How to predict relapse in leukaemia using time series data: A comparative in silico study.

Helene Hoffmann[1*], Christoph Baldow[1*], Thomas Zerjatke[1], Andrea Gottschalk[1], Sebastian Wagner[1], Elena Karg[1], Sebastian Niehaus[1,2], Ingo Roeder[1,3], Ingmar Glauche[1,5] and Nico Scherf[1,4,5]

[1] Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, School of Medicine, TU Dresden, 01307 Dresden, Germany
[2] AICURA medical GmbH, 12103 Berlin, Germany
[3] National Center of Tumor Diseases (NCT), Partner Site Dresden, 01307 Dresden, Germany
[4] Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany

* These authors contributed equally.

[5] Correspondence: nico.scherf@tu-dresden.de and ingmar.glauche@tu-dresden.de

Summary (max. 150 words)

Risk stratification and treatment decisions for leukaemia patients are regularly based on clinical markers determined at diagnosis, while measurements on system dynamics are often neglected. However, there is increasing evidence that linking quantitative time-course information to disease outcomes can improving the predictions for patient-specific treatment response.

We analyzed the potential of different computational methods to accurately predict relapse for chronic and acute myeloid leukaemia, particularly focusing on the influence of data quality and quantity. Technically, we used clinical reference data to generate in-silico patients with varying levels of data quality. Based hereon, we compared the performance of mechanistic models, generalized linear models, and neural networks with respect to their accuracy for relapse prediction. We found that data quality has a higher impact on prediction accuracy than the specific choice of the method. We further show that adapted treatment and measurement schemes can considerably improve prediction accuracy. Our proof-of-principle study highlights how computational methods and optimized data acquisition strategies can improve risk assessment and treatment of leukaemia patients.

Introduction

Leukaemia describes blood cancers in which immature, dysfunctional cells progressively displace functional blood cells. Myeloid leukaemias are characterized by aberrations affecting the proliferation and maturation of myeloid progenitor cells. They are further subdivided into chronic myeloid leukaemia (CML), typically presenting with a disease-specific *BCR-ABL1* fusion gene (Zhou and Xu, 2015), and acute myeloid leukaemia (AML), which is a highly heterogeneous disease with a variety of mutational profiles involved (Cancer Genome Atlas Research Network et al., 2013). Although the overall treatment strategy aims towards achieving sustainable remission, the available drugs, and the heterogeneity of the phenotypes lead to very different therapeutic approaches. For CML, tyrosine kinase inhibitors (TKI) have been established as a targeted therapy leading to molecular remission in most patients under continuous drug administration (Hochhaus et al., 2017). AML is treated by cyclic induction chemotherapy, usually combined with subsequent cycles of maintenance therapy. Molecular monitoring of disease-specific markers is currently established as the method of choice to quantify the leukemic burden.

A recurrence of the disease after treatment-induced remission is a significant and life-threatening risk for leukaemia patients. For AML, relapse usually occurs after completion of intensive chemotherapy treatment (Oliva et al., 2018). In contrast, in CML molecular recurrence is commonly observed in about 50% of the patients once TKI administration is terminated to probe treatment-free remission (Cerveira et al., 2018; Mahon et al., 2010; Nagafuji et al., 2019). In any case, the ability to prospectively predict the risk and timing of relapse or molecular recurrence is of outstanding importance to optimize and adjust individual therapies.

Currently, treatment decisions are based on the recommended risk stratification schemes. Those risk assessments are commonly based on *static* measurements from single time points, often at diagnosis (Döhner et al., 2017; Othus et al., 2016). In contrast, treatment response dynamics, such as the speed of initial remission, are only rarely evaluated for risk stratification. However, others and we have shown that molecular disease dynamics indeed correlate with therapy response and future relapse occurrence (Branford et al., 2014; Hoffmann et al., 2019, 2020; Saussele et al., 2018; Shanmuganathan et al., 2020). We reason that the direct integration of molecular response dynamics in the form of time-series data, which are increasingly available from standard disease monitoring, is a crucial element to improve the patient-specific risk stratification.

There are several, conceptually different approaches to integrate time-series data of molecular monitoring into risk assessment. Three common approaches represent the pillars of this methodological spectrum:

- Mechanistic models (MM) describe the molecular disease dynamics as a functional consequence resulting from the interaction between relevant system components (such as drugs, cell types, cytokines etc.). Based on the model's fit to patient time-series data, the further course of the disease can be simulated. Although MMs require considerable expert knowledge about the underlying mechanisms, the results of these models are readily interpretable as the model parameters typically carry explicit biological meaning.
- On the other end of the spectrum, there are deep learning approaches (Fawaz et al., 2018; Goodfellow et al., 2016; Zhang, 1994) in which generic neural network models (NN) are trained to classify time-series data by implicitly identifying characteristic features that correlate with future outcomes. Those methods require no *a priori* knowledge, but they are not suitable to directly interpret underlying biological mechanisms.
- Classical statistical models, in particular, generalized linear models (GLM) (McCullagh and Nelder, 1989) like logistic regression classifiers are applied to estimated distribution parameters that describe population characteristics to classify predefined features of the time-course data. Herein, prior knowledge about general treatment dynamics is directly incorporated as an explicit feature of the GLM, while no understanding of the underlying biological mechanisms is required. Although GLMs are typically easier to interpret than neural networks (as the influence of parameters on the

prediction can be assessed (Lundberg and Lee, 2017)) this probabilistic approach does not allow for explicit mechanistic interpretations as it is the case for MMs.

In this work, we will systematically compare these three methods. In particular, we study the influence of data size, sampling density and measurement error on their prediction accuracy. As available data sets of relevant molecular time courses for AML and CML are currently limited, we use established mechanistic *in-silico* models of those diseases (Hähnel et al., 2020; Hoffmann et al., 2020) to generate fully annotated artificial patient data. Based on this reference data, we are further able to suggest alternative disease surveillance schemes that may enhance the predictive power (Fig.1).

