PERMIT
PERsonalised MedicIne Trials

AI for patient stratification: Challenges and recommendations

Enrico Glaab, University of Luxembourg

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

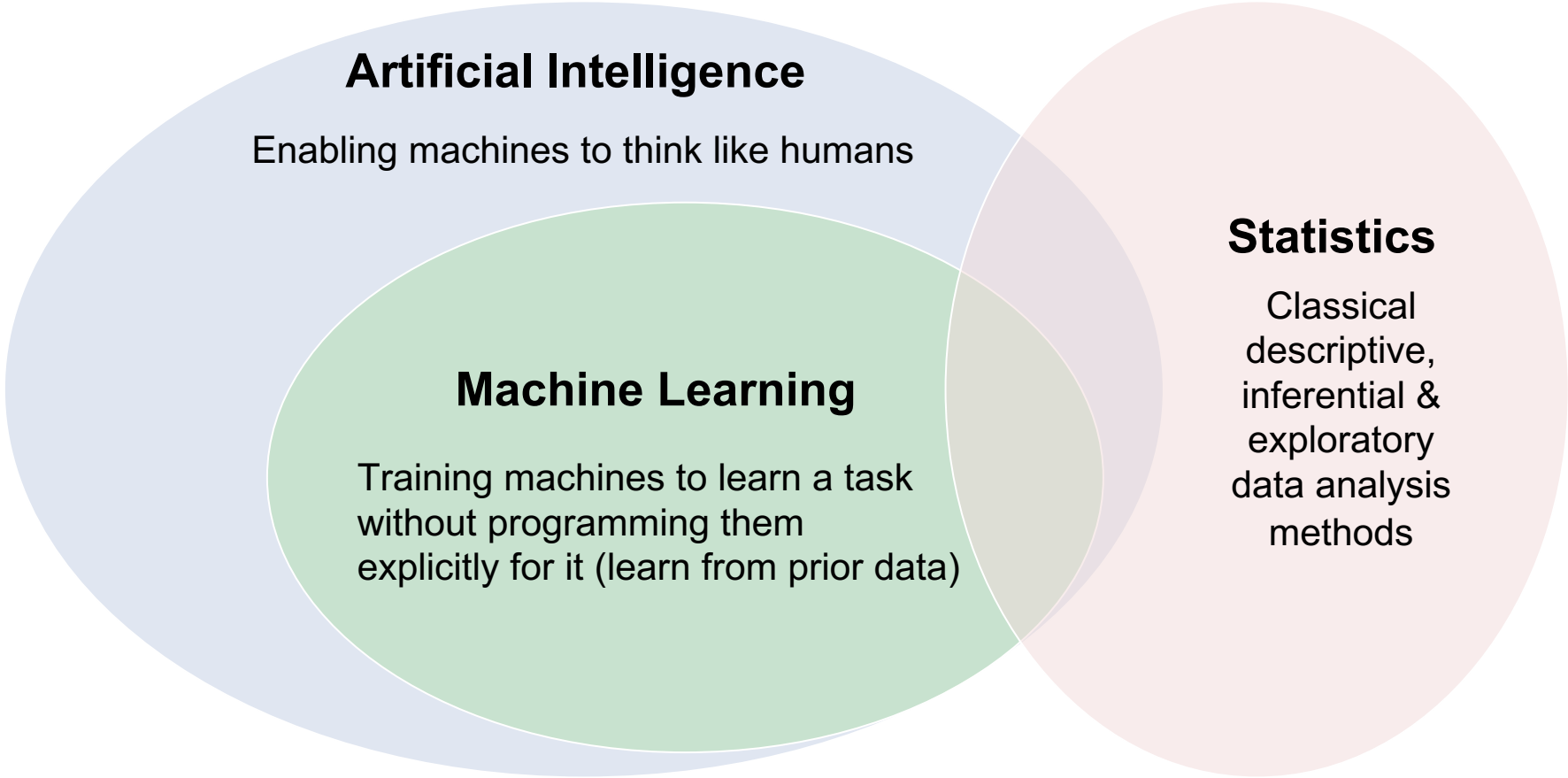
European Commission

Overview

- **Introduction**: AI for stratification – definitions, applications & workflows
- **Challenges**: Gaps and limitations in AI-based patient stratification
- **Recommendations 1**: Study design & planning
- **Recommendations 2**: Discovery & optimization
- **Recommendations 3:** Validation & interpretation
- **Example use cases / success stories**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Definition: AI vs. classical statistics



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

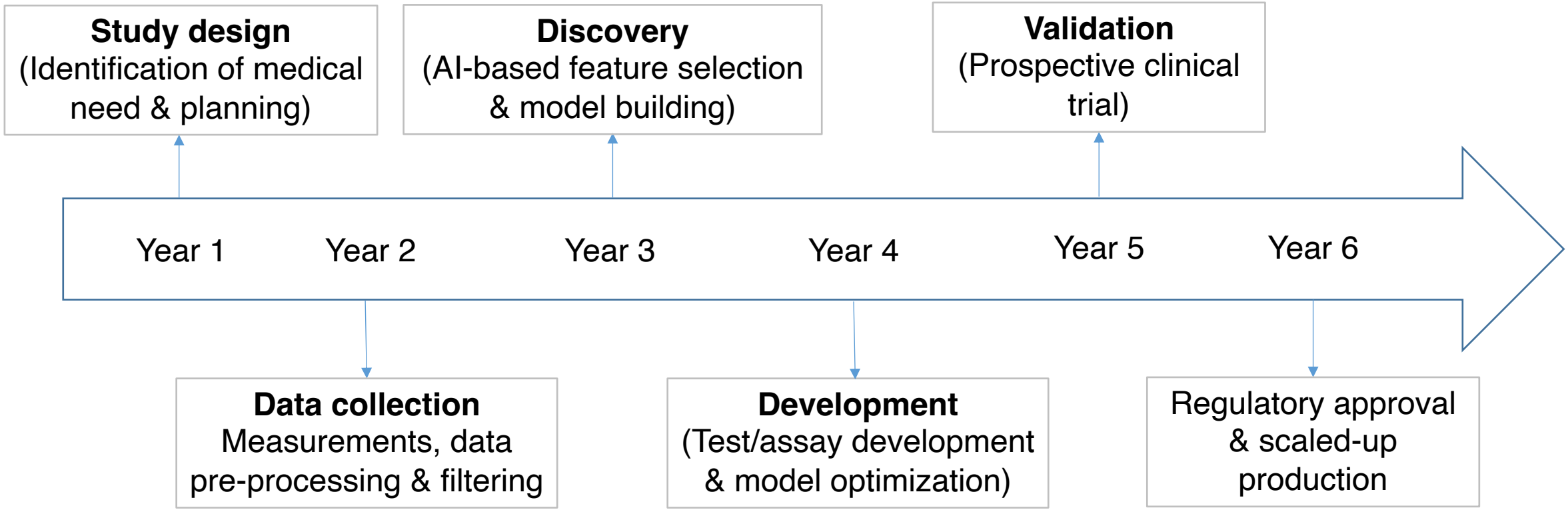
Applications: AI for patient stratification

AI algorithms have several applications in biomedical stratification:

- **Risk stratification** → differentiate between risk categories 
- **Diagnostic stratification** → differentiate between diseases & sub-types 
- **Prognostic stratification** → predict future diseases trajectories & outcomes 
- **Treatment / trial placement** → recommend the right treatment / clinical trial for each patient 

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Typical workflow



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Gaps and challenges

(1) study design and sample size selection

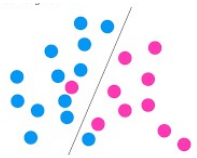
→ underpowered studies, imbalanced study groups, dropouts

