Alexander Oeser, Anne Sophie Kubasch, Tim Meschke, Nora Grieb,
Lukas Schmierer, Uwe Platzbecker & Thomas Neumuth

KAIT

Knowledge-Driven and Artificial Intelligence-Based Platform for Therapy Decision Support in Hematology

White Paper

May 2021

K A I T

Knowledge-Based and AI-Driven Platform for Therapy Decision-Support in Hematology

White Paper

May 2021

Alexander Oeser, Anne Sophie Kubasch, Tim Meschke, Nora Grieb, Lukas Schmierer, Uwe Platzbecker & Thomas Neumuth

Technical Contact

Prof. Dr. Thomas Neumuth
thomas.neumuth@medizin.uni-leipzig.de

Alexander Oeser
alexander.oeser@medizin.uni-leipzig.de

Medical Contact

Prof. Dr. Uwe Platzbecker
uwe.platzbecker@medizin.uni-leipzig.de

Dr. Anne Sophie Kubasch
annesophie.kubasch@medizin.uni-leipzig.de

Editorial Note

This white paper covers the basic aspects of the development and operation of the KAIT platform for providing AI-supported therapy decision support in hematology. It is intended to provide the reader with an introduction to its functional and methodologic approaches. This document is a snapshot and the authors reserve the right to make functional and methodological changes in the course of the project should they become necessary.

ISBN 978-3-9822619-4-2

1 Background and Status Quo

The field of haematology is characterized by heterogeneous diseases and considerably varying patient disease courses. However, clinical trial design, drug development and the subsequent therapeutic decision have mostly relied on the administration of the same therapeutic regimen to an utterly diverse patient population. This one-size-fits-all treatment approach predisposes patients with hematologic diseases towards suboptimal response rates. Further progress in understanding the pathophysiology of complex haematological malignancies like myelodysplastic syndromes (MDS), acute myeloid leukaemia (AML) and multiple myeloma (MM) has changed treatment approaches in the last years. Hematologists now require an understanding of a rapidly evolving treatment paradigm that is increasingly nuanced, complex and patient-directed. The underlying heterogeneous disease biology demands differences in personalized therapeutic decisions, making individualized patient treatment a core objective in the hematologic field.

Until now, therapeutic decision-making still depends on whether the treating physician has the relevant therapeutic experience and access to novel therapies. The implementation of KAIT, an artificial intelligence (AI)[1] assisted therapy decision-support system[2] for patients with MDS, AML, and MM will guide patient-specific, personalized treatment and will overcome these challenges. KAIT will provide the treating physician with suggestions for individualized therapeutic strategies, even from a vast pool of candidate compounds, offering an optimal design for mono- vs. combination therapy with increased efficacy and safety in each therapy line.

2 Introducing the KAIT Platform

2.1 Motivation and Goals

KAIT will provide hematologists with highly innovative solutions to support medical case evaluation and therapy decision-making for patients with MDS, AML and MM. The platform enables a multi-layer view of the individual medical case in comparison with a comprehensive medical knowledge base. This advanced assessment enables a detailed case study that reflects the combined experience and expertise of multiple individual experts. Thus, the bundled

[1] **Artificial Intelligence (AI)** is a discipline of computer science to simulate human intelligence in machines. Typical application areas are robotics, computer vision, natural language processing, or machine learning.

[2] A **therapy decision-support system** enables the evaluation of medical patient information through a computer with the intention to suggest an optimal treatment option for the individual patient. It comprises the application of specialized methodological approaches (e.g. rule-based systems or machine learning) to make those assessments.

competence and the collected knowledge will be available for each patient with MDS, AML or MM at any time, allowing for an additional objective and purely data-driven perspective to be incorporated into the therapy decision-making process.

The overarching goal of the KAIT platform is to provide all MDS, AML or MM patients with therapy strategies that are specifically tailored to their individual characteristics and needs and are based on the latest state of available medical knowledge. The approach considers all aspects from the general physical constitution down to the variable molecular and genetic characteristics of the disease. In addition, KAIT will address the demanding aspects of process management and optimization in terms of preparation, planning and supervision of medical meetings and councils with the goal to enable easier access to meaningful expert panels, e.g. multicentric tumor boards and valuable information exchange across the platform community.

Finally, KAIT will provide a comprehensive platform for the collection, processing and delivery of medical knowledge as well as the latest findings in clinical research. As a perfect symbiosis of innovative and state-of-the-art methods deriving from the field of AI and a team of hematologists, the result is a tool that supports the derivation of complex therapeutic decisions in hematology in the long-term.