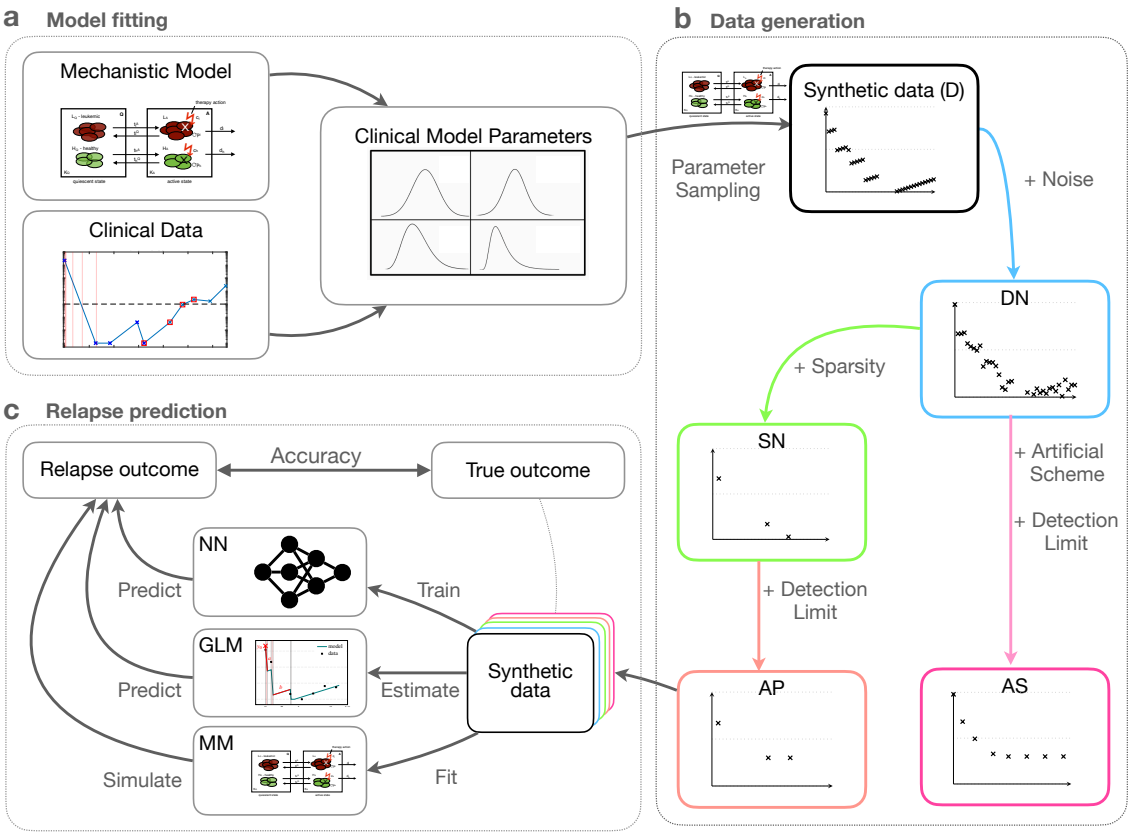


Fig. 1: Conceptual overview of our methodological approach: (a) We developed computational models (MM) for both AML and CML from mechanistic and empirical knowledge (Hähnel et al., 2020; Hoffmann et al., 2020). The models are first fit to actual patient data to obtain realistic parameters distributions. (b) We sampled from this empirical parameter distribution to simulate dense, synthetic data (D) with our MM. We gradually reduced the data quality to mimic actual clinical measurements by introducing noise (dense-noisy, DN), undersampling (sparse-noisy, SN) and a minimum detection limit (artificial patient data, AP). Additionally, we introduced a more informative scheme (artificial scheme, AS), in which the temporal measurements are optimally spaced (AML) or a period of reduced treatment dose precedes the cessation (CML). (c) We systematically compared the performance of our mechanistic model (MM), a generalized linear model (GLM) and a neural network (NN) to predict the outcome (relapse/no relapse) of our virtual patient data with varying quality.

Results

(i) Artificial patient data that closely resemble clinical time courses provide an excellent basis to systematically analyze the performance of predictive, computational models.

We apply two mechanistic models to simulate the dynamics of AML and CML(Hähnel et al., 2020; Hoffmann et al., 2020) thereby creating sets of synthetic response data. To make sure that the synthetic data resemble real patient time-courses as close as possible, we fitted the models to respective data sets obtained from 275 AML patients carrying a traceable NPM1-mutation (consisting of a total of 1567 measurements quantifying the relative amount of NMP1-mut transcript [8] over time on a log10-scale) and 21 CML patients (with in total 478 measurements (Hähnel et al., 2020) quantifying the relative amount of BCR-ABL1 transcripts over time on a log10-scale). Fig. 1a illustrates corresponding model fits, while we report on the overall fitting quality in Suppl. Fig. 1. The fitted model parameters are used to simulate synthetic time courses (Fig. 1b). To assess the influence of data quality, we gradually degraded the fully sampled, noise-free time series. We used estimates of the measurement frequencies and measurement errors obtained from the patient data to adjust the corresponding sampling density and noise level for the synthetic data (see supplementary methods). In total, we created four different datasets with 5000 time-courses from each model to systematically study the influence of data quantity and quality: (i) a dense (D) data set consisting of weekly (AML) or monthly (CML) measurements of the leukemic burden free of any measurement error. (ii) For the dense-noisy (DN) data we added noise (see Experimental Procedures) to all data points of D to match the measuring error (AML) or the residuals observed between real data and their corresponding model fits (CML). (iii) In a third step, we reduced the total number of measurements per patient, creating a sparse-noisy (SN) data set that matches the measurement frequency in the real data. (iv) Finally, to make the data as realistic as possible, we also added a detection limit for very low measurements, thereby creating a set of artificial patient (AP) data. Example time-courses for all data sets can be found in Suppl. Fig. 2 and 3.

To verify that the created artificial patient data (AP) sets are indeed similar to the real patient data, we derived characteristic features to quantitatively compare them. Those characteristic features refer to typical time scales and remission levels of the patient's response (see Suppl. Figure 4) and are explained in detail in the Exp. Procedures. The features are computed separately for the AP data and the actual input data. The visual comparison in Suppl. Figure 5 indicates that the median values of the characteristic features are very similar between AP and real data. It appears, that especially for the case of CML, the synthetic data sets yields a larger variance compared to the real data. A closer look at the data reveals that this is effect, at least partially, results from a sampling effect, as the variance measurement is only based on a small data set (n=21) of real patients.

Similar to the clinical presentation, we classified the synthetic time-courses as whether they show a relapse or not. For both CML and AML, we define disease recurrence by an increase of the leukemic burden (measured in terms of relative transcript abundance) within a predefined period above a given threshold. We then systematically compared the accuracy of relapse predictions between the three general approaches: (1) fitting the mechanistic model (MM) to the data and simulating the outcome, (2) feeding the previously derived explicit features of the time-course into a GLM classifier or (3) using an end-to-end learning approach with a neural network (NN) model (Fig. 1c).

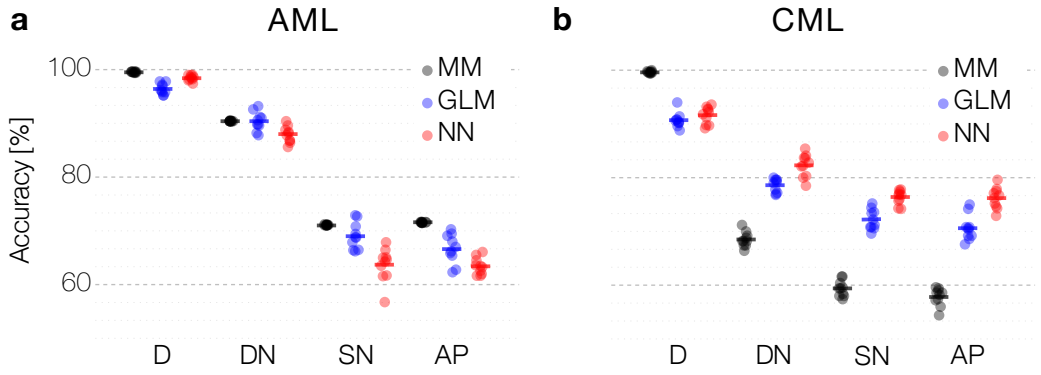


Fig. 2: Prediction accuracy across data quality and computational models: (a, b) Comparison of performance between mechanistic model (MM), generalized linear model (GLM) and neural network (NN) to predict relapse in synthetic data for AML (a) and CML (b) using 10-fold cross-validation. Data quality gradually decreases from fully sampled, noise-free data (D), to noisy (DN), sparse and noisy (SN), and artificial patient data (AP) (see main text for details).

(ii) Data quality has a strong influence on prediction accuracies, but the drop in performance considerably differs between models and use-cases.

Next, we analyzed how well the different approaches (MM, GLM, NN) can predict the outcome for the virtual patient data and how model performance changes with varying data quality (Fig. 1b,c, and Experimental Procedures). The results of the 10-fold cross-validation of the model performance are depicted for AML (Fig. 2a) and CML (Fig 2b). As expected, the prediction accuracy (see Methods for the mathematical definition) declines for all approaches when the data quality decreases. We point out that the decrease in data quality differs between use-cases and models. In the case of AML, the introduction of sparsity leads to a relatively sharp drop in model performance. This drop illustrates the dependency on data size (here data points per time series) as we have a median of only 4 measurements in the SN and AP data, compared to the original 39 measurements in the dense data set (D) set. In line with this, we observe a more gradual decline in performance when comparing the effect of introducing noise and sparsity in the CML case. Here, we have a median of 25 measurements in the SN and AP data, compared to the original median of 93 measurements.

Interestingly, the difference in model performance is not consistent across the two use-cases. For the sparser AML data, all models perform similarly on the dense (D) and noisy data (DN). However, when introducing more sparsity into the data, a mechanistic model performs more robustly than the generic NN model (a difference in the accuracy of 6.3 and 7.4 percentage points for the SN and AP) and the GLM model performance is in between MM and NN. This result reflects the importance of introducing prior knowledge (or inductive bias) when dealing with very few data (Fig. 2a).

We observe a different situation in the CML case. Here, the prediction accuracy for the mechanistic model drops down substantially more compared to the statistical GLM model and the generic NN when data quality decreases (a difference in accuracy between MM and NN of 19.7% for SN and 19.8% for AP, respectively). We recall that the noise-free data (D) was generated by the very same mechanistic model (compare Fig. 2b). The high prediction accuracy for this data indicates that the correct (generative) MM can truly be identified. However, given the higher number of free parameters (n=7) in the CML case, a reduction of data quality (either resulting from noisy or sparse measurements) more strongly effects the identifiability of the correct MM, while the GLM and the NN appear more robust.