(2) data pre-processing, filtering and normalization

→ inadequate choice of approaches, lack of standards

(3) model building (algorithm selection, parameter choice/optimization)

→ modeling approach not suitable for input data, overfitting or underfitting



(4) model optimization & calibration

→ biased parameter selection procedures, missing calibration step

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Gaps and challenges

(5) integration of prior biological knowledge

→ relevant prior data ignored, ineffective data integration methods

(6) model performance assessment

→ inadequate evaluation methods & performance metrics, lack of robustness

(7) validating model performance

→ cohort-specific biases, choice of suitable validation schemes

(8) ensuring model interpretability and biological plausibility

→ use of black-box instead of white-box modeling methods

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Recommendations (1)

1) Planning phase

Challenge/Risk/Gap	Recommendations
Insufficient sample size / study underpowered	• Pilot study for prior sample size estimation • Algorithmic biospecimen matching & selection methods • Integration of complementary biological data to increase power
Imbalanced study groups	• Detailed prior plan for further subject recruitment • Address class imbalance in the modeling (e.g., weighting, under-sampling)
Dropouts in longitudinal studies	• Detailed prior plan for further subject recruitment • Address dropouts in the modeling phase (e.g., bias checks) • Carefully consider possible causes of missing data with domain experts

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Recommendations (2)

2) Discovery & modeling phase

Challenge/Risk/Gap	Recommendations
Inadequate data pre-processing	• Apply quality control analyses before and after data pre-processing • Assess distribution assumptions using statistical tests • Apply pre-processing techniques tailored specifically to observed distributions
Modeling approach is not suitable	• Compare multiple modeling approaches using a cross-validation • Consider combining multiple learning approaches (ensemble learning)
Model is too complex or too simple (overfitting or underfitting)	• Adjust model complexity (regularization) and optimize using a cross-validation • Combine feature selection methods with subsequent learning algorithms

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Recommendations (3)

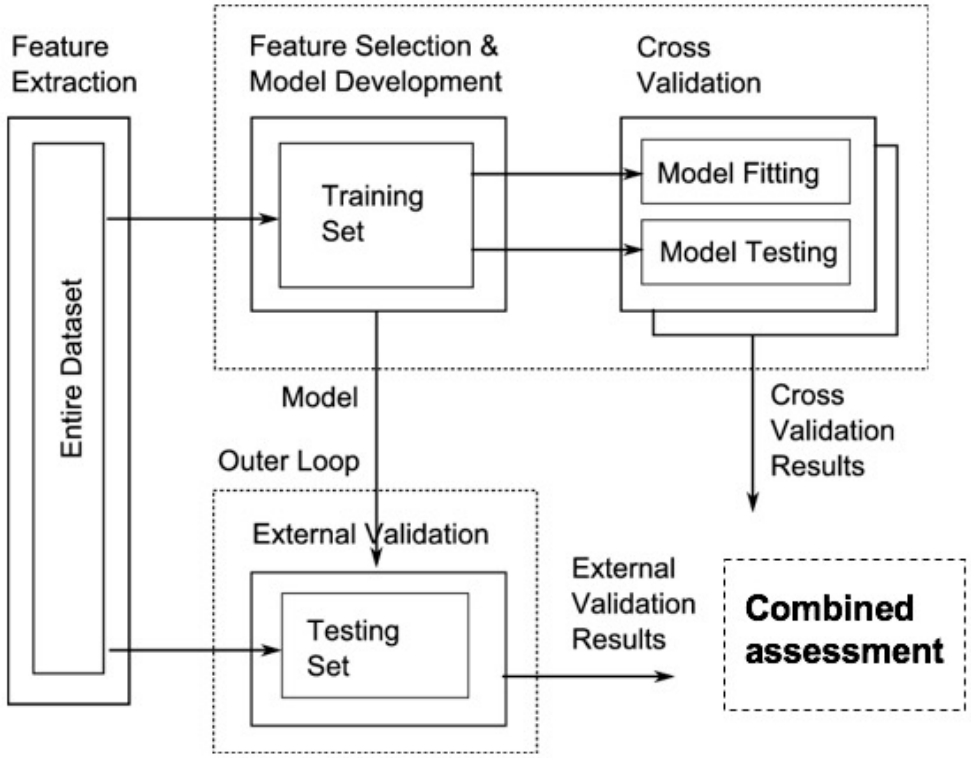
3) Validation phase

Challenge/Risk/Gap	Recommendations
Validation is not robust enough	• Consider both the discovery and validation study in the sample size estimation • Use robust cross-validation methods and multiple performance metrics
The predictive model does not generalize across different cohorts / populations	• Consider a meta-analysis of datasets from other cohorts for feature selection • Plan an external validation on a distinct cohort / population
Insufficient model interpretability	• If model interpretability is required, choose "white-box" learning algorithms • Use structured machine learning approaches guided by prior biological knowledge from cellular pathways and networks to build interpretable models

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Recommendations (4)

→ Use robust and reproducible model building & validation frameworks



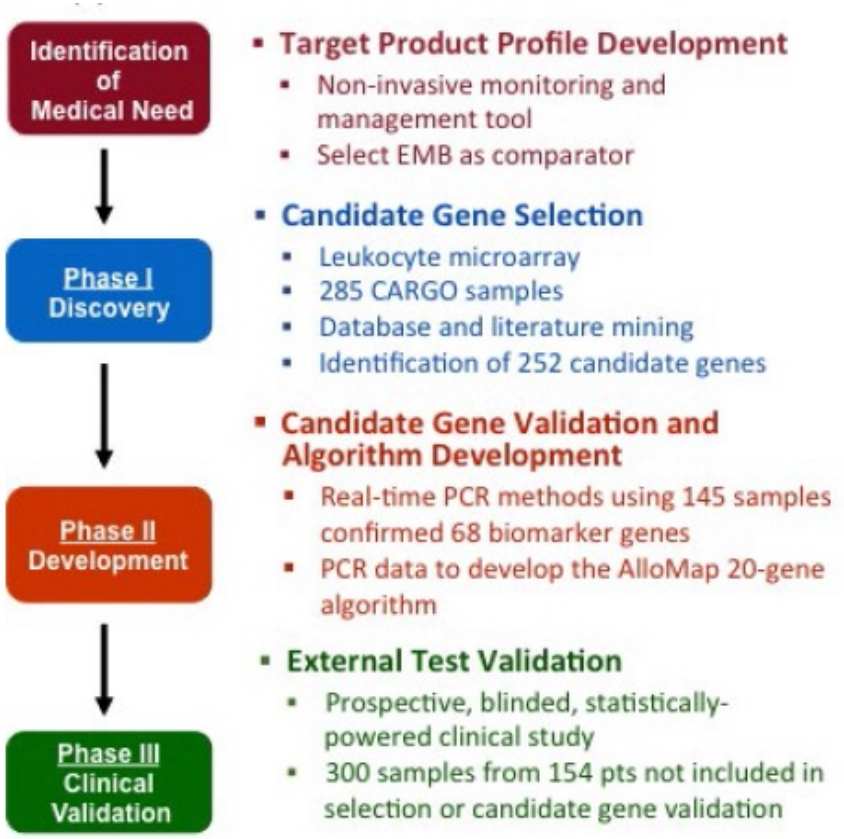
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Example use cases (1)

AlloMap® signature: Predict risk of heart transplant rejection

- Knowledge-guided biomarker discovery: combine prior pathway knowledge with statistical analyses

- Rigorous multi-stage validation: sensitive rtPCR validation + statistically powered external testing



(source: Deng, J Clin Transl Res, 2016)

