

1  
2 Data integration between clinical research and patient care: a  
3 framework for context-depending data sharing and in silico  
4 predictions

5  
6  
7  
8 Katja Hoffmann<sup>1,2\*</sup>, Anne Pelz<sup>1</sup>, Elena Karg<sup>1</sup>, Andrea Gottschalk<sup>1</sup>, Thomas Zerjatke<sup>1</sup>, Silvio  
9 Schuster<sup>1</sup>, Heiko Böhme<sup>2</sup>, Ingmar Glauche<sup>1</sup>, Ingo Roeder<sup>1,2</sup>

10  
11  
12  
13 <sup>1</sup>Institute for Medical Informatics and Biometry, Faculty of Medicine Carl Gustav Carus,  
14 Technische Universität Dresden, Dresden, Germany

15 <sup>2</sup>National Center for Tumor Diseases (NCT), Partner Site Dresden, Dresden, Germany

16  
17  
18 \*Corresponding author

19 E-mail: [katja.hoffmann@tu-dresden.de](mailto:katja.hoffmann@tu-dresden.de) (KH)

20  
**NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**

## 21 **Abstract**

22 Usually, it takes quite some time until new insights from basic or clinical research are ultimately  
23 transferred into clinical routine. On the other hand, there are still many hurdles to directly  
24 provide and use routine data in the context of basic and clinical research. Specifically, no  
25 coherent software solution is available that allows a convenient and immediate bidirectional  
26 transfer of data between concrete treatment contexts and research settings.

27 Here, we present a generic framework that integrates health data (e.g., clinical, molecular) and  
28 computational analytics (e.g., model predictions, statistical evaluations, visualizations) into a  
29 clinical software solution which simultaneously supports both patient-specific healthcare  
30 decisions and research efforts, while also adhering to the requirements for data protection and  
31 data quality. Specifically, we emerge from a recently established generic data management  
32 concept, for which we designed and implemented a web-based software framework that  
33 integrates data analysis, visualization as well as computer simulation and model prediction  
34 with audit trail functionality and a regulation-compliant pseudonymization service. Within the  
35 front-end application, we established two tailored views: a *clinical (i.e., treatment context)*  
36 *perspective* focusing on patient-specific data visualization, analysis and outcome prediction,  
37 and a *research perspective* focusing on the exploration of aggregated, but pseudonymized  
38 data.

39 We illustrate the application of our generic framework by two use-cases from the field of  
40 haematology/oncology. Our implementation demonstrates the feasibility of an integrated  
41 generation and backward propagation of data analysis results and model predictions at an  
42 individual patient level into clinical decision-making processes.

## 43 **Author summary**

44 Patient-oriented research is based on comprehensive, quality-assured medical data that is  
45 visualized and analysed to gain knowledge. Based hereon, computer models can be

46 developed, which e.g., calculate risk scores or predict treatment success. Such approaches  
47 can be used for risk staging or for selecting the optimal therapy for a specific patient. In recent  
48 years, a lot of efforts have been made to develop generic concepts for data processing and for  
49 providing the data in the research context. What has been missing so far is a suitable software  
50 infrastructure to facilitate the direct backward propagation of scientific results into everyday  
51 clinical practice to support the treating clinicians in their decision-making processes. To close  
52 this gap, we designed a generic software framework into which, in principle, any computational  
53 model or algorithm can be integrated. For demonstration purposes, we developed a web  
54 application that integrates two mathematical models from the field of haematology, specifically  
55 relating to chronic myeloid leukaemia (CML). Both models calculate the leukaemia recurrence  
56 probability of a specific patient, after the intended stopping of the applied therapy. The  
57 particular prediction is based on patient-specific molecular diagnostic data and can be used  
58 for personalized treatment adaptation.