Overall, our analysis confirms that predictive computational methods are promising tools to objectively and reproducibly predict relapse in myeloid leukaemias. For the real patient data (and the corresponding artificial patient samples AP) these methods achieve up to 70% accuracy (compare Figure 2a,b). While this indicates that the computational methods can in principle identify predictive, nontrivial patterns in time series data obtained during treatment, the resulting prediction accuracy might not adhere to the expected standards for clinical decision support. Our systematic study suggest that especially the scarcity and limited accuracy of available measurements per patient appears as a limiting factor for the overall prediction accuracy for relapse occurrence. Given those constrains on the data side, we doubt that structural changes to the computational method s(e.g. by refining the neural network architecture) can substantially improve the overall performance. However, we see great potential in optimizing the measurement process to yield more informative sampling schemes.

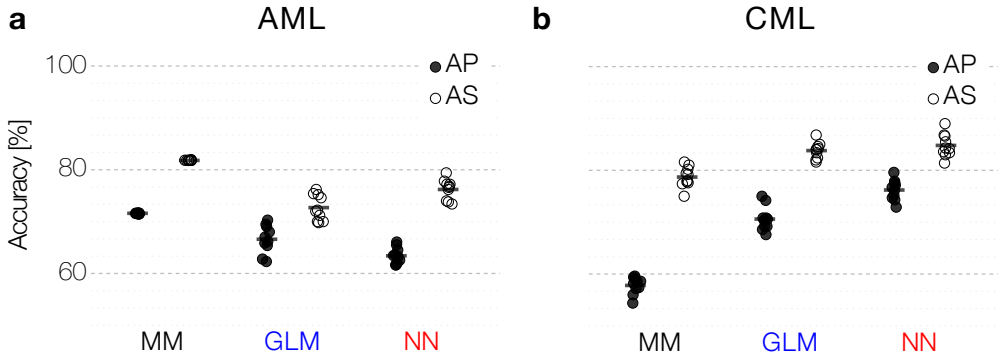


Fig. 3: Dedicated measurement schemes: (a, b) A dedicated measurement scheme (AS) improves prediction performance with the same number of data points for all models compared to the AP data both for AML (a) and CML (b) data.

(iii) A refined measurement and treatment scheme leads to improved prediction accuracies

As outlined above, a significant limitation for the prediction accuracy results from the sparsity of the available data, in particular for the case of AML. Here, molecular diagnostics and especially bone marrow aspirates are limited resources in the clinical setting. As only increasing the sampling frequency is not an option in many cases, we wondered whether an optimized timing of the measurements could lead to better predictions while the overall number of measurements remains the same. To investigate this question, we created an additional set of artificial patients (AS) with consistent measurement intervals during the nine-month treatment period (i.e. the first day of each therapy cycle and every six weeks during the treatment-free phase, typically 4 to 8 (median = 7) measurements per patient). Fig. 3a indicates that for this amended sampling regimen, we can already increase the accuracy of all prediction approaches (MM and NN by up to 12% and for GLM less pronounced). This finding strongly suggests that an adapted sampling scheme can considerably contribute to better relapse predictions, e.g. using methods from optimal experimental design (Chaloner and Verdinelli, 1995; Goodwin, 1977; Seeger, 2008; Walter and Pronzato, 2010).

The DESTINY trial implemented a study protocol for CML patients, in which patients in molecular remission reduced their TKI dose to 50% of the original dose for 12 months before TKI was finally stopped (Clark et al., 2019). Motivated by this study, we simulated a corresponding data set in which a 12-month dose reduction is explicitly modelled (AS dataset). Training the prediction approaches to explicitly integrate this perturbation, we found a substantial increase in the prediction accuracy of up to 19.1% (Fig. 3b). We argue that probing the system's response to perturbation (such as dose reduction) provides additional information about control mechanisms that cannot be obtained from ongoing monotherapy (Gottschalk et al., 2020; Hähnel et al., 2020; Roeder and Glauche, 2020).

Discussion

In summary, we could show that qualitatively different computational approaches, ranging from machine learning approaches to mechanistic models, are in principle suited for predicting relapse occurrence based on time-series data of leukaemia remission levels. To this end, we employed simulated time course data generated by mechanistic mathematical models, which we previously developed to describe disease and treatment dynamics in CML and AML. It is the advantage of this approach that we obtain highly controlled, although idealized, remission curves from which we can abstract different levels of sampling density and measurement error. The simulated data always allows us to refer to the ground truth of the underlying generative model. Applying this technique, we could also demonstrate that data quality in terms of measurement frequency and measurement error has a more substantial influence on the accuracy of the prediction than the employed prediction method, which is particularly evident in the AML data. Our results for the CML case indicate that it can be harder to fit a more complex mechanistic model (in terms of the number of model parameters) to noisy data than it is to fit a statistical predictor like a GLM or a generic, black-box neural network.

Our analysis illustrates that generic methods, such as neural networks work well for the prediction of disease recurrence if frequent measurements are available (as in the CML data). For diseases with sparse measurements and limited data on the other hand (exemplified in the AML data), neural networks (and representation learning in general) is less suited for identifying the critical factors underlying the disease dynamics. In such cases, it is beneficial to incorporate prior knowledge to yield better predictions using either mechanistic models of the disease, if available, or statistical approaches based on explicit (phenomenological) features. In our current study, we used an LSTM neural network as the standard approach for analyzing sequential data. An interesting next step is to assess if more complex neural network models (Chen et al., 2018; De Brouwer et al., 2019) can even improve upon the LSTM results, although we suspect that data quality is the dominant limiting factor.

Regardless of the exact choice for a predictive computational method, our study indicates that the optimization of measurement schemes and clinical protocols is a promising strategy to improve the overall prediction accuracy without necessarily requiring more measurements per patients. In our predictions for AML recurrence, we could reach a level of accuracy of about 80% for the prognosis of relapse occurrence within two years after diagnosis. This result would already exceed the prediction accuracy for relapse-free survival after 12 months in the study by (Othus et al., 2016). As our results are based on synthetic data, this comparison should be treated with caution. Still, our findings indicate that standardized measurement schedules adds critical leverage to improve the ability for predicting relapse no matter what computational methods are used in the end. Our artificial measurement schemes indicated a clear improvement, while we did not even apply formal optimization criteria to obtain most suitable regimes that maximizes accuracy while minimizing the number of measurements. This finding opens a clear perspective for future research on optimized measurement strategies that balance a maximized gain of information from clinical data with an economical use of resources. We argue that such refined schedules can contribute to reaching a level of prediction accuracy, which indeed supports clinical decision making.

In this work, we focused on the accuracy of relapse prediction employing three different, prototypic computational approaches working on time-series data. However, their implementation in a decision-making context also requires an intuitive understanding of how the method works. Although NNs do not require any prior knowledge and can achieve excellent prediction accuracies, it is not trivial to identify which aspects of the data are causative for a particular prediction (Arras et al., 2017; Shrikumar et al., 2017). In other words, the "black box" nature of NNs does intrinsically not reveal the key features of the data on which a decision is based. There is a general, ongoing scientific discussion about how severe this apparent lack of interpretability is (Esteva et al., 2019; Rudin, 2019). Consequently, there is still a level of reluctance and discomfort with decision-makers and regulatory authorities to consider such methods for integration into clinical routines. Mechanistic models represent the other side of the "interpretability spectrum" as they

superimpose a principal understanding of the underlying interactions onto the final observations. It appears tempting to favour this type of approach. However, it comes with other limitations: it is highly specific and not easily transferable to other disease entities, and it cannot be guaranteed that all essential interactions are indeed mapped (compare (Hoffmann et al., 2020)). GLMs represent a middle ground and balance the pros and cons of NN and MM approaches. They can be helpful if detailed mechanistic knowledge is missing while important features of the response characteristics can readily be named, estimated and also interpreted. However, their overall performance depends strongly on the choice of those hand-crafted features and is also vulnerable to missing critical aspects.

The increasing availability of diagnostic methods to track molecular remission in different cancer types over extended time periods will establish a rich data source to explore further how this dynamic information can be correlated with the future course of treatment and disease. Obtaining a systematic understanding of how different computational methods can be used to exploit this data is of crucial importance to provide usable predictions and potentially integrate them into decision making in the clinical context.

