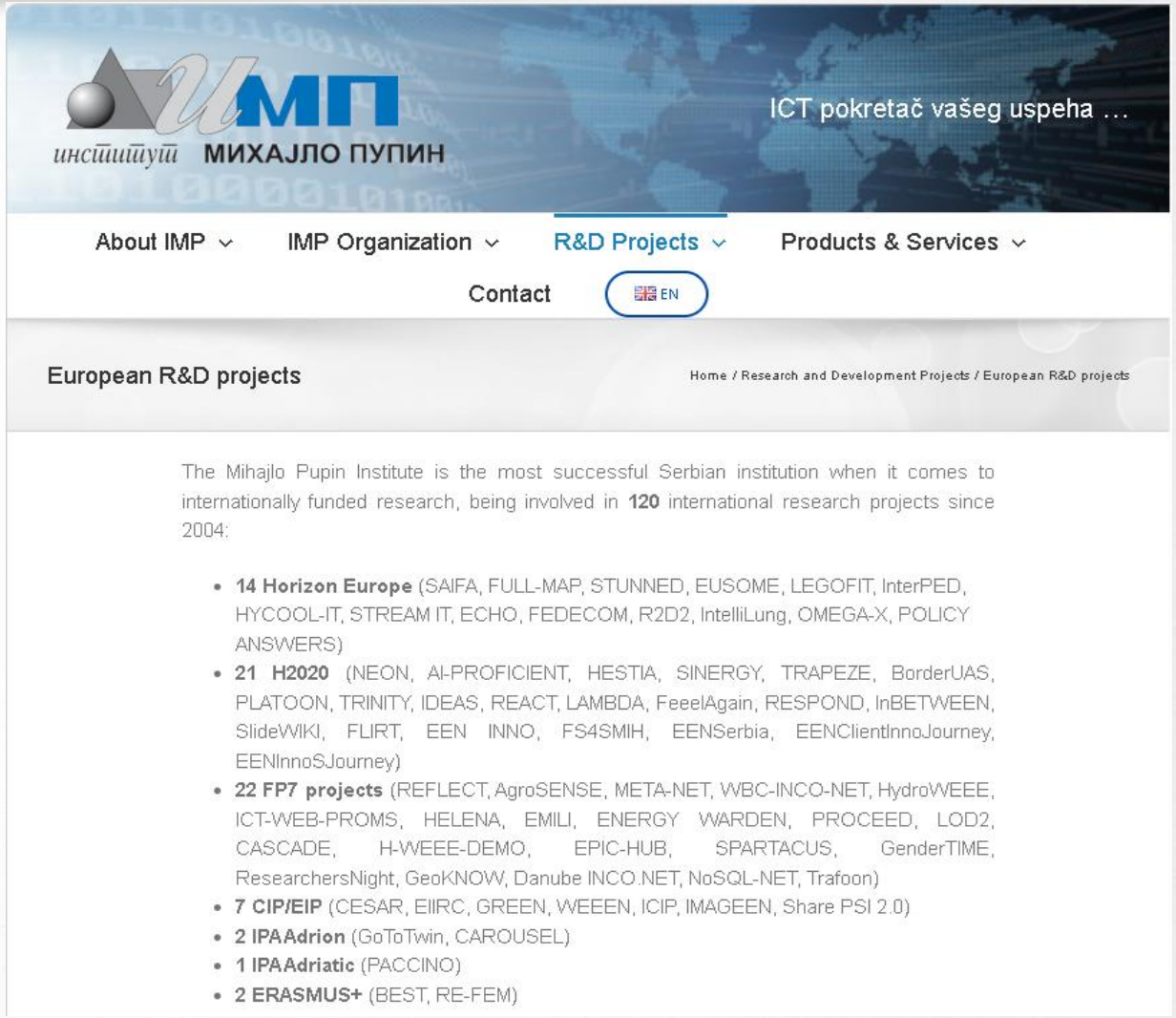
Session-2:
Explainable AI Demo on HPC

Valentina Janev, valentina.janev@pupin.rs
Miloš Nenadović, milos.nenadovic@pupin.rs
Dejan Paunović, dejan.paunovic@pupin.rs

institut MIHAJLO PUPIN

INSTITUTE MIHAJLO PUPIN

- Leading Serbian R&D institution in information and communication technologies (ICT)
- The biggest and oldest (1946) R&D Institute in ICT area in whole Southeastern Europe
- World Bank SEE Knowledge Economy report (2011, 2016): “Internationally competitive Institute”
- EU Commissionaire – “Pupin as the best practice example for bridging academia and industry”
- 90% of turnover via Technology Transfer



https://www.pupin.rs/en/research-and-development-projects/european-rd-projects/

The Institute Scientific and Exploitation Objectives

- The Institute has been involved in development of intelligent systems since 1980
- In May 2020, the Institute developed the first Serbian respirator (mechanical ventilation, MV)
- On the way to market, the device has to go through a comprehensive testing
- For AI-assisted MV, automatic AI-based control of the Respirator, we need a thorougly testing of the AI-based system including provision of explainability

https://www.pupin.rs/en/2020/05/the-first-serbian-respirator/

OVERVIEW

- Regulatory Framework (Trustworty AI Guidelines, 2019; AI Act, 2024)
  - AI Regulatory sandboxes as Frameworks for testing AI systems (August 2026)
  - Explainable AI (high-risk systems)
- Motivation (Healthcare)
- AI-DSS Introduced
- Explainable AI component in AI-DSS
- AI-DSS Testing Framework
- AI-DSS Demo
- Discussion

Policy Framework

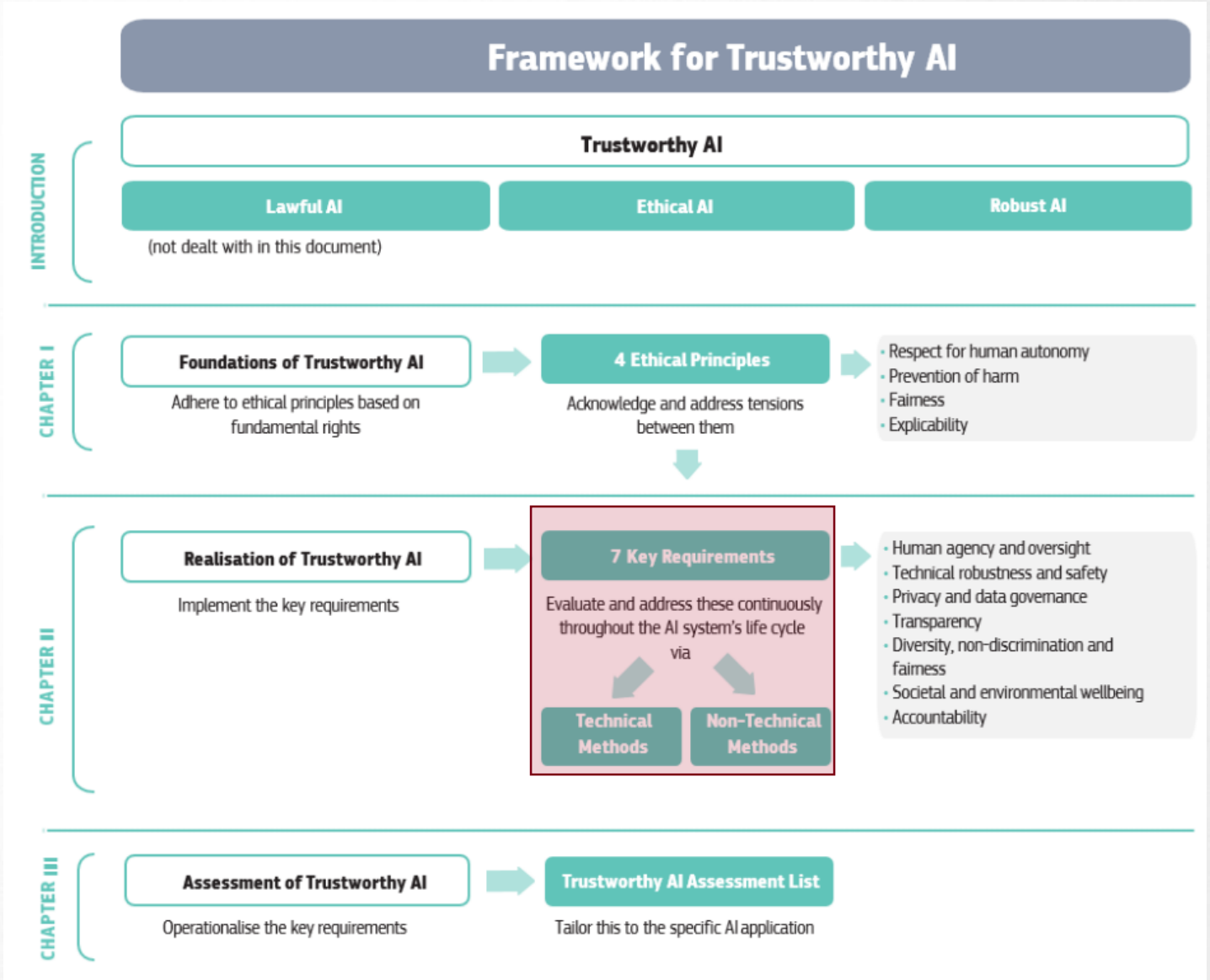
EU Policy Framework

- Ethical Guidelines for Trustworthy AI (2019)
  - four ethical principles (Respect for human autonomy, Prevention of harm, Fairness, Explicability)
- The AI Act (Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence) is the first-ever comprehensive legal framework on AI worldwide. The rules aim to foster trustworthy AI in Europe.