59

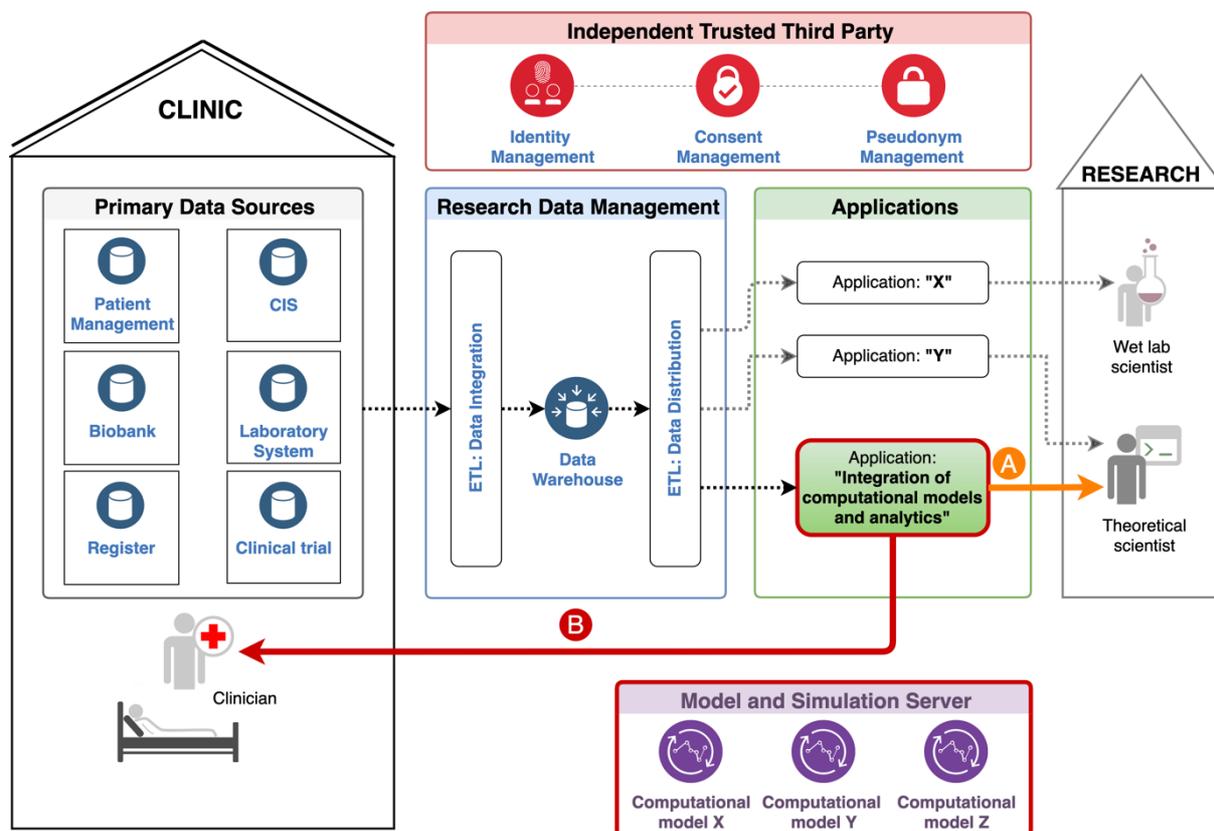
## 60 Introduction

61 Mathematical models and computational algorithms have already proven useful in analysing  
62 data, explaining functional relationships of (patho-)physiological mechanisms, and guiding  
63 clinical research. Along this line, also computational applications (“apps”) are increasingly used  
64 to support clinical decision-making (1-3). Prominent examples are risk scores, which provide  
65 patient-specific risk estimates by integrating relevant clinical parameters based on  
66 mathematical/statistical models (4-6). Furthermore, mathematical models and computer  
67 simulations for disease progression and treatment response have been developed for several  
68 diseases (7-14). As these models are intrinsically based on clinical data, there is an increasing  
69 need to provide such data in a structured and secure way. To fully realize the translational  
70 potential of both, models and data, it is necessary to establish interfaces that allow the  
71 integration of clinical data and corresponding model predictions and provide them in a hospital  
72 setting.

73 In our previous work (15) we focused on how patient-specific model predictions can be  
74 provided to clinicians within a particular clinical information system (CIS), while we have not  
75 considered the full life cycle of clinical and simulation data. However, due to privacy and  
76 security aspects, we need to ask how health data from multiple decentralized data sources  
77 can be consistently integrated and made available for computer simulations and analytics in  
78 general and how analytic and simulation results, e.g., generated to predict an individual  
79 patient’s future behaviour, can be transferred back to the routine clinical practice. Currently,  
80 we are missing technically applicable concepts to integrate the results of data analyses as well  
81 as computer simulations and mathematical model predictions in a hospital setting to support  
82 decision-making at an individual patient level. Therefore, we here describe a generic concept  
83 of integrating data analysis as well as computer simulation results and mathematical model  
84 predictions for scientific research and to support decision-making at an individual patient level  
85 in a common hospital setting. The described implementation goes beyond our previously  
86 published framework (15), specifically in terms of centralised and privacy-compliant

87 management of patient-identifying and medical data and thus for simultaneous use in clinical  
 88 and research settings.

89 Based on the conceptional works on data integration in scientific research (16), we investigated  
 90 how quality-control, data protection, and ethical requirements can be guaranteed when  
 91 integrating medical data in the research and clinical context in parallel. We took advantage of  
 92 the fact that hospitals implement data management concepts for the efficient and legally  
 93 compliant provision and use of health data with permanent traceability of all data sources with  
 94 a clear origin, time stamp and authorship. For this purpose, many hospitals are establishing  
 95 Research Data Management Systems for data pre-processing, storage and provision,  
 96 sometimes coupled with an Independent Trusted Third Parties (TTP) to centrally regulate  
 97 personal identifying data and to manage consents, revocations, and pseudonym assignments.  
 98 Fig 1 shows an example of a generic configuration of data management components from the  
 99 primary data sources to the applications in patient-oriented research.



100  
 101 **Fig 1. Example of a generic configuration of data management from the primary data sources to the applications:** Patient  
 102 data from various, decentralized, and heterogeneous data sources (Primary Data Sources) are extracted, transformed to

103 *harmonized data and loaded into a Data Warehouse (Extract-Transform-Load (ETL): Data Integration). Data transfer points*  
104 *pseudonymize these data and make them available for analysis and specific applications (ETL: Data Distribution). The*  
105 *identifying data, consents, and the pseudonym linkages are managed by an Independent Trusted Third Party or another*  
106 *pseudonymised service/provider. This institution/service also verifies the existence of the required consent before any*  
107 *processing of personal data. Our specific application “Integration of computational models and analytics”, which is*  
108 *additionally connected to a Model and Simulation Server, is intended to be used in clinical research (A) and also in clinical*  
109 *practice (B). Both components (outlined in bold red) complement the existing concept of data management in scientific*  
110 *research.*

111 Such a typical configuration comprises the selective extraction of heterogeneous patient data,  
112 which are stored in various primary data sources, the transformation into a unique target  
113 structure, and the loading into a Data Warehouse (cf. Fig 1 “ETL: Data Integration”). This  
114 procedure is often denoted as ETL (**Extract-Transform-Load**) process. The TTP verifies the  
115 existence of informed consent for the usage of the required data and provides application-  
116 specific pseudonyms. This ensures that the pseudonymized medical data do not allow any  
117 conclusions about a patient's identity. Eventually, data transfer points (cf. Fig 1 “ETL: Data  
118 Distribution”) provide the pseudonymized medical data requested for specific applications,  
119 which can be used by authorized researchers (e. g. clinical scientist, data analyst, modeler).

120 In contrast to other tools and workflows that have been already developed to provide health  
121 data for research (e.g. (17), (18)), the main focus of our approach is on the backward  
122 propagation of research findings and data analysis results into the clinical process, illustrated  
123 as the “B” arm in Fig 1. Specifically, we investigated how patient-specific predictions based on  
124 mathematical models and corresponding computer simulations can be integrated directly into  
125 the clinical workflows to support actual clinical decision-making for treatment optimizations.

126 Our work emerges from a previously established demonstrator software (15) that integrates  
127 time course data of residual disease levels for CML patients to provide estimates of their future  
128 behaviour. While the focus on CML is a particular example application, the demonstrator itself  
129 is flexible to integrate other types of data and underlying models. The application is primarily  
130 intended to be used by clinicians as a clinical decision support tool, complementary to existing

131 primary data sources to optimize treatment strategies on a patient-specific level (see Fig 1B).

132 In addition, it can also be used by researchers in the context of scientific projects (see Fig 1A).

133 For both use cases, the software application has to fulfil the following criteria:

134 1. Collection, processing, and provision of quality-assured pseudonymized data  
135 compliant with Data Protection Laws.

136 2. Provision of mathematical model predictions and data analyses.

137 3. Permanent traceability of all data and computational results with a unique origin, time  
138 stamp, and authorship (i.e., audit trail functionality).

139 4. Interactive visualization of medical data and simulation results.

140 5. Assignment of individual clinical data and model predictions and/or analyses to  
141 individual patients in clinical practice (reidentification of pseudonymized data).

142 6. User and access control by role-based right management.