Experimental Procedures

Mechanistic models

To generate the synthetic data, we used our recently published mechanistic models for AML (Hoffmann et al., 2020) and CML (Hähnel et al., 2020). The models use ordinary differential equations to describe leukemic cell populations and their respective drug responses and mutual interactions. For the AML models, patient-specific differences in the disease characteristics are represented by two free parameters and varying treatment details (length, number and interval of chemotherapy cycles), while for the CML models we are estimating seven free parameters to describe a patient's response optimally. Details of the model setup are provided in the Supplementary Materials.

Patient data

For the generation of a set of realistic parameters, we fitted the respective mechanistic model to previously published time course data reflecting the patient's tumour remission during and after therapy. In particular, we used the time courses of 275 NPM1-mut AML patients, in which the level NPM1-mut/ABL abundance is used as a measure of leukaemia load (median follow-up time of 10 months, the median number of 5 measurements (Hoffmann et al., 2020)). Furthermore, we integrated data sets from 21 CML patients reflecting both their BCR-ABL1/ABL remission levels under TKI therapy and after therapy cessation (median follow-up time of 84 months, the median number of 28 measurements (Hähnel et al., 2020)). Examples of model fits to patient data, and the mean absolute error for each fitted patient can be found in Suppl. Fig. 1.

Generation of artificial data

To generate artificial patient data, we sampled from the set of parameters that we derived from fitting our models to the patient data.

In the case of AML, we sampled a random parameter combination from the empirically observed parameters. We added a small random variation to the parameters (see Supplementary Material for details) and sampled one clinical chemotherapy schedule from our pool of patient data. We then simulate an artificial time course of nine months length with our mechanistic model and the sampled parameters.

For the corresponding artificial CML time-courses, we sampled the seven model parameters from the distribution of empirical estimates in the available data basis, maintaining their mutual correlations (for details see Supplementary information). The therapy cessation time was sampled based on kernel density

estimates from the cessation time of the real patients. This information was then used to generate time-courses with the mechanistic CML model.

For each disease model, we generated the following 5 data sets with varying levels of data quality:

- Dense data (D): with weekly (AML) or monthly (CML) exact measurements, respectively.
- Dense-noisy data (DN): where white noise was added to each measurement, according to the noise level found in the real data.
- Sparse-noisy data (SN): generated from the DN data set by reducing the number of data points to reflect the measurement frequency in real patients.
- Artificial-Patient data (AP): by adding a detection limit to the SN data as found in the real data.
- Artificial scheme data (AS): Similar to AP data but using an improved sampling scheme compared to the clinical data. For AML measurements are made at the end of each chemotherapy cycle and every six weeks afterwards. For CML, the treatment dose is reduced to half of the usual dose 12 months before therapy cessation.

Example time courses of all data sets for each use-case can be found in Suppl. Fig. 2 and 3.

Using this synthetic data, we define the relapse prediction task as follows: for AML, we use a time window of 9 months after diagnosis (covering the treatment phase) to predict whether a patient will relapse within the subsequent 15 months. For CML, we use a time window from treatment start to cessation (avg of 92 months with a standard deviation of 28.2 months) to predict whether a patient will relapse within the subsequent ten years, as a CML can evolve very slowly, especially for a low number of tumour cells at treatment cessation. To obtain the model predictions in the case of MM, we fitted the model parameters to the available time course then simulated the future behaviour using the fitted model for each dataset individually. In contrast, both GLM and NN are initially optimized on a separate, labelled training set for which the respective outcome of relapse occurrence was given as a target value.

Explicit features of time series for GLM analysis

As the Generalized Linear Model, we use a logistic regression classifier. The model uses explicit features that describe characteristics of the time-course data. We took the two characteristics of AML time-courses defined in our previous work (Hoffmann et al., 2019): the elimination slope α, describing the speed of decrease of leukemic burden over the time of treatment and the lowest measured leukemic burden after treatment n. In this work, we further added three additional features: the leukemic burden at diagnosis (y_0, the following decreasing slope during the times of treatment (a) and the increasing slope of the leukemic burden in between treatment cycles (b) (Suppl Fig. 4A).

For CML, we defined seven features from fits of a bi-exponential function that described the decrease of the leukemic burden after treatment start. These features include the bi-exponential parameters (A, α, B, β), the corresponding deviation of the fit and the data (σ), the cessation time and the BCR-ABL1 value before cessation or half dose. For the AS data, we expand these features with the behaviour of the leukemic burden during the time of dose reduction including linear function parameter (γ), the deviation during half dose (C) and the last measured value before cessation. (Suppl Fig. 4B).

Neural Network

To predict the occurrence of relapse from the time series data, we used a bidirectional Long-short-term-memory (LSTM) network as the default architecture to handle sequence data with varying length. The model consists of a bidirectional LSTM layer followed by a fully connected feature extractor and a binary classification output. We use the respective cross-entropy loss to train the network. We implemented the network in Python using the Keras library (Chollet). To get a robust estimate of the model performance, we conducted 10 training runs on the same dataset and chose the network with the highest validation

accuracy. We then did 10-fold cross-validation for the entire experiment to assess the average and the variability of the results. Further details about the network architecture and training can be found in the Supplementary Materials and Methods.

Accuracy

We use the traditional definition of accuracy as the ratio of the number of correct predictions over the total number of predictions: $acc = \frac{\#correct}{\#total} = \frac{TP+TN}{TP+FP+TN+FN}$ where TP, TN, FP, and FN are true positives, true negatives, false positives and false negatives respectively.

Lead Contact,

nico.scherf@tu-dresden.de

Data and Code Availability.

The patient data for the AML patients are published in (Hoffmann et al., 2020) and can be found here: https://doi.org/10.6084/m9.figshare.12871777.v1
The CML patient data will be provided upon request to the correspondent author (Ingmar Glauche).
Source code is available at https://zenodo.org/record/4293490#.X8DznMtKg-Q

Acknowledgements

This work was supported by the Technische Universität Dresden, Faculty of Medicine Carl Gustav Carus, MeDDrive grant no. 60470 to NS. We thank clinical partners for providing the patient data within the originals works (Hähnel et al., 2020; Hoffmann et al., 2019, 2020) on which this simulation study is based.

Author Contributions

HH generated the AML data, carried out the analyses with the mechanistic model and the GLM for the AML use-case, conducted the numerical studies with the neural networks and drafted the manuscript. CB generated the CML data, carried out the analyses with the mechanistic model and the GLM for the CML use-case, and drafted the manuscript. TZ and AG were involved in creating a concept and a corresponding prototype. They critically revised the manuscript. SW, SN and NS revised the implementation of the neural network and improved the parametrization. EK contributed to the parameterization of the CML models. IG and IR contributed conceptually to the development and application of the GLM and the mechanistic models. IR critically revised the manuscript. IG and NS gave feedback during the analyses and drafted the manuscript.

References

Arras, L., Montavon, G., Müller, K.-R., and Samek, W. (2017). Explaining Recurrent Neural Network Predictions in Sentiment Analysis.

Branford, S., Yeung, D.T., Parker, W.T., Roberts, N.D., Purins, L., Braley, J.A., Altamura, H.K., Yeoman, A.L., Georgievski, J., Jamison, B.A., et al. (2014). Prognosis for patients with CML and >10% BCR-ABL1 after 3 months of imatinib depends on the rate of BCR-ABL1 decline. Blood *124*, 511–518.

Cancer Genome Atlas Research Network, Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A.G., Hoadley, K., Triche, T.J., Jr, Laird, P.W., et al. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N. Engl. J. Med. *368*, 2059–2074.

Cerveira, N., Loureiro, B., Bizarro, S., Correia, C., Torres, L., Lisboa, S., Vieira, J., Santos, R., Pereira, D., Moreira, C., et al. (2018). Discontinuation of tyrosine kinase inhibitors in CML patients in real-world clinical

practice at a single institution. BMC Cancer *18*, 1245.

Chaloner, K., and Verdinelli, I. (1995). Bayesian Experimental Design: A Review. Stat. Sci. *10*, 273–304.

Chen, R.T.Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. (2018). Neural Ordinary Differential Equations.

Chollet, F. keras (Github).

