PERMIT Workshop

Scoping review on machine learning methods for stratification

Enrico Glaab
Luxembourg Centre for Systems Biomedicine
Scoping Review Objectives

GOALS:
- Inventory AI methods for omics-based patient stratification and their validation (supervised and unsupervised ML approaches)
- Identify limitations, challenges, gaps and existing recommendations

<table>
<thead>
<tr>
<th>MACHINE LEARNING</th>
<th>SUPERVISED LEARNING</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CLASSIFICATION</td>
<td></td>
</tr>
<tr>
<td></td>
<td>REGRESSION</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CLUSTERING</td>
<td></td>
</tr>
</tbody>
</table>
Research Questions

Machine learning methods for stratification:
• What are the main types of supervised and unsupervised ML methods for omics-based stratification? What are the recommended workflows?
• What are the specific strengths/weaknesses of different stratification approaches?

Validation methods:
• Which validation methods are available to assess accuracy, robustness and biomedical relevance? What are their strengths/weaknesses?

Applications:
• Which practical utility has been demonstrated in real-world settings (success/failure stories, lessons learned)?
Scoping Review Results: PRISMA flow diagram

Records identified through initial database search (n = 1164)

Records after duplicates removed (n = 1079)

Records screened (n = 1079)

Records excluded (n = 502)

Full-text articles assessed for eligibility (n = 577)

Full text articles were excluded:
- Lack of adequate statistical validation or insufficient details on validation: 67
- Insufficient sample size (at least 50 samples for the main conditions studied, or demonstrated power calculation results): 251
- No omics-scale data used: 19
- Redundant article (already covered): 2
- Unable to retrieve full-text article: 4

Studies included (n = 234)
Results: Used ML methodologies

- **Over-represented approaches:** Tree- and kernel-based ML methods
- **Under-represented approaches:** Probabilistic, prototype-based and neural network based ML methods
Results: Used validation methodologies

- **Common methods**: Training/test set split + cross-validation on training data (LOOCV, 10-fold CV), metrics: accuracy, AUC
- **Less frequent**: External cohort validation, robust bootstrapping and & bolstered CV approaches, metrics: F1, MCC, PR-AUC
Main gaps & limitations identified (1)

Study design and documentation related issues:

(1) study group design and sample size selection
   (underpowered, imbalanced)

(2) statistical evaluation
   (robustness/completeness, multiple hypothesis testing)

(3) clarity of clinical applications
   (primary/secondary outcomes)

(4) study documentation
   (settings/parameters, reproducibility)
Main gaps & limitations identified (2)

Issues affecting model reliability, robustness and interpretability:

(5) Sampling & blocking design  
(batch effects and biases)

(6) data pre-processing, filtering and normalization  
(lacking standards)

(7) integration of prior biological knowledge  
(pathway/network knowledge, multi-omics analyses)

(8) Ensuring model interpretability and biological plausibility  
(black-box vs. white-box models)
Gaps & limitations: Example

Common error: Global feature selection

- Data
- Supervised selection
- Filtered data
- Training
- Test
- Model
- Score
- Evaluation

Correct approach

- Data
- Supervised selection
- Filtered training data
- Training
- Filtered test data
- Test
- Model
- Score
- Evaluation
- transfer
Recommendations from the scoping review / literature

Data pre-processing, filtering & normalization:
→ use cross-validation to check if pre-processing leads to information loss
→ compare or combine multiple pre-processing approaches

Integration of prior knowledge & multi-omics analyses:
→ check prior literature on the cost/benefit of multi-omics analyses for the studied conditions / cell types, or conduct pilot analyses
→ use existing software & frameworks for integrative biological data analysis

Ensuring model interpretability & biological plausibility:
→ use dedicated methods to build interpretable models (e.g. rule learning)
→ use cellular pathway/network analysis & literature mining to guide modeling
Previous success stories

- Multiple omics-derived biomarker signatures already clinically validated
- Most tests are for cancer diseases, but first non-cancer applications exist

<table>
<thead>
<tr>
<th>Name</th>
<th>Test approval (FDA-cleared and/or LDT)</th>
<th>Purpose</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlloMap Heart</td>
<td>FDA-cleared, LDT</td>
<td>identifying heart transplant recipients with risk of cellular rejection</td>
<td>Yamani et al., J Heart Lung Transplant, 2007</td>
</tr>
<tr>
<td>Prosigna Assay / PAM50</td>
<td>FDA-cleared, LDT</td>
<td>breast cancer risk of distant recurrence prediction</td>
<td>Nielsen et al., BMC Cancer, 2014</td>
</tr>
<tr>
<td>Oncotype DX</td>
<td>LDT</td>
<td>breast cancer risk-of-recurrence assessment</td>
<td>Kelley et al., Cancer, 2010</td>
</tr>
<tr>
<td>Decipher</td>
<td>LDT</td>
<td>prostate cancer metastatic risk prediction</td>
<td>Marrone et al., PLoS Curr., 2015</td>
</tr>
</tbody>
</table>
Previous success stories – Main conclusions

Shared characteristics of prior success stories as a guideline:

• **Early filtering:**
  rigorous statistical, clinical and biological filtering criteria applied (strict inclusion/exclusion criteria; multiple layers of statistical and ML-based feature selection; integration of prior knowledge)

• **Continuous technological improvements:**
  transition from cheap, low-sensitivity to high-sensitivity measurements (e.g. from microarray technology to deep sequencing, RT-PCR and digital PCR)

• **Robust validation schemes:**
  multi-level cross-validation, bootstrapping and external validation involving multiple performance metrics, large sample sizes, and multiple cohorts
Summary

Main gaps & limitations:

→ **study design**: many studies are underpowered, imbalanced

→ **statistical validation**: often incomplete, lacking robustness or even incorrect

→ **study documentation**: lack of details, irreproducible

Main proposed recommendations:

→ follow existing study design & documentation guidelines (e.g. NCI check list)

→ use robust validation schemes & early filtering

→ exploit prior biological knowledge & existing data integration frameworks
References