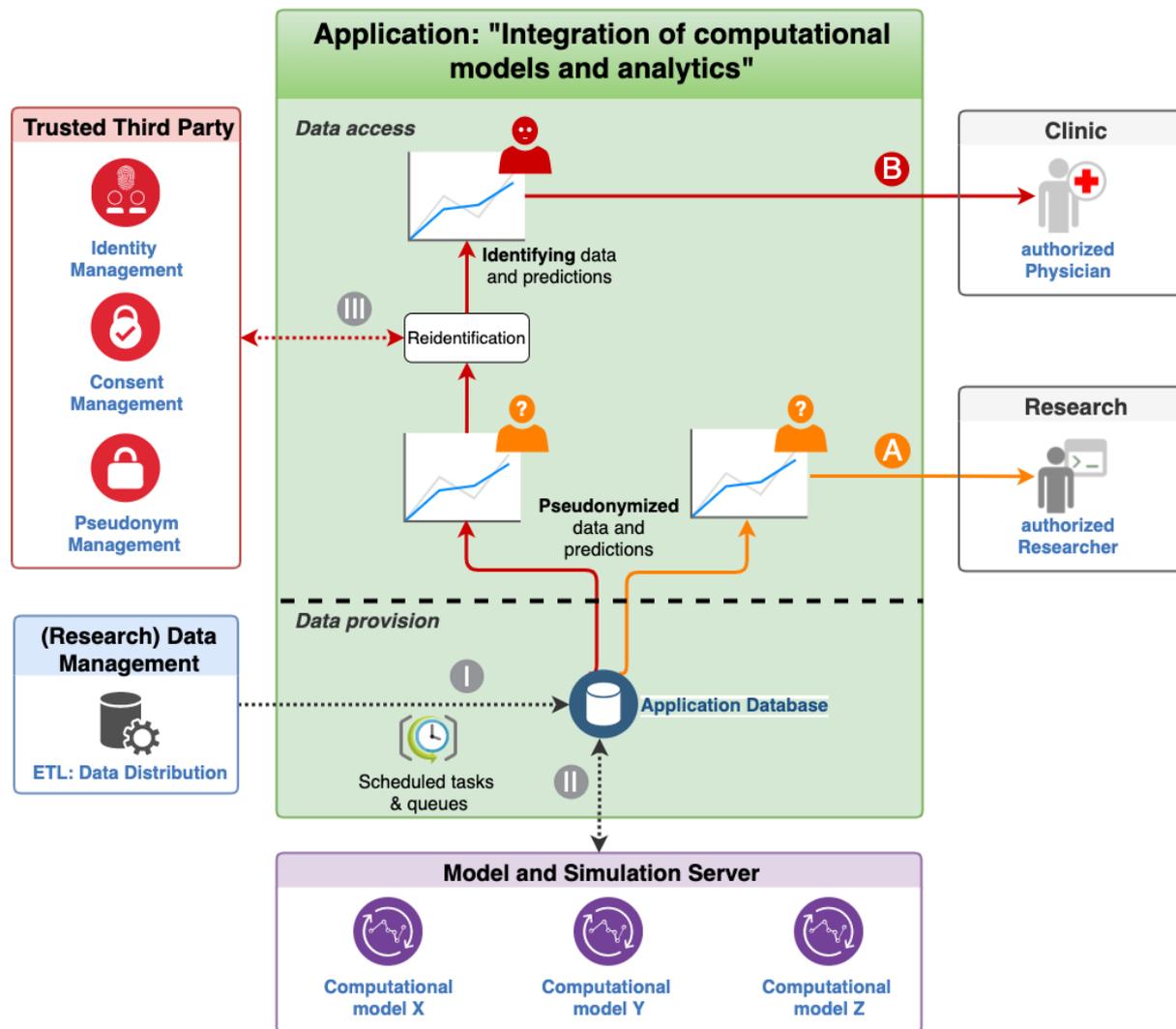
143 In the following, we describe our generic software framework (see Fig 2) that ensures all the  
144 specified requirements. We illustrate how algorithmic results (e.g., model predictions, analytic  
145 results) can be made available to researchers and clinicians directly in their working  
146 environment in an accurate, privacy compliant and transparent way. Finally, we present a  
147 newly developed web application (cf. the resulting demonstrator) to exemplify the generic  
148 software framework for a particular clinical application from the field of haematology/CML.

149

## 150 **Results**

### 151 **Generic overall design and data flows**

152 Based on the generic concept of data management in patient-oriented research (c.f. Fig 1), we  
153 developed a software framework that demonstrates how mathematical model predictions, and  
154 - in a broader context - any algorithmic/computational procedure can be integrated into clinical  
155 practice to support decision-making. Most importantly, our solution strictly complies with  
156 standards for data security and pseudonymization and allows audit trail compatible  
157 reconstruction of all of the provided analysis results/predictions (i.e., full transparency). To  
158 demonstrate the functionality of the software framework, we use the newly developed web  
159 application (the resulting demonstrator) as a particular computational application. It is  
160 complemented by a model and simulation server, a TTP/pseudonymization server, and an  
161 interface to the (Research) Data Management of the hospital. Fig 2 shows the overall design  
162 of the framework and illustrates the processes of data provision and access, that are described  
163 below.



164

165 *Fig 2. Data provision and access of the generic software framework: The Application Database contains exclusively*  
 166 *pseudonymized medical data transferred by the data transfer point of the hospital's (Research) Data Management (I) and*  
 167 *pseudonymized model predictions of particular computational models calculated by a model and simulation server for*  
 168 *modelling and analysis (II) through scheduled tasks. Authorized researchers can retrieve the pseudonymized medical data and*  
 169 *model predictions (A). For authorized clinicians, who have to be able to retrieve identifying medical data and model predictions*  
 170 *(B) in clinical practice, a reidentification step (III) through the Trusted Third Party has been integrated.*

171 Medical data, as input for data analysis and the calculation of model predictions, are provided  
 172 by a (Research) Data Management (cf. Figs 1 and 2I). This can be any database or data  
 173 warehouse storing medical data from (various) primary data sources, or also the primary data  
 174 source itself (cf. Fig 1). Only a suitable data distribution has to be established. Here, the  
 175 (Research) Data Management transfers exclusively pseudonymized data. No patient-  
 176 identifying data is stored in the Application Database. After the provision of new or updated

177 medical data, new computer simulations are scheduled periodically or event-driven. The  
178 results are provided by the Model and Simulation Server and transferred to the Application  
179 Database (Fig 2II). In this way, we ensure that the Application Database contains up-to-date  
180 simulation results/model predictions based on the most recent pseudonymized medical data.  
181 The data access is regulated by a role-based Access and Rights Management to establish  
182 different views on the data, e.g., trial-specific *roles* or department-specific *roles*. Members of  
183 the trial-specific roles can survey a whole spectrum of *pseudonymized* patients from one or  
184 more specific clinical trials, potentially involving different sites (see section Research view).  
185 Members of the department-specific role (e.g., haematology) have permission to access to the  
186 *full data* set of his/her patients treated in the particular center (see section Clinic view). In the  
187 treatment context, the reidentification of pseudonymized medical and simulation data is  
188 performed by the TTP (Fig 2III).