Medical Device Regulation 2017/745
https://eur-lex.europa.eu/eli/reg/2017/745/

Ethical Guidelines for Trustworthy AI

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai



Ethical Guidelines for Trustworthy AI

Explainability. Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency).

Transparent AI Design
"Ethics by design"

Explainable AI Challenges

- The goal – to make the AI models interpretable and transparent, understandable for human users
  - Neural networks (black-box), reinforcement learning process and decisions (actions)
  - However, we can not reduce the model complexity, but rather improves interpretability

- Challenge / Barriers
  - Balancing between model performance and interpretability
  - Bias detection and mitigation
  - Transferability of models (e.g. is the model developed on data from Western Europe applicable in China)
  - Privacy and security issues, especially when we put the Explainability component in practice, explanations might reveal sensitive information
  - How to build the Explainability layer (text explanations, visual explanations, instructions to control devices, etc.) to allows human users to comprehend and trust the results and output created by machine learning algorithms
  - Knowledge transfer and skills of users

TRUST

AI Act – Compliance Requirements and Business Risks

- Section III regulates High-risk AI systems
- Providers of High-risk AI systems must ensure compliance with the requirements set out in Articles 8–15 throughout the system's lifecycle, including the establishment of a documented risk management system, robust data governance measures, detailed technical documentation, automatic logging, appropriate human oversight, and safeguards for accuracy, robustness, and cybersecurity. Prior to placing a system on the market or putting it into service, providers must carry out the applicable conformity assessment, draw up an EU declaration of conformity, affix the CE marking, and register the system in the EU database. Ongoing obligations include corrective actions, cooperation with competent authorities, and maintaining a quality management system that enables continuous compliance.

- Deployers of High-risk AI systems are required to use such systems strictly in accordance with the provider's instructions and to implement appropriate technical and organisational measures, including assigning trained and competent human oversight.

https://www.legalnodes.com/article/eu-ai-act-2026-updates-compliance-requirements-and-business-risks

Frameworks for testing AI systems (August 2026)

- Mandatory & Regulatory Frameworks
  - **EU AI Act (High-Risk Requirements):** Effective August 2, 2026, this is the primary regulatory framework. It mandates that high-risk systems undergo strict testing for accuracy, robustness, cybersecurity, and bias mitigation. It requires documented, reproducible, and traceable evaluation, with fines up to 7% of global annual turnover or EUR 35 million.
  - **EU AI Regulatory Sandboxes:** By August 2, 2026, each EU Member State must establish at least one national regulatory sandbox to provide a controlled environment for testing and validating AI systems before market entry.
    Sectorial AI Testing and Experimentation Facilities under the Digital Europe Programme

Serbian AI Factory
https://www.hpc.rs/news/serbian-artificial-intelligence-antenna-factory

Intelligent Lung Support for Mechanically Ventilated Patients in the Intensive Care Unit (IntelliLung)

Call: HORIZON-HLTH-2021-Disease-04

Grant Agreement Number: 101057434

Artificial Intelligence Decision Support System (AI-DSS)

on behalf of the IntelliLung Consortium

Funded by the European Union

Facts and Figures

15 medical and technical partners

Intelligent Lung Support for Mechanically Ventilated Patients in the Intensive Care Unit (IntelliLung)

Call: HORIZON-HLTH-2021-Disease-04

Grant Agreement Number: 101057434

Coordinator: TUD University of Technology Dresden

Funding Period: 1 September 2021 - 31 August 2027

Total Funding: around € 5,98 million

Principal Investigator: Dr. med. Jakob Wittenstein (TUD, University Hospital Dresden)

Follow us on LinkedIn **@IntelliLung**

visit **intelliLung-project.eu**

Funded by the European Union

TECHNISCHE UNIVERSITÄT DRESDEN

Universitätsklinikum Carl Gustav Carus
THE DRESDENERS.

European Society of Anaesthesiology and Intensive Care

Cleveland Clinic

EKFZ Digital Health

InfAI
Institut für Angewandte Informatik

IMP institute MIHAILO PUPIN

Fundación Pública Andaluza Progreso y Salud
Consejería de Salud y Consumo

EPC European Project Center

Hospital Universitario de La Princesa
SaludMadrid

DEMCON | MACAWI RESPIRATORY SYSTEMS

Parc Taulí
Institut d'Investigació i Innovació I3PT

KKS Dresden

bettercare

DEMCON | ADVANCED MECHATRONICS

INSELSPITAL
UNIVERSITÄTSSPITAL BERN
HÔPITAL UNIVERSITAIRE DE BERNE

Università di Genova

OSPEDALE POLICLINICO SAN MARTINO
Sistema Sanitario Regione Liguria

Was hab' ich?

The National Institute of Medicine of the Ministry of Interior and Administration

Funded by the European Union

IntelliLung Consortium

intellilung-project.eu

Objectives - AI Decision Support System

- **Background:** Optimizing mechanical ventilation (MV) is complex and prone to errors. With rising ICU demand and staff shortages by 2030, inappropriate settings risk lung damage and increased mortality.
- **Objective:** Develop AI-based decision support (AI-DSS) system to provide recommendations for MV settings to reduce ventilator-induced lung injury (VILI) and ventilator time.
- **Significance:** The AI system improves ICU care by reducing ventilator time, improving survival, reducing complications, and lowering healthcare costs.

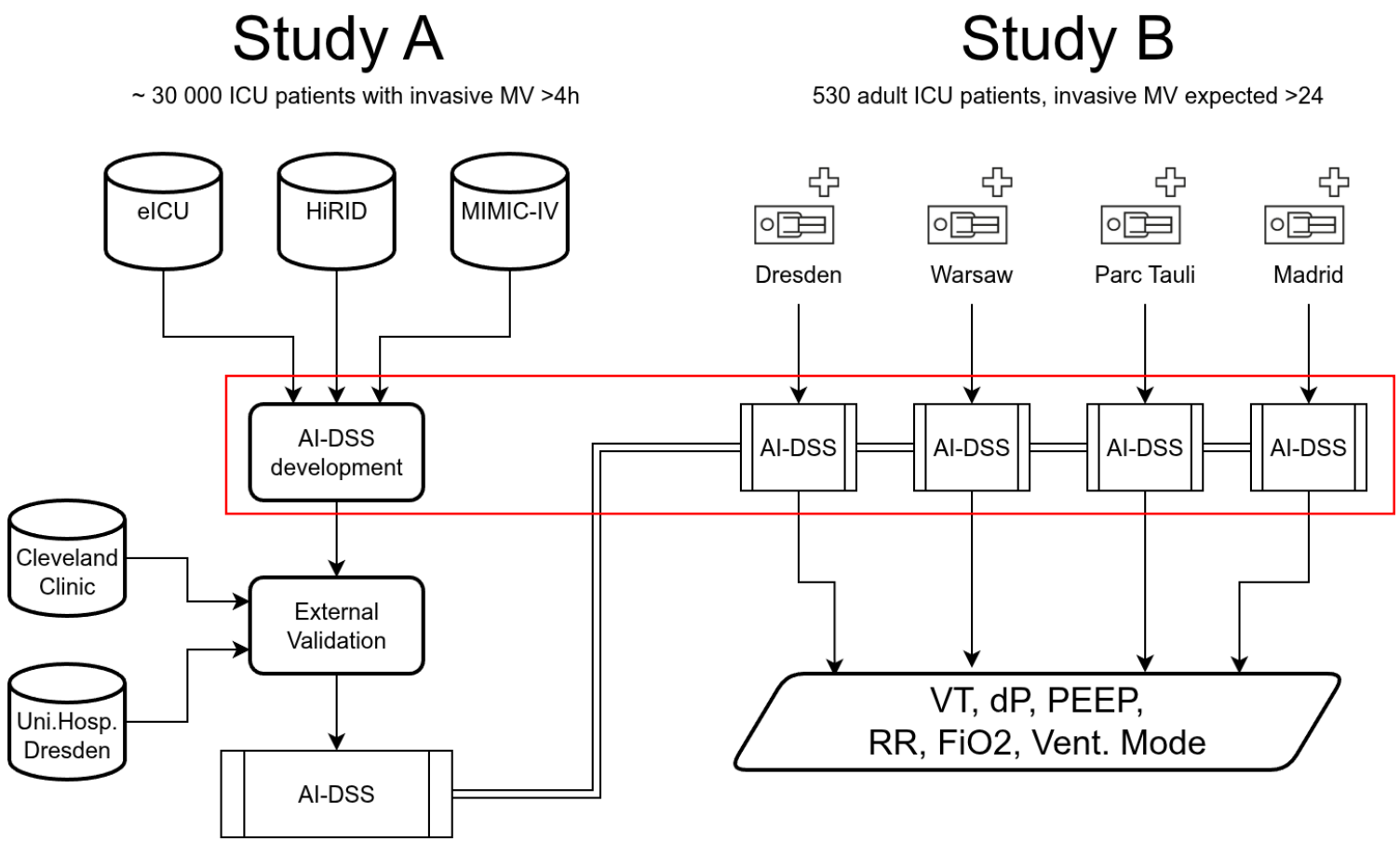
Methods

Overall Goal:

