

Towards Robust Machine Learning for Health Applications

DISSERTATION

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
M. SC. LISA EISENBERG
aus Göppingen

Tübingen
2022

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 18.11.2022

Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Nico Pfeifer
2. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
3. Berichterstatterin:	Prof. Dr. Caroline Friedel

Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel

“Towards Robust Machine Learning for Health Applications”

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Unterschrift Lisa Eisenberg:

Contents

Abstract	vii
Kurzfassung	ix
Acknowledgments	xi
List of Publications	xiii
1 Introduction	1
1.1 Challenges and Potential of Machine Learning for Health Applications	1
1.2 Approaches to Robust Machine Learning	5
2 Objectives	9
3 Results and Discussion	11
3.1 Methods for unsupervised domain adaptation	11
3.1.1 Partially blind domain adaptation for age prediction from DNA methylation data (Manuscript 1)	13
3.1.2 Weighted elastic net for unsupervised domain adaptation (Manuscript 2)	14
3.1.3 Discussion	15
3.2 Robust model-based analyses for HIV research	20
3.2.1 Combination therapy with anti-HIV-1 antibodies maintains viral suppression (Manuscript 3)	21
3.2.2 Safety and antiviral activity of combination HIV-1 broadly neutralizing antibodies in viremic individuals (Manuscript 4)	23
3.2.3 Discussion	26

3.3	Robust models for transplantation medicine	28
3.3.1	Time-dependent prediction of mortality and cytomegalovirus reactivation after allogeneic hematopoietic cell transplantation using machine learning (Manuscript 5)	29
3.3.2	XplOit: An ontology-based data integration platform supporting the development of predictive models for personalized medicine (Manuscript 6)	33
3.3.3	Discussion	34
4	Integrated discussion and conclusions	39
	Bibliography	45
A	Publications	61
A.1	Manuscript 1	63
A.2	Manuscript 2	71
A.3	Manuscript 3	83
A.4	Manuscript 4	103
A.5	Manuscript 5	115
A.6	Manuscript 6	133

Abstract

Machine learning has enabled striking technological advances over the last decades and has the potential to transform many aspects of our lives. Its application is especially promising in the health domain, where it can improve our understanding of increasingly complex health data, accelerate processes such as diagnosis or risk assessment while also making them more objective, and enable a more personalized approach to medicine. At the same time, machine learning for health faces particular challenges. Health data is often temporal and heterogeneous, distributed across many institutions, and accessible only in modest amounts for a specific machine learning application. Consequently, machine learning for health requires generally robust methods capable of handling heterogeneous and limited data and models that are well-tailored to the task at hand. This thesis contributes to both of these aspects. It includes new methods for unsupervised domain adaptation, which were designed for high-dimensional molecular health data and improved prediction across heterogeneous datasets. As a concrete application example, these methods were applied to the problem of age prediction from DNA methylation data across tissues, where they improved age prediction on a tissue not used for model training compared to a non-adaptive reference model. In addition, this thesis includes robust models for the analysis of data from an early clinical trial evaluating the use of broadly neutralizing antibodies for the treatment of HIV, which were suitable to account for heterogeneity between patient groups despite a limited sample size. Another application-specific contribution was the development of robust models for the time-dependent prediction of mortality and early cytomegalovirus reactivation after hematopoietic cell transplantation. These models were validated in a prospective non-interventional clinical trial and demonstrated similar performance as experienced physicians in a pilot comparison. Finally, this thesis supported the development of the Xploit platform, a software platform that facilitates robust machine learning for health by semantically integrating heterogeneous datasets.

Kurzfassung

Methoden des maschinellen Lernens haben über die letzten Jahrzehnte beeindruckende technologische Fortschritte ermöglicht und haben das Potenzial, viele Aspekte unseres Lebens nachhaltig zu verändern. Besonders vielversprechend ist maschinelles Lernen im Gesundheitsbereich. Hier kann es unser Verständnis immer komplexerer Gesundheitsdaten vertiefen, Prozesse wie Diagnostik und Risikoeinschätzung beschleunigen sowie deren Objektivität erhöhen, und eine personalisiertere medizinische Versorgung ermöglichen. Zugleich steht maschinelles Lernen im Gesundheitsbereich vor besonderen Herausforderungen. Gesundheitsdaten sind häufig zeitabhängig und heterogen, über mehrere Institutionen verteilt und nur in begrenztem Umfang für spezifische Modellierungsanwendungen zugänglich. Infolgedessen erfordert das maschinelle Lernen für den Gesundheitsbereich grundsätzlich robuste Methoden, die für heterogene und im Umfang begrenzte Daten geeignet sind, sowie besonders auf die jeweilige Anwendung zugeschnittene Modelle. Diese Dissertation umfasst Beiträge zu beiden dieser Aspekte. Sie enthält neue Methoden zur unüberwachten Domänenadaptation, die speziell für hochdimensionale molekulare Gesundheitsdaten entwickelt wurden und eine genauere Vorhersage über heterogene Datensätze hinweg ermöglichen. Als konkretes Anwendungsbeispiel wurden diese Methoden auf das Problem der Altersvorhersage basierend auf DNA-Methylierungsdaten über Gewebe hinweg angewandt. Im Vergleich zu einem nicht-adaptiven Referenzmodell verbesserten sie hierbei die Vorhersage auf einem Gewebe, das nicht zum Trainieren der Modelle verwendet wurde. Zusätzlich enthält diese Dissertation robuste Modelle zur Analyse von Daten einer frühen klinischen Studie, die die Verwendung von breitneutralisierenden Antikörpern zur Behandlung von HIV untersuchte. Hier wurden Modelle und Methoden gewählt, die trotz des begrenzten Stichprobenumfangs Heterogenität zwischen Patientengruppen berücksichtigen konnten. Ein weiterer anwendungsspezifischer Beitrag war die Entwicklung robuster Modelle zur zeitabhängigen Vorhersage der Mortalität sowie einer Cytome-

galievirus-Reaktivierung nach hämatopoetischer Stammzelltransplantation. Diese Modelle wurden in einer prospektiven, nicht-interventionellen klinischen Studie validiert und generierten in einem Pilot-Vergleich eine ähnliche genaue Vorhersage wie die Einschätzung erfahrener Kliniker. Zusätzlich unterstützte diese Dissertation die Entwicklung der XplOit-Plattform, einer Software-Plattform, die robustes maschinelles Lernen für den Gesundheitsbereich durch die semantische Integration heterogener Daten erleichtert.