## 189 **The resulting demonstrator**

190 To demonstrate the generic software framework (cf. Generic overall design and data flows),  
191 we:

- 192 a) integrated a specific implementation of a model and simulation server (termed MAGPIE (19))  
193 to manage and execute computational models.
- 194 b) equipped MAGPIE with two mathematical models from the field of haematology (chronic  
195 myeloid leukemia, CML).
- 196 c) integrated the MOSAIC TTP server for identity, consent and pseudonym management.
- 197 d) added fictitious patients to the identity management module (E-PIX) of the MOSAIC TTP server  
198 for illustration purposes.
- 199 e) equipped the consent management module (gICS) of the MOSAIC TTP server with informed  
200 consents for various CML trials (Research view) and the data usage within the haematological  
201 treatment context (Clinic view).

202 f) generated pseudonyms for each informed consent form using the pseudonym management  
203 module (gPAS) of the MOSAIC TTP server.

204 g) designed ETL processes for transferring the medical data into the “Application database” that  
205 are needed for the specific analyses and model predictions.

206 h) implemented a role-based user and access management and established roles for the “Clinic  
207 view” and “Research view” presentation.

208 The resulting demonstrator is available at <https://tu-dresden.de/med/demoserver>. A detailed  
209 description of the demo server can be found in the following sections and the demo server  
210 walkthrough video (S1 File).

## 211 **Access control**

212 The demonstrator offers a role-based user and access management and can be adjusted to  
213 reflect department-specific roles (e.g., to be accessible for the Haematology Department) and  
214 trial-specific roles (to be limited to units involved in a given clinical trial e.g., the “Demonstrator  
215 CML Test Trial”). Each role corresponds to a consent implemented in the gICS module of the  
216 MOSAIC TTP server. This ensures that only data from patients who gave their written consent  
217 to the respective use are finally provided. Depending on the specific role, the user interface  
218 menu provides access to patient-specific data within a treatment context (here within the  
219 haematological treatment context; see section Clinic view) and/or to pseudonymized data of  
220 particular trials (see section Research view). For testing purposes, the resulting demonstrator  
221 provides guest accounts with the ability to switch between the department specific and the  
222 trial-specific roles.

## 223 **Clinic view**

224 Depending on the user role and the corresponding consent of the gICS module of the MOSAIC  
225 TTP server, the user can search for identifying patient data via a particular search form. For  
226 testing and demonstration purposes, the identity management module of the MOSAIC TTP

227 server is equipped with a selection of fictitious patients, inspired by but not identical to patient  
228 time courses from the DESTINY trial (NCT01804985, (20)).

229 Once a particular patient had been identified, the gPAS module of the MOSAIC TTP server  
230 provides the pseudonym, which is then used to retrieve the medical data and/or the computer  
231 simulation results from the Application Database (*re-identification*, see Fig 2III).

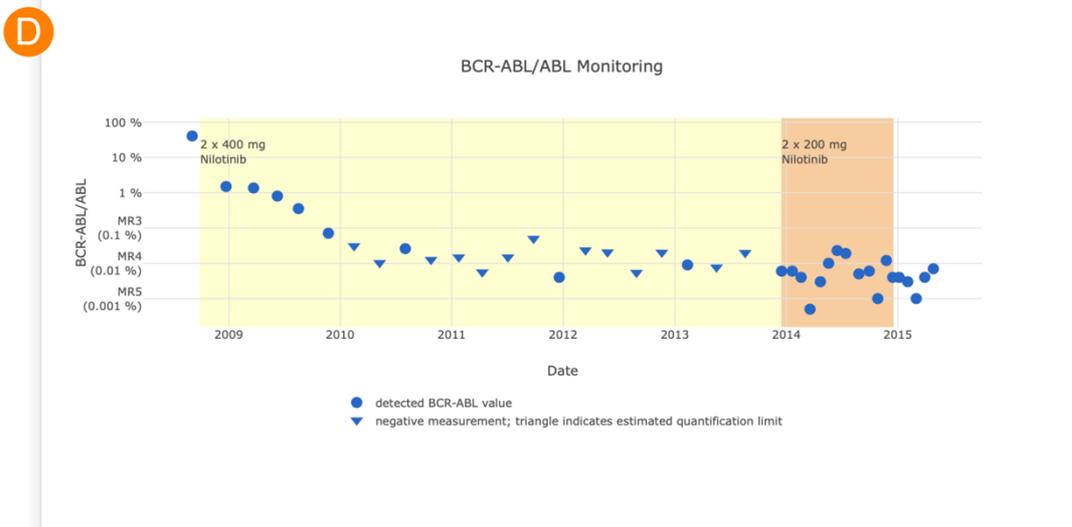
232 The “Clinic view” (Fig 3) presents the current medical data (D) and the latest model predictions  
233 (E) of the requested patient visualized in a dashboard (C1). The prediction of the recurrence  
234 probability after stopping the TKI treatment based on the model “Molecular monitoring during  
235 dose reduction” (cf. Computational models and analytics of the resulting demonstrator), is  
236 presented in a simplified and optionally in an expert view (E<sub>1</sub> and E<sub>2</sub>). The current medical data  
237 can be also retrieved and downloaded in form of a table view (Fig 3C2). Finally, the complete  
238 history of calculated computer simulation/model prediction can be followed (Figs 3C3 and 4),  
239 which allows an audit trail functionality of the computational results. Fig 4 shows a model-  
240 predicted immune classification based on the “Estimation of immunological leukaemia control”  
241 (cf. Computational models and analytics of the resulting demonstrator). For further details  
242 about the computational models see section Materials and Methods and the implementation  
243 of the demo server at <https://tu-dresden.de/med/demoserver>.

**A** Step 1: Search patient / Step 2: Retrieve patient-specific medical data and model predictions

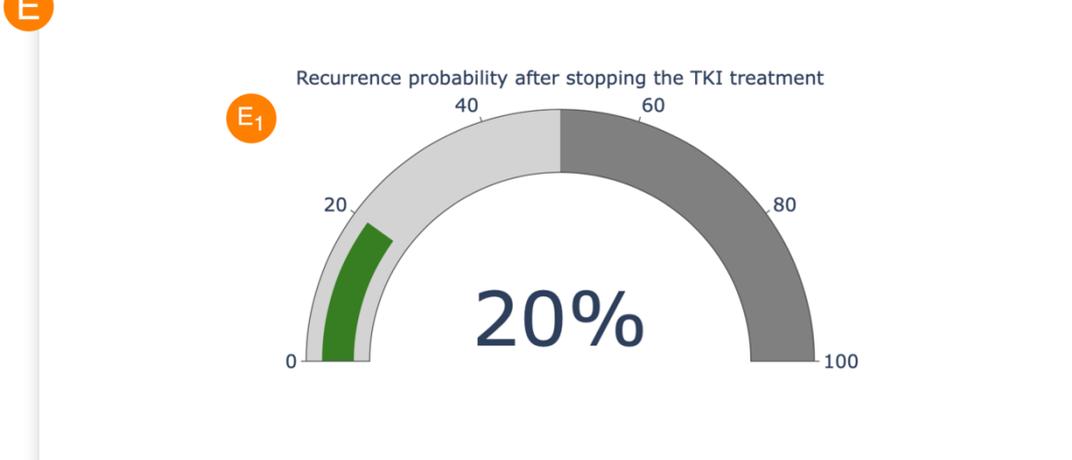
**B** Patient: **Mustermann, Max** born on 2000-01-01 in (M)