2.2 The KAIT Foundation: Holistic and Expert-Curated Disease Models

The platform utilizes multiple approaches for automatic reasoning[3] to provide results with clinical and methodological significance. To offer the highest amount of validity in KAIT's suggestions, the development process of its internal algorithms combines not only findings from extensive data analysis but also human competence of leading medical experts from the fields of hematology and biomedical research. For the decision-making process, our assessment algorithms consider entity-specific clinical practice guidelines (CPG)[4], medical publications and studies, as well as the experience gained through the consideration and integration of several thousand clinical cases.

To comply with the requirements for making complex decisions in the medical domain, we will apply the principle of »*traceable AI*[5]«, which makes the conclusions drawn by the system entirely understandable for the physicians. In

[3] The process of **reasoning**, in this context, describes the adaptation of the human ability to draw conclusions from a set of input signals through an IT system.

[4] A **clinical practice guideline (CPG)** is a document that contains best practices for the treatment of a patient based on current evidence.

[5] The concept of **traceable AI** includes the ability to trace all the individual conclusions that the system has drawn through a causal, intuitive representation. Thus, the user must be able to follow the path of the decision the system has made.

contrast to the concept of »*explainable AI (XAI)*«, which aims to deliver insights into the black box principles of machine learning methods, our concept is based on the prior integration of process-dependent causality which is aimed at reproducing human considerations rather than pure mathematical calculations. In this way, KAIT will offer a transparent secondary case assessment in the interdisciplinary therapeutic decision-making process, thus, ensuring an additional level of quality control and objectiveness.

The foundation for KAIT's extensive knowledge base is provided by specific disease models, which describe the characteristics of the integrated clinical pathologies in the form of causal models and thus make them compatible for computer-aided calculation using AI. These models are subject to a constant review of new clinical knowledge and take it into account after thorough examination by leading medical experts.

3 Platform Architecture

The KAIT platform consists of various encapsulated components which comprise distinct tasks and feature sets. The services provided by these independent components are exposed by a single holistic graph-based data interface[6] to provide a streamlined data access point for efficient server-client communication. In general, the platform and its components can be classified into three fundamental processes:

- **Patient data management and validation** - new patient records are fed into the central database. The data is validated and transformed to fit into a predefined schema that is explicitly modelled for the KAIT platform.
- **Knowledge extraction and modeling** - using statistical, expert-based and machine-learning (ML)[7] approaches, medical knowledge is gathered, processed and formalized into distinct disease models which represent a mathematical and IT-compatible representation.
- **AI-assisted reasoning and decision-support** - those knowledge models are precisely mapped to the integrated patient records to (1) allow for the derivation of novel insights about their respective biomedical impacts through data analysis and (2) use those insights to reason, simulate and predict suggestions, risks and outcomes for the individual patient.

[6] The KAIT **graph-based data interface** is powered by GraphQL. All of the platform's integrated information formalized as a mathematical graph structure. This structure contains all data entities as well as their respective relationships, which allows all of the technical system components a unified information access.

[7] **Machine learning** is a sub-field of AI which focuses on self-optimizing in regard to a specific problem by learning from data.