Acknowledgments

Like any big undertaking, this doctoral thesis would not have been possible without the support and encouragement of many people around me.

First of all, I would like to thank my advisor Nico Pfeifer for his unwavering support, his advice going far beyond this thesis, and for the opportunity to work on exciting and diverse interdisciplinary research projects. I also want to thank Oliver Kohlbacher and Tjeerd Dijkstra for helpful discussions and for agreeing to be on my thesis advisory committee, and all co-authors and collaborators of the projects I worked on. In particular, I would like to thank Ralf Eggeling for his helpful critical feedback and Amin Turki for the constructive interdisciplinary collaboration.

I am grateful to the German Federal Ministry of Education and Research and the Ministry of Science, Research, and the Arts of Baden-Württemberg for funding my research via the consortium project XplOit and the Centre of Innovative Care, respectively. Moreover, I would like to thank the International Max Planck Research School for Intelligent Systems for hosting both scientific and social events where I had the pleasure to meet many inspiring people that I am proud to call my friends. Thanks especially to Katja, Nico, Lennart, Alex, and all other regulars at B12 for everything from shared sorrows to uplifting bouldering sessions.

Throughout my PhD years, I have been part of multiple institutes and I would like to thank my colleagues at each of them. They are too many to name them all, but I want to mention Aaron and Tim for their mentorship during my time in Ulm and after, Anna, Nora, Michael, Markus, Peter, Dilip, Azim and all other members of our coffee group at the MPI for Informatics, as well as Jacqui, Ali, Nurhan and Agnes. Thank you for your company, interesting conversations, and support, the journey would have been lonely without you.

Last but certainly not least, I want to thank my friends and family for reminding me of what else is important in life. Most of all I want to thank my husband, Jan, for his love and support and for his patience with me through challenging times.

List of Publications

This cumulative doctoral thesis is based on the following six manuscripts, which are ordered by topic rather than chronologically. All manuscripts have previously been published (some under my maiden name, Lisa Handl), five of them in peer-reviewed international journals. They are reprinted here with permission.

Each manuscript was a collaboration with multiple co-authors, and my contributions in relation to the entire project are outlined on the subsequent pages.

1. **Handl, L.**, Jalali, A., Scherer, M. & Pfeifer, N. Partially blind domain adaptation for age prediction from DNA methylation data. *Workshop “Machine Learning for Health” at the Thirtieth Conference on Neural Information Processing Systems* (2016), arXiv:1612.06650 [q-bio.QM].
2. **Handl, L.**, Jalali, A., Scherer, M., Eggeling, R. & Pfeifer, N. Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data, *Bioinformatics* **35** (2019), i154–i163.
3. Mendoza, P., Gruell, H., Nogueira, L., Pai, J. A., Butler, A. L., Millard, K., Lehmann, C. Suárez, I., Oliveira, T. Y., Lorenzi, J. C. C., Cohen, Y. Z., Wyen, C., Kümmerle, T., Karagounis, T., Lu, C.-L., **Handl, L.**, Unson-O’Brien, C., Patel, R., Ruping, C., Schlotz, M., Witmer-Pack, M., Shimeliovich, I., Kremer, G., Thomas, E., Seaton, K. E., Horowitz, J., West Jr, A. P., Bjorkman, P. J., Tomaras, G. D., Gulick, R. M., Pfeifer, N., Fätkenheuer, G., Seaman, M. S., Klein, F., Caskey, M. & Nussenzweig, M. C. Combination therapy with anti-HIV-1 antibodies maintains viral suppression, *Nature* **561** (2018), 479–484.
4. Bar-On, Y., Gruell, H. Schoofs, T., Pai, J. A., Nogueira, L., Butler, A. L., Millard, K., Lehmann, C., Suárez, I., Oliveira, T. Y., Karagounis, T., Cohen, Y. Z., Wyen, C., Scholten, S., **Handl, L.**, Belblidia, S., Dizon, J. P.,

- Vehreschild, J. J., Witmer-Pack, M., Shimeliovich, I., Jain, K., Fiddike, K., Seaton, K. E., Yates, N. L., Horowitz, J., Gulick, R. M., Pfeifer, N., Tomaras, G. D., Seaman, M. S., Fätkenheuer, G., Caskey, M., Klein, F. & Nussenzweig, M. C. Safety and antiviral activity of combination HIV-1 broadly neutralizing antibodies in viremic individuals, *Nature Medicine* **24** (2018), 1701–1707.
5. **Eisenberg, L.**, XplOit Consortium, Brossette, C., Rauch, J., Grandjean, A., Ottinger, H., Rissland, J., Schwarz, U., Graf, N., Beelen, D. W., Kiefer, S., Pfeifer, N. & Turki, A. T. Time-dependent prediction of mortality and cytomegalovirus reactivation after allogeneic hematopoietic cell transplantation using machine learning, *American Journal of Hematology* **97** (2022), 1309–1323.
6. Weiler, G., Schwarz, U., Rauch, J., Rohm, K., Lehr, T., Theobald, S., Kiefer, S., Götz, K., Och, K., Pfeifer, N., **Handl, L.**, Smola, S., Ihle, M., Turki, A. T., Beelen, D. W., Rissland, J., Bittenbring J. & Graf, N. XplOit: An ontology-based data integration platform supporting the development of predictive models for personalized medicine, *Studies in Health Technology and Informatics* **247** (2018), 21–25.

