

COVER PAGE
BETTER LEUKEMIA DIAGNOSTICS THROUGH AI (BELUGA)

**Long title: A Case-Control Study To Determine The Suitability Of Artificial Intelligence
For Leukemia Diagnostics**

Submission ID: MLL_001

Sponsor of the trial: Prof. Dr. Dr. Torsten Haferlach, Munich Leukemia Laboratory (MLL) –
MLL Münchner Leukämielabor GmbH

Address: MLL Munich Leukemia Laboratory, Max-Lebsche-Platz 31, 81377 MUNICH
(GERMANY)

Document Date (Date of initial Submission): 6TH OF JULY 2020 (07—06—2020)

Date of Submission of the revised Document: 9th OF JULY 2020 (07—09—2020)

NCT Number: NCT ID not yet assigned

Submitting investigator: Adam Wahida, MD

Contact Email: adam.wahida@mll.com

Protocol Title: A Case-Control Study To Determine The Suitability Of Artificial Intelligence For Leukemia Diagnostics

Running Title: BETTER LEUKEMIA DIAGNOSTICS THROUGH AI (BELUGA)

Principal Investigator (PI): Prof. Dr. med. Dr. phil. Torsten Haferlach (TH)

Telephone: +49 (0)89 99017-100

Fax: +49 (0)89 99017-109

E-Mail: torsten.haferlach@mll.com

Co-Investigator(s): Prof. Dr. med. Wolfgang Kern (WK), Adam Wahida (AW)

Sponsor: Prof. Dr. Dr. Torsten Haferlach, Munich Leukemia Laboratory (MLL) – *MLL Münchner Leukämielabor GmbH*

Address: MLL Munich Leukemia Laboratory, Max-Lebsche-Platz 31, 81377 MUNICH (GERMANY)

Telephone: +49 (0)89 99017-100

GCP Statement: BELUGA will be performed in compliance with the good clinical practice directive from the European Union (2005/28/EC)

Funding: MLL Munich Leukemia Laboratory, Max-Lebsche-Platz 31, 81377 MUNICH (GERMANY)

SUMMARY

To the best of our knowledge, BELUGA will be the first prospective trial investigating the usefulness of deep learning-based hematologic diagnostic algorithms. Taking advantage of an unprecedented collection of diagnostic samples consisting of flow cytometry datapoints and digitalized blood-smears, categorization of yet undiagnosed patient samples will prospectively be compared to current state-of-the-art diagnosis at the Munich Leukemia Laboratory (hereafter MLL). In total, a collection of 25,000 digitalized blood smears and 25,000 flow cytometry datapoints will be prospectively used to train an AI-based deep neuronal network for correct categorization. Subsequently, the superiority will be challenged for the primary endpoints: sensitivity and specificity of diagnosis, most probable diagnosis, and time to diagnose. The secondary endpoints will compare the consequences regarding further diagnostic work-up and, thus, clinical decision making between routine diagnosis and AI guided diagnostics. BELUGA will set the stage for the introduction of AI-based hematologic diagnostics in a real-world setting.

Type of Research: observational study, diagnostic study, hematology, leukemia

Intervention: none

APPROVAL

This study will be conducted with the utmost respect for individual patients in accordance with the requirements of this diagnostic trial protocol and especially per the following:

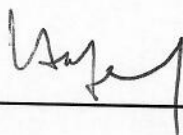
- Good clinical practice directive (European Union) (2005/28/EC)
- The ethical principles according to the Declaration of Helsinki
- International Conference on Harmonization (ICH) E6 Good Clinical Practice: consolidated guideline
- Guidelines for Good Clinical Practice (Deutsche Forschungsgemeinschaft DFG)
- All other applicable laws and regulations, including, data privacy laws, clinical trial disclosure laws, and regulations

The information provided in this document is confidential and proprietary to the MLL Munich Leukemia Laboratory. The information in this document shall not be disclosed to any third party, in any form, without the prior written consent of the MLL