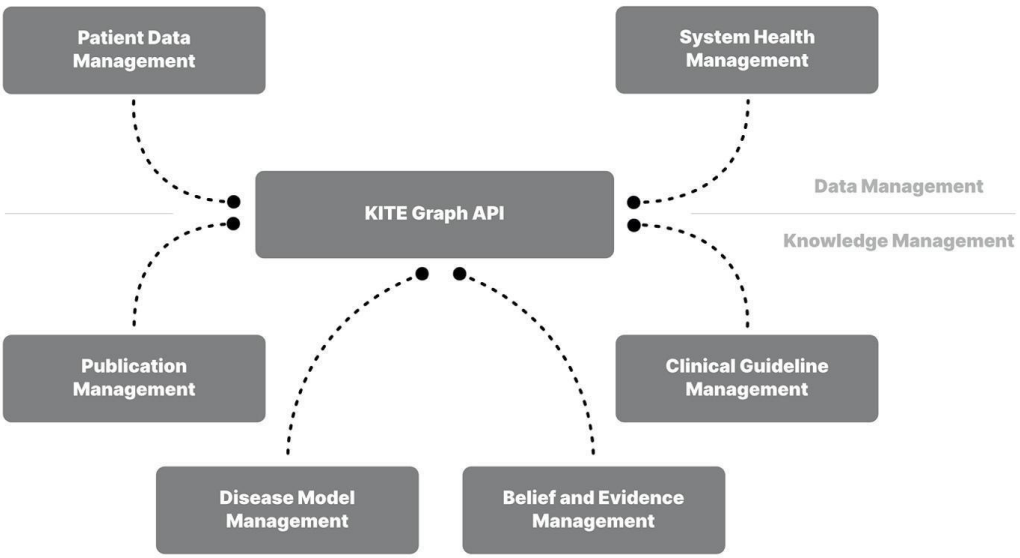

Figure 1 – Topology of KAIT's service-based architecture

3.1 Patient Data Management and Validation

KAIT's central source of knowledge is its extensive patient database, which will feature a granular collection of several thousand real-world medical cases, each represented by detailed medical patient profiles, disease progressions and respective outcomes. This data is already prepared and validated accordingly through manual curation as well as automated quality assurance features inside the platform. KAIT will implement the HL7 FHIR[8] standard for health care data exchange in combination with medical terminology standards such as SNOMED CT[9] and LOINC[10]. Patient data is therefore structured in resources such as observations or conditions that can be further specified by binding it to specific terminologies and codes or linking it to other resources to form a network. This ensures integrity, (semantic) interoperability, accuracy and consistency in the documentation of medical cases. To build this extensive collection of real-world medical cases, KAIT will integrate data from:

1. **Medical registries and clinical studies** - data from several medical registries focusing on the targeted hematologic entities is gathered, prepared and mapped to the required data schemas. This will supply KAIT's algorithms with significant amounts of data input right from the start to provide objective and considerable output.

[8] Fast Healthcare Interoperability Resources (http://hl7.org/fhir/)

[9] Systematized Nomenclature of Medicine: Global standards for health terms (https://www.snomed.org/)

[10] Logical Observation Identifiers Names and Codes: International standard for identifying health measurements, observations, and documents (https://loinc.org/)

2. **Independent participating medical institutions** - established medical experts will be able to register their local patients in the KAIT system. On the one hand, they contribute to an expansion of the database and thus to the improvement of KAIT's knowledge models. On the other hand, KAIT supports the physician in the documentation and management of the patient's disease progress, enables AI-assisted therapy decision-support and provides a platform for cross-institutional data exchange, which is considered a win-win situation.

Due to the fact that the database is continuously growing every time new case data is provided, KAIT's internal algorithms are also subject to constant improvement. This new data is automatically validated, pre-processed and appropriately structured before entering the system. To comply with data security and privacy regulations, medical data is handled primarily on the client side and within the local IT infrastructure of the treating physician. While some of KAIT's features require the transmission of selected data entities, these are selected with the principle of data economy in mind and are encrypted and pseudonymized by default.

3.2 Knowledge Extraction and Modeling

KAIT will utilize a hybrid process for building its internal disease models. This features classic knowledge engineering processes[11], carried out by medical and IT domain experts as well as machine learning (ML) to complement human expertise with novel insights from real-world data. The platform will integrate medical knowledge through formalized mathematical models implemented as graph-based data structures. Those reflect the causal relationships between individual information entities, derived either from measured (e.g. laboratory findings) or calculated values represented by multi-factor assessments (e.g. medical scores or classifications).

Just as with the patient data, KAIT will gather disease-related data from multiple data sources to form its knowledge base. Those include electronic medical records[12] (EMR) of actual patients, clinical trial data, patient reports and respective outcomes as well as significant studies and publications from the field of hematology. All evidence is initially curated by our team of domain experts to ensure the highest amount of quality control.

---

[11] A **knowledge engineering process** is a method to formalize specific expertise and background information about a certain application. It usually involves the interdisciplinary collaboration of modelling and domain experts.

[12] An **Electronic Medical Records (EMR)** is a digital container that includes patient- and disease-related information for a medical case. This might involve various documents (e.g. radiologic images, laboratory findings, etc.) in different digital formats.

Since the platform is intended to always consider the most current and relevant findings from the hematological domain, the knowledge base is designed as a dynamic component. New results can be contributed by the community of experts, which are then evaluated according to the highest professional standards in a platform-assisted peer-review process. In combination with technical procedures and custom metrics for the evaluation of updates, the system is subject to continuous growth and always oriented to represent the current state of the art.