Clark, R.E., Polydoros, F., Apperley, J.F., Milojkovic, D., Rothwell, K., Pocock, C., Byrne, J., de Lavallade, H., Osborne, W., Robinson, L., et al. (2019). De-escalation of tyrosine kinase inhibitor therapy before complete treatment discontinuation in patients with chronic myeloid leukaemia (DESTINY): a non-randomised, phase 2 trial. Lancet Haematol *6*, e375–e383.

De Brouwer, E., Simm, J., Arany, A., and Moreau, Y. (2019). GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. ArXiv [Cs.LG].

Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F.R., Büchner, T., Dombret, H., Ebert, B.L., Fenaux, P., Larson, R.A., et al. (2017). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood *129*, 424–447.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. Nat. Med. *25*, 24–29.

Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2018). Deep learning for time series classification: a review.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning (The MIT Press).

Goodwin (1977). Dynamic System Identification: Experiment Design and Data Analysis (Academic Press).

Gottschalk, A., Glauche, I., Cicconi, S., Clarke, R.E., and Roeder, I. (2020). Molecular dynamics during reduction of TKI dose reliably identify molecular recurrence after treatment cessation in CML. Blood.

Hähnel, T., Baldow, C., Guilhot, J., Guilhot, F., Saussele, S., Mustjoki, S., Jilg, S., Jost, P.J., Dulucq, S., Mahon, F.-X., et al. (2020). Model-Based Inference and Classification of Immunologic Control Mechanisms from TKI Cessation and Dose Reduction in Patients with CML. Cancer Res. *80*, 2394–2406.

Hochhaus, A., Larson, R.A., Guilhot, F., Radich, J.P., Branford, S., Hughes, T.P., Baccarani, M., Deininger, M.W., Cervantes, F., Fujihara, S., et al. (2017). Long-Term Outcomes of Imatinib Treatment for Chronic Myeloid Leukemia. N. Engl. J. Med. *376*, 917–927.

Hoffmann, H., Thiede, C., Glauche, I., Kramer, M., Röllig, C., Ehninger, G., Bornhäuser, M., and Roeder, I. (2019). The prognostic potential of monitoring disease dynamics in NPM1-positive acute myeloid leukemia. Leukemia *33*, 1531–1534.

Hoffmann, H., Thiede, C., Glauche, I., Bornhaeuser, M., and Roeder, I. (2020). Differential response to cytotoxic therapy explains treatment dynamics of acute myeloid leukaemia patients: insights from a mathematical modelling approach. J. R. Soc. Interface *17*, 20200091.

Lundberg, S.M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc.), pp. 4765–4774.

Mahon, F.-X., Réa, D., Guilhot, J., Guilhot, F., Huguet, F., Nicolini, F., Legros, L., Charbonnier, A., Guerci, A., Varet, B., et al. (2010). Discontinuation of imatinib in patients with chronic myeloid leukaemia who have maintained complete molecular remission for at least 2 years: the prospective, multicentre Stop Imatinib (STIM) trial. Lancet Oncol. *11*, 1029–1035.

McCullagh, P., and Nelder, J.A. (1989). Generalized Linear Models (Chapman & Hall/CRC Monographs on Statistics and Applied Probability) (Chapman and Hall/CRC).

Nagafuji, K., Matsumura, I., Shimose, T., Kawaguchi, T., Kuroda, J., Nakamae, H., Miyamoto, T., Kadowaki, N., Ishikawa, J., Imamura, Y., et al. (2019). Cessation of nilotinib in patients with chronic myelogenous leukemia who have maintained deep molecular responses for 2 years: a multicenter phase 2 trial, stop nilotinib (NILSt). Int. J. Hematol. *110*, 675–682.

Oliva, E.N., Franek, J., Patel, D., Zaidi, O., Nehme, S.A., and Almeida, A.M. (2018). The Real-World Incidence of Relapse in Acute Myeloid Leukemia (AML): A Systematic Literature Review (SLR). Blood *132*, 5188–5188.

Othus, M., Wood, B.L., Stirewalt, D.L., Estey, E.H., Petersdorf, S.H., Appelbaum, F.R., Erba, H.P., and Walter, R.B. (2016). Effect of measurable ('minimal') residual disease (MRD) information on prediction of relapse and survival in adult acute myeloid leukemia. Leukemia *30*, 2080–2083.

Roeder, I., and Glauche, I. (2020). Overlooking the obvious? On the potential of treatment alterations to predict patient-specific therapy response. Exp. Hematol.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence *1*, 206–215.

Saussele, S., Richter, J., Guilhot, J., Gruber, F.X., Hjorth-Hansen, H., Almeida, A., Janssen, J.J.W.M., Mayer, J., Koskenvesa, P., Panayiotidis, P., et al. (2018). Discontinuation of tyrosine kinase inhibitor therapy in chronic myeloid leukaemia (EURO-SKI): a prespecified interim analysis of a prospective, multicentre, non-randomised, trial. Lancet Oncol. *19*, 747–757.

Seeger, M.W. (2008). Bayesian Inference and Optimal Design for the Sparse Linear Model. J. Mach. Learn. Res. *9*, 759–813.

Shanmuganathan, N., Pagani, I.S., Ross, D.M., Park, S., Yong, A.S., Braley, J.A., Altamura, H.K., Hiwase, D.K., Yeung, D.T., Kim, D.-W., et al. (2020). Early BCR-ABL1 kinetics are predictive of subsequent achievement of treatment-free remission in chronic myeloid leukemia. Blood.

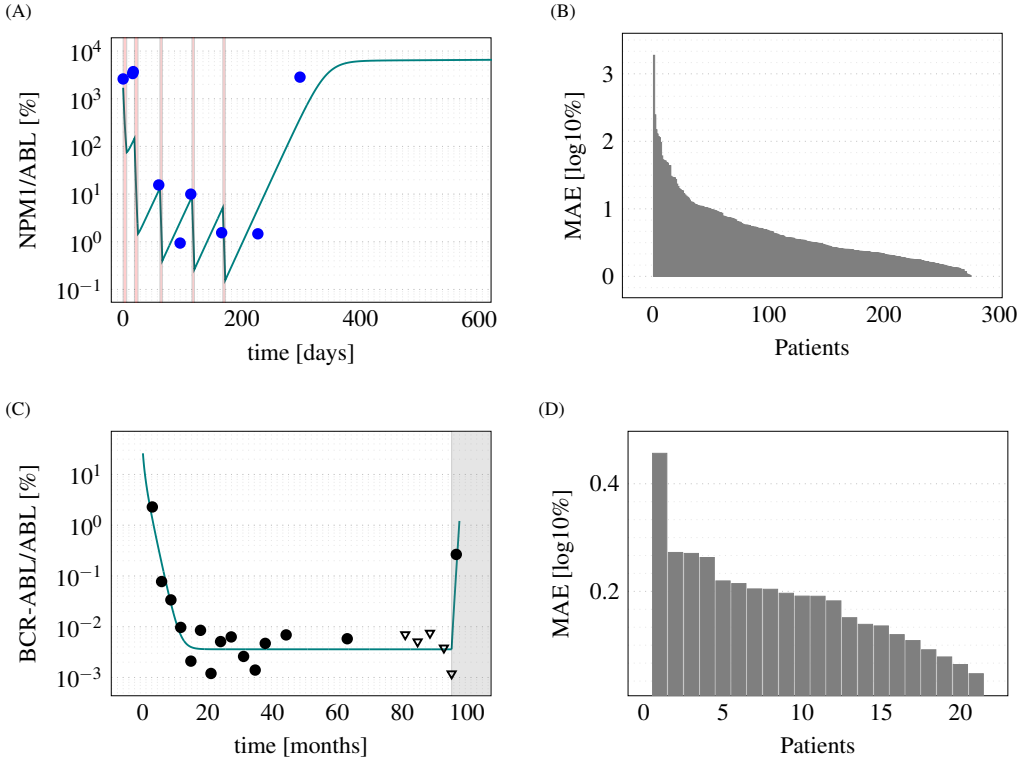
Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences.

Walter, E., and Pronzato, L. (2010). Identification of Parametric Models: from Experimental Data (Springer London).