SIGNATURES

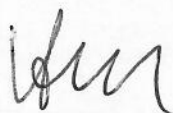
Prof. Dr. Dr. Torsten HAFERLACH

Date: 2020-07-05



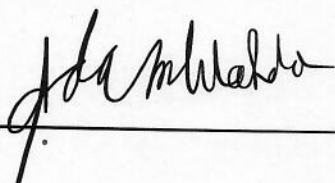
Prof. Dr. Wolfgang KERN

Date: 2020-07-05



Adam WAHIDA

Date: 2020-07-05



STUDY SUMMARY

Protocol Title	A Case-Control Study To Determine The Suitability Of Artificial Intelligence For Leukemia Diagnostics
Running Title	BETter LeUkemia diaGnostics through AI (BELUGA)
Main Aim	Prospectively determine the superiority of pattern recognition in the classification of hematological disorders
Coordinating Investigator	Prof. Dr. med. Dr. phil. Torsten Haferlach Telephone: +49 (0)89 99017-100 Fax: +49 (0)89 99017-109 E-Mail: torsten.haferlach@mll.com
Primary Objective	Assessing sensitivity and specificity of pattern recognition in performing hematologic diagnostics in comparison to current state-of-the-art algorithms
Secondary Objective	<ul style="list-style-type: none">▪ comparison of clinical consequences▪ predictive diagnostic value▪ turn-around-time▪ enumerate entity-specific benchmarks (e.g., blast count)
Explorative Objectives	Determine the relationship between AI and non-AI supported diagnostic tools for hematologic diagnostics
Study Center	Munich Leukemia Laboratory (MLL)
Trial Site	Munich Leukemia Laboratory (MLL)
Number of Subjects	The estimated number of patients throughout the study duration: 25,000
Diagnosis and Main Inclusion Criteria	Peripheral blood smears for morphology and/or peripheral blood or bone marrow for immunophenotyping from all hematological malignancies recognized by current WHO classification (Swerdlow et al. 2017)
Duration of Study	The study will be conducted throughout twelve (12) months from the 1st of August 2020 until the 31st of July 2021.

TABLE OF CONTENTS

INTRODUCTION.....	- 7 -
BACKGROUND AND RATIONALE.....	- 7 -
OBJECTIVES.....	- 8 -
STUDY DESIGN.....	- 10 -
TYPE OF STUDY.....	- 10 -
DURATION OF STUDY	- 11 -
QUALITY CONTROL	- 11 -
PRIMARY STUDY ENDPOINTS	- 12 -
<i>Endpoint assessment</i>	- 12 -
<i>Primary endpoint</i>	- 12 -
SECONDARY ENDPOINTS	- 13 -
SUBJECT SELECTION AND WITHDRAWAL	- 14 -
NUMBER OF SUBJECTS	- 14 -
GENDER AND AGE	- 14 -
INCLUSION CRITERIA	- 14 -
EXCLUSION CRITERIA	- 14 -
SUBJECT IDENTIFICATION & RECRUITMENT	- 15 -
LOCATION	- 15 -
STATISTICAL PLAN	- 16 -
SAMPLE SIZE	- 16 -
STATISTICAL METHODS	- 16 -
RISKS AND BENEFITS	- 16 -
POTENTIAL DIRECT BENEFITS TO SUBJECT	- 16 -
DATA HANDLING AND RECORD.....	- 17 -
STUDY MONITORING, AUDITING, AND INSPECTION.....	- 17 -
FINANCIAL CONSIDERATIONS.....	- 17 -
CONFLICT OF INTEREST	- 18 -
PUBLICATION PLAN.....	- 18 -
ARCHIVING	- 19 -
TRIAL MASTER FILE.....	- 19 -
REFERENCES.....	- 20 -

INTRODUCTION

Background and Rationale

In numerous recent studies, deep neuronal networks (DNN) have been leveraged to examine the usefulness of artificial intelligence (AI)-based DNN for diagnostic purposes (summarized by Topol et al. 2019). In essence, they have successfully proved to recapitulate state-of-the-art diagnoses currently performed by humans.