- Optimize invasive ventilation of critically ill patients in the intensive care unit (ICU)

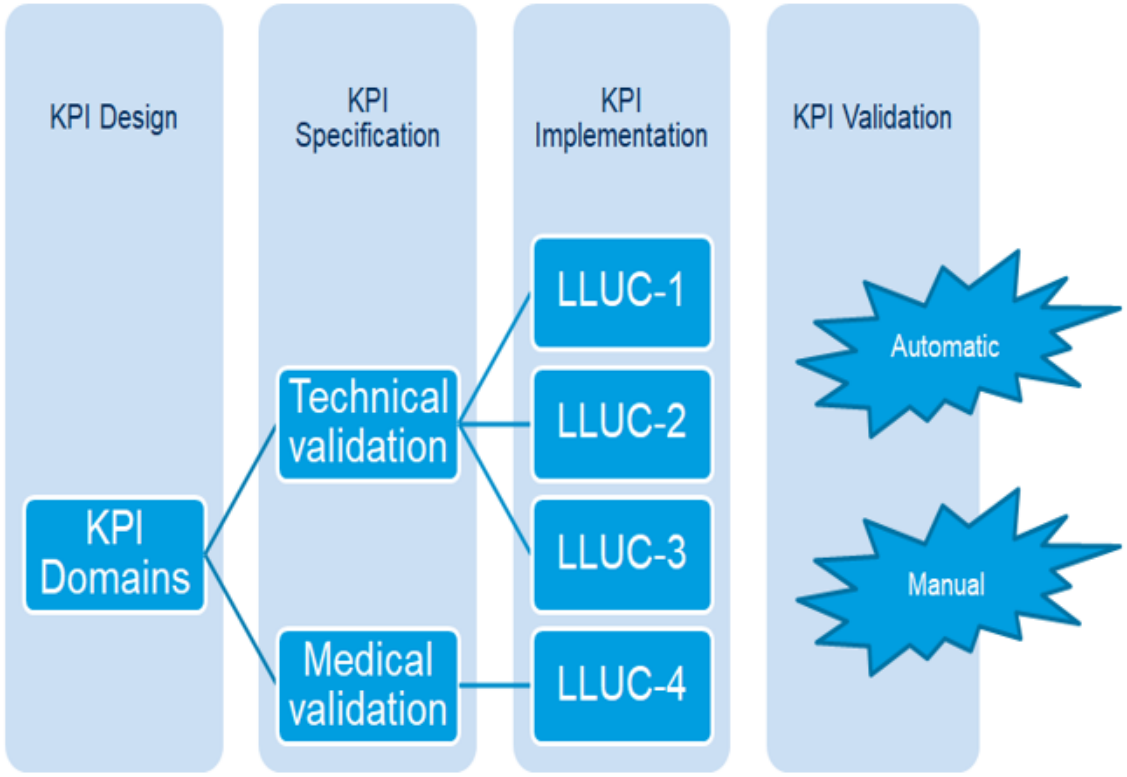
Methods:

- Study A, Study B
- Development of an AI-based decision support system that provides the care team with recommendations for the setting of invasive MV.



...invasive mechanical ventilation (MV) stands out as one of the most frequently employed life-saving treatments...
...the significance of MV has been especially underscored during the COVID-19 pandemic in 2020...

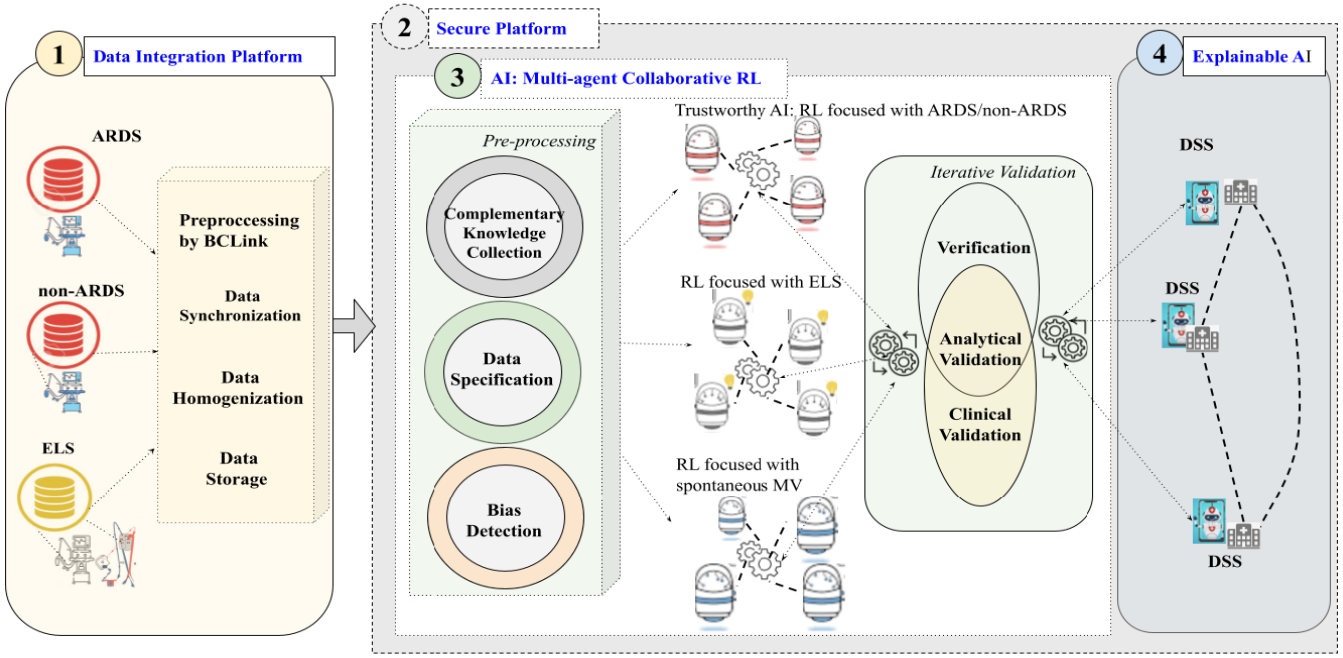
Trustworthiness Evaluation Framework



- **Four-phase methodology** (KPI design, KPI specification, KPI implementation, KPI validation).
- The methodology takes into consideration the
  - "**Ethics Guidelines for Trustworthy AI**", by High-Level Expert Group on Artificial Intelligence, 2019
  - **Medical Device Regulation 2017/745**

V. Janev *et al*., "IntelliLung AI-DSS Trustworthiness Evaluation Framework," *2024 32nd Telecommunications Forum (TELFOR)*, Belgrade, Serbia, 2024, pp. 1-4, doi: 10.1109/TELFOR63250.2024.10819068.

Transparent AI Design, “Ethics by design”