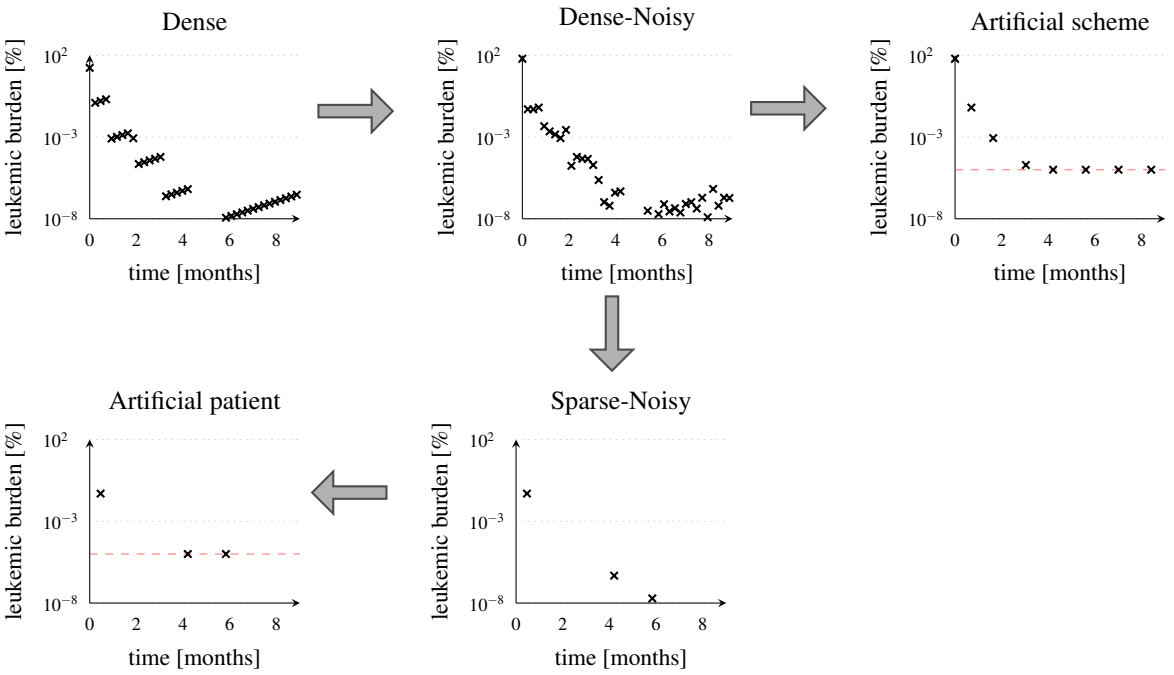
Zhang, X. (1994). Time series analysis and prediction by neural networks. Optim. Methods Softw. *4*, 151–170.

Zhou, H., and Xu, R. (2015). Leukemia stem cells: the root of chronic myeloid leukemia. Protein Cell *6*, 403–412.

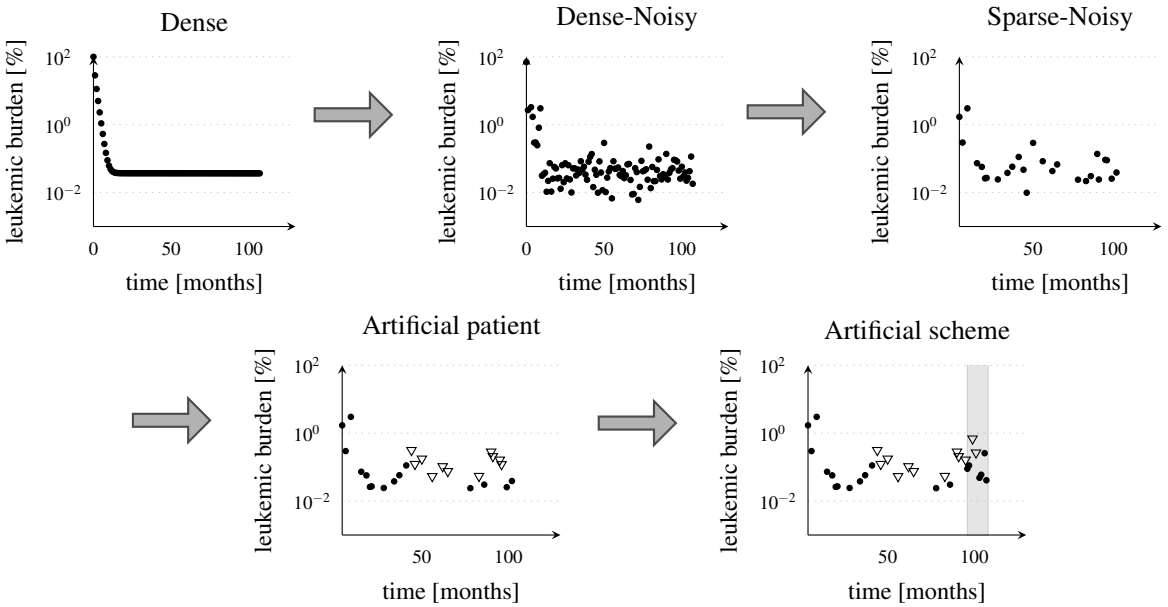
Supplement



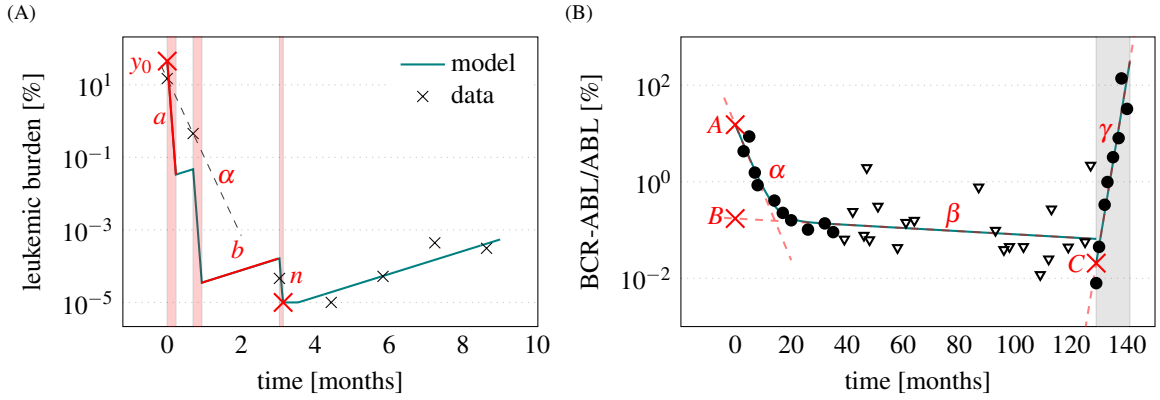
Suppl Fig. 1: Mechanistic model fit to patient data: (A) Example relapse of an AML patient and respective model fit. Red areas show time of chemotherapy administration. (B) Mean absolute error (MAE) of all 275 fitted AML patients. (C) Example of an CML patient and respective model fit, showing a relapses after TKI stop. (D) MAE of all 21 fitted CML patients.



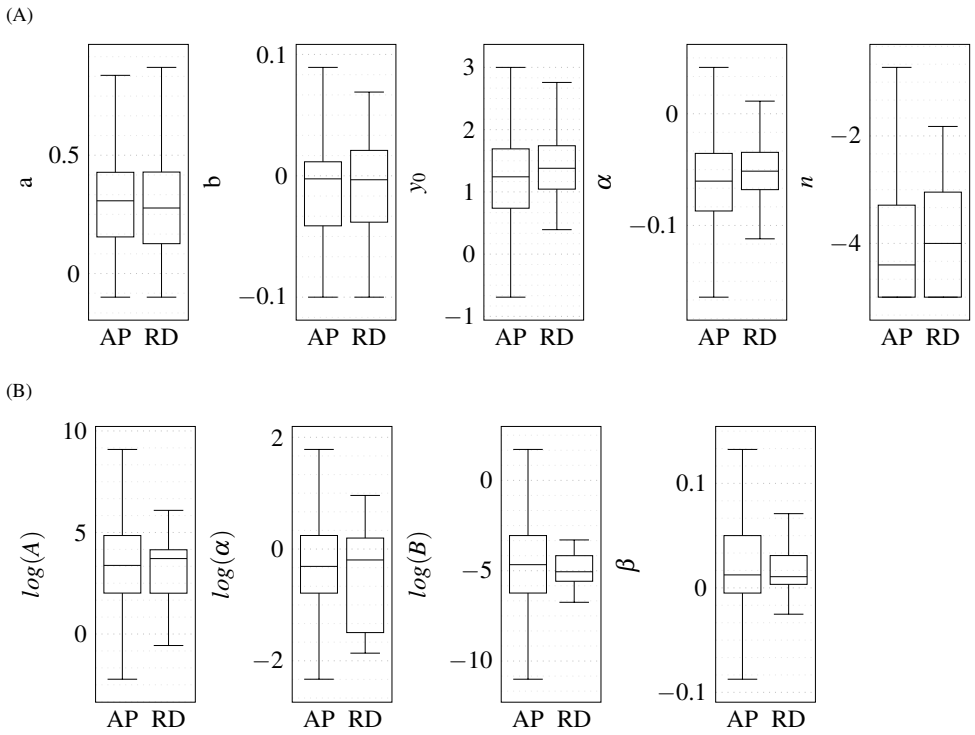
Suppl Fig. 2: Overview of artificial AML data sets: Dense data (D) with weekly exact measurements. Dense-noisy data (DN) where white noise was added to each measurement. Sparse-noisy data (SN) was generated from the DN data set, by reducing the number of data points to meet the measurement frequency in real patients. Artificial patient data (AP) is the data set most similar to the real patient data, which differs from the SN data set only by the inclusion of a detection limit, as it is found in the real data. Artificial scheme data (AS) is a data set, close to real data, with a measurement scheme, where measurements are made at the end of each chemotherapy cycle and every 6 weeks afterwards.