Specifically, the use of artificial intelligence for pattern recognition showed that DNN could categorize complex and composite data points, chiefly images, with high fidelity to a specific pathogenic condition or disease. The majority of these studies are primarily based on extensive training sample collections that were categorized a priori. Subsequently, this "training" provided the necessary input to classify newly delivered specimens into the correct subgroups, frequently even outperforming independent human investigators. So far, these studies have thus provided the rationale for the use of DNN in real-world diagnostics. However, the prerequisite for using DNN in a real-world setting, where specimen sampling and analysis would need to outperform human diagnosis prospectively, would be a blinded and prospective trial. Currently, there is a lack of prospective data, therefore still challenging the notion that DNN can outperform state-of-the-art human-based diagnostic algorithms. Here we want to investigate the validity and usefulness of AI-based diagnostic capabilities prospectively in a real-world setting.

Hematologic diagnostics heavily rely on multiple methodically distinct approaches, of which phenotyping aberrant blood or bone marrow cells from affected patients represents a cornerstone for all subsequent methods, such as chromosomal or molecular genetic analyses. At the MLL, five different diagnostic pillars are required to provide diagnostic evidence for a specific malignant

blood disorder faithfully: cytomorphology and immunophenotyping first, guiding more specific methods such as cytogenetics, FISH, and a diversity of molecular genetic assays.

Objectives

Phenotyping of blood cells is primarily based on two distinct challenges; (1) the morphological appearance and abundance of specific cell types and (2) the presence of particular lineage markers detected by flow cytometry. These two methods are critical for each subsequent decision-making process and, thus ultimately, the final diagnosis. Simultaneously, these two methods are ideally suited for automated analysis by DNN due to their inherent image-based nature. This has been recently illustrated by a publication by Marr and colleagues (Matek et al., 2019)

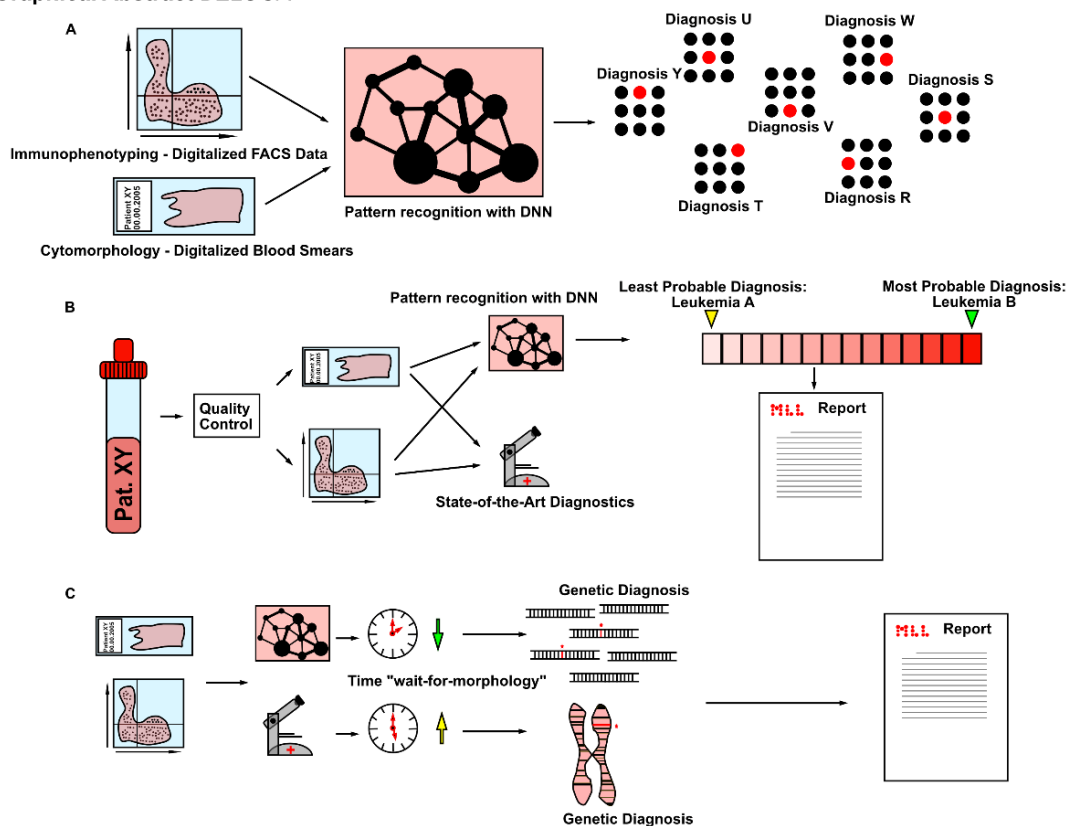
In BELUGA, we want to investigate whether the automated analysis of blood (from peripheral blood and bone marrow aspirates) smears and flow-cytometry-based analyses can provide a benefit for diagnostic quality and, ultimately, patient care. Moreover, BELUGA will provide evidence for the cooperative nature of image-based diagnostic tools for other pillars of hematologic diagnostic decision making such as genetic and molecular genetic characterization.