**C<sub>1</sub>** **Dashboard** | TableView "Treatments" | **TableView "BCR-ABL / ABL"** | All Predictions **C<sub>3</sub>**

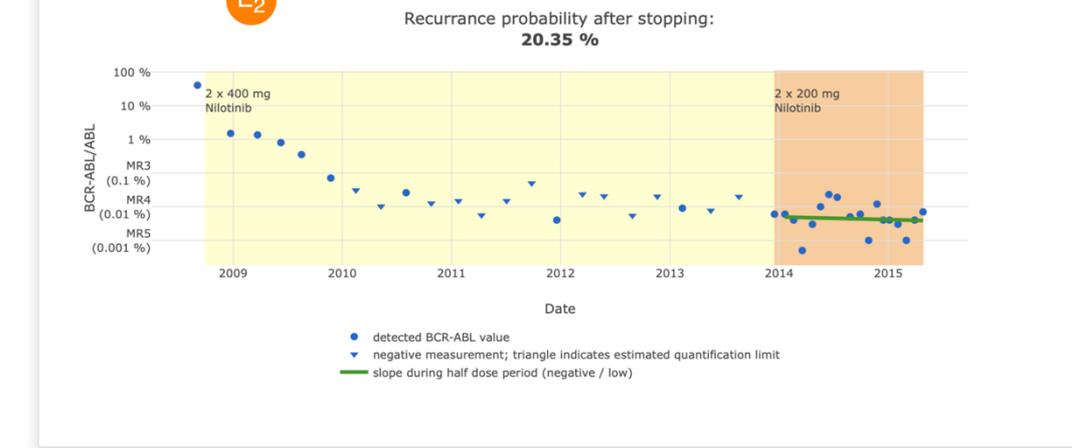
**C<sub>2</sub>** Latest clinical data



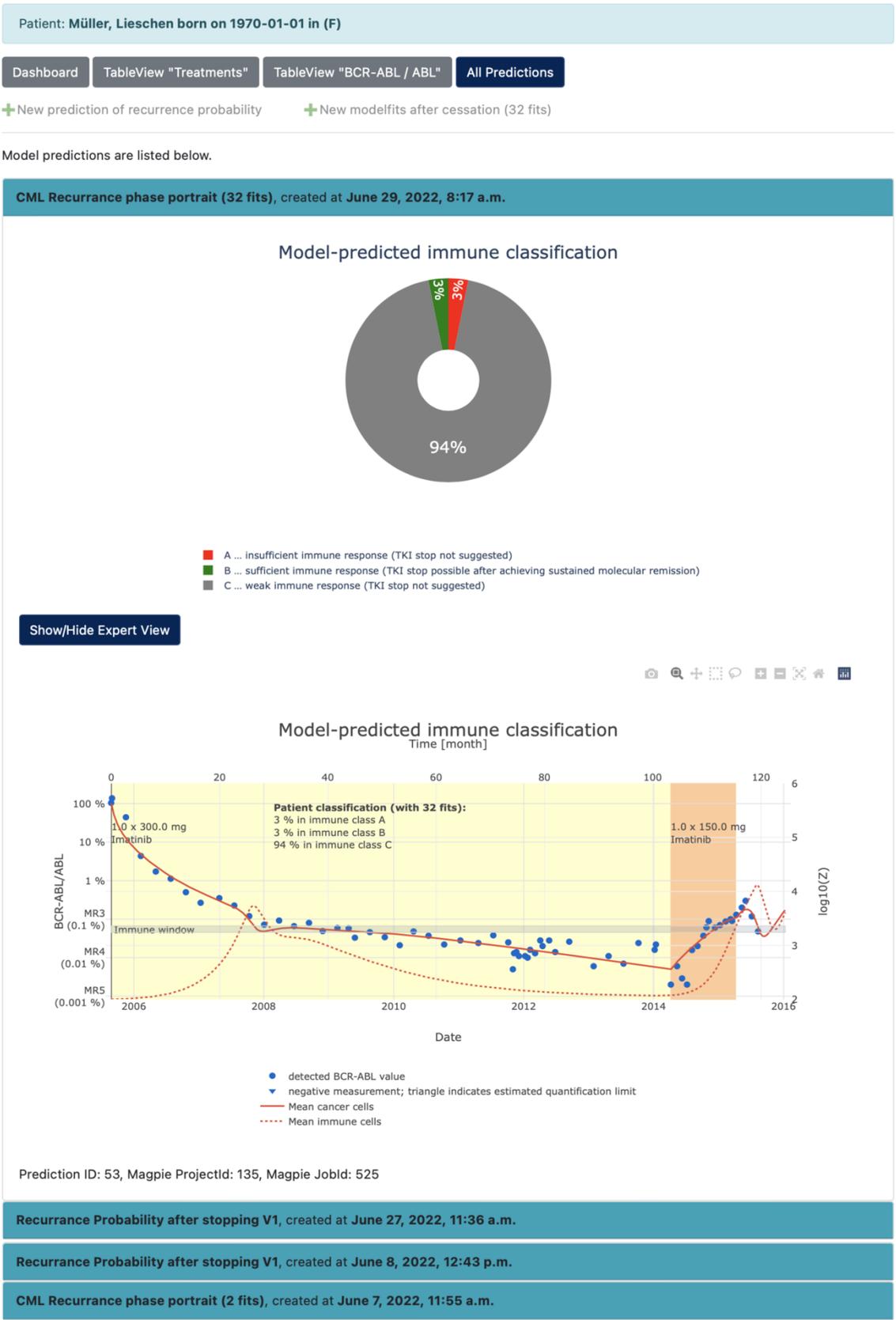
Latest prediction from model: "Recurrence probability after stopping V1"



Show/Hide Expert View **E<sub>2</sub>**



245 **Fig 3. Clinic view:** After searching for a patient by identifying data (A) and re-identification (B) of the pseudonymized medical  
246 data and simulation results, the latest medical data (D) and model predictions (E) are visualized as simplified and/or expert  
247 view ( $E_1$  and  $E_2$ ) in a dashboard (C1). Furthermore, the current medical data can be displayed and downloaded in table form  
248 (C2) and all previous predictions can be retrieved, too (audit trail compatibility, (C3) and Fig 4).