Suppl Fig. 3: Overview of artificial CML data sets: Dense data (D) with monthly exact measurements. Dense-noisy data (DN) where white noise was added to each measurement. Sparse-noisy data (SN) was generated from the DN data set, by reducing the number of data points to meet the measurement frequency in real patients. Artificial-Patient data (AP) is the data set most similar to the real patient data, which differs from the SN data set only by the inclusion of a detection limit, as it is found in the real data. Artificial scheme data (AS) is a data set, close to real data, with an additional 12-month period of half-dose TKI treatment (shown in gray).



Suppl Fig. 4: Derived features of the time-courses: (A) AML features with y_0 the leukemic burden at diagnosis, a the decreasing slope during treatment and b the increasing slope in treatment free intervals. (B) CML features with A, B and C being the intersections of the fitted line to the first and the second part of the biexponential fit and to the increase of the leukemic burden during half-dose periods, respectively. α, β and γ are the respective slopes.



Suppl Fig. 5: Similarity of artificial patients and real patients: a distribution comparison of statistical parameters: Comparison of distribution of parameters describing the course characteristics between artificial patient data (AP) and real data (RD).

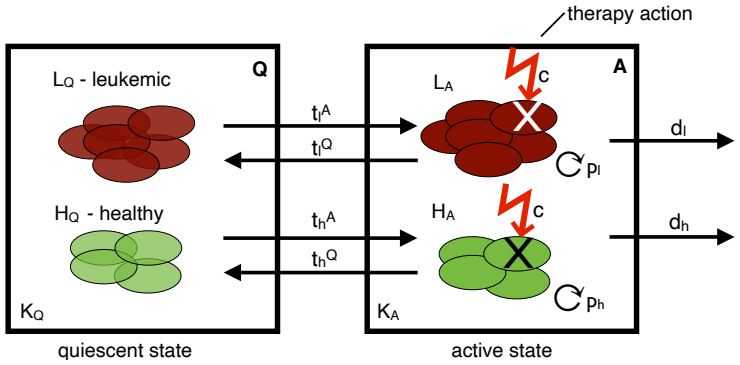
Supplementary Materials and Methods

AML

Clinical data

To be able to generate data with the mechanistic AML models that were as similar as possible to real clinical data, we used a patient data set for comparison. The data published together with the mechanistic model was used (Hoffmann2020). 61 of these 275 patients relapsed within two years after therapy initiation. The data consists of time course measurements of the relative tumour load of NPM1-mutated patients, by qPCR measurements of the amount of NPM1-mut transcripts relative to the amount of the reference genes transcripts (ABL).

Mechanistic model



t^A	activation rate	day^{-1}
t^Q	inactivation rate	day^{-1}
d	differetiation rate	day^{-1}
p	proliferation rate	day^{-1}
c	kill rate	day^{-1}
K	carrying capacity	*cells*
L	LIC	*cells*
H	HSC	*cells*

Our mechanistic model of the molecular disease dynamics of AML, already published in (Hoffmann et al., 2020), describes the dynamics of healthy stem cells (H) and leukemia-initiating cells (L) in the bone marrow of an AML patient. Each cell can adopt one of the two differential states: a quiescent state (Q) and an active state (A). The cells can reversibly switch between states with transition rates $t^A_{L/H}$ and $t^Q_{L/H}$. Cells in active state (Q_A/L_A) proliferate with proliferation rate $p_{L/H}$ and are sensitive to chemotherapeutic treatment with the kill rate c. Active cells can also differentiate into other states and therefore leave the two observed states. Individual chemotherapy schedules as well as patient-specific transition rates from quiescent to active state of the leukemic cells (t_L^A) and proliferative potential of leukemic cells (p_L) are taken into account to adapt the model to individual patients. The model equation are as follows:

$$\frac{dH_Q}{dt} = t_h^Q \cdot \left(1 - \frac{L_Q + H_Q}{K_Q}\right) \cdot H_A - t_h^A \cdot \left(1 - \frac{L_A + H_A}{K_A}\right) \cdot H_Q \quad (1)$$

$$\frac{dL_Q}{dt} = t_l^Q \cdot \left(1 - \frac{L_Q + H_Q}{K_Q}\right) \cdot L_A - t_l^A \cdot \left(1 - \frac{L_A + H_A}{K_A}\right) \cdot L_Q \quad (2)$$

$$\frac{dH_A}{dt} = t_h^A \cdot \left(1 - \frac{L_A + H_A}{K_A}\right) \cdot H_Q - t_h^Q \cdot \left(1 - \frac{L_Q + H_Q}{K_Q}\right) \cdot H_A + \left(p_h \cdot \left(1 - \frac{L_A + H_A}{K_A}\right) - c - d_h\right) \cdot H_A \quad (3)$$

$$\frac{dL_A}{dt} = t_l^A \cdot \left(1 - \frac{L_A + H_A}{K_A}\right) \cdot L_Q - t_l^Q \cdot \left(1 - \frac{L_Q + H_Q}{K_Q}\right) \cdot L_A + \left(p_l \cdot \left(1 - \frac{L_A + H_A}{K_A}\right) - c - d_l\right) \cdot L_A \quad (4)$$

For individual patient fitting we minimized the residual sum of squares.

Artificial data

For data generation the chemotherapy information of real patients was used, as well as the fitted parameters from the patient set, published in Hoffmann2020, which were slightly varied with white noise. The following 5 data sets with 5000 time-courses each, half of them relapsing, half of them non-relapsing within two years, and a measurement period of 9 months were generated:

- Dense (D) data: one measure per week starting with the first day of chemotherapy.
- Dense-noisy (DN) data: added noise to each measurement from D.
- Sparse-noisy (SN) data: Measurement frequency was adjusted to the number and intervals of measurements in the real data.
- Artificial patient (AP) data: all values of SN below the detection limit of -5 [log10] were set to -5 [log10].
- Artificial scheme (AS) data: using DN measurement time points at the beginning of each therapy cycle and every 6 weeks thereafter. All values below detection limit were set to -5 [log10].

Only patients which reached remission after therapy and did not relapse within 9 months were included in the data set. If the leukemic burden exceeded the relapse threshold of 1% at two years after treatment start, the time series was classified as a relapse.

Generalized linear model

For training a generalized linear model (GLM), in more depth a logistic regression classifier, we derived 5 features. Two were taken from the previous description of the main characteristics of the time course in AML (Hoffmann2019): the elimination slope α, giving a measure of the decrease of leukemic burden during therapy, until the relapse threshold of 10^-3 is reached and the minimal leukemic burden after therapy (n). The other three features were derived from a simple model describing the time course with a starting point (y0) at the beginning of treatment, a linear decrease with slope a of the leukemic burden during the time of treatment and a linear increase with slope b during treatment-free periods (see Suppl. Figure 4A). The starting points (y0) and the two slopes (a and b), together with the two characteristics were then handed to the GLM. When fitting the GLM using the parameters it became clear that two of them could be left out without losing accuracy. Therefore, the final GLM was fitted to predict the relapse based only on the minimal leukemic burden after therapy (n), the decreasing slope during therapy (a) and the increasing slope during treatment free periods (b). A 10-fold cross validation was used to estimate the variation of the estimated accuracy.