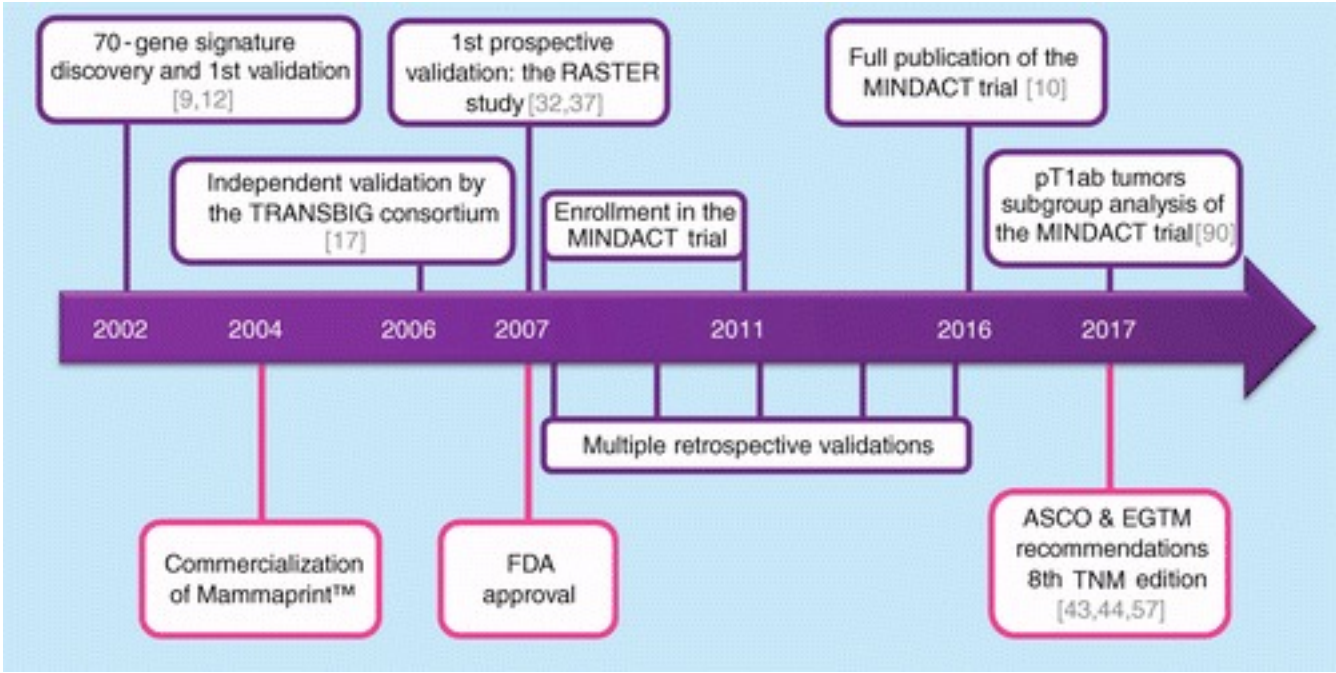
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Example use cases (2)

MammaPrint® signature:
Estimate breast cancer risk of recurrence

- Early and stringent filtering procedure: Reducing candidate predictive features from 25k to a signature of 70 genes

- Robust external validation: Several independent validation studies on external cohorts with large sample sizes



(source: Brandão et al., Future Med., 2019)

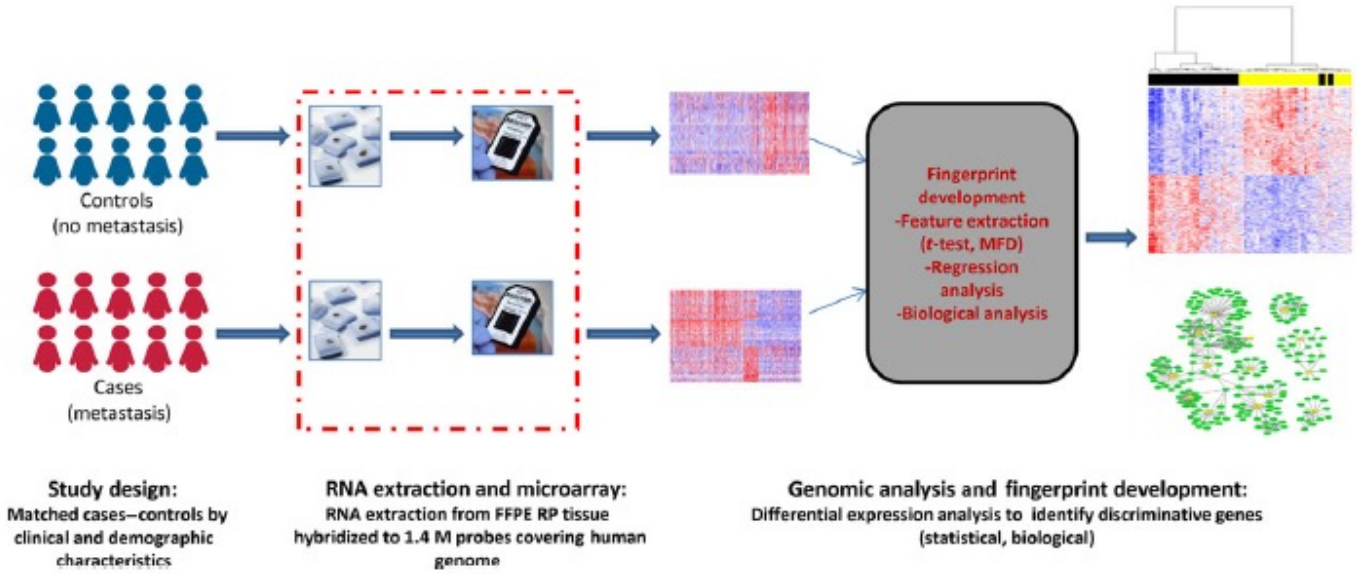
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Example use cases (3)

Decipher® signature: Predict prostate cancer metastatic risk

- Combined statistical and bioinformatics analyses: Statistical + AI-based selection of predictive genes & biological filtering (pathway enrichment + network analyses)

- Robust discovery & validation: High statistical power & multiple distinct cohorts involved



(source: Alshalalfa et al., Biol. Cell, 2015)