3.3 AI-Assisted Reasoning and Decision-Support

Apart from considering clinical practice guidelines (CPG) and relevant research publications for assistance in the decision-making process, KAIT will feature an unique practice-based reasoning component, which utilizes a semi data-driven approach to draw knowledge from its inherent patient and medical evidence database. Purposefully selected ML and AI technologies are used to individually evaluate therapeutic options, offering best-of-class assessments while taking into account possible outcomes, quality of life and a multitude of other relevant disease-specific factors for an individual patient. While the approach features a high degree of automatization, it is not entirely data-driven as features will be classified in accordance to current medical knowledge provided by verifiable external sources, e.g. the experience from several medical experts, other medical knowledge bases and CPG's as well as current medical research publications.

The approach solves a multi-label classification problem[13,14] based on tabular data derived from various sources, e.g. patient data, therapy, process, omics, etc. To simplify this rather complex calculation problem, multiple sub-models (operators), each solving a binary or multi-class classification problem, will be utilized and causally chained to enable the prediction of, for instance, a patient state or risk assessment necessary for an even higher-level decision problem. Therefore, AI is applied to gather medical insights about the patient directly from the database. Thus, the practice-based reasoner forms a network of such classifier chains, with layers ranging from measurable evidence (e.g. lab values or individual symptoms) as the initial input to a tailor-made therapy decision as the final output. In short, KAIT's data-driven decision-support approach consists of multiple chained classification tasks with the following properties:

---

[13] Predicting multiple targets from one sample is a multi-label classification. Therefore, it can be understood as a combination of binary or multi-class classification problems. For example, predicting the age and diagnosis of a patient using one classifier represents a **multi-label classification**, whereas separately classifying them would be considered as binary (male/female) and multi-class (any diagnosis) classification problems.

[14] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. In: International Journal of Data Warehousing and Mining (IJDWM) 3.3 (2007), pp. 1-13.

- nominally and ordinally scaled features and targets,
- binary and multi-class classification problems (usually $k \leq 3$, with k being the number of classes of the target variable).

In addition to the guideline- and practice-based reasoning approaches, KAIT will also feature case-based reasoning functionality. Thus, whenever a physician enters a new patient case, it is automatically compared to all other known cases in the platform. From a methodological view, KAIT will utilize data-driven patient similarity metrics to calculate a numerical score which expresses case comparability. If, in this way, a patient with a high degree of similar features and disease progression is found, contact information about the treating physician is provided to the current user, and he or she is able to schedule an internal meeting to further discuss the medical case.

3.3.1 Ensemble Learning

The main challenges of chaining KAIT's operators are limited feature sets[15] and possible error propagations[16]. While the issues related to limitations will be handled by applying domain knowledge to sufficiently select features for each task, error propagation will be avoided by improving the confidence of each model. KAIT considers two ways to achieve this goal: (1) using state of the art AI methods to maximize accuracy and minimize variance and (2) selecting the most fitting machine learning model for each task.

The most promising method to achieve both is *»ensemble learning*[17]«. In this approach, multiple weak ML models are built on top of the training dataset or subsets of it, and are then ensembled to accumulate one robust learning model. While the weak models are often not better than random guessing, the ensembled model usually outperforms any single model on a given dataset. Ensemble models tend to fit the data better, leading to more accurate classifications. Also, they reduce bias and variance[18,19,20]. Another advantage of using ensemble learning methods is their efficiency even if the complexities of the weak models are low[4]. Considering KAIT's approach to chain operators, a high performance of the practice-based reasoner is expected.

---

[15] Referring to machine learning, a **feature set** is the set of measurable properties of an object used to estimate a specific target. For example, if you want to estimate the best therapy for a patient, medical data describing the patient could be used as features.

[16] **Error propagations** happen when a classification fails and delivers its false output to another classification problem.

[17] In machine learning, various learning algorithms can be ensembled to improve the performance of a prediction task. Common **ensemble learning** methods are boosting, bagging, or stacking.

[18] Zhi-Hua Zhou. Ensemble methods: foundations and algorithms. CRC Press, 2012.

[19] Omer Sagi and Lior Rokach. Ensemble learning: A survey. In: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8.4 (2018), e1249.

[20] Thomas G Dietterich et al. Ensemble learning. In: The handbook of brain theory and neural networks 2 (2002), pp. 110-125.