BELUGA, therefore, consists of three parts (Fig. A-C). In A, we want to train a DNN with an unprecedented collection of blood smears and flow-cytometry-based data points collected during the course of 15 years. These samples consist of all hematological malignancies currently identified and recognized by the current WHO classification for hematologic malignancies. Due to the varying incidences of these entities, the total number of training items varies from 1,000 to 20,000 for 15 years. However, we deem this discrepancy a benefit to this trial's overall aims, because this diverse spectrum will inform us on the number of training items needed for outperforming the state-of-the-art diagnostics in cytomorphology or flow cytometry.

In part B, we will compare the overall performance of our trained DNN prospectively to new yet undiagnosed samples arriving at our laboratory (see the main section for details). The superiority of DNN based categorization will be challenged based on the pre-defined outcome parameters accuracy with respect to state-of-the-art diagnostics, mismatch-rate, and time needed to provide a diagnostic probability.

Lastly, in C, we will investigate the effects on faster and more accurate diagnostic power by leveraging our trained DNN to aid downstream diagnostic methodologies such as chromosomal analysis or panel sequencing of patient samples.

Graphical Abstract BELUGA



Graphical Abstract BELUGA. A. Training set using retrospective data. B. A prospective trial investigating phenotyping analysis by state-of-art diagnostics versus DNN based clustering C. Downstream effect of DNN based clustering on downstream diagnostic methods, chiefly chromosomal analysis, molecular genetics, and panel sequencing.

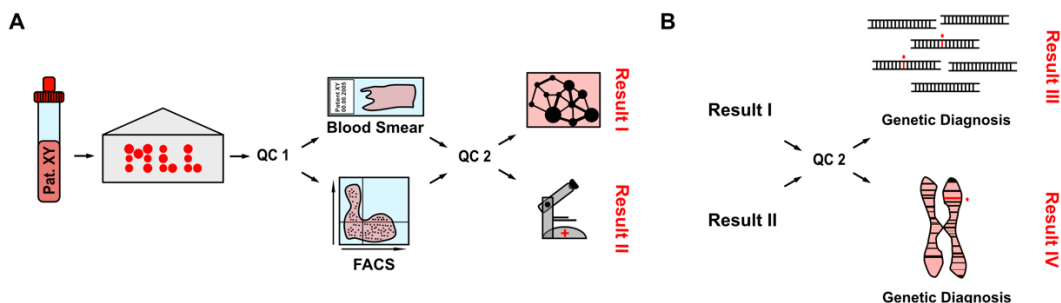
STUDY DESIGN

Type of Study

BELUGA is conducted as a monocentric prospective nested case-control study. The study population consists of all patients with suspected hematological disorders for which samples are provided to the MLL between 01.08.2020 and 31.07.2021. BELUGA is a non-interventional study without therapeutic consequences for direct patient care.