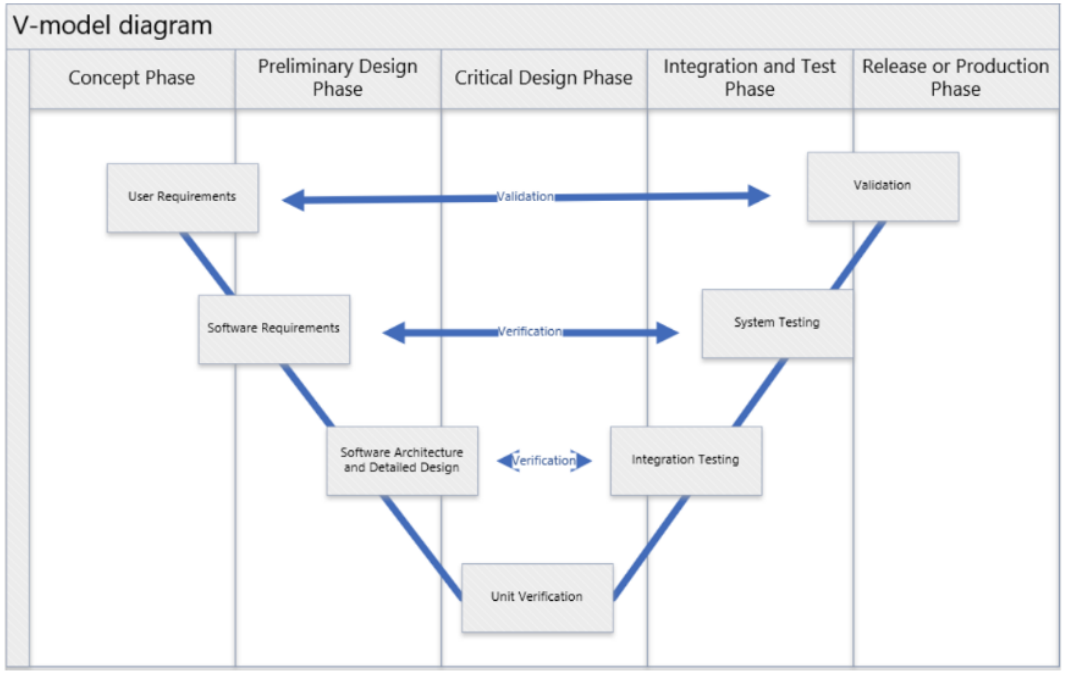


Requirements Elicitation Methodology

IEC 62559-2:2015 Use Case Methodology

- LLUC-01 Live streaming and Integration with Data Integration System; Pre-processing and workflow optimisation
- LLUC-02 Integration for AI, Workflow orchestration and data exchange
- LLUC -03 Trustworthy AI Customisation and Explainable AI
- LLUC -04 Trustworthy AI Validation.

Software Development Process & Trustworthiness Evaluation

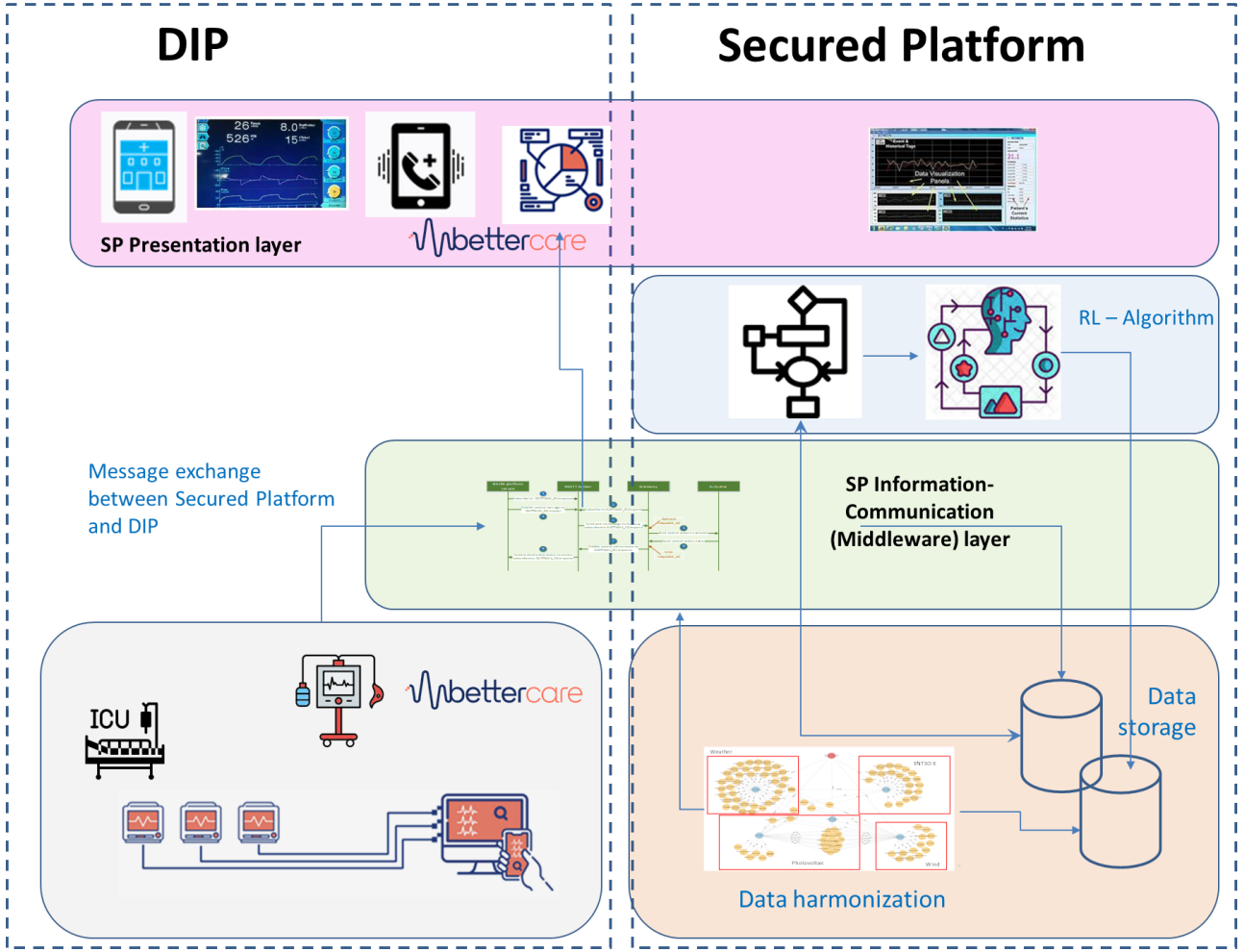


Software Unit ID	Specification Document
SRS 5.2.1_001 Data Integration Platform (DIP)	*Test Cases for* • KPI 1.1 - Data acquisition • KPI 1.2 - License validation • KPI 1.3 - Data harmonisation • KPI 1.4 - Data storage • KPI 1.5 - API post verification
SRS 5.2.1_002 Pre-processing Component	*Test Cases for* • KPI 2.1 Completeness of input data • KPI 2.2 Validity of input data • KPI 2.3 Time efficiency of the preprocessing pipeline • KPI 2.4 Adaptability of the preprocessing algorithm • KPI 2.5 Reliability of preprocessing
SRS 5.2.1_004 AI Algorithm	*Test Cases for* • KPI 3.1 Algorithm Performance • KPI 3.2 Qualitative Analysis • KPI 3.3 Error Analysis • KPI 3.4 Optimality Convergence • KPI 3.5 Sample Efficiency • KPI 3.6 Cross-Validation • KPI 3.7 Off-Policy Evaluation (OPE)

Table 4

AI Decision Support System

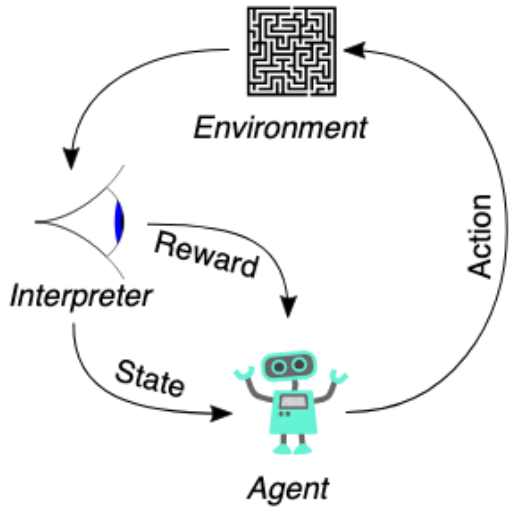
- **Goal:** reduce ventilator time, improve survival, reduce complications, and lower healthcare costs.



HL7 International

Reinforcement Learning (RL) Algorithm

- **Multi-agent reinforcement** learning involves multiple agents interacting with each other and the environment, learning to achieve individual or collective goals through reinforcement learning techniques.
- Through in-depth analysis, we have concluded that a **single-agent RL framework** is currently not only sufficient but also advantageous for achieving our objectives.



The main goal in reinforcement learning is to learn an **optimal policy** that maximizes cumulative rewards over time.
"Policy" refers to the strategy that an agent employs to **determine its actions based on the current state of the environment**. It defines the behavior of the agent by mapping states to actions.
The learning process often involves exploring different actions and exploiting known successful actions to improve the policy progressively.
State: variables and their ranges, priorities, id's and other metaparameters
Actions: variables and their bin ranges
Reward (teaching signal in Algorithm**)**: the required variables, their safe ranges or conversion formulas

Benefits from Explainability (the IntelliLung case)

- **Feedback from medical doctors in the beginning of the project (2023)**
  - “While the model may produce seemingly accurate output through that process, the rationale of the computation cannot be provided to the end users in most cases. This might lead to resistance of AI models into daily practice, as clinicians fear that performing unnecessary interventions or changing a treatment strategy without supporting scientific evidence could do more harm than help.” [TUD]

- **SHAP (SHapley Additive exPlanations)-based visualizations (2024)**
  - Visualizations facilitate a direct comparison between the decision-making processes of clinicians and the algorithm
  - Help to uncover and understand which data feature influence the RL policy suggestion
  - Open opportunities for interpretation of recommendations for MV as a decision support system
  - Ensures the results generated by the model are both medically meaningful and possess clinical validity
  - X-Vent: ICU Ventilation with Explainable Model-Based Reinforcement Learning

Explainability vs Interpretability [IBM]

- https://www.ibm.com/think/topics/explainable-ai

- Explainable AI is one of the key requirements for implementing responsible AI, a methodology for the large-scale implementation of AI methods in real organizations with fairness, model explainability and accountability
- Interpretability is the degree to which an observer can understand the cause of a decision. It is the success rate that humans can predict for the result of an AI output, while explainability goes a step further and looks at how the AI arrived at the result.