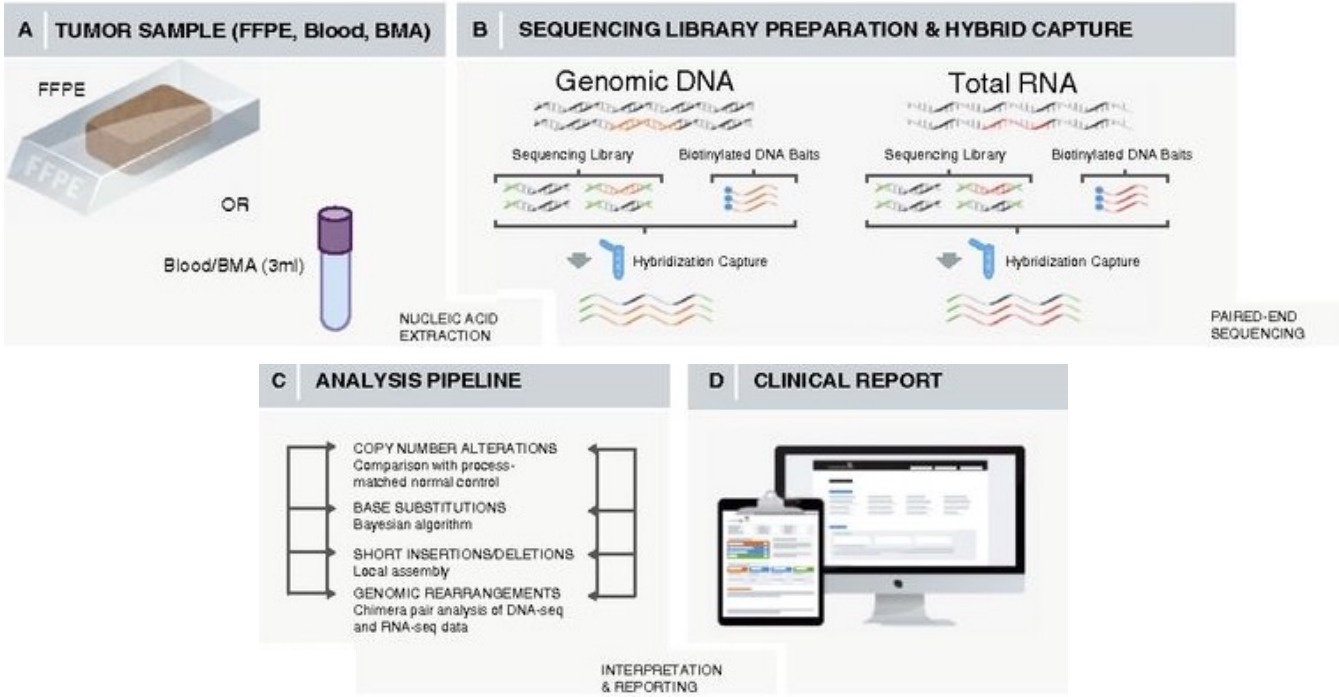
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Example use cases (4)

FoundationOne™ Heme test:
Detect malignancies or solid tumours

- Integrating complementary information sources:
  Combines data from both RNA and DNA sequencing

- Result interpretability:
  Prior knowledge and data used to facilitate test result interpretation



(adapted from: He et al., Blood, 2016)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Summary & Conclusion

Common gaps & limitations:

→ study design phase: many studies are underpowered, imbalanced, suffer from dropouts

→ model building phase: inadequate choice of methods, overfitting or underfitting

→ validation phase: external evaluation often missing or lacks robustness

Main recommendations:

→ involve interdisciplinary expertise (experimental, computational, clinical) in the study design

→ exploit prior biological knowledge & existing data integration frameworks

→ use early filtering & robust validation schemes

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Thank you for your attention!

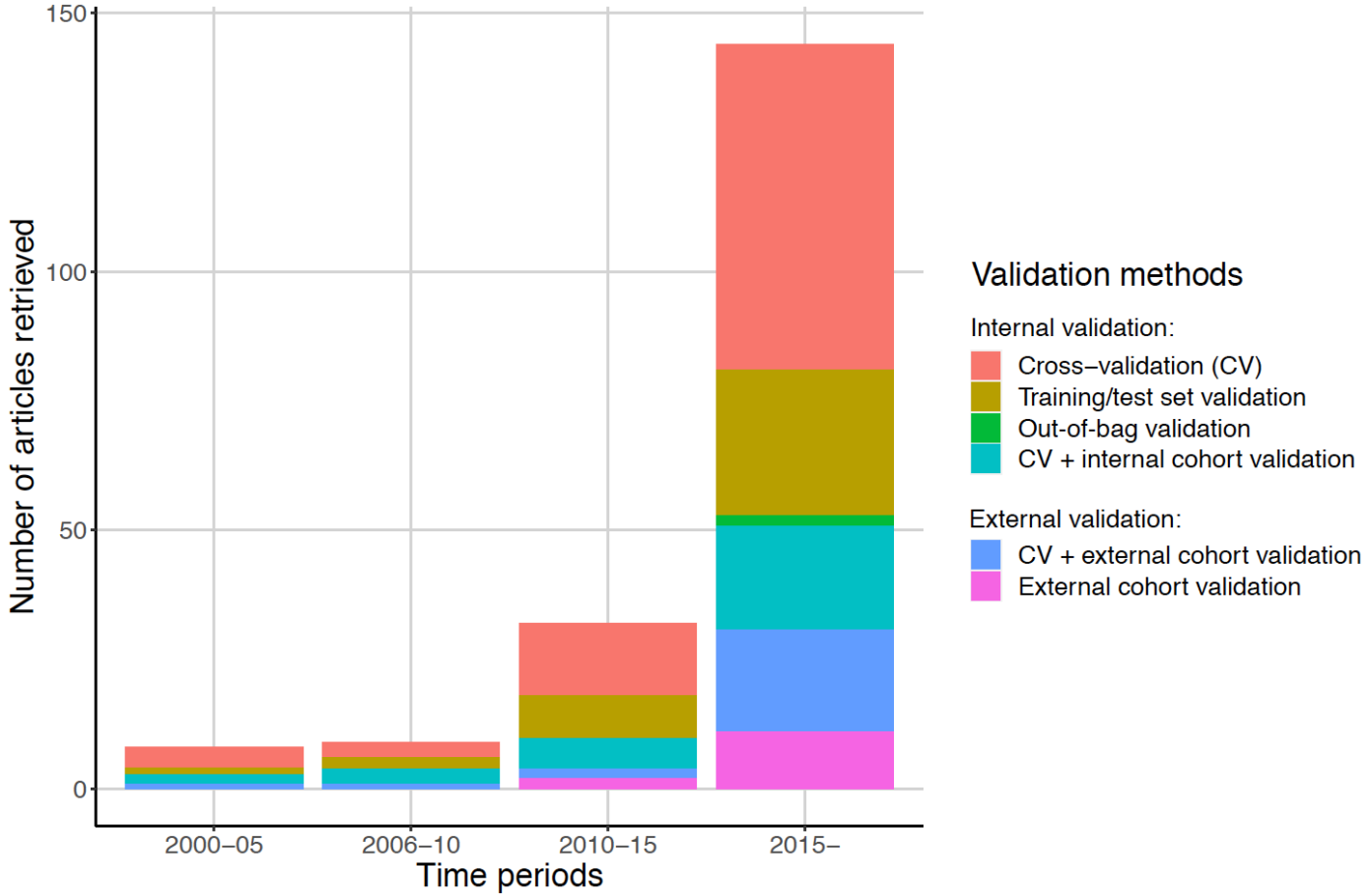
Online quiz:
https://tinyurl.com/permedquestions

PERMIT
PERsonalised Medicine Trials

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

European Commission

Gaps and challenges: Validation



→ Only a minority of published biomarker studies includes an external validation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