Our technology will also be of particular interest for studies of pathophysiological precursors of disease, since we will be able to obtain discrete and continuous readouts. This will enable us to scrutinize *a priori* diagnosis with internal diagnosis and, at the same time, compare the most probable diagnostic decision with the actual patient disease course over time.

Schedule of Events BELUGA



Duration of Study

Retrospective analysis reveals that the MLL currently provides a state-of-the-art diagnosis for 60% of all hematological disorders detected in the Federal Republic of Germany. Specifically, the MLL provides a therapy-consequent diagnosis for all current hematological malignancies recognized by the WHO classification (Swerdlow et al. 2017). Thus, the number of samples according to their respective entity reflects incidence in the general population. Diagnosis is provided to external caregivers in a seven-day workflow, with new samples arriving at the MLL facility every day.

The mean number of daily samples processed amounts to 280. From these, 50% are primary diagnoses and 50% follow-up assessments of disease, most importantly MRD diagnosis in CML.

BELUGA will be conducted over the entire course of 52 weeks; therefore, each entity statistically will be present with at least 50 novel diagnoses.

Quality Control

Due to the heterogeneity in the quality of blood samples arriving at the MLL, BELUGA is preceded by a rigorous quality assessment of every patient sample potentially included in the present study. We deem this measure essential to exclude as many bias sources as possible.

Subsequently, to prevent any shortcoming of diagnosis to direct patient care, every sample will receive the routine diagnostic work-up routinely performed at our institution.

Primary Study Endpoints

Endpoint assessment

Due to this trial's exploratory nature, statistical testing will be performed at the one-sided 5% level. However, if the direction of the hypothesis is not entirely self-evident, or two-sided, two-sided tests will be performed.

Subsequently, a noninferiority test will be performed at the one-side 5% level to test whether AI aided is non-inferior to gold-standard testing. This will be instrumental in understanding the value of AI guided diagnosis in a real-world setting.

Quantitative variables will be described with the number of non-missing values, mean, standard deviation (SD), median, and minimum/maximum values. Qualitative variables will be defined with the number and percentage of patients with each qualitative characteristic. Missing values will not be included in the calculation of percentages.

The efficacy data will be descriptively summarized and used for exploratory purposes only.

Primary endpoint

As a primary endpoint, we will examine the ability of DNN to classify disorders according to (after initial assessment disease/healthy) to the gold-standard diagnosis. The gold-standard diagnosis is defined as an integrated diagnosis, including cytomorphology, flow cytometry, cytogenetics, FISH, and molecular genetics. DNN will independently provide a bi-directional (probabilistic) diagnosis, with the most probable diagnosis. The primary analysis will include a direct comparison between the human cytomorphological examination and the pattern recognition software.

Secondly, this result will be provided to downstream diagnostic departments to assess phenotypic diagnosis's usefulness for genetic characterization. We hypothesize that the turn-around time will be significantly enhanced, further providing quality at sooner timepoint.

Secondary Endpoints

As secondary endpoints, we will examine disease-specific measurements with respect to clinical decision making based on either the AI guided diagnosis or the internal gold standard. Due to the holistic nature of DNN in assessing the overall "status-quo" of the disease, we envisage a stratification change with disease subtypes. Moreover, disease-specific items such as blast count in acute myeloid leukemia will be compared between automated DNN-based diagnosis and classical cytomorphology.