Intelligent Lung Support for Mechanically Ventilated Patients in the Intensive Care Unit (IntelliLung)

Call: HORIZON-HLTH-2021-Disease-04

Grant Agreement Number: 101057434

Trustworthiness Evaluation in AI applications: Example from the IntelliLung project

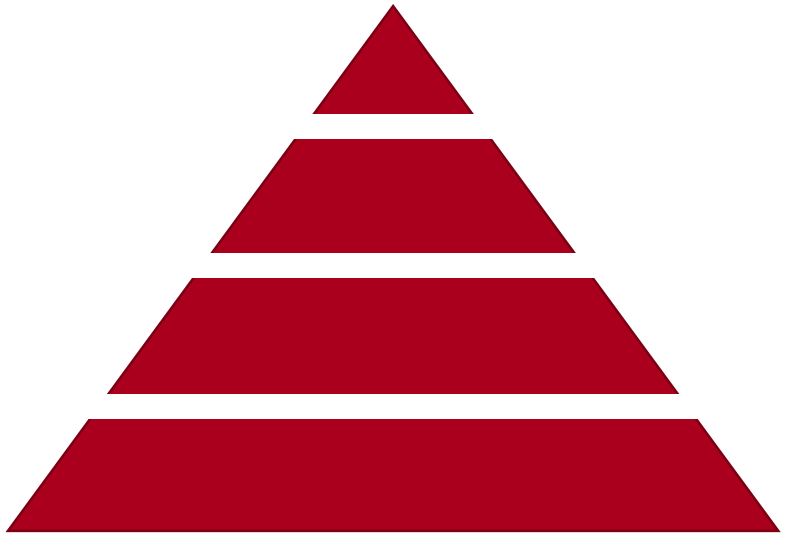
Institute Mihajlo Pupin Belgrade, Serbia

on behalf of the IntelliLung Consortium

Funded by the European Union

Scope of Testing and Validation

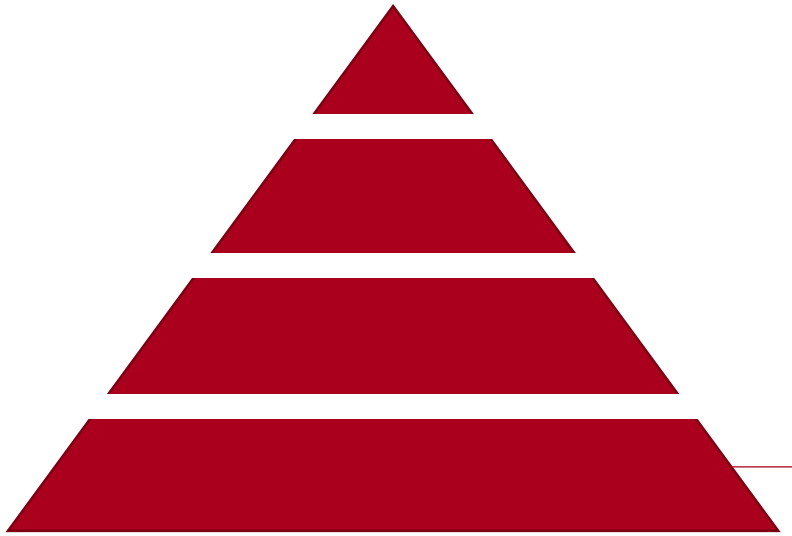
AI-DSS is not an autonomous system



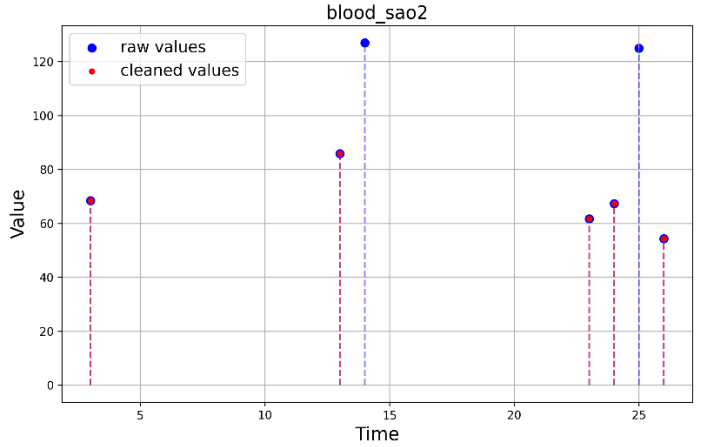
The scope of required testing for a ***Class IIb*** **medical device** can be split into 4 sections.

Class IIb: Class IIb devices fall between Class IIa and III, posing a significant risk but not as high as life-supporting or long-term implantable devices.

Scope of Testing and Validation



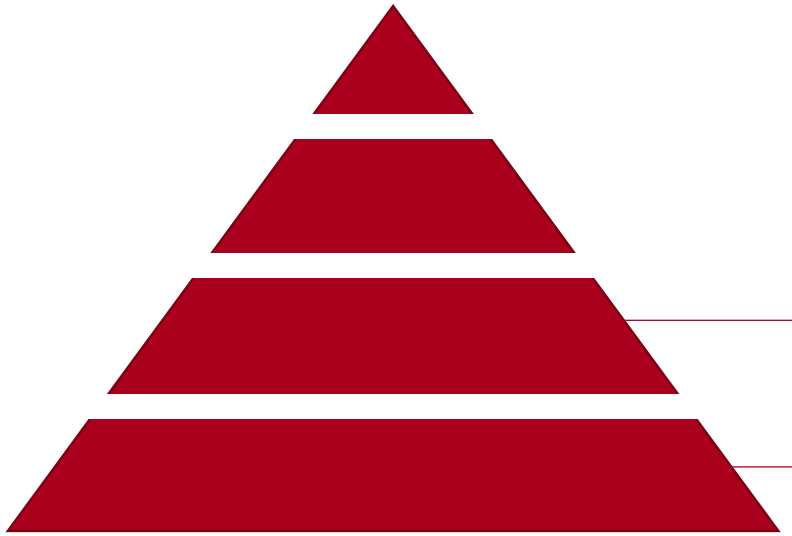
Unit level




```
configfile: pyproject.toml
collected 6 items

test/test_preprocessing.py::test_cohort_selection[cohort0] PASSED       [ 16%]
test/test_preprocessing.py::test_cohort_selection[cohort1] PASSED       [ 33%]
test/test_preprocessing.py::test_cohort_selection[cohort2] PASSED       [ 50%]
test/test_preprocessing.py::test_cohort_selection_fail[cohort0] PASSED  [ 66%]
test/test_preprocessing.py::test_cohort_selection_fail[cohort1] PASSED  [ 83%]
test/test_preprocessing.py::test_cohort_selection_fail[cohort2] PASSED  [100%]

============================== 6 passed in 0.25s ===============================
```

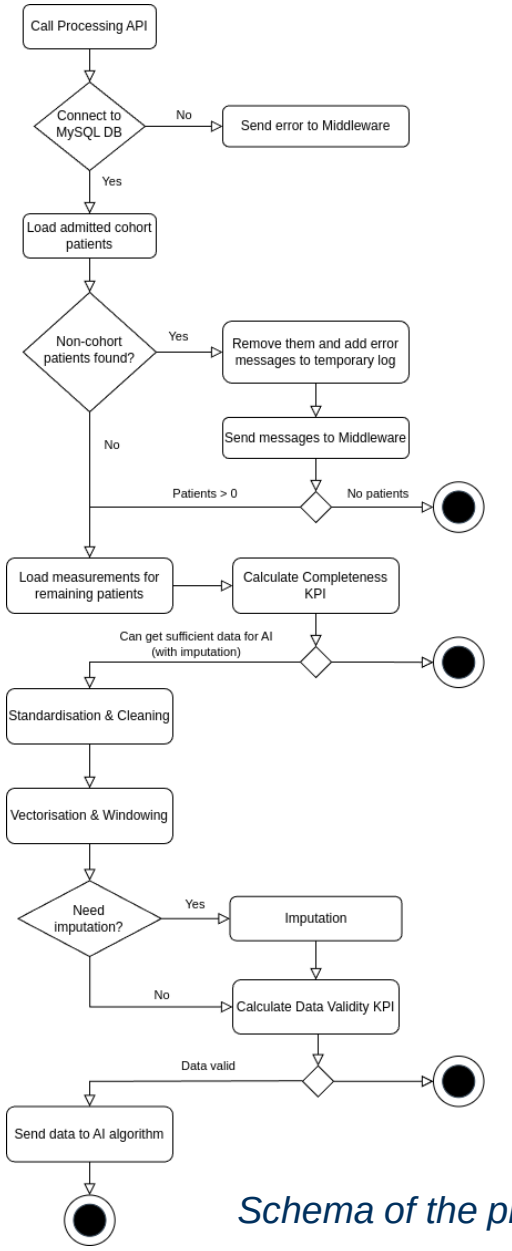
```
D:\Development\Intellilung\Intellilung-Secured-Platform\Middleware>npm test

> intellilung-middleware@0.6.6 test
> jest

 PASS  test/user.controller.test.js
 PASS  test/hospital.controller.test.js
 PASS  test/export.controller.test.js
 PASS  test/recommendationfeedback.controller.test.js
 PASS  test/catalog.controller.test.js
 PASS  test/error.controller.test.js
 PASS  test/measurement.controller.test.js
 PASS  test/patient.controller.test.js
 PASS  test/recommendation.controller.test.js

Test Suites: 9 passed, 9 total
Tests:       12 passed, 12 total
Snapshots:   0 total
Time:        1.76 s, estimated 2 s
Ran all test suites.
```