Contributions

1. **Handl, L.**, Jalali, A., Scherer, M. & Pfeifer, N. Partially blind domain adaptation for age prediction from DNA methylation data. *Workshop “Machine Learning for Health” at the Thirtieth Conference on Neural Information Processing Systems* (2016), arXiv:1612.06650 [q-bio.QM].

Contributions:

This extended abstract describes an early version of my work on unsupervised domain adaptation, which was later refined and extended in Manuscript 2. The project was initiated by N.P. and A.J., building on previous work by M.S., who had assembled the DNA methylation dataset and trained an initial reference model. During the course of the project, N.P. and I jointly developed the main scientific ideas. I wrote all code for training the adaptive models, ran the experiments, and analyzed and interpreted the results. I also wrote most of the manuscript with additional input from N.P. and M.S.

2. **Handl, L.**, Jalali, A., Scherer, M., Eggeling, R. & Pfeifer, N. Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data, *Bioinformatics* **35** (2019), i154–i163.

Contributions:

This manuscript describes the domain adaptation method *wenda*, which builds on the early experiments described in Manuscript 1. Again, N.P. and A.J. initiated the project and M.S. provided the dataset for the application to age prediction and an initial reference model. Here, I contributed most scientific ideas for the method, including the core idea to use feature weights in the regularization term of an elastic net model for domain adaptation, with additional input from N.P. I also implemented these ideas, wrote the code for model training, simulations and analysis, and ran the experiments. With advice from N.P. and valuable critical feedback from R.E., I wrote most of the manuscript and managed its submission and revision.

This work was accepted for presentation at the combined international conference ISMB/ECCB 2019, which took place July 21–25 in Basel. I prepared

and gave the presentation, which was recorded and is currently still available on YouTube (<https://www.youtube.com/watch?v=JKvglb2QFCc>).

3. Mendoza, P., Gruell, H., Nogueira, L., Pai, J. A., Butler, A. L., Millard, K., Lehmann, C. Suárez, I., Oliveira, T. Y., Lorenzi, J. C. C., Cohen, Y. Z., Wyen, C., Kümmerle, T., Karagounis, T., Lu, C.-L., **Handl, L.**, Unson-O'Brien, C., Patel, R., Ruping, C., Schlotz, M., Witmer-Pack, M., Shimeliovich, I., Kremer, G., Thomas, E., Seaton, K. E., Horowitz, J., West Jr, A. P., Bjorkman, P. J., Tomaras, G. D., Gulick, R. M., Pfeifer, N., Fätkenheuer, G., Seaman, M. S., Klein, F., Caskey, M. & Nussenzweig, M. C. Combination therapy with anti-HIV-1 antibodies maintains viral suppression, *Nature* **561** (2018), 479–484.

Contributions:

This study evaluated if combination treatment with two potent broadly neutralizing antibodies, 3BNC117 and 10-1074, can maintain suppression of HIV-1 during an intentional interruption of the standard antiretroviral therapy. To demonstrate the benefit of combination therapy, the study cohort was compared to two cohorts from previous studies, one receiving monotherapy with only 3BNC117 and one receiving no intervention during the interruption of antiretroviral therapy. I performed a statistical comparison of these three cohorts to detect confounders and adjusted for detected confounders in a final analysis of treatment effects. Supported by advice and feedback from N.P., I wrote the code for the analysis, interpreted the results and provided the main authors with text modules describing these methods and results. Given the scale of the study and the large number of authors, this analysis was only a small part of the entire project. The contributions of all other authors are outlined in the contributions statement of the published article.

4. Bar-On, Y., Gruell, H. Schoofs, T., Pai, J. A., Nogueira, L., Butler, A. L., Millard, K., Lehmann, C., Suárez, I., Oliveira, T. Y., Karagounis, T., Cohen, Y. Z., Wyen, C., Scholten, S., **Handl, L.**, Belblidia, S., Dizon, J. P., Vehreschild, J. J., Witmer-Pack, M., Shimeliovich, I., Jain, K., Fiddike, K., Seaton, K. E., Yates, N. L., Horowitz, J., Gulick, R. M., Pfeifer, N., Tomaras, G. D., Seaman, M. S., Fätkenheuer, G., Caskey, M., Klein, F. & Nussenzweig,

M. C. Safety and antiviral activity of combination HIV-1 broadly neutralizing antibodies in viremic individuals, *Nature Medicine* **24** (2018), 1701–1707.

Contributions:

This study evaluated if combination treatment with the two broadly neutralizing antibodies 3BNC117 and 10-1074 can achieve viral suppression in patients with active HIV-1 infection who do not (yet) receive antiretroviral therapy. I performed a model-based analysis of the viral load over time and a statistical comparison of the effects of combination therapy and monotherapy (with either 3BNC117 or 10-1074) on the viral load. With advice and feedback from N.P., I wrote the code for these analyses, interpreted the results and contributed to writing the manuscript by describing the results of my analyses and writing the corresponding part of the methods section. Again, many authors contributed to this large-scale study and my work was one of multiple analyses presented in the manuscript. The contributions of all other authors are outlined in the contributions statement of the published article.

5. **Eisenberg, L.**, XplOit Consortium, Brossette, C., Rauch, J., Grandjean, A., Ottinger, H., Rissland, J., Schwarz, U., Graf, N., Beelen, D. W., Kiefer, S., Pfeifer, N. & Turki, A. T. Time-dependent prediction of mortality and cytomegalovirus reactivation after allogeneic hematopoietic cell transplantation using machine learning, *American Journal of Hematology* **97** (2022), 1309–1323.