SUBJECT SELECTION AND WITHDRAWAL

Number of Subjects

The training cohort of BELUGA consists of 50,000 annotated samples for which cytomorphological smears (25,000 samples) and immunophenotyping (25,000 samples) data points have been collected. This cohort serves as a foundation for the DNN to perform training. Our test cohort will consist of all samples for which cytomorphology and immunophenotyping will be performed for one year. Due to BELUGA's holistic nature, we want to ensure sufficient statistical power for each distinct entity throughout the study.

Gender and Age

Patients samples from both sexes will be used (male and female). Only samples from adult patients (i.e. 18 years or older) will be used.

Inclusion Criteria

- Patients having been diagnosed with a suspected hematological disorder:
- The suspected diagnoses constitute a primary diagnosis
- Only samples of patients min. 18 years of age will be used
- Samples must suffice quality attributes control which are denoted in “Exclusion Criteria”

Exclusion Criteria

- The sample is not fit for state-of-the-art diagnosis, fails initial quality control. For quality insurance, we will exclude samples in heparin- instead of EDTA. Samples with damage due to atmospheric reasons (freeze-thaw damage or elevated temperature) will be excluded.
- Samples with too scarce material jeopardizing routine gold-standard diagnosis will be excluded.

- Bone marrow aspirates without sufficient material to assess malignant or healthy hematopoiesis.

Subject Identification & Recruitment

A unique **MLL_identifier** routinely identifies patients. This ID is not traceable to the patient's identity, and thus full anonymization/pseudonymization is ensured.

Location

BELUGA will be conducted as a monocentric trial at the MLL.

STATISTICAL PLAN

Sample size

A sample size of N approximately 25,000 samples per method is expected to achieve sufficient power for statistical significance testing; preliminary data suggest that machine learning-based classification achieves an accuracy of 95 – 98% compared to conventional diagnostics. This leads us to estimate the power of 1.00 for an $\alpha=0.05$ for testing conventional and DNN categorization.

Statistical Methods

Quantitative variables will be described with the number of non-missing values, mean, standard deviation, median, and minimum/maximum values. Qualitative variables will be expressed as a number and percentage of patients with each qualitative characteristic. The missing values are not intended to be included in the calculation of percentages. Sensitivity and specificity will be assessed specific to each method, with respect to internal gold standard diagnostic work-up at two different time points, first in comparison to our initial phenotypic diagnosis and lastly, in comparison to our final assessment based on the genetic and chromosomal analysis.

RISKS AND BENEFITS

Potential Direct Benefits to Subject

Conducting BELUGA will not bear any risks for patients enrolled in the study. Routine state-of-the-art diagnosis is provided for each sample, and prioritization of sample material in favor of current diagnostic material and reporting is performed to prevent jeopardization of gold standard diagnosis.

DATA HANDLING AND RECORD

Data management documents will be generated under the responsibility of the sponsor. A management plan will be issued before data collection begins and will describe all functions, processes, and specifications for data collection, cleaning, and validation.

The data management documents will describe analysis methods and individuals who are authorized to enter the data, decisions about ownership of data, source data storage, the origin and destination of the data, and who will get access to the data at all times.

Data Management Responsibilities are primarily handled by co-investigator Wolfgang Kern (WK)

Upon request of external researchers, the sponsor will provide these investigators with additional data relating to the trial, duly anonymized and protected according to applicable requirements.

STUDY MONITORING, AUDITING, AND INSPECTION

The Investigator will make all the trial-related source data and records available at any time to quality assurance auditor(s) mandated by the sponsor, or to domestic/foreign regulatory inspectors or representatives who may audit/inspect the trial.

The primary purposes of an audit or inspection are to assess compliance with the trial protocol and the principles of ICH-GCP, including the Declaration of Helsinki and all other relevant regulations.

FINANCIAL CONSIDERATIONS

The MLL is the sole sponsor of this trial.