On the other hand, a common problem with ensemble learning methods is their proneness to overfitting[21]. Basically, this can be exemplified as the model remembering all the training data without being able to generalize and draw knowledge from it. As a result, the model performs almost perfectly on data it has seen, but poorly on new samples. In ensemble learning, various weak models learn on potentially intersecting subsets of the training data. Therefore, if the size of the dataset is low, many learning steps are processed on the same data before validating the model on new samples[22,23]. In this regard, too, selecting small feature sets and building fewer complex models on KAIT's sufficiently large patient database helps to avoid overfitting.

3.3.2 Gradient Boosting Machines[24] and XGBoost[25]

KAIT will focus on the application of gradient boosting techniques for ensuring better precision in its prediction models. In this approach, pseudo-residuals will be calculated for each model learning iteration using a loss function. After the weak model is fit to these pseudo-residuals, its prediction multiplied with the learning rate will be added to the overall model function. So, the model is boosted by sequentially updating it by every weak model's performance. This means, after each iteration the overall model can be tested for performance, giving the opportunity of intercepting an overfitting process, or finishing the model learning prematurely if performance is satisfactory[26].

When it comes to ordinal data, which most of the patient database will consist of (e.g. feature values such as good / bad, low / standard / high), XGBoost[27] is the best performing gradient boosting method in recent years[28]. XGBoost is a gradient tree boosting algorithm, meaning that it is an ensemble of decision tree models. While maintaining high efficiency due to its scalability to multiple tasks, it also produces state of the art classification performances. Also, because of its basing on decision trees, XGBoost models are interpretable by estimating feature importance based on model metrics, such as information gain of tree nodes.

---

[21] **Overfitting** happens when a machine learning model is too attuned to its training data and can't be applied to new data

[22] Robi Polikar. Ensemble based systems in decision making. In: IEEE Circuits and systems magazine 6.3 (2006), pp. 21-45.

[23] Zeinab Khatoun Pourtaheri and Seyed Hamid Zahiri. Ensemble classifiers with improved overfitting. In: 2016 1st Conference on Swarm Intelligence and Evolutionary Computation (C SIEC). IEEE. 2016, pp. 93-97.

[24] **Gradient boosting** is a specific ensemble learning and boosting method. In contrast to other boosting methods, ensembled weak models don't boost each other directly by learning from the previous model's mistakes, but by optimizing an arbitrary loss function, which generalizes the ensembled model's performance.

[25] **XGBoost** is a gradient boosting framework, which outperforms other gradient boosting methods in terms of computation time and prediction performance by making use of features like regularization, parallelized model construction, distributed computing, or handling missing values.

[26] Jerome H Friedman. Stochastic gradient boosting. In: Computational statistics & data analysis 38.4 (2002), pp. 367-378.

[27] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016, pp. 785-794.

[28] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. CatBoost: gradient boosting with categorical features support. In: arXiv preprint arXiv:1810.11363 (2018)

XGBoost's high efficiency and performance is due to many optimizations it provides for the gradient tree boosting method. Some of them are its handling of sparse data due to a different tree learning algorithm, its improved parallelization techniques, and it is enabling regularization, which reduces overfitting. Thus, it will be the default classification model for each sub task of KAIT's practice-based reasoner. To optimize model performance further, automatic hyperparameter tuning will be applied, for instance, to include and set the regularization parameter and reduce overfitting as a result, or to lower the tree depth for lower computational[29]. Furthermore, if tasks differ from the standard, other gradient boosting methods can be used, such as CatBoost for categorical instead of ordinal data, or LightGBM, if features are frequent and data is sparse[30].

3.4 The KAIT Concept of Traceable AI

Computer-aided decision-support is only relevant for clinical practice if the associated recommendations are fully comprehensible. Thus, only the implementation of traceable decision pathways will allow a clinical expert to monitor and individually evaluate the automatic conclusions drawn by the algorithm. To meet this requirement, KAIT uses a federated reasoning approach to provide AI-assisted assessments. As introduced in section 3.3, the process breaks down the complex problem of treatment selection at different stages (e.g. first-line, second-line) into its atomic components and summarizes them into consolidated operators. Each operator contains a specific amount of input evidence (e.g. measured laboratory blood values) and distinct output stages of the respective decision. In essence, each operator has its own optimized AI processing engine, which makes an assessment based on the integrated data by applying classification and optimization methods. KAIT's extensive database serves as a continuous training set that trains the weights of the classifiers.

[29] Sayan Putatunda and Kiran Rama. A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. In: Proceedings of the 2018 International Conference on Signal Processing and Machine Learning. 2018, pp. 6-10.