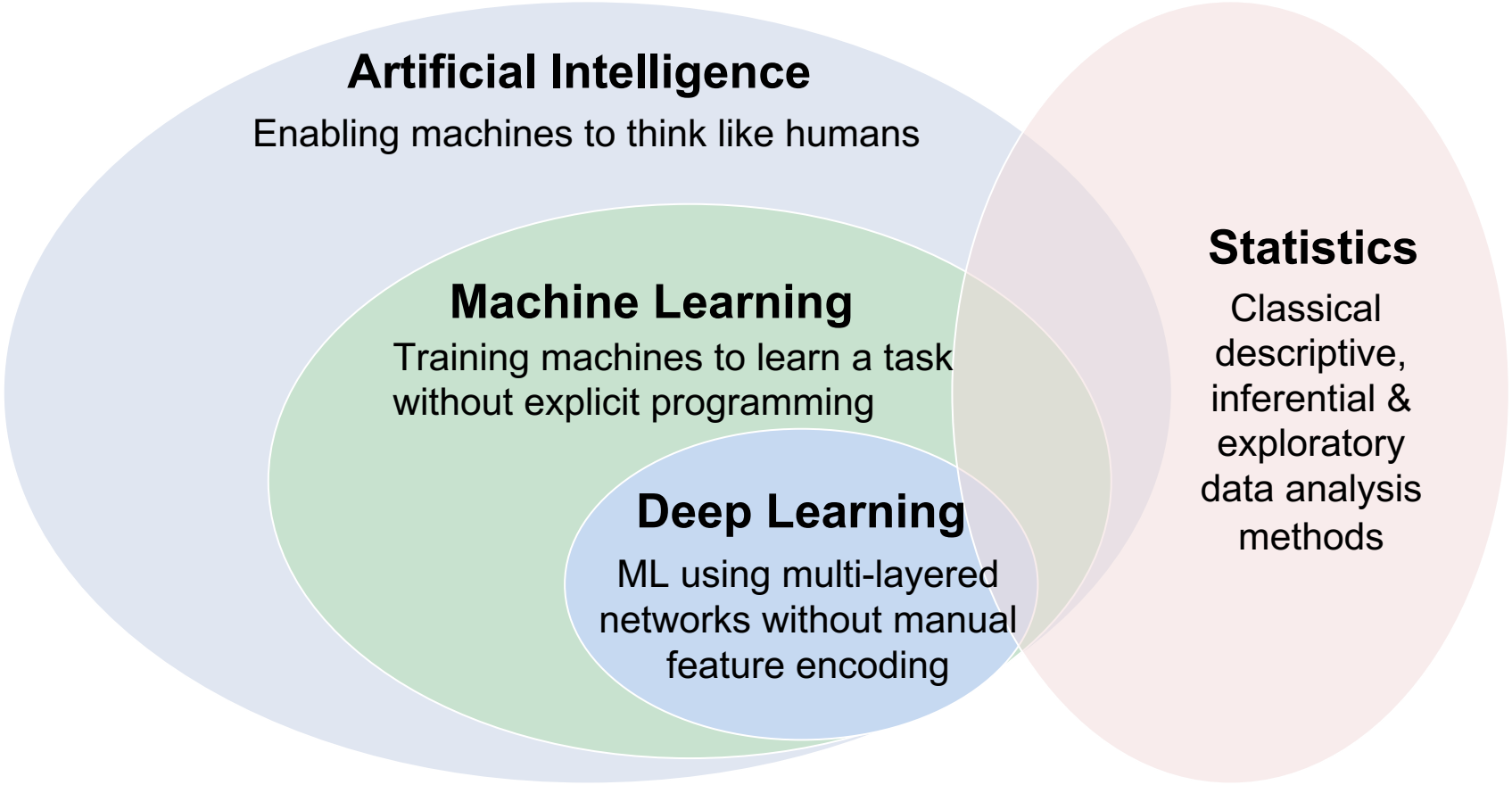
Example use cases

- Multiple omics-derived biomarker signatures already clinically validated
- Following best practices for computational modeling & analysis contributed to the study success

Name	Test approval (FDA-cleared and/or LDT)	Purpose	References
AlloMap® Heart	FDA-cleared, LDT	identifying heart transplant recipients with risk of cellular rejection	Yamani et al., J Heart Lung Transplant, 2007
MammaPrint®	FDA-cleared, LDT	breast cancer risk-of-recurrence assessment	Van't Veer et al., Nature, 2002
Prosigna® Assay / PAM50	FDA-cleared, LDT	breast cancer risk of distant recurrence prediction	Nielsen et al., BMC Cancer, 2014
Decipher®	LDT	prostate cancer metastatic risk prediction	Marrone et al., PLoS Curr., 2015
FoundationOne® Heme	LDT	test for haematologic malignancies, sarcomas or solid tumours	He et al., Blood, 2016

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Definition: AI vs. classical statistics



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Computational analysis & modeling stage

- What this stage comprises
  - This stage covers the computational pre-processing, quality control, statistical and machine learning analysis of the collected data for patient stratification
- How this stage fits in the pipeline
  - Preparations for this stage are already required during the early study design (e.g. to conduct a sample size calculation, define the analysis plan)
  - During a project, this stage follows after the biological data collection
  - This stage lays the ground for the experimental validation of a candidate biomarker model for stratification, derived from the computational analyses

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Methodology

Collection and structuring of material & discussion points for the working sessions

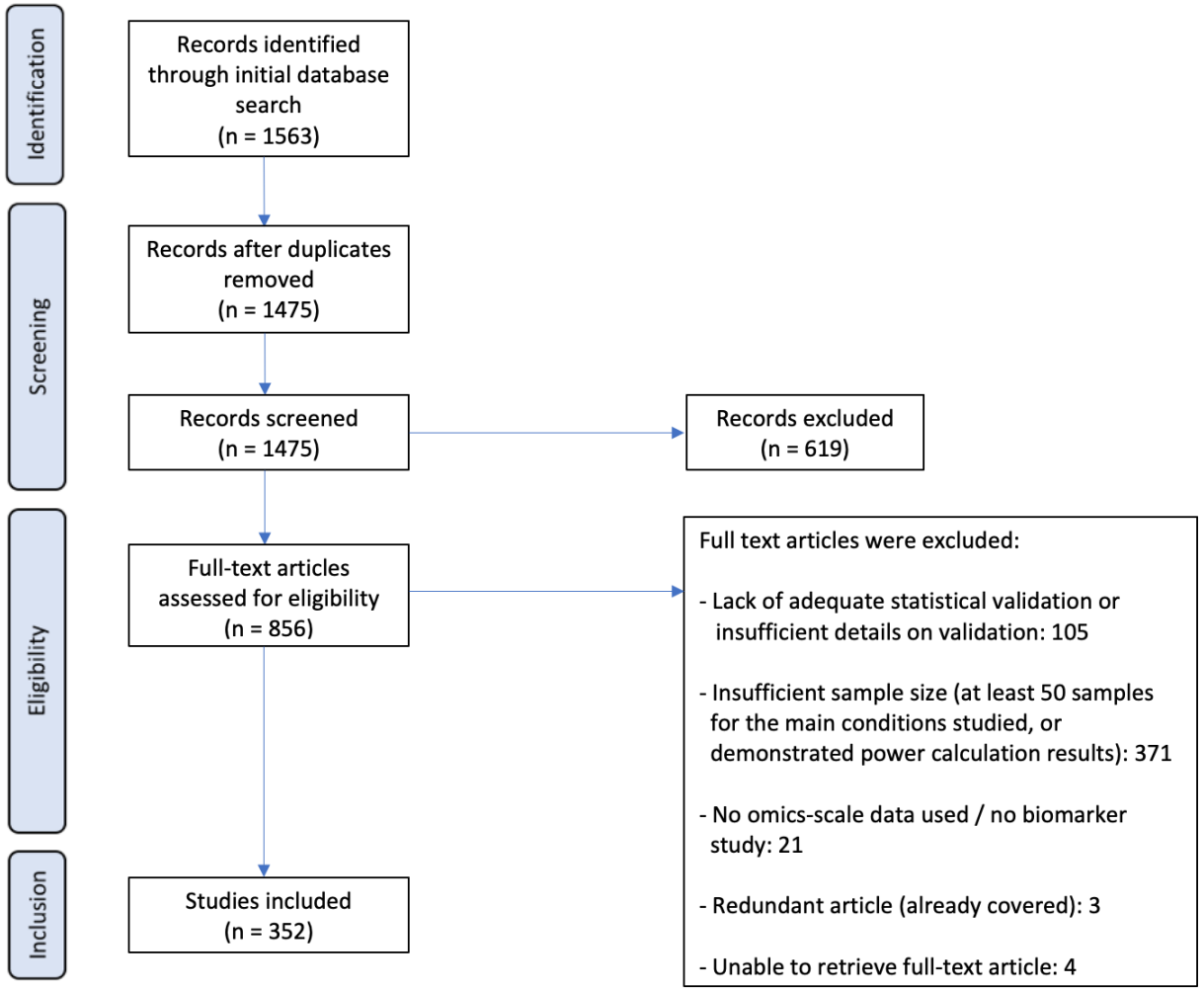
Discussion, extension and revision of collected material on machine learning & validation methods