Neural Network

The neural network consists of a bidirectional Long-short-term-memory (LSTM) layer with 32 hidden units (using ReLU activation), followed by a fully connected layer with 64 nodes with a ReLU activation and a single, sigmoid output layer.

We use a vector of log10 values of NPM1/ABL values as input. Missing values (i.e. no measurement available at this particular time point) were encoded as -1 and a respective masking-layer was introduced to the network. For the non-detectable datapoints, we used the detection limit for the clinical data (see details about Clinical Data). The time points of chemotherapy were given as a second channel input in the case of AML.

To prevent overfitting, 10-fold cross validation was used. We saved the model with the highest validation (not training) accuracy. The entire training process was repeated 10 times on each dataset, to report a more robust estimate of the network performance as we regularly observed numerical instabilities in the

training process leading to failure of model fitting. Networks were trained with binary cross-entropy loss using the Adam optimizer with a learning rate of 1e-3 and a learning rate decay of 1e-5. The size of the validation set was chosen as 15 % of the training set. We used a batch size of 128 and trained for 100 epochs using an early stopping checkpoint to stop the training if the validation loss did not decrease in 20 epochs.

CML

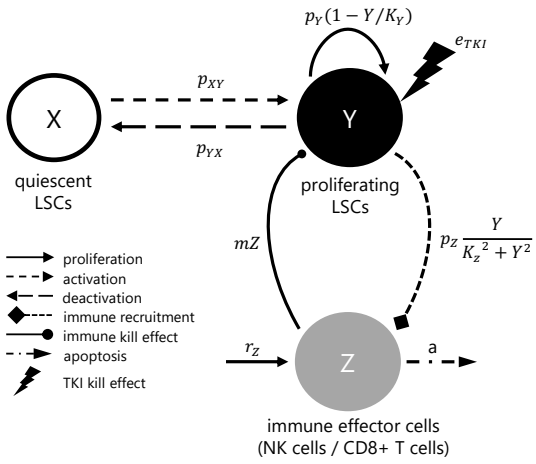
Clinical data

To generate in-silico CML patients as similar to real patients as possible, we utilized a cohort of 21 CML patients based on (Hähnel et al., 2020).

Each measurement can be either a detectable or an undetectable value. If a measure is detectable a corresponding BCR-ABL1 ratio on a log scale (LRATIO) is given. An undetectable measure is defined by a quantification limit (lQL) dependent on how much of the reference gene ABL1 was found in the sample. The more ABL1 is in the sample, the lower is the quantification limit.

For CML, typically a biphasic course in the time series is observed. A first, rapid decline symbolizes the fast clearout of the majority of tumor cells in blood. This phase usually takes between 6 and 12 months. After this initial phase, a second phase with a moderate decline begins, symbolizing the slow tumor degradation in bone marrow. During the treatment the tumor load measurable in blood becomes lower, thus the number of non-detectable measures increases over time.

Mechanistic model



The mechanistic model simulating the CML cells consists of 3 compartments. X defines the quiescent, non replicating cells, Y defines the active, proliferating cells and Z defines the immune cells specific to the CML cells described by X and Y. Cells change from a quiescent state into an active state and vice versa with certain probabilities defined by the rates p_{XY} and p_{YX}, respectively. Furthermore, Y cells proliferate with a logistic growth defined by a maximal rate p_Y. The TKI-effect is described by a constant rate TKI killing the corresponding proportion of leukemic cells in Y. Immune cells in Z get activated by the number of active cells in Y. At the same time the immune cells kill proportional cells from Y depending on its number. However, the activation function of Z, depending on Y, defines a so-called immune window. In other words,

depending on the parametrization immune effector cells are suppressed on very high and very low numbers of active leukemic cells. In between the immune cells rise and get effective (Hähnel et al., 2020).

Mechanistic parameters

To fit the model to clinical and in-silico data, we used the following model parameters:

- maximum activation rate of Z (p_Z)
- immune window suppressing constant (K_Z)
- cell switch rate from X to Y (p_{XY})
- cell switch rate from Y to X (p_{YX})
- maximum activation rate of Y (p_Y)
- TKI kill rate (e_{TKI})
- initial tumor burden (initRatio)

The following model parameters were fixed among all mechanistic simulations:

- maximum number of tumor cells in Y (K_Y = 1e^6)
- kill rate of active leukemic stem cells by immune cells (m = 1e-4)
- constant additive influx of immune cells (r_Z = 200)
- apoptosis rate of immune cells (a = 2.0)

Fitting Mechanistic Model to Data

We fitted the mechanistic parameters of all clinical as well as in-silico patients using a genetic algorithm. The fitness function is defined as the sum of the distance of all measurements with the following rules:

(1) detectable measures: quadratic distance between patient data and in-silico data
(2) undetectable measures: left censoring of values meaning if the simulation value is higher than the given lQL value (see Clinical Data) we use a quadratic distance, in case it is equal or lower, we use a 0-distance.

Artificial data

We generated the in-silico parameters using copulas to sample from the fitted patient parameters, such that we receive a highly similar correlation structure:

(1) Using Copulae, describing a functional dependency between the marginals of multiple independent variables and their corresponding joint probability distribution, we:
    (a) create a normal Copula with the correlation matrix reflecting the correlations of the empirical random variables,
    (b) generate the random observations with the Copula targeting marginal uniform distributions,
    (c) and transform it into the observed empirical distribution.

From this, we sampled 5000 in-silico patient parameters, whereas half are relapsing and half are non-relapsing patients. Additionally, we are taking the information of the properties of cessation time, measurement noise, density and frequency of undetectable values from the patient data. As properties differ at specific treatment phases, e.g. sample frequency is usually higher in the beginning, we are using the following sampling time intervals: 0-6, 6-12, 12-24, 24-36, 36-(cessation time).

(1) Dense Data (D): This dataset represents the raw simulation data taking one measurement per month. All measurements are noise-less and detectable.

(2) Dense-noisy Data (DN): We introduce noise, assuming it is solely represented by the distance between clinical data and the corresponding simulation fits. We sample from D using the corresponding error distributions, ending up with the newly defined DN dataset.
(3) Sparse-noisy (SN): Using the clinical data, we calculate for each time interval a distribution of the time of the first observed sample and a distribution of time intervals by taking the differences of consecutive time points. Sampling from these distributions, we can generate the sparse SN dataset, starting from DN.
(4) Artificial Patient (AP): For each interval a probability that a measurement is not-detectable is calculated. These probabilities are then applied to the in-silico patients from SN to receive the final in-silico patient data.
(5) Artificial Scheme (AS): We are simulating patients with a scheme inspired by a study called DESTINY (Clark et al., 2019). Therefore, we set the TKI dose in the model to half 12 months before cessation time. To generate this AS dataset, we started all over from (1) -(4) with the only difference in the time intervals. As patients in DESTINY during half dose are monitored very closely, we split the last time interval using 36 - (half dose start) and (half dose start) - (cessation time) intervals.

In all our CML simulations, we classify whether an in-silico patient relapses by simulating 10 years ahead and check whether the tumor burden is above MR1.

Generalized linear model

To predict the relapse outcome, we trained a logistic regression classifier (glm) using

- the first slope α,
- the corresponding intercept A,
- the second slope β,
- the corresponding intercept B,
- the standard deviation during the biexponential phase σ,
- the cessation time and
- the BCR-ABL1 ratio at cessation time,

whereas α, A, β, B and σ result from the corresponding biexponential fits.

In the case of the artificial scheme AS, the additional predictors were added:

- the third slope g during half dose,
- the corresponding intercept C,
- the first point in time of the half-dose and
- the BCR-ABL1 ratio at half dose time.

We allowed first-order interactions. However, resulting models were simplified based on the AIC until a minimum is reached. We used a 10-fold cross-validation.

Neural Network

We used mostly the same configuration as described above for the AML data. As input we used the vector of log10 values of BCR-ABL values. In case the value was not detectable, we used the lQL (see Clinical Data). To account for the more complex dynamics in the CML case and the higher number of data points within one time-series we increased the maximum number of epochs to 2000 with an early stopping if the validation loss does not decrease for 100 epochs.