Contributions:

This manuscript describes the development and prospective validation of machine learning models for time-dependent risk assessment after hematopoietic stem cell transplantation. These models are the main contribution of the component project *Statistical and Machine Learning Based Model Development for Transplantation Medicine*, which was headed by N.P. and executed by me within the *XplOit* consortium project.

Based on initial ideas by N.P. and with input from A.T.T., I developed the precise modeling scenario, selected appropriate model types and tools for model inspection and evaluation, and developed ideas for the special-

ized preprocessing of time-dependent laboratory measurements. I performed most data preprocessing, with the exception of initial format adjustments performed within the *XplOit* platform and the natural language processing of unstructured medical documents, wrote the code for model training and analysis, and ran the experiments. Jointly with A.T.T., J. Rissland, S.K. and N.P., I developed ideas for model validation and the prospective *XplOit* study. With advice from N.P., A.T.T. and I jointly interpreted the results and wrote the manuscript. Many more people contributed to the manuscript through their work in the *XplOit* project or in the prospective validation study. All names of the consortium members and their contributions are provided in the published article.

6. Weiler, G., Schwarz, U., Rauch, J., Rohm, K., Lehr, T., Theobald, S., Kiefer, S., Götz, K., Och, K., Pfeifer, N., **Handl, L.**, Smola, S., Ihle, M., Turki, A. T., Beelen, D. W., Rissland, J., Bittenbring J. & Graf, N. XplOit: An ontology-based data integration platform supporting the development of predictive models for personalized medicine, *Studies in Health Technology and Informatics* **247** (2018), 21–25.

Contributions:

This article describes the software platform that was developed in the consortium project *XplOit: Semantic Support for Predictive Modeling in Systems Medicine* with the goal to support data harmonization, model development and model validation in health care. While I did not contribute to the implementation of the platform itself, I was the only active member of the consortium developing machine learning models (with N.P. providing supervision in the background). In this role, I took part in monthly web conferences and biannual consortium meetings, and continuously supported the platform development through suggestions, testing and critical feedback from the perspective of a machine learning researcher. K.G., K.O. and T.L. provided similar feedback from the perspective of model developers using systems biology approaches. J.R. and K.R. implemented most of the platform, while U.S. and M.I. contributed specific components. G.W. and S.K. coordinated the consortium project. D.W.B., A.T.T., J.R., J.B. and S.S. were clinical

consortium partners who contributed data and medical expertise, and S.T. supported data pseudonymization and platform deployment. G.W. wrote most of the manuscript, which was proofread by all authors including myself.

1 Introduction

We are living in the information age. Digital communication and storage of information have grown exponentially over the last decades, and so has our capacity to process digital data [1]. This development has enabled analyses and data-driven decision-making on a scale that has not previously been possible and promises to transform many aspects of our society.

Machine learning is at the center of this transformation. Merely storing large amounts of data does not unlock its potential; we need algorithms to process it in order to draw conclusions from this data and turn it into actionable information. Machine learning can extract knowledge from data by discovering previously unknown associations between variables and can identify patterns that allow the prediction of future observations [2]. It has already become a standard tool for data-intensive research across scientific fields [3, 4, 5] and increasingly permeates our daily lives, whether in social media [6], automated translations [7], or recommender systems [8]. Applying machine learning in situations with real-life consequences is not without risks [9, 10] and is viewed with a mix of optimism and skepticism [11]. Nevertheless, doing so in a responsible way inarguably has the potential to improve our lives.

1.1 Challenges and Potential of Machine Learning for Health Applications

One area where machine learning could have a highly beneficial impact is health care. On the one hand, many processes in health care involve the interpretation of complex high-dimensional data by clinical experts, for example, for diagnosis or as a basis for treatment decisions. Humans, however, have only a limited capacity to process and accurately judge high-dimensional data, especially when time is

limited [12]. On the other hand, many individual factors contribute to a person's health and response to treatments they receive, including genetic factors, lifestyle choices, age, and disease history. Our understanding of these factors is still incomplete, and even if their impact is known, they cannot be accounted for in classical standardized treatments. Machine learning can advance health care concerning both issues. Models trained to assign patients to diagnosis labels or risk groups, for example, can support physicians in quickly drawing the right conclusions from high-dimensional data. And by incorporating individual health factors, models can also enable the prediction of patient-specific risks and outcomes, which may allow for more personalized treatments. In cases where diagnosis or treatment decisions can be fully automated, machine learning models could even improve access to highly specialized care [13]. Once an automated system has been trained using expert knowledge, it could be transferred to places where this knowledge would otherwise not be available.

In pursuit of these goals, machine learning has already become an essential component of health research. Several promising models have been developed to diagnose medical conditions [14, 15, 16] or predict individual patient risks [17, 18] based on medical image data or electronic health records, although translation to clinical practice is still challenging [19]. In health research, machine learning is also used as a supporting tool to study a host of other fascinating questions. Combined with bioinformatics approaches, model-based analyses contribute to studying genetic causes and molecular mechanisms underlying diseases [20, 21] and differing responses to treatment [22, 23, 24]. In addition, machine learning techniques are employed to support and accelerate research on drug discovery and development [25] or to explore possibilities for drug repurposing [26]. Recently, machine learning has, for the first time, enabled the accurate prediction of three-dimensional protein structures from amino acid sequences, which could catalyze these efforts [27]. Overall, machine learning in health research is a fast-evolving field and will need to continue to adapt as we gain the ability to measure biological processes with ever increasing throughput and level of detail.

However, while machine learning on health data is promising and opens new opportunities for health research, it also faces specific challenges. Here, health data refers to any information relating to the health, disease, and medical treatment of human beings, irrespective of the data type (e.g., unstructured medical documents,

image data, genetic data, blood measurements, vital signs, etc.) and of the location and purpose of its recording. There are many different kinds of health data, including routine clinical data recorded in hospitals or doctor’s offices, research data generated in clinical trials, and molecular data collected in biobanks [28] or large-scale projects like The Cancer Genome Atlas (TCGA) [29] or The Genotype-Tissue Expression (GTEx) Consortium [30]. The challenges associated with machine learning for health applications arise from the characteristics of both health data itself and its management.