Scope of Testing and Validation



Component level

Unit level

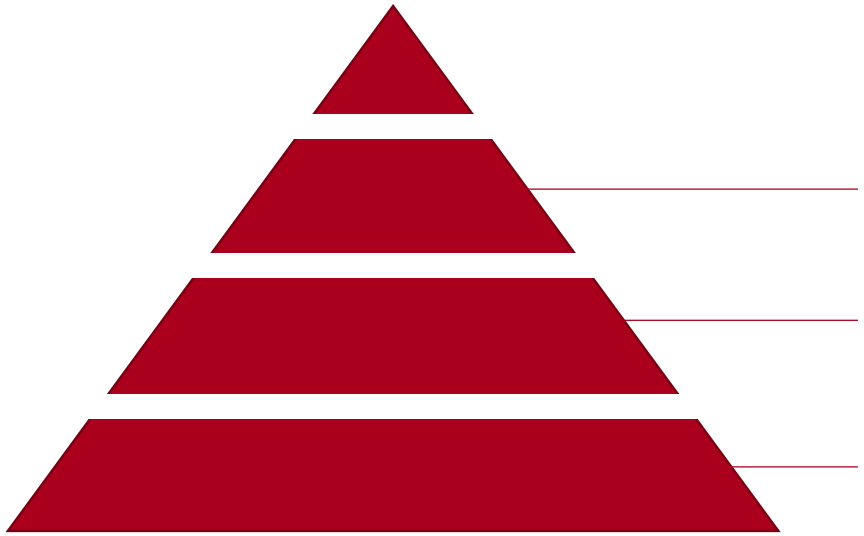


Schema of the preprocessing algorithm

Component testing handles individual larger components of the AI-DSS:

- Data Collection
- DIP API
- Middleware
- Preprocessing API
- Preprocessing
- AI algorithm

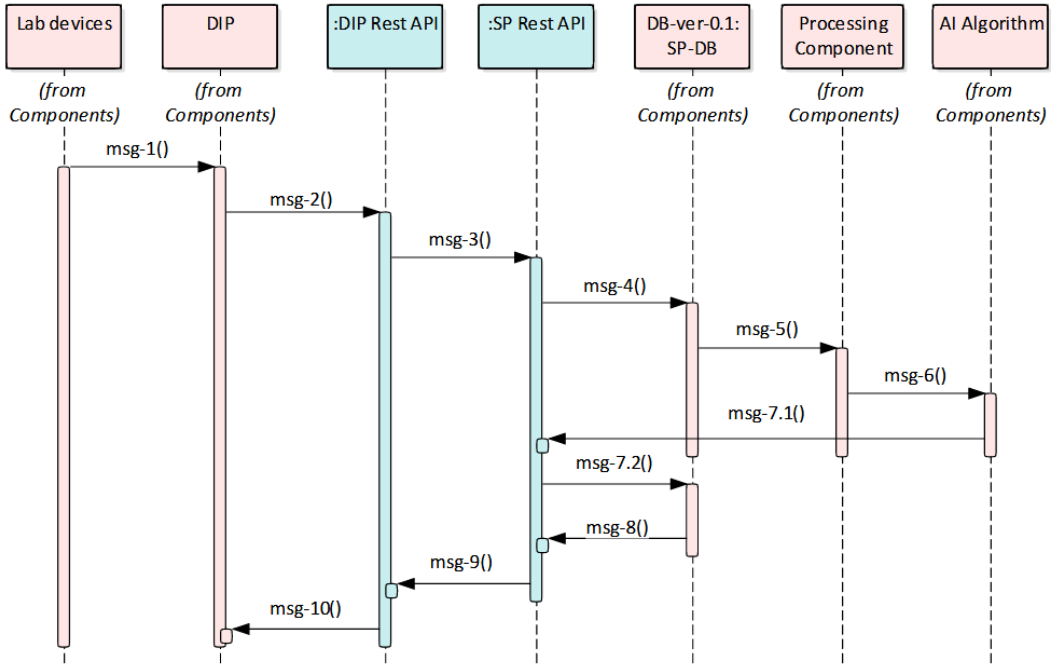
Scope of Testing and Validation



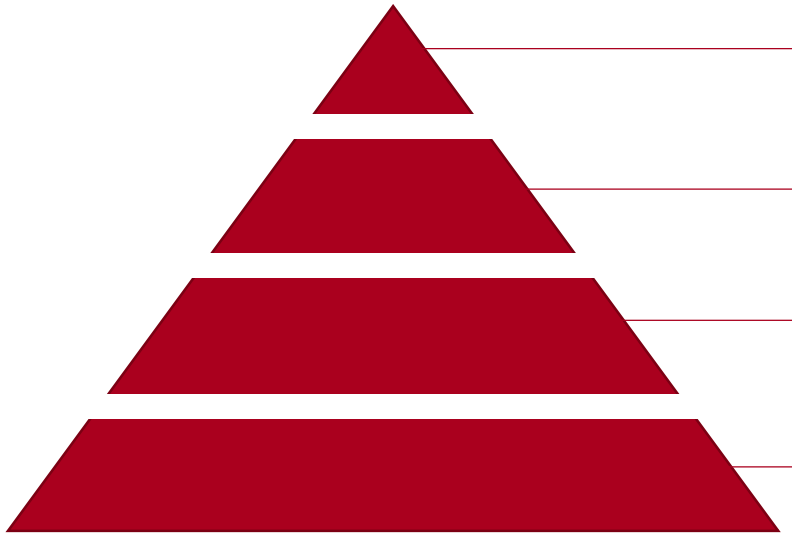
Integration level

Component level

Unit level



Scope of Testing and Validation

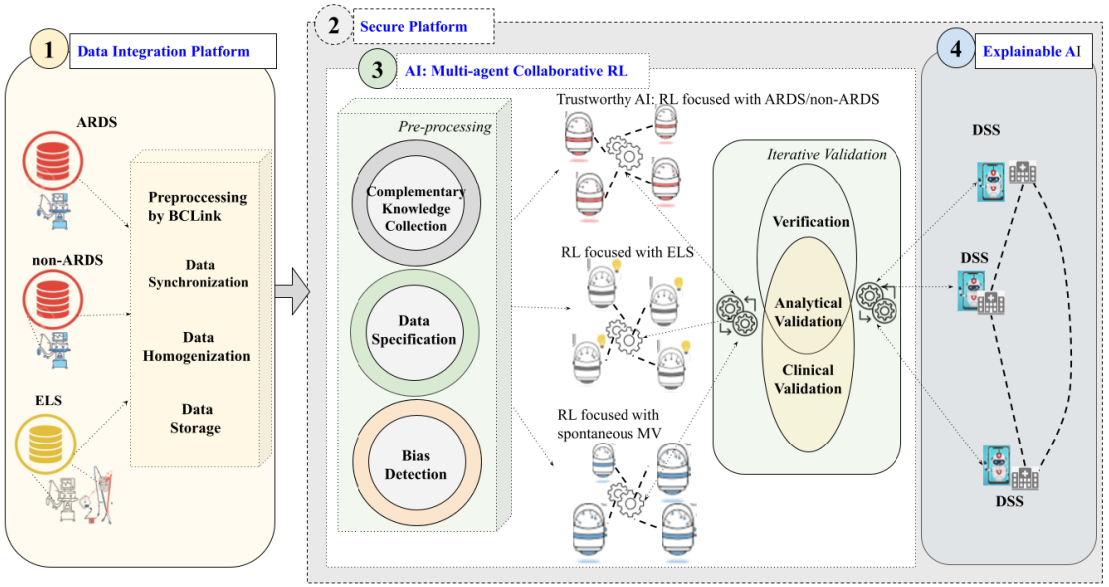


System level

Integration level

Component level

Unit level



IntelliLung Validation Methodology

IntelliLung Validation Methodology

Meaningful selection of indicators based on their contribution to either technical or medical validation.