CONFLICT OF INTEREST

Prof. Dr. med. Dr. phil. Torsten Haferlach and Prof. Dr. med. Wolfgang Kern are part owners of the Munich Leukemia laboratory. Adam Wahida receives a research fellowship from the Torsten-Haferlach Leukemia Diagnostics Foundation.

PUBLICATION PLAN

Results from BELUGA will be generated and analyzed according to the guidelines of the Good Scientific Practice of the German Science funding agency (DFG). Results that will be interesting for the scientific community will be submitted and subsequently published in peer-review based journals, according to the appropriate scope and audience.

At the end of the trial, one or more manuscripts for joint publication may be prepared in collaboration between the Investigator(s) offered authorship and the sponsor. The sponsor reserves the right to be the last author(s) in all publications related to this trial. In the event of any disagreement in the content of any publication, both the Investigator's and the sponsor's opinion will be fairly and sufficiently represented in the publication.

ARCHIVING

The PI is responsible for maintaining all the records (protocol and protocol amendments, relevant correspondence, and all other supporting documentation), which enable the conduct of the trial at the site to be fully understood, in compliance with ICH-GCP.

The study site should plan on retaining such documents for ten years after study completion. These documents should be retained for a more extended period if required by the applicable regulatory requirements or the hospital, institution, or private practice in which the study is being conducted. Patient identification codes (patient names and corresponding study numbers) will be retained for this same period.

Trial Master File

The sponsor will archive the Trial Master File in accordance with ICH-GCP and applicable regulatory requirements.

REFERENCES

- Investigators of this study are denoted in **Bold**

Haferlach T. 2019. Taschenatlas Hämatologie: Mikroskopische und klinische Diagnostik für die Praxis. *Georg Thieme Verlag*.

Haferlach T. 2020. Hämatologische Erkrankungen. *Springer* <https://doi.org/10.1007/978-3-662-59547-3>.

Krappe S, Benz M, Wittenberg T, **Haferlach T**, and Münzenmayer C. 2015. “Automated Classification of Bone Marrow Cells in Microscopic Images for Diagnosis of Leukemia: A Comparison of Two Classification Schemes with Respect to the Segmentation Quality.” *Medical Imaging 2015: Computer-Aided Diagnosis*. <https://doi.org/10.1117/12.2081946>.

Krappe S, Wittenberg T, **Haferlach T**, and Münzenmayer C. 2016. “Automated Morphological Analysis of Bone Marrow Cells in Microscopic Images for Diagnosis of Leukemia: Nucleus-Plasma Separation and Cell Classification Using a Hierarchical Tree Model of Hematopoiesis.” *Medical Imaging 2016: Computer-Aided Diagnosis*. <https://doi.org/10.1117/12.2216037>.

Matek, C, Schwarz S, Spiekermann K, and Marr C. 2019 “Human-Level Recognition of Blast Cells in Acute Myeloid Leukemia with Convolutional Neural Networks.” *Nature Machine Intelligence*. <https://doi.org/10.1101/564039>.

Swerdlow, S. H., Elias Campo, N. Lee Harris, E. S. Jaffe, S. A. Pileri, H. Stein, J. Thiele, et al. 2017. “WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues (Revised 4th Edition).” *IARC: Lyon* 421.

Zhao M, Mallesh N, Höllein A, Schabath R, Haferlach C, **Haferlach T**, Elsner F, Lüling H, Krawitz P, and **Kern W**. “Hematologist-Level Classification of Mature B-Cell Neoplasm Using Deep Learning on Multiparameter Flow Cytometry Data.” *Cytometry*, June 2020. <https://doi.org/10.1002/cyto.a.24159>.

Matek, C, Krappe S, Münzenmayer C, **Haferlach T**, Marr C “Highly accurate differentiation of physiological and pathological bone marrow cell morphologies using deep residual networks” *paper sent out for formal peer review 5/2020*