For instance, from a machine learning perspective, the sample size in health datasets is often relatively small. For some research questions, sample sizes are naturally small, for example, when studying rare diseases or novel treatments. Data from routine clinical practice, on the other hand, is theoretically abundant but distributed across many institutions. Combining these datasets is still challenging, mainly because of privacy concerns and a lack of standardization [19]. When working with molecular health data, even a considerable sample size may seem small compared to the vast number of features, and correlations between features further complicate the development of meaningful models.

In addition, health data typically has high variability, originating from both biological and technical factors. Biological variability can be introduced by the subjects or patients (through individual health factors as mentioned previously), by their environment, or by high variability in the studied disease, for example, through genetic diversity in cancer cells or viruses. Technical variability can additionally arise from measurement errors, documentation errors, or differences in measurement techniques and treatment protocols across institutions. This high variability makes it difficult for a model to separate signal from noise, i.e., to distinguish the variability related to the phenomenon of interest from all variability caused by other factors, especially when the dataset available for model training is small.

Similar factors can cause health data to be heterogeneous, meaning that it contains subsets with distinctly different statistical properties. For example, patients may be sampled inhomogeneously from multiple medical centers following different treatment protocols or focusing on different diagnoses or age groups. In supervised machine learning, heterogeneity within a dataset can give rise to confounding factors, which are statistically associated with the outcome of interest but have no biologically meaningful relationship with it. If the goal is inference, such con-

founding factors may distort the conclusions drawn from a model unless they are controlled in the study design or accounted for during data analysis. In models built for a prediction task, confounding factors may lead to an overestimation of model performance and poor generalization. A prominent recent example are models that were trained to diagnose COVID-19 based on lung computed-tomography scans using datasets that combined images from different sources of COVID-19 positive and negative patients, respectively, sometimes using images of pediatric patients as controls [31]. In such settings, a model can learn to recognize the image source or patient age rather than true signs of COVID-19 [31, 32]. Even if the source of heterogeneity itself is not directly associated with the model output, it can make prediction tasks more challenging. For instance, the input-output relationship may differ between data subsets, requiring more complex model types and larger sample sizes for model training than in a homogeneous setting.

The heterogeneity of health data may also lead to distribution differences between the data used for model development and data on which the model is later applied. Even though a model learned a biologically meaningful relationship between input and output, this relationship might not remain valid, e.g., if the inputs are measured with a different device or if the model is applied to a distinct patient population. The distribution of health data may also change over time as health practices evolve. This problem is known as domain shift or dataset shift [33] and can be an obstacle to the validation and broad applicability of machine learning models. Dataset shift may also pose an ethical challenge, e.g., if model performance is not stable across patients of different ethnicities or genders [34, 35].

Another challenge in health data is its temporal aspect [36]. Participants in clinical trials or patients who undergo intense or long-term treatment are either continuously monitored or receive regular follow-up examinations. If a dataset contains longitudinal measurements, individual observations are no longer statistically independent, which needs to be taken into account during model development. Time may also be an additional source of variability and heterogeneity. For instance, there may be heterogeneity between groups of patients because they have different temporal trends or heterogeneity over time if a treatment (or recovery from it) has different phases.

In response to these challenges, machine learning models for health data need to be particularly robust. Models for data analysis need to be able to extract useful

information from data with high variability, even if the sample size for model training is modest, and should account for external sources of heterogeneity. Models trained for prediction tasks should ideally be robust to differences in data distributions across patient groups or institutions. And in both situations, models should be able to integrate longitudinal data despite these challenges.

1.2 Approaches to Robust Machine Learning

Depending on the context and the exact challenge at hand, there are different ways to approach robust machine learning.

In this thesis, I have focused on supervised learning, where the goal is to learn a relationship between an input X and an output Y that generalizes well to new data. Typically, X is a random vector and Y is a random variable, which follow some unknown joint probability distribution $P(X, Y)$, and we aim to estimate a function f such that $f(X)$ is a good approximation of Y , using a set of labeled training examples drawn from $P(X, Y)$. Although this is often done by minimizing the differences between predicted and observed output in the training set, the goal is that this difference will be small for new, independent samples drawn from $P(X, Y)$. Robustness in this context means that a model should still be able to achieve this goal even if it is confronted with less than ideal conditions.

If a small sample size and high variability of the training data are the main challenges, one approach to robust machine learning is to train models with low capacity. Models with low capacity are limited regarding which functional relationships f they can represent, which may introduce bias if the true relationship between X and Y is not part of this restricted class. However, they also have low variance with respect to changes in the training data, and this reduction in variance may outweigh the error introduced by a small bias [2]. A model with low capacity is, e.g., a parametric model with only a few free parameters. Such models make strong assumptions regarding the relationship between X and Y , but if these assumptions are well justified and guided by prior knowledge, they allow for accurate models based on small datasets [2, 37].

Alternatively, regularization can reduce the capacity and variance of a generally flexible model type. In high-dimensional settings, even a linear regression model

trained by ordinary least squares has many degrees of freedom and may have high variance. A standard regularization technique is to add a penalty term to the loss function to penalize large regression coefficients, typically some norm of the coefficient vector. For instance, ridge regression uses the L_2 norm to shrink coefficients towards zero [38], and the LASSO uses the L_1 norm to perform feature selection [39]. The elastic net combines the advantages of both methods by using a convex combination of L_1 and L_2 norm [40]. It produces a sparse coefficient vector like the LASSO but handles correlated features similar to ridge regression and is still applicable if the number of features exceeds the number of samples.