IntelliLung Validation Methodology

KPI Specification → KPI Implementation → KPI Validation

Step by step description on how to calculate the defined KPIs.

GENERAL INFORMATION					
ID	KPI 2.1		**Name**	Completeness of input data	
Related LLUC	LLUC-1☒	LLUC-2☒	LLUC-3☐	LLUC-4☐	N/A☐
Description	Completeness measures the amount of information available with respect to the total information that could be available given the capture process and data format. Data unavailable in the dataset are called "missing".				
Formula	Number of missing variables/numbers of all input variables * 100 • For vital parameters • For laboratory parameters				
Monitoring frequency	• Start monitoring at month one at the Hospital with 1-minute resolution				
Units	%		**Related KPI**	[KPI Name/ID]	
Reporting	**Data upload rate**		Minute	**"Other" upload rate**	
	Information display		Cumulated value☒	Trend☐	N/A☐
CALCULATION/EXTRACTION METHODOLOGY					

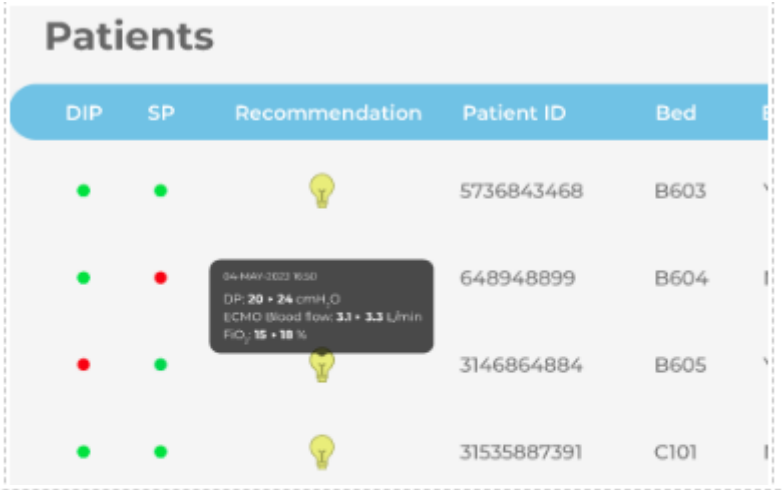
IntelliLung Validation Methodology

KPI Implementation

Writing additional software tools and services required for running the tests for the defined KPIs and storing the results.

KPI Validation

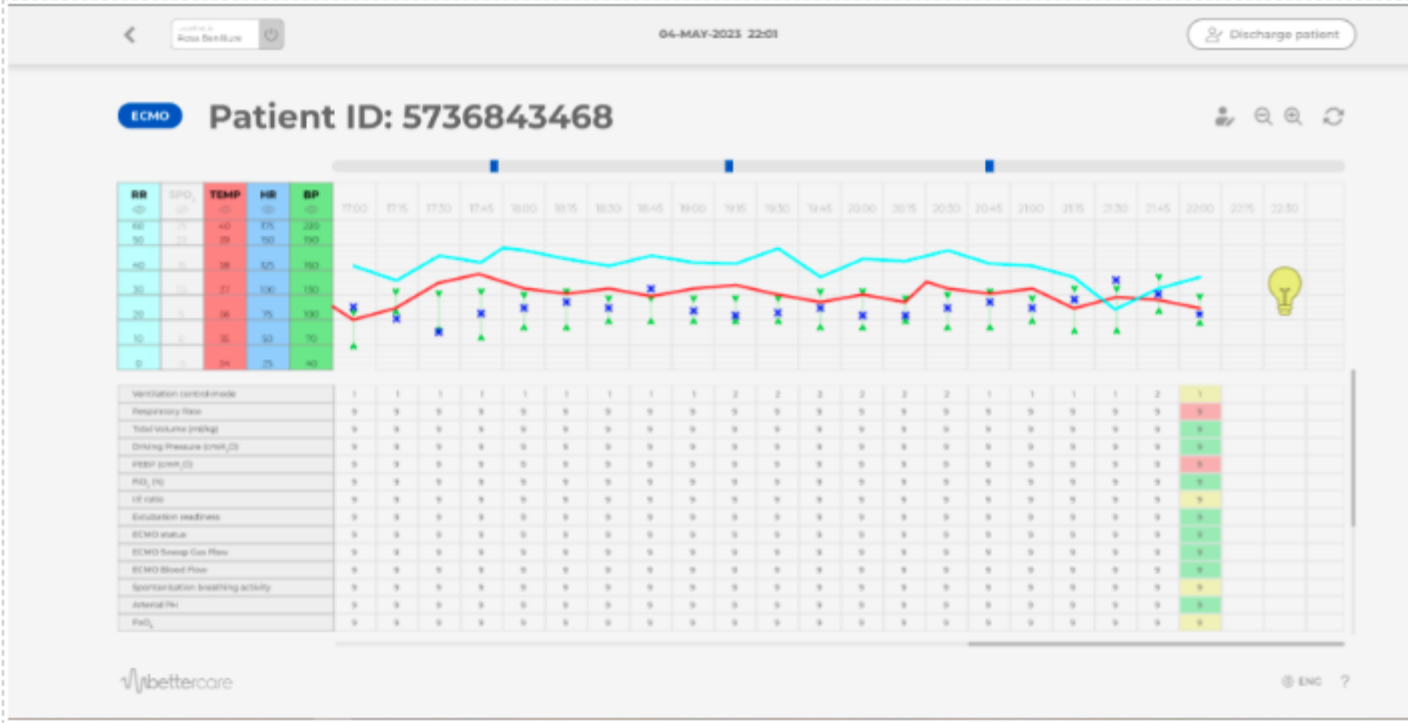
IntelliLung Validation Methodology



KPI Validation

Final phase where experts analyses the KPI testing results or use dashboards for

- Continuous monitoring
- Tracking progress
- Gaining insights into factors contributing to performance changes
- Identifying areas for improvement



Technical Robustness, Safety and Security

Cybersecurity Measures

Hospital systems protect AI-DSS against unauthorized access and ensure system availability. Measures include strong authentication, network segmentation, and intrusion detection.

Data Completeness

The Data Integration Platform (DIP) ensures comprehensive data collection. It records the reason for missing data, such as errors or unchanged variables.

Privacy and Data Governance

Role-Based Access

Only authorized personnel can access AI-DSS data through strict role-based access controls.

Data Quality & Integrity

Key performance indicators like *data validity* and *reliability* of preprocessing ensure data accuracy and prevent malicious data alteration.

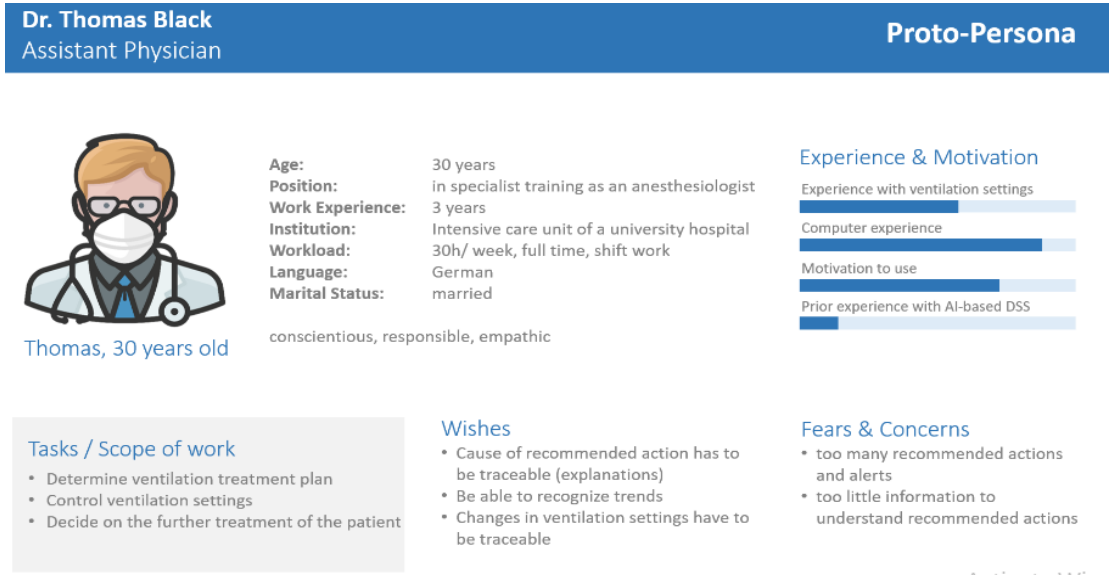
Privacy and Data Governance

Role-Based Access

Only authorized personnel can access AI-DSS data through strict role-based access controls.
e.g. Doctor

Data Quality & Integrity

Key performance indicators like *data validity* and *reliability* of preprocessing ensure data accuracy and prevent malicious data alteration.



