fine-tuned performance of the three best-performing ML models to predict the first five PCs. Intriguingly, the individual PCs can, on average, be predicted with an absolute error close to or even well below the within-person variability. Ridge regression performs best on average with the lowest variability, so all further analyses are based on this model. Figure 2 (b) illustrates true and predicted values of PCs 1 and 2 for four randomly selected individuals.

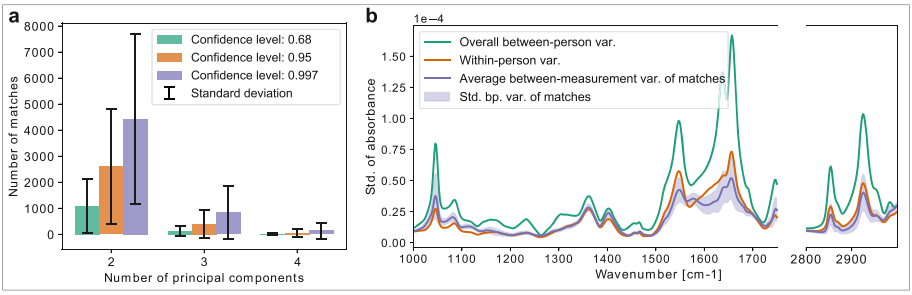

Fig. 1. (a) Number of matching measurements per individual at different confidence levels. (b) Average within-person variability, between-person variability of the full population and between-measurement variability of matching measurements. Matching measurements are determined using 3 PCs and a confidence level of 0.95.

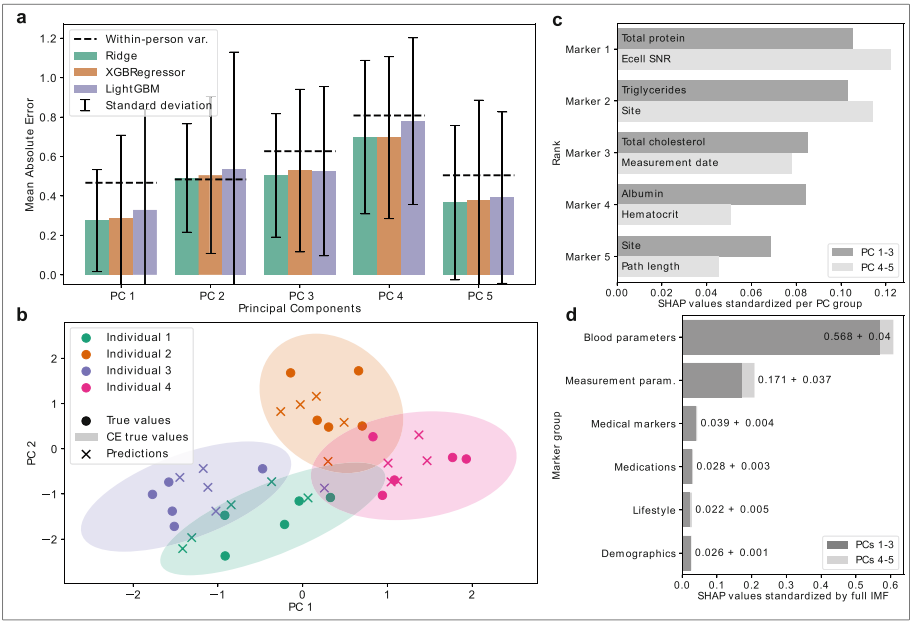

Fig. 2. (a) Mean absolute error of predicting PCs using different ML models. (b) Illustration of true values, predictions, and confidence intervals. (c) SHAP values of highest scoring features for PCs 1–3 and 4–5 relative to the respective PCs. (d) SHAP values per category relative to the full IMFs' variability.

Explaining Subgroups. To explain the models' predictions, the five most influential factors for PCs 1–3 and PCs 4–5 are investigated. Predictions for PCs 1–3 mainly rely on blood parameters, while PCs 4–5 are mainly based on technical IMF measurement parameters. This dominance of measurement parameters in PCs 4–5 explains the lack of matches when including PCs 4–5 and their lim-

ited relevance for reducing intra-group variability. Weighting SHAP values by the explained variance of each PC shows that blood parameters have by far the highest contribution (57.2 %), followed by measurement parameters (20.8 %) and medical markers (4.3 %). Medications, lifestyle, and demographics contribute close to 3 % each.

4 Discussion and Outlook

We showed that subgroups with variability close to the within-person variability exist within IMFs and found a low-dimensional representation to identify such groups of matching measurements. We further showed that this representation can be predicted with surprisingly low error from standard clinical data. Future work should jointly evaluate the PC predictions and assess the resulting intra-group variability. To fully design a precision medicine setting for diagnostics or health monitoring using IMFs, a framework must be developed to move from matching measurements, as investigated here, to matching individuals while accommodating patients with potential disease. Whether the IMF subgroups based on the prediction of the first PCs yield a meaningful reduction in the reference range leading to enhanced diagnostic sensitivity must be explored in future work. Overall, our findings encourage the further investigation of IMFs for personalized medicine by showing that narrowed reference classes exist and can be identified.

Disclosure of Interests. The authors have no competing interests to declare relevant to this article's content.

References

1. Kent, D.M., Steyerberg, E., Van Klaveren, D.: Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ **363** (2018)
2. Huber, M., et al.: Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer. eLife **10**, e68758 (2021)
3. Huber, M., et al.: Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring. Nat. Commun. **12**(1), 1511 (2021)
4. H4H Study Homepage. https://h4h.hu/en/. Accessed 02 Feb 2025
5. Lipkovich, I., Dmitrienko, A., D'Agostino, R.B.: Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat. Med. **36**(1), 136–196 (2017)