Overall, a broad range of models and methods are available to reduce variance and enable generalization, and the best choice is application-specific. Some models include hyperparameters that directly control their capacity [41, 42] or consist of an ensemble of multiple predictors that are averaged to reduce variance [43, 44]. Even deep neural networks, which are at the high-capacity end of the scale, benefit from regularization [45]. Here, techniques range from classical penalty-based methods such as weight decay [46] to neural-network-specific methods like dropout [47] or parameter sharing, e.g., in convolutional neural networks [48]. However, in the context of deep learning, robust machine learning refers mainly to the robustness against security risks, such as adversarial attacks [49] and data poisoning [50], and has become a research field of its own [51, 52]. Robust deep learning and its security aspects are not the focus of this thesis.

A more difficult challenge to address is heterogeneity between data used for model training and data observed during model deployment. Here, domain adaptation aims to explicitly design models that are robust to dataset shift [53, 54]. Domain adaptation is a branch of transfer learning and considers the same prediction task in two domains, a *source* domain and a *target* domain, with related but different underlying distributions, $P_S(X, Y)$ and $P_T(X, Y)$. The goal is to predict well on data following the target domain distribution $P_T(X, Y)$, while training mostly on data drawn from $P_S(X, Y)$. This is only possible if the source and target domain distributions are sufficiently similar to allow for some transfer of knowledge [53].

Domain adaptation approaches can be categorized based on how much information from the target domain is available for model training. In supervised domain adaptation, a small number of labeled examples from the target domain are available [55], which allows adjusting the model parameters to optimize target domain

performance directly. In the more challenging setting of unsupervised domain adaptation, only unlabeled examples from the target domain are available at training time [56]. Blind domain adaptation goes even further and aims to perform the same task without using any target domain data [57]. Unsupervised and blind domain adaptation have no way of estimating target domain performance directly and require assumptions on the similarity and differences between $P_S(X, Y)$ and $P_T(X, Y)$. A common choice is, e.g., the *covariate shift* assumption, where the domains differ only in the marginal distributions $P_S(X)$ and $P_T(X)$, while the conditional distributions $P_S(Y | X)$ and $P_T(Y | X)$ are the same [58]. It is worth noting that in the context of domain adaptation, the terms *supervised* and *unsupervised* refer only to the domain adaptation setting and not to the machine learning task. All variants of domain adaptation use labeled source domain data and aim to perform the supervised task of predicting Y based on X in the target domain.

Domain adaptation has many potential applications in the health domain. It can improve the transferability of models between medical centers and patient collectives, and may even allow transferring knowledge between related diseases when little data is available on the scenario of interest. Especially methods for unsupervised domain adaptation are attractive since they do not require expensive labeling for every new target domain. Unsupervised domain adaptation has been studied extensively for deep neural networks [59, 60, 61] and for classification tasks like sentiment classification in natural language processing [62, 63] or object recognition from digital images [64, 65, 66]. Yet, fewer methods exist for regression tasks and model types with a lower capacity, and these focus predominantly on supervised domain adaptation [67, 68]. Overall, applications of unsupervised domain adaptation to health-related questions are still rare.

Another way to approach robust machine learning is to consider the uncertainty of a model and its predictions. Under challenging circumstances, an accurate prediction may not always be possible, and in such cases, a robust model should indicate that it is uncertain. Probabilistic machine learning aims to estimate the conditional distribution $P(Y | X)$ instead of making hard predictions [69]. In classification, this means predicting class probabilities that accurately reflect observed frequencies in the real world, which is more challenging than requiring only that the correct class obtained the highest score [70]. Estimating and communicating uncertainty is especially relevant in the health domain, where physicians or patients receive and

interpret model output, and wrong decisions may have severe consequences [71]. For instance, a physician using a decision support system must be able to judge the confidence of a prediction to decide whether to follow or overrule it. Similarly, if an early-warning system predicts some critical event, a physician needs to know its probability in order to judge the risk to the patient and decide how to react.

Interpretable or explainable models may additionally allow physicians to judge the plausibility of the model output based on expert knowledge. Models may either be interpretable directly, if they have a simple and human-understandable structure, or explanations may be generated post-hoc by analyzing or approximating an otherwise non-interpretable model [72]. Such explanations can range from global descriptions of feature importance [43, 73] to local explanations for individual predictions [74, 75, 76]. Although forcing models for health applications to be interpretable or explainable has also been discussed critically, especially if it requires further approximating the original predictions [77], accurate explanations can help to reveal artifacts in the data and may allow physicians to act as a safeguard against implausible predictions [71, 72]. Thus, both probabilistic and interpretable machine learning can be seen as extending the robustness of a model to its interactions with humans.

Ensuring model robustness is not only a challenge for method development, but also an essential responsibility of scientists developing models. Addressing challenges such as confounders or heterogeneity within a dataset requires careful analysis and algorithms tailored to the specific application. If confounding factors are known, it may be possible to account for them during model training or to remove them through normalization or curation of the dataset. Yet, to discover unexpected confounders or heterogeneity, scientists need to inspect trained models, compare model performance in data subsets and discuss the interpretation of their results with experts in the application domain. Ultimately, the robustness of a model can only be judged once it has been challenged. Verifying model robustness requires validation in a realistic application scenario, ideally on independent data from a different source and in a prospective setting to account for potential dataset shift over time. While it is difficult to meet all these requirements in the limited scope of a research project, doing so is crucial for the successful application of any model in practice.

2 Objectives

The overarching goal of my studies was to develop robust machine learning methods and models that are tailored to characteristic challenges in health data. My work towards this goal can be divided into three topics spanning three application areas in health research.

As a first topic, I aimed to develop new methods for unsupervised domain adaptation that are suitable for high-dimensional health data. Although this goal is methodological in nature, it was motivated by the problem of age prediction from DNA methylation data across different tissues. Here, previous work had shown poor prediction performance of standard models on data from a tissue that was not included in the training set [78]. A common challenge associated with molecular data like DNA methylation is a large number of correlated input features that exceeds the number of samples. To address both challenges in combination, I aimed to develop an unsupervised domain adaptation method for the elastic net [40], a popular robust model type that is well suited for high-dimensional input data.