[30] Guolin Ke et al. Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in neural information processing systems. 2017, pp. 3146-3154.

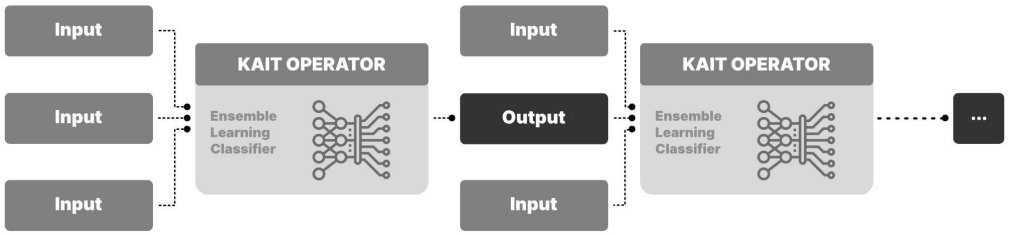

Figure 2 – Schematic signal flow across a network of KAIT operators

Through a causal interconnection of the operators, which formally map the cognitive decision process for treatment selection, a network is created from which individualized decision paths can be derived. Based on Figure 2, the raw data (e.g. measured values) from the patient profile are fed into the network. The operators determine all necessary calculations (e.g. multifactorial scores or estimates) along the causal pathway and forward the respective output signals to the next operator in line. Thus, the signal is automatically routed through the network towards the actual therapy selection endpoint. The resulting path can be easily traced, because the required input (in form of the raw data) as well as the operator-based estimation in the output is entirely traceable. The method is firmly based on the actual cognitive processes during human decision-making and imitates them in a genuinely rational and objective manner.

A basic premise for model interpretability is the quality of the decision-making algorithm. Selected metrics will be automatically analysed and prepared visually to show the physician the confidence the platform has in any individual decision. To reinforce this, biases and uncertainties shall be identified and eliminated by bias assessments and precise machine learning models for each homogenous patient cohort. To further ensure interpretability across operators and decision pathways, model-independent feature analysis will be applied. This allows for straight comparisons of distinct multifactorial decisions made by the KAIT platform.

3.5 Benefits and Differentiation

One of KAIT's most exclusive features is its structured and pseudonymized patient database which provides the solid foundation for its primarily data-driven approaches. This internal data collection is accompanied by a rigorous technical policy on data validation right from the start to ensure content related as well as technical and methodological value. By setting the primary focus on the establishment of a purely data-centric approach, severe barriers introduced by the heterogeneous landscape of IT systems and documentation standards within

the medical domain can be overcome. However, KAIT is not intended to be just another proprietary and closed infrastructure but will focus on the consideration of established methods and interfaces for data exchange to further emphasize flawless communication, e.g. HL7 FHIR.

Another significant benefit of the KAIT platform is represented by the custom AI-assisted reasoning engine and the way that medical knowledge is handled and processed. Based on the fragmentation of the primary therapy decision problem, KAIT can calculate and provide extensive assessments of all relevant information entities along a patient's journey. This also includes the consideration of patient-reported outcome measures (PROM)[31] to integrate individual self-assessments concerning adverse events, quality of life and other valuable metrics. Beyond case-specific review of those useful insights, KAIT will utilize the knowledge derived from these feedback mechanisms to improve its own suggestions continuously.

Just like any other successful IT platform, KAIT will rely heavily on a diverse and vibrant community of professional users. They will utilize its inherent benefits and contribute to further development in a variety of ways. Thus, the platform will offer plenty of ways for active participation on a technical level. This includes the assisted integration of novel insights and knowledge, peer-review mechanisms to supervise and curate the entity-specific knowledge bases, access to patient data to support medical research, and so much more.

4 Summary

The KAIT platform represents the most ambitious take on AI-assisted and completely traceable hematologic clinical decision-support to date. It will act as a comprehensive example of how the most recent advancements in the field of ML and AI can complement the daily clinical routine efficiently and sustainably.

KAIT is built with continuous growth and progress in mind. Thus, it will be our ubiquitous goal to ensure that its inherent benefits are accessible to as many users as possible to establish the most reliable and trustworthy platform for therapeutic decision-support and knowledge management in hematology.

[31] **Patient-reported outcome measures (PROM)** are specific reports which are handed out to the patient to capture their subjective evaluation about certain medical factors (e.g. quality of life, pain assessment, etc.).