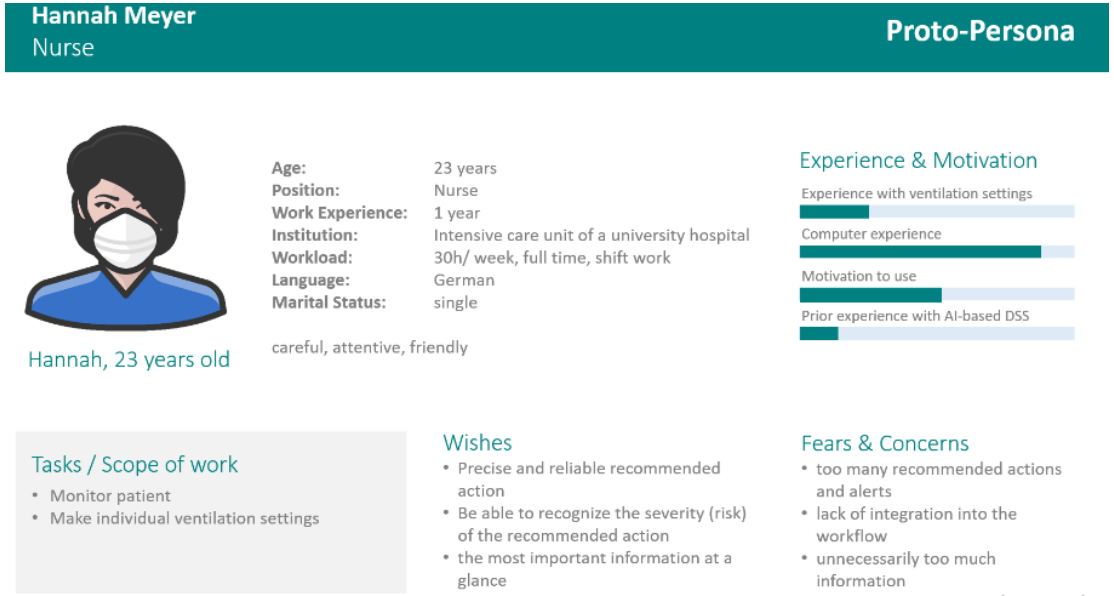
Privacy and Data Governance

Role-Based Access

Only authorized personnel can access AI-DSS data through strict role-based access controls.
e.g. Nurse

Data Quality & Integrity

Key performance indicators like *data validity* and *reliability* of preprocessing ensure data accuracy and prevent malicious data alteration.



Privacy and Data Governance

Role-Based Access

Only authorized personnel can access AI-DSS data through strict role-based access controls.
e.g. Nurse

Data Quality & Integrity

Key performance indicators like *data validity* and *reliability* of preprocessing ensure data accuracy and prevent malicious data alteration.

Metric Name	Unit	Metric Name	Unit
Completeness	Percentage of completed data fields in the specific dataset.	Consistency	Percentage of values that match across multiple sources.
Validity	Percentage of data fields whose values are within the domain of acceptable values.	Timeliness	Percentage of data that can be obtainable within a certain period (e.g. seconds).
Relevancy	Percentage of data that is relevant to the specific use case.	Integrity	Percentage of data that remained the same across multiple systems after being moved.
Auditability	Data and processes that allow tracking e.g. how data is used / overwritten / misused.	Uniqueness	Percentage of unique entries in the dataset.

Transparency and Traceability

Data Provenance

Provenance management tracks data collection and processing for full auditability and explainability of AI decisions.

Data Management

Data is anonymized, linked to patient/hospital IDs, and managed with a shared metadata catalogue across systems.

Explainability

Explainable RL framework enhances model transparency, aligning with medical validation and established practices.

Transparency and Traceability

Data Provenance

Provenance management tracks data collection and processing for full auditability and explainability of AI decisions.

Data Management

Data is anonymized, linked to patient/hospital IDs, and managed with a shared metadata catalogue across systems.

Explainability

Explainable RL framework enhances model transparency, aligning with medical validation and established practices.

Transparency and Traceability

Data Provenance

Provenance management tracks data collection and processing for full auditability and explainability of AI decisions.

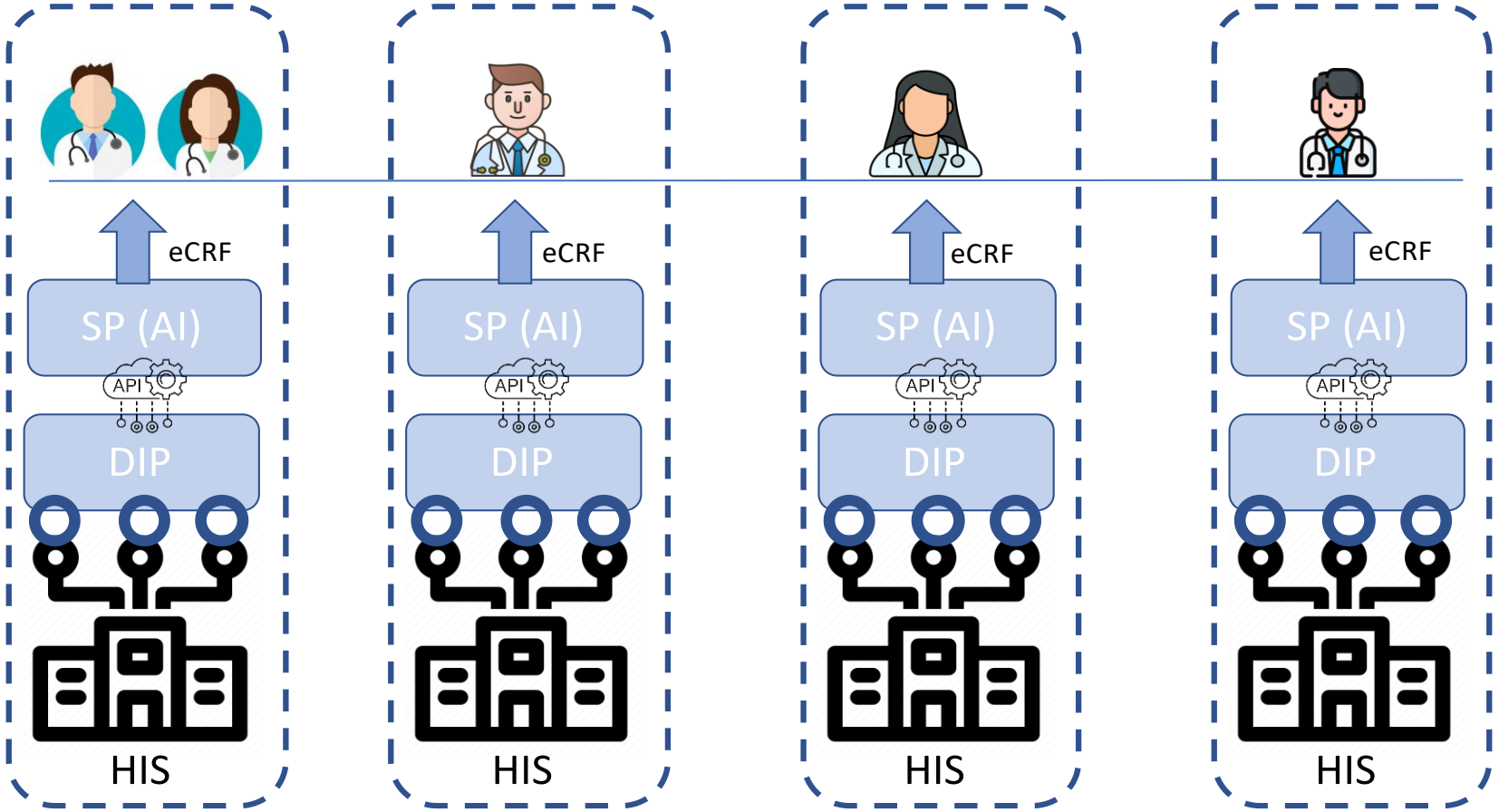
Data Management

Data is anonymized, linked to patient/hospital IDs, and managed with a shared metadata catalogue across systems.

Explainability

Explainable RL framework enhances model transparency, aligning with medical validation and established practices.

Compliance with Standards



Legend:

- **AI** Artificial Intelligence
- **SP** Secured Platform
- **DIP** Data Integration Platform
- Data connector (HL7)
- **HIS** Hospital Information System
- **eCRF** electronic Case Report Form
- **API** Application Programming Interface

Conclusion

- Explainable AI could be used to describe an AI model, its expected impact and potential biases.
- It helps characterize model accuracy, fairness, transparency and outcomes in AI-powered decision making.
- Explainable AI is crucial for building trust and confidence when putting AI models into production.

Follow us on LinkedIn
@IntelliLung

Visit
intelliLung-project.eu

Conclusion

- Dataset preparation took longer than planned.
- Different types of RL-based algorithms were explored and validated; Findings: trained policies improved performance vs. clinician actions under evaluation metrics, especially on safety-related objectives.
- Deployment finished in 3 hospitals in Europe; the clinical study is in place; the results are expected in 2027.

Follow us on LinkedIn
@IntelliLung

Visit
intelliLung-project.eu