The second topic involved the model-based analysis of data from current clinical HIV research. Here, my goal was to contribute robust models for the analysis of data from an early clinical trial studying the safety and efficacy of a combination of two broadly neutralizing antibodies, 3BNC117 and 10-1074, in the treatment of HIV-1. The main objective of my analyses was to compare outcome measures between the current trial participants and multiple patient groups from previous clinical trials who received only one or neither of the two broadly neutralizing antibodies. In particular, I aimed to detect heterogeneity between these patient groups and to account for relevant differences during analysis. Other methodological challenges were a high variability combined with a small number of participants and longitudinal data.

The third topic was embedded in the consortium project *XplOit: Semantic Support for Predictive Modeling in Systems Medicine* and focused on allogeneic

hematopoietic stem cell transplantation (HCT) as the area of application. Here, my goal was to train and evaluate robust models for the prediction of critical events after HCT based on routine clinical data. More precisely, I aimed to combine recipient and donor baseline data with longitudinal laboratory measurements to predict patient-specific risks of death and cytomegalovirus (CMV) reactivation at multiple time points after HCT. The main challenges were a high variability and heterogeneity over time as the distribution and frequency of laboratory measurements changed considerably over time after HCT. As a secondary aspect, I aimed to contribute to the collective goal of the consortium to develop a software platform that supports the provision and harmonization of health data as well as the development and validation of prediction models based thereon. Here, the models I developed served as a use case and first application of the platform to obtain feedback on the user experience from the perspective of a model developer.

3 Results and Discussion

In this chapter, I will first provide an overview and discussion of my work separately for each topic and area of application. I will summarize the key ideas and results of the manuscripts on each topic, emphasizing how they relate to each other and to the overarching goal of robust machine learning for health applications. Since each manuscript has multiple authors, I will switch between the use of *we* and *I* depending on whether I discuss our work as a whole or my individual contributions and views. A detailed contributions statement for each manuscript is included in the frontmatter of this thesis.

In Chapter 4, I will subsequently provide a joint perspective and connection points between these three topics in an integrated discussion.

3.1 Methods for unsupervised domain adaptation

In Manuscripts 1 and 2, we developed methods for unsupervised domain adaptation, which are tailored to challenges in molecular health data. Our methods are designed for regression tasks with high-dimensional input and modest sample size, and are adaptations of an elastic net model with improved robustness to heterogeneity between datasets.

In unsupervised domain adaptation, labeled source domain data and unlabeled target domain data are available for model training, yet the aim is to predict the output well in the target domain. If the input is high-dimensional, the heterogeneity between source and target domain may be caused by only some of the features while other features behave similar in both domains. The shared core idea of the methods we developed is to identify features which behave similar, and to train an elastic net model on source domain data which relies mostly on these robust features. To identify how reliable features are, we compare the dependencies between features in source and target domain. This idea is based on the assumption that if a feature

has the same dependencies with other features in both domains, it will likely also have the same relationship with the output we aim to predict. A model using these features should therefore be able to robustly predict the output across domains.

In Manuscript 1, we first explored this idea by performing feature selection for each target domain sample individually. This approach is close to blind domain adaptation since the only target domain information used for each prediction is the input for which the prediction is made. In Manuscript 2, we built on these early experiments to develop the method *wenda* (weighted elastic net for unsupervised domain adaptation). *Wenda* combines information from multiple target domain inputs to estimate the reliability of each feature more robustly, and prioritizes features instead of performing strict feature selection.

To evaluate these methods on real-world data, we applied them to the problem of age prediction from DNA methylation data across different tissues. Although age is typically measured chronologically, it describes biological changes, which do not necessarily progress at the same rate for every individual or every cell [79, 80]. In search of an indicator of biological age, researchers have aimed to predict age from molecular data, and elastic net models based on epigenetic DNA methylation proved to be particularly accurate [80, 81, 82]. These models were trained to predict a tissue donor’s chronological age, yet their predictions can be interpreted as a biological *epigenetic age* of the donor or tissue. An acceleration of epigenetic aging in comparison to chronological aging is associated with a higher risk for age-related diseases [83, 84] and a higher all-cause mortality [85], indicating that epigenetic age could be a more informative descriptor of a person’s health status than chronological age. However, DNA methylation patterns and age-associated changes within them are, at least in part, tissue-specific [86, 87, 88]. Applying an age prediction model on other tissues than it was trained on is therefore a domain adaptation problem. Methods which robustly predict age across tissues can enable age prediction on tissues for which labeled training data is scarce or unavailable, but can also serve as an example for future prediction tasks based on tissue-specific molecular data.

In our experiments, we used DNA methylation data from TCGA [89] and GEO [90] with an intentionally mismatched tissue composition in training and test set. The focus of our evaluation was on cerebellum samples, which were part of the test set but not represented in the training data. This tissue is known to be biologically

distinct even from other brain tissues [87, 91] and its age was poorly predicted with the model type of a popular age prediction model based on a standard elastic net [80], which we used as a reference model.

3.1.1 Partially blind domain adaptation for age prediction from DNA methylation data (Manuscript 1)

In Manuscript 1, we explored the idea to strictly select only the most reliable features for each prediction. Using a model-based approach, we estimated a confidence score for each feature of a target domain sample, measuring how well this feature matches the input dependencies observed in the source domain. We then trained an elastic net model to predict the output for this specific sample, using only those features with the highest confidence scores.

To estimate confidence scores, we first trained Gaussian process models to capture the conditional distribution of each feature given all other features in the source domain. Building on previous work by Jalali and Pfeifer [92], we then defined the confidence score of each feature based on how likely its observed value is according to the distribution predicted by the Gaussian process models.