249

250 *Fig 4.: Audit trail functionality: Availability of every executed computer simulation/model prediction*

## 251 Research view

252 Controlled by the access management, the resulting demonstrator additionally offers the ability  
253 to access pseudonymized data (e.g., from clinical trials) and corresponding, aggregated  
254 analyses and simulation/model results (see Fig 5). The demo server yields clinical data sets  
255 from several trials from the field of haematology, specifically recent CML trails. Depending on  
256 the particular permission status, the user can or cannot access the detailed information of a  
257 given clinical trial.

	Name	Code	ClinicalTrials.gov	EudraCT	Disease
<a href="#">Details</a>	Demonstrator CML Test Trial	Trial_CML			CML
<a href="#">Details</a>	DESTINY	DESTINY	NCT01804985	2012-004025-24	CML
<a href="#">Details</a>	ENEST1st	ENEST1st	NCT00519090	2009-017775-19	CML
<a href="#">Details</a>	NordCML006	NordCML006	NCT00852566	2008-004106-13	CML

258  
259 **Fig 5. Research view – List of collected trials:** *Dependent on permission, the user can retrieve the detailed information of a*  
260 *trial, here the guest researcher is only able to retrieve the pseudonymized data and computer analyses of the “Demonstrator*  
261 *CML Test Trial”.*

262 The “Detail View” of a specific clinical trial, here the “Demonstrator CML Test Trial” (see Fig  
263 6), provides general information (A), the ability to access and download the pseudonymized  
264 medical data (B) and finally, the summary statistics and graphical illustrations for aggregated  
265 data (C). Further information about the analyses is provided in the section Example statistics.  
266 We wish to emphasized that the specific data analytics serve as an example to demonstrate  
267 the general ability of the framework, which can be flexibly extended or adapted to other

268 further/different computational analytics and graphical representations according to the needs  
269 of the physician and/or researcher.

[Home](#) [For Clinic](#) [For Research](#) [Logged in as guest researcher](#)

[List trials](#) / [Details](#) / [Documents](#)

## Trial details of trial Trial\_CML

[Edit](#) [Delete](#)

**A** **Trial name:** Demonstrator CML Test Trial  
**Study code:** Trial\_CML  
**Description:** Test trial for prediCt demonstrator

**ClinicalTrials.gov:**  
**EutraCT:**  
**Disease:** CML  
**Created At:** April 11, 2021, 7:49 p.m.  
**Created by:** su

### Access to research data

The trial contains **38 patients**.

- Diagnostic values
- Treatments

### Data visualization

Select plot:

Median and interquartile range (with disabled patient-specific time courses) x

**C**

#### BCR-ABL/ABL Aggregation from trial Demonstrator CML Test Trial

Legend:

- median of Demonstrator CML Test Trial(both genders)
- demo\_cml\_trial\_046156
- demo\_cml\_trial\_046144
- demo\_cml\_trial\_046143
- demo\_cml\_trial\_046110
- demo\_cml\_trial\_046108
- demo\_cml\_trial\_046101
- demo\_cml\_trial\_046086
- demo\_cml\_trial\_046080
- demo\_cml\_trial\_046069
- demo\_cml\_trial\_046061
- demo\_cml\_trial\_046055
- demo\_cml\_trial\_046054
- demo\_cml\_trial\_046046

**Select gender:**

both genders  only males  only females

**Include patients from other trials:**

Show trials separately (if false: jointly include trials)

Show interquartile ranges

270

271 *Fig 6. Research view – Details of a trial: This detailed view consists of general information (A), the ability to retrieve and*  
272 *download the current pseudonymized medical data (B) and the visualization of particular medical data and analysis on*  
273 *aggregated data (C).*

## 274 **Materials and Methods**

### 275 **Software environment of the demonstrator**

276 We used the high-level Python Web framework Django (21, 22) to build the frontend web  
277 application. Django provides robust security features, a user authentication system, and a fully  
278 featured admin interface.

279 The computer simulations (e.g., model predictions) are generated using an instance of the  
280 model and simulation server MAGPIE, which is a software framework, designed for publishing  
281 and executing computational models. MAGPIE also offers a full reproducibility of the calculated  
282 model predictions (i.e., audit trail functionality). For parameterization and execution of  
283 computational models and simulations, the “MAGPIE-API-R” package (19) is used. For a  
284 detailed description of the MAGPIE framework and implementation, we refer to Baldow et al.  
285 2017 (23).

286 The Python framework Plotly/Dash (24) is used to build the interactive visualizations of medical  
287 and simulation results, which are seamlessly embedded in the frontend application.

288 For identity, consent, and pseudonym management, which are accomplished by the TTP, the  
289 MOSAIC Toolbox (25), consisting of E-PIX (Identity Management Module), gICS (Consent  
290 Management Module), gPAS (Pseudonym Management Module) and the TTP dispatcher  
291 (TTP Workflow Manager Module) developed within the MOSAIC project (26) are used. The  
292 reidentification of pseudonymized data is performed by accessing the TTP Dispatcher (18)  
293 using the representational state transfer protocol (REST).

294 The demonstrator is based on an Apache HTTP Server (27) and uses PostgreSQL (28) as a  
295 database backend.

## 296 **Computational models and analytics of the demonstrator**

297 In the described demonstrator, two mathematical models for predictions on a patient level and  
298 a statistical analysis on aggregated data (see below for details) have been implemented for  
299 demonstration purposes. The framework itself is not restricted to these two particular models.  
300 It allows the deployment of different mathematical models and other algorithms, as long as  
301 they are registered in the MAGPIE model database, and feeding generated model predictions  
302 into the described workflow. There is no general restriction, neither on the model type nor on  
303 the particular implementation/programming language.

### 304 **Example model 1: Molecular monitoring during dose reduction in CML patients**

305 The DESTINY trial (#NCT01804985) investigated whether a reduction of tyrosine kinase  
306 inhibitor (TKI) dose in CML patients prior to treatment stop can lead to better treatment-free  
307 remission rates compared to full dose treatment (29). Based on a reanalysis of these molecular  
308 response data, we showed that an increasing BCR-ABL level during the dose reduction phase  
309 is indicative for CML recurrences after TKI stop and can be used to risk-stratify patients prior  
310 to treatment cessation (5). Given the individual molecular monitoring data during dose  
311 reduction, the statistical model implemented within the resulting demonstrator calculates the  
312 local slope parameter of the BCR-ABL dynamics and provides a prediction of the patient-  
313 specific recurrence probability (5).