Structuring and revision of challenges and suggested recommendations

Write-up of machine learning & validation method recommendations

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

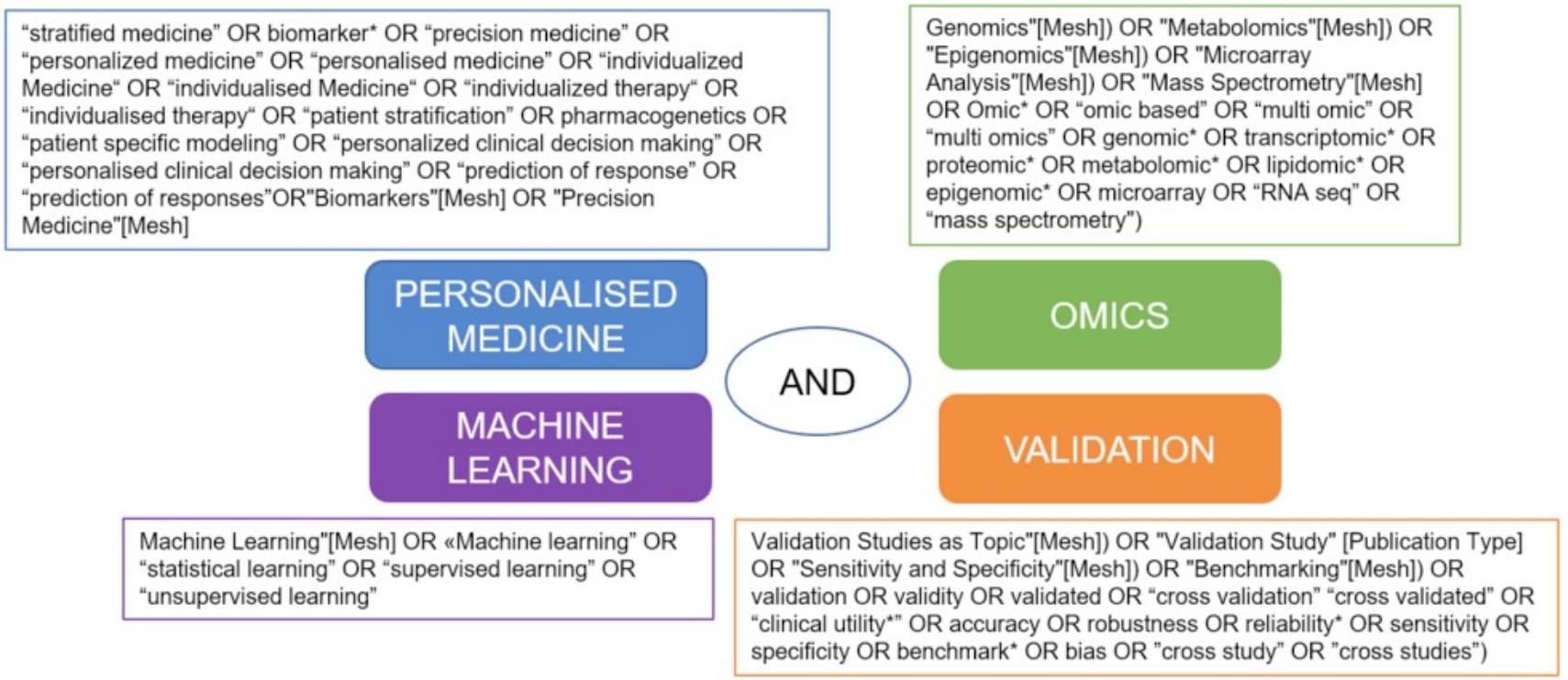
Methodology: Scoping Review



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

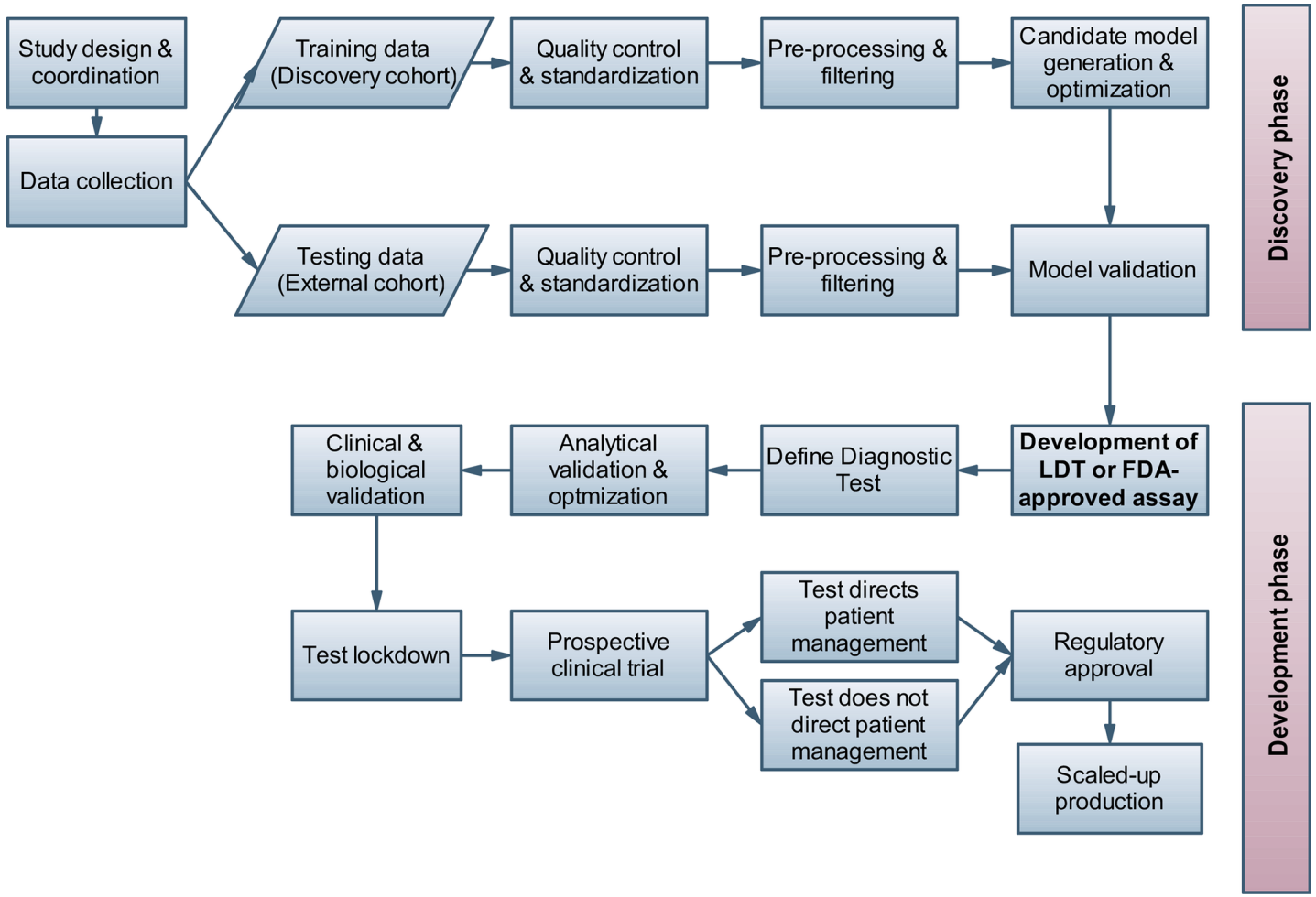
Methodology: Scoping Review

Literature search using Medical Subject Headings (MeSH) term keywords:



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

AI for stratification: Typical workflow



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Methodology: Recommendation structure

- Challenges/risks & associated recommendations are grouped by study phase:
  (1) Planning, (2) Discovery & Modeling, (3) Validation
- Tabular information collection format:

Challenge/Risk	**Likelihood** (low, medium or high)	**Impact** (low, medium or high)	**Recommendations / Mitigation strategies**
insufficient sample size	high	medium or high, depending on the study type	• Prior power estimation • Algorithmic biospecimen selection • Integration of complementary data (e.g. multi-omics data)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Gaps and challenges: Overview

(1) study design and sample size selection

→ underpowered studies, imbalanced study groups, dropouts

(2) data pre-processing, filtering and normalization

→ inadequate choice of approaches, lack of standards

(3) model building

→ modeling approach not suitable for input data, overfitting and underfitting

(4) model optimization & calibration

→ biased parameter selection procedures, missing calibration

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Gaps and challenges

(5) integration of prior biological knowledge

→ relevant prior data ignored, ineffective data integration methods

(6) model performance assessment

→ inadequate evaluation methods & performance metrics, lack of robustness

(7) validating model performance

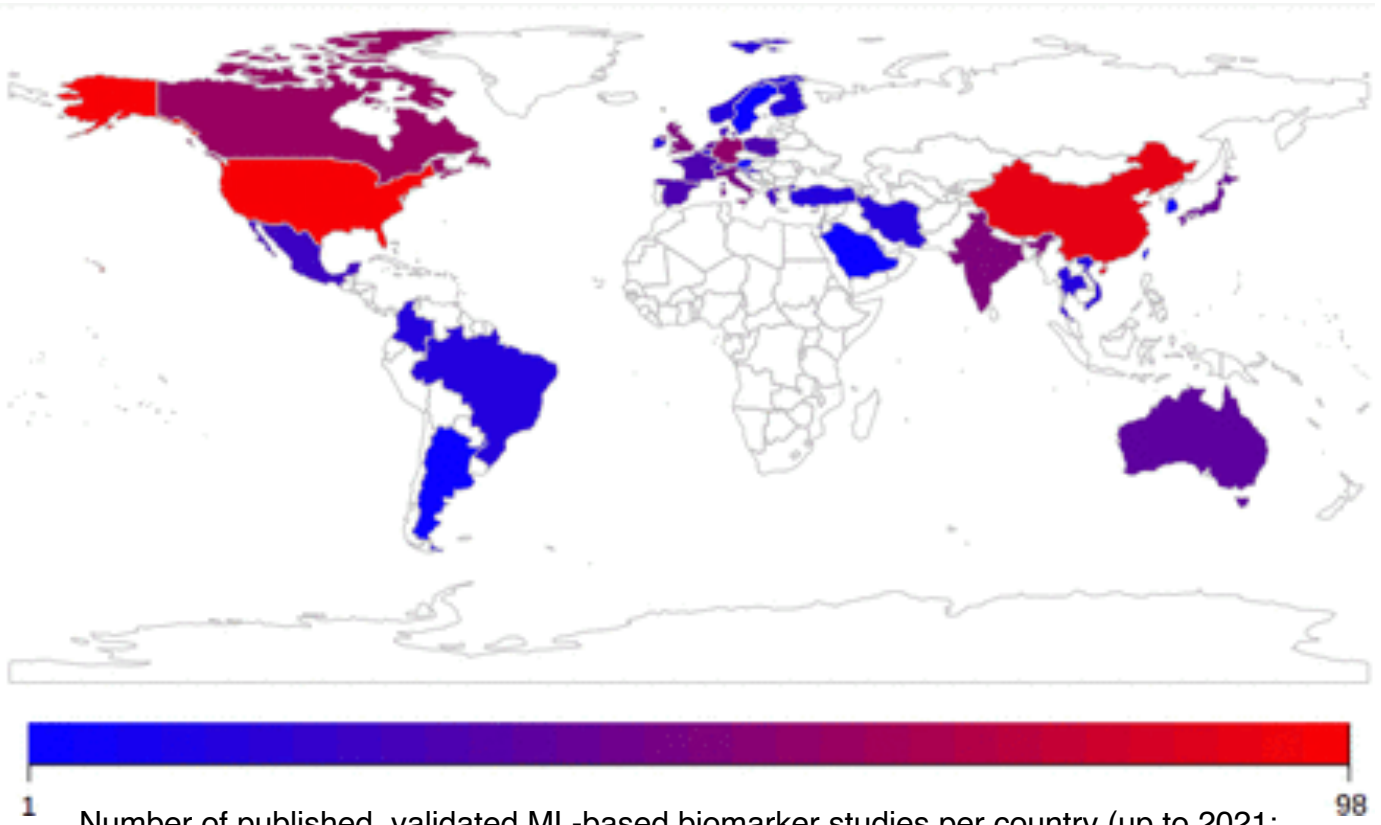
→ cohort-specific biases, choice of suitable validation schemes

(8) ensuring model interpretability and biological plausibility

→ use of black-box instead of white-box modeling methods

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

Gaps and challenges: Country representation



Number of published, validated ML-based biomarker studies per country (up to 2021; Glaab et al., BMJ Open, 2021)

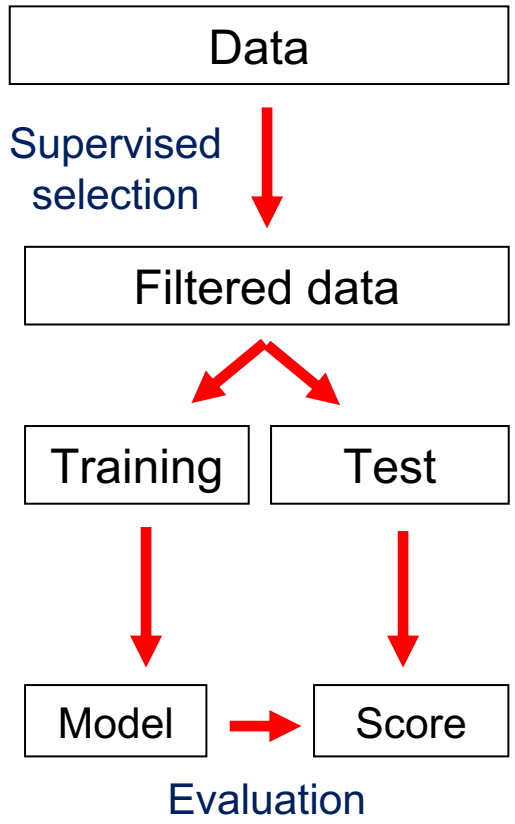
→ Great imbalances in country representation among published studies on validated, machine learning derived biomarker signatures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