This method can be seen as partially blind domain adaptation. It uses some target domain information for training since it trains a separate elastic net for each target domain input using only the selected high-confidence features. However, just like in blind domain adaptation, each prediction is only based on the target domain input for which the prediction is made, in addition to labeled source domain data.

In our experiments, we used a fixed percentage to define high-confidence features, and varied it between the top 10% and the top 40%. Applied to age prediction from DNA methylation data, our method reduced the prediction error for cerebellum samples compared to the non-adaptive reference model. This reduction was stronger if the feature set was narrowed down more severely. However, there was a clear trade-off between the error on cerebellum samples and the error on the full test set. While the restriction to a narrow set of high-confidence features reduced the error on samples with a distribution mismatch compared to the source domain data, it also removed features which would have been useful for prediction on samples with no such distribution mismatch.

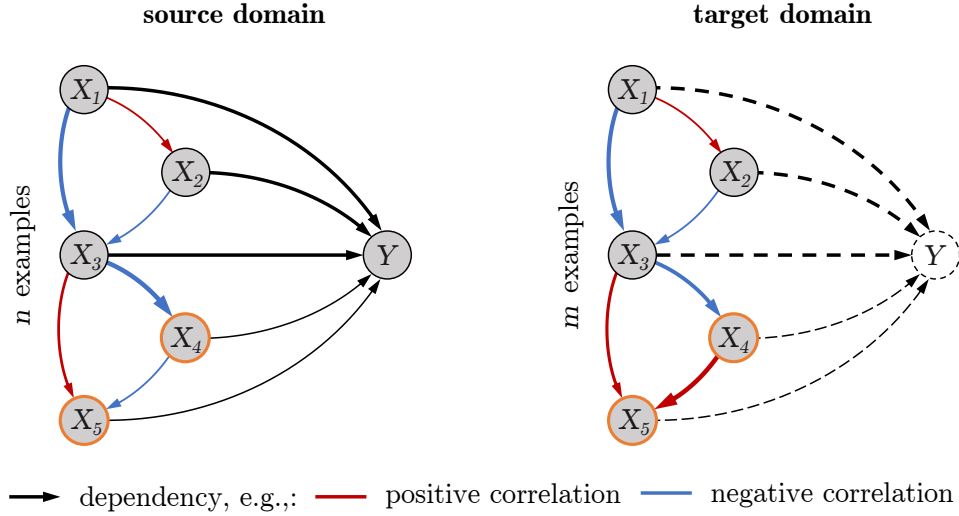


Figure 3.1: Schema illustrating the main idea of *wenda*. In this example, the aim is to predict Y based on five features X_1, \dots, X_5 in the target domain. While features X_1, X_2 and X_3 have the same dependency structure in source and target domain, the relationship between X_4 and X_5 (highlighted in orange) is different. To predict Y in the target domain, *wenda* trains an elastic net on the source domain examples while enforcing a stronger regularization on the coefficients of X_4 and X_5 , reducing their impact on the predictions. This method is based on the assumption that features with different input dependencies in source and target domain will likely also have a different relationship with Y in the two domains.

3.1.2 Weighted elastic net for unsupervised domain adaptation (Manuscript 2)

In Manuscript 2, we built on these early experiments to develop the unsupervised domain adaptation method *wenda*. Instead of strict feature selection, *wenda* prioritizes features by weighting their contributions to the elastic-net regularization penalty. It places larger weights on features with a low confidence score, which encourages the model to prefer high-confidence features without excluding any features completely. Figure 3.1 shows a schema illustrating this idea. In contrast to Manuscript 1, *wenda* averages confidence scores over the target domain inputs, estimating only one set of feature confidences for the whole target domain.

We introduced two versions of *wenda*, which take different approaches to selecting the regularization parameter of the elastic net: *wenda-pn* and *wenda-cv*. Since *wenda* couples domain adaptation and regularization, this parameter influences not only the overall strength of regularization, but also the degree to which high-

confidence features are prioritized. Its optimal value can therefore depend on the size of the distribution mismatch between domains. In *wenda-pn*, we proposed a way to use prior knowledge on the size of this mismatch to select the regularization parameter. For instance, in the application to age prediction across tissues, we utilized estimates of tissue similarity published by the GTEx consortium [91] as prior knowledge. Alternatively, *wenda-cv* uses cross-validation on the labeled source domain data to select it. The approach taken by *wenda-cv* determines an optimal value for the source domain rather than the target domain, which is not ideal for domain adaptation, but is still applicable if prior knowledge is not available.

To study the behavior of *wenda* in a controlled setting, we simulated data with a known distribution mismatch. We used directed acyclic graphs to represent dependencies between features, and modeled the distribution of each child node as a linear combination of its parent nodes with additive Gaussian noise. To model the distribution mismatch in the target domain, we selected a subset of features and altered their dependencies with other features as well as their effect on the output. On simulated data, *wenda-pn* led to considerable improvements in two out of three simulated scenarios, in which prior knowledge could be utilized more easily, while *wenda-cv* performed similar to or only slightly better than a standard elastic net. These results show that cross-validation does indeed not select the optimal regularization parameter for the target domain, and that prior knowledge can improve parameter selection.

Applied to age prediction from DNA methylation data, both *wenda-pn* and *wenda-cv* substantially reduced the error on cerebellum samples and performed similar to the non-adaptive reference model on the full test set. In contrast to our experiments with feature selection in Manuscript 1, which showed a trade-off between the performance on cerebellum samples and on the full test set, *wenda* performed well on all tissues, regardless of whether they had a distribution mismatch with the source domain data or not. Figure 3.2 shows a direct comparison of the performance of *wenda-pn* and the reference model on multiple test tissues.

3.1.3 Discussion

Summarizing Manuscripts 1 and 2, we have shown that feature selection and prioritization can be powerful tools for unsupervised domain adaptation in high-

3 Results and Discussion