### 314 **Example model 2: Estimation of immunological leukaemia control**

315 We also developed a mathematical model of CML-immune interaction (30), which can be fitted  
316 to available time course data of TKI-treated CML patients and after therapy stop. We showed  
317 that qualitatively different dynamic treatment response patterns can be characterized in the  
318 context of a mathematical model by different stable steady states, describing the individual  
319 immune response (31). The different steady states can be interpreted as “immune classes”.  
320 Specifically, we distinguish between an insufficient immune response (class A) for which a TKI  
321 stop does not seem to be appropriate, a sufficient immune response for which one can expect

322 leukaemia control after a TKI stop (class B), and a weak immune response (class C) for which  
323 the effectiveness of the immune response cannot be clearly predicted. For a given set of data  
324 (as provided within the demonstrator framework), this model can be used to infer patient-  
325 specific immune parameters and, therefore, to estimate a probability for his/her assignment to  
326 the three immune classes. As the estimated immune parameter are correlated with the  
327 patients' recurrence behaviour after treatment cessation, this classification can be used to  
328 support the decision-making with respect to TKI stopping.

### 329 **Example statistics: Estimating median time courses**

330 To evaluate particular patient-specific disease dynamics e.g., in comparison with an overall  
331 response pattern (e.g., across different studies), one can aggregate data from patient groups  
332 or trials. As an example, we calculate and visualize the median and interquartile ranges of the  
333 quantitative molecular measurements at different time intervals during treatment (see (32)).  
334 This feature is based on pseudonymized data and can be used also outside the treatment  
335 context (research view) for all data that are accessible for the particular user (cf. access  
336 control).

337

## 338 Discussion

339 We have investigated how data analytics, specifically mathematical model predictions and  
340 computer simulation results can be integrated into a clinical routine environment to support  
341 clinical decision-making at an individual patient level and, simultaneously, support scientific  
342 research at the level of pseudonymized data. In particular, we examined how quality-assured  
343 and harmonized health data from multiple decentralized clinical data sources can be provided  
344 to multiple users (in different contexts) for analyses processing (e.g., modelling/simulation) and  
345 how analytic results (e.g., mathematical model predictions) can be returned for direct  
346 application in clinical practice, ensuring data protection and ethical requirements.

347 As a result, a generic software framework (c.f. Fig 2) was designed that allows for seamless  
348 integration of computational applications into a common concept of Data Management in  
349 hospitals. Specifically, we demonstrate the integration and use of a particular application,  
350 denoted as “Integration of computational models and analytics”. The application can also be  
351 regarded as a generic “container” of a variety of different computational sub-  
352 application/algorithms.

353 In our particular example, we integrated two mathematical models, which provide different  
354 simulation-based predictions for disease and treatment dynamics in individual patients, in the  
355 context of TKI-treated CML. Our approach assumes that the data needed for these particular  
356 models are available in a (Research) Data Management System. In our specific examples,  
357 these are BCR-ABL/ABL time courses as molecular diagnostic data as well as TKI therapy  
358 information. We also assume that a pseudonymization service is established to provide and  
359 manage the particular patient-identifying data, application-specific consents and pseudonyms.

360 Our work continuous earlier efforts to demonstrated how mathematical model predictions and  
361 analytics can be integrated into a specific clinical information system (CIS) (15). In contrast to  
362 these previous studies, we are now presenting a stand-alone Clinical Decision Support System  
363 (CDSS) that is intrinsically connected to a (Research) Data Management System (c.f. Fig 1)

364 to retrieve all required medical data. The novel aspect of our current approach lies in the  
365 simultaneous provision and usage of health data in both clinical and research contexts  
366 (depending on the particular user context). It is not limited to the outlined molecular time course  
367 data but can be used for a wide range of health data (e.g., from various CIS, laboratory  
368 systems, biobanks) to be integrated for joined analyses, simulations and visualizations. This  
369 simultaneous usage is particularly beneficial for clinician scientists in the treatment context  
370 since all relevant medical information is available from various primary data sources within a  
371 unified and dedicated framework. Our general concept is also flexible in terms of the  
372 computational models to be integrated. Future research should therefore investigate what  
373 content and requirements are necessary for a user-friendly CDSS within a medical speciality.  
374 Furthermore, usability studies can be used to improve the layout and functionalities.

375 Another important advancement of the presented framework is the integration of an  
376 independent Trusted Third Party/pseudonymization service as a core element of the research  
377 data management infrastructure. Like others (17, 18, 33), we illustrate how health care data  
378 can be used for scientific research compliant with data protection. However, to our knowledge,  
379 this is the first example to illustrate that this service can also be used to provide patient-specific  
380 findings in a backwards manner to directly support clinical decision-making. As of now we are  
381 only aware of two pseudonymisation services (25, 34) while only one of them (the MOSAIC  
382 Tools) offers the necessary functionalities within our framework (i.e. identity, consent and  
383 pseudonym management). Advantages of other pseudonymization services should be  
384 addressed in future research projects.

385 In summary, this paper illustrates and describes how data analytic results, particularly in silico  
386 predictions, can be seamlessly integrated into workflows for simultaneous use in research and  
387 clinical care routines. We believe that the presented software framework is an ideal basis for  
388 further, even commercial developments and the transfer into clinical practice. While we  
389 extensively discussed the joined provision of both data and models in a clinical context, we did  
390 not elaborate on how the model predictions can actually be compared with future treatment

391 outcomes. Such comparison is another step that iteratively contributes to the validation and  
392 continuous improvement (possibly automated by artificial intelligence) of the models.

## 393 **Acknowledgements**

394 Not yet applicable.

395

## 396 References

- 397 1. Jing X, Himawan L, Law T. Availability and usage of clinical decision support systems  
398 (CDSSs) in office-based primary care settings in the USA. *BMJ Health Care Inform.*  
399 2019;26(1).
- 400 2. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An  
401 overview of clinical decision support systems: benefits, risks, and strategies for success. *npj*  
402 *Digital Medicine.* 2020;3(1):17.
- 403 3. Harada T, Miyagami T, Kunitomo K, Shimizu T. Clinical Decision Support Systems for  
404 Diagnosis in Primary Care: A Scoping Review. *International Journal of Environmental*  
405 *Research and Public Health.* 2021;18(16):8435.
- 406 4. Joseph PV, Wang Y, Fourie NH, Henderson WA. A computational framework for  
407 predicting obesity risk based on optimizing and integrating genetic risk score and gene  
408 expression profiles. *PLOS ONE.* 2018;13(5):e0197843.
- 409 5. Gottschalk A, Glauche I, Cicconi S, Clark RE, Roeder I. Molecular  
410 monitoring during dose reduction predicts recurrence after TKI cessation in CML. *Blood.*  
411 2020;135(10):766-9.
- 412 6. Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, et al.  
413 BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and  
414 nongenetic risk factors. *Genetics in Medicine.* 2019;21(8):1708-18.
- 415 7. Cook LM, Araujo A, Pow-Sang JM, Budzevich MM, Basanta D, Lynch CC. Predictive  
416 computational modeling to define effective treatment strategies for bone metastatic prostate  
417 cancer. *Scientific Reports.* 2016;6(1):29384.
- 418 8. Kheifetz Y, Scholz M. Individual prediction of thrombocytopenia at next chemotherapy  
419 cycle: Evaluation of dynamic model performances. *British Journal of Clinical Pharmacology.*  
420 2021;87(8):3127-38.
- 421 9. Huang S, Hu P, Lakowski TM. Predicting breast cancer drug response using a multiple-  
422 layer cell line drug response network model. *BMC Cancer.* 2021;21(1):648.
- 423 10. Chmielecki J, Foo J, Oxnard GR, Hutchinson K, Ohashi K, Somwar R, et al.  
424 Optimization of Dosing for EGFR-Mutant Non-Small Cell Lung Cancer with  
425 Evolutionary Cancer Modeling. *Science Translational Medicine.* 2011;3(90):90ra59-90ra59.
- 426 11. Jain HV, Clinton SK, Bhinder A, Friedman A. Mathematical modeling of prostate cancer  
427 progression in response to androgen ablation therapy. *Proceedings of the National Academy*  
428 *of Sciences.* 2011;108(49):19701-6.
- 429 12. Mumenthaler SM, Foo J, Leder K, Choi NC, Agus DB, Pao W, et al. Evolutionary  
430 Modeling of Combination Treatment Strategies To Overcome Resistance to Tyrosine Kinase  
431 Inhibitors in Non-Small Cell Lung Cancer. *Molecular Pharmaceutics.* 2011;8(6):2069-79.
- 432 13. Noble SL, Sherer E, Hannemann RE, Ramkrishna D, Vik T, Rundell AE. Using adaptive  
433 model predictive control to customize maintenance therapy chemotherapeutic dosing for  
434 childhood acute lymphoblastic leukemia. *Journal of Theoretical Biology.* 2010;264(3):990-  
435 1002.

- 436 14. Russo G, Palumbo GAP, Salvatore VD, Maimone D, Pappalardo F, editors.  
437 Computational models to predict disease course and treatment response in multiple sclerosis.  
438 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET);  
439 2021 9-10 Dec. 2021.
- 440 15. Hoffmann K, Cazemier K, Baldow C, Schuster S, Kheifetz Y, Schirm S, et al. Integration  
441 of mathematical model predictions into routine workflows to support clinical decision making in  
442 haematology. *BMC Medical Informatics and Decision Making*. 2020;20(1):28.
- 443 16. IT-Infrastrukturen in der patientenorientierten Forschung: Aktueller Stand und  
444 Handlungsbedarf 2015 / verf. und vorgelegt vom IT-Reviewing Board der TMF-Technologie-  
445 und Methodenplattform für die Vernetzte Medizinische Forschung e.V. [Hrsg. Johannes  
446 Drepper ; Sebastian C. Semler]2016.
- 447 17. Bahls T, Pung J, Heinemann S, Hauswaldt J, Demmer I, Blumentritt A, et al. Designing  
448 and piloting a generic research architecture and workflows to unlock German primary care  
449 data for secondary use. *Journal of Translational Medicine*. 2020;18(1):394.
- 450 18. Bialke M, Penndorf P, Wegner T, Bahls T, Havemann C, Piegsa J, et al. A workflow-  
451 driven approach to integrate generic software modules in a Trusted Third Party. *Journal of*  
452 *Translational Medicine*. 2015;13(1):176.
- 453 19. Baldow C. MAGPIE-API\_R. 2017.
- 454 20. Clark RE, Polydoros F, Apperley JF, Milojkovic D, Pocock C, Smith G, et al. De-  
455 escalation of tyrosine kinase inhibitor dose in patients with chronic myeloid leukaemia with  
456 stable major molecular response (DESTINY): an interim analysis of a non-randomised, phase  
457 2 trial. *Lancet Haematol*. 2017;4(7):e310-e6.
- 458 21. Django. 3.1.6 ed: Django Software Foundation; 2021.
- 459 22. Python Language. 3.6 ed: Python Software Foundation.
- 460 23. Baldow C, Salentin S, Schroeder M, Roeder I, Glauche I. MAGPIE: Simplifying access  
461 and execution of computational models in the life sciences. *PLOS Computational Biology*.  
462 2017;13(12):e1005898.
- 463 24. Collaborative data science. Plotly Technologies Inc.; 2015.
- 464 25. Bialke M, Bahls T, Havemann C, Piegsa J, Weitmann K, Wegner T, et al. MOSAIC – A  
465 Modular Approach to Data Management in Epidemiological Studies. *Methods Inf Med*.  
466 2015;54(04):364-71.
- 467 26. The MOSAIC-Project 2014 [Available from: <http://mosaic-greifswald.de/>].
- 468 27. Foundation AS. Apache HTTP Server.
- 469 28. PostgreSQL: a powerful, open source object-relational database system. The  
470 PostgreSQL Global Development Group.
- 471 29. Clark RE, Polydoros F, Apperley JF, Milojkovic D, Rothwell K, Pocock C, et al. De-  
472 escalation of tyrosine kinase inhibitor therapy before complete treatment discontinuation in  
473 patients with chronic myeloid leukaemia (DESTINY): a non-randomised, phase 2 trial. *Lancet*  
474 *Haematol*. 2019;6(7):e375-e83.

- 475 30. Hähnel T, Baldow C, Guilhot J, Guilhot F, Saussele S, Mustjoki S, et al. Model-Based  
476 Inference and Classification of Immunologic Control Mechanisms from TKI Cessation and  
477 Dose Reduction in Patients with CML. *Cancer Research*. 2020;80(11):2394-406.
- 478 31. Karg E, Baldow C, Zerjatke T, Clark RE, Roeder I, Fassoni AC, et al. Modelling of  
479 immune response in chronic myeloid leukemia patients suggests potential for treatment  
480 reduction prior to cessation. *medRxiv*. 2022:2022.06.28.22277004.
- 481 32. Glauche I, Kuhn M, Baldow C, Schulze P, Rothe T, Liebscher H, et al. Quantitative  
482 prediction of long-term molecular response in TKI-treated CML – Lessons from an imatinib  
483 versus dasatinib comparison. *Scientific Reports*. 2018;8(1):12330.
- 484 33. Bruland P, Doods J, Brix T, Dugas M, Storck M. Connecting healthcare and clinical  
485 research: Workflow optimizations through seamless integration of EHR, pseudonymization  
486 services and EDC systems. *International Journal of Medical Informatics*. 2018;119:103-8.
- 487 34. Lablans M, Borg A, Ückert F. A RESTful interface to pseudonymization services in  
488 modern web applications. *BMC Medical Informatics and Decision Making*. 2015;15(1):2.

## 489 **Supporting information**

490 S1 File. Walkthrough video of the resulting demonstrator