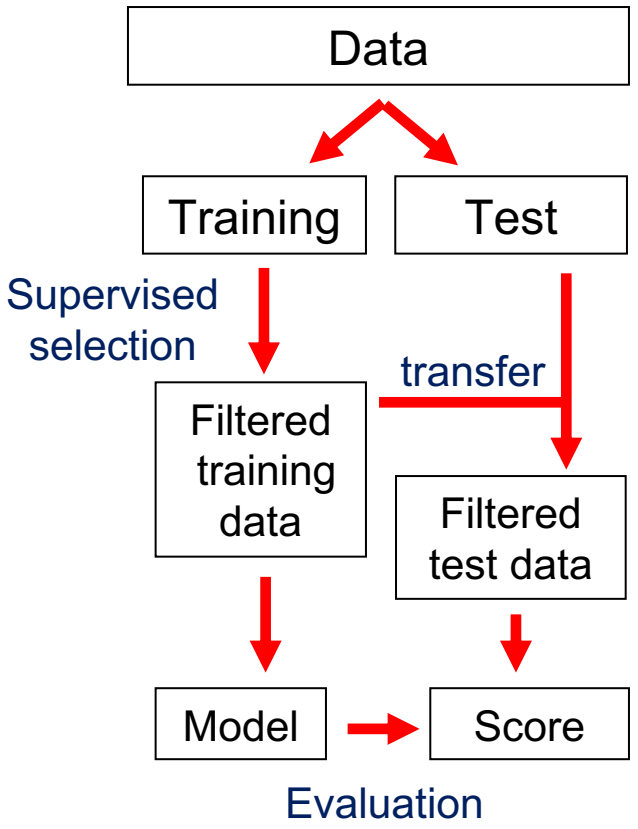
Gaps and challenges: Workflows

PERMIT
PERsonalised MedicIne Trials

Error: Global feature selection



Suggested approach



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

European Commission

Example use cases – Main conclusions

Shared characteristics of prior successful uses cases:

- Early and rigorous filtering:
  Statistical, clinical and biological filtering criteria applied in the initial model development (strict inclusion/exclusion criteria; multiple layers of statistical and ML-based feature selection; integration of prior knowledge)

- Integration of diverse data types & measurement technologies:
  exploiting pathway & network data & technological advances (e.g., progressing from microarray technology to deep sequencing, RT-PCR and digital PCR)

- Robust validation schemes:
  internal multi-level cross-validation + independent external validation involving multiple performance metrics, large sample sizes, and multiple cohorts

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825

References

1. A. Rauschenberger, Z. Landoulsi, M. A. van de Wiel, E. Glaab. *Penalized regression with multiple sources of prior effects*, Bioinformatics (2022), 39(12), doi: 10.1007/s12035-022-02985-2.
2. M. Ali, O. Uriarte Huarte, T. Heurtaux, P. Garcia, B. Pardo Rodriguez, K. Grzyb, R. Halder, A. Skupin, M. Buttini, E. Glaab. *Single-Cell Transcriptional Profiling and Gene Regulatory Network Modeling in Tg2576 Mice Reveal Gender-Dependent Molecular Features Preceding Alzheimer-Like Pathologies*, Mol Neurobiol (2022), doi:10.1007/s12035-022-02985-2.
3. A. Rauschenberger, E. Glaab. *Predicting Dichotomised Outcomes from High-Dimensional Data in Biomedicine*, Journal of Applied Statistics, (2023), doi: 10.1080/02664763.2023.2233057.
4. L. C. Tranchevent, R. Halder, E. Glaab. *Systems level analysis of sex-dependent gene expression changes in Parkinson's disease*, NPJ Parkinson's Disease, (2022), 9, 8.
5. A. Rauschenberger, E. Glaab, *Predicting correlated outcomes from molecular data*, Bioinformatics (2021), 37(21), 3889–3895
6. R. Diaz-Uriarte, E. Gómez de Lope, R. Giugno, H. Fröhlich, P. V. Nazarov, I. A. Nepomuceno-Chamorro, A. Rauschenberger, E. Glaab, *Ten Quick Tips for Biomarker Discovery and Validation Analyses Using Machine Learning*, PLoS Computational Biology (2022), doi:10.1371/journal.pcbi.1010357
7. E. Glaab, J.P. Trezzi, A. Greuel, C. Jäger, Z. Hodak, A. Drzezga, L. Timmermann, M. Tittgemeyer, N. J. Diederich, C. Eggers, Integrative analysis of blood metabolomics and PET brain neuroimaging data for Parkinson's disease, Neurobiology of Disease (2019), Vol. 124, No. 1, pp. 555
8. S. Köglsberger, M. L. Cordero-Maldonado, P. Antony, J. I. Forster, P. Garcia, M. Buttini, A. Crawford, E. Glaab, *Gender-specific expression of ubiquitin-specific peptidase 9 modulates tau expression and phosphorylation: possible implications for tauopathies*, Molecular Neurobiology (2017), 54(10), pp. 7979
9. N. Vlassis, E. Glaab, *GenePEN: analysis of network activity alterations in complex diseases via the pairwise elastic net*, Statistical Applications in Genetics and Molecular Biology (2015), 14(2), 221
10. E. Glaab, *Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification*, Briefings in Bioinformatics (2015), 17(3), pp. 440
11. E. Glaab, R. Schneider, *Comparative pathway and network analysis of brain transcriptome changes during adult aging and in Parkinson's disease*, Neurobiology of Disease (2015), 74, 1-13
12. E. Glaab, R. Schneider, *RepExplore: Addressing technical replicate variance in proteomics and metabolomics data analysis*, Bioinformatics (2015), 31(13), pp. 2235
13. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. Extending pathways and processes using molecular interaction networks to analyse cancer genome data, BMC Bioinformatics, 11(1):597, 2010
14. E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, A. Valencia. EnrichNet: network-based gene set enrichment analysis, Bioinformatics, 28(18):i451-i457, 2012
15. E. Glaab, A. Rauschenberger, R. Banzi, C. Gerardi, P. Garcia, J. Demotes-Mainard, and the PERMIT Group, Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review, BMC Open (2021), 11, e053674
16. D. M. Hendrickx, P. Garcia, A. Ashrafi, A. Sciortino, K. J. Schmit, H. Kollmus, N. Nicot, T. Kaoma, L. Vallar, M. Buttini, E. Glaab, A new synuclein-transgenic mouse model for early Parkinson's reveals molecular features of preclinical disease, Molecular Neurobiology (2020), 58, 576-602
17. C. Brzenczek, Q. Klopfenstein, T. Hähnel, H. Fröhlich, E. Glaab, *Integrating digital gait sensor data with metabolomics and clinical data to predict clinically relevant outcomes in Parkinsons disease*, npj Digital Medicine (2024), 7, 235
18. S. Le Bars, E. Glaab, *Single-Cell Cortical Transcriptomics Reveals Common and Distinct Changes in Cell-Cell Communication in Alzheimer's and Parkinson's Disease*, Molecular Neurobiology (2024), 10.1007/s12035-024-04419-7
19. E. Gómez de Lope, … , R. Krüger, E. Glaab, *Comprehensive blood metabolomics profiling of Parkinson's disease reveals coordinated alterations in xanthine metabolism*, npj Parkinson's Disease (2024), 10, 68
20. M. Ali, P. Garcia, L.P. Lunkes, A. Sciortino, M. Thomas, T. Heurtaux, K. Grzyb, R. Halder, D. Coowar, A. Skupin, L. Buée, D. Blum, M. Buttini, E. Glaab, *Single cell transcriptome analysis of the THY-Tau22 mouse model of Alzheimer's disease reveals sex-dependent dysregulations*, Cell Death Discovery (2024), 10, 119
21. R.T.J. Loo, O. Tsurkalenko, J. Klucken, G. Mangone, F. Khoury, M. Vidailhet, J.-C. Corvol, R. Krüger, E. Glaab, *Levodopa-induced dyskinesia in Parkinson's disease: Insights from cross-cohort prognostic analysis using machine learning*, Parkinsonism & Related Disorders (2024), Vol. **126**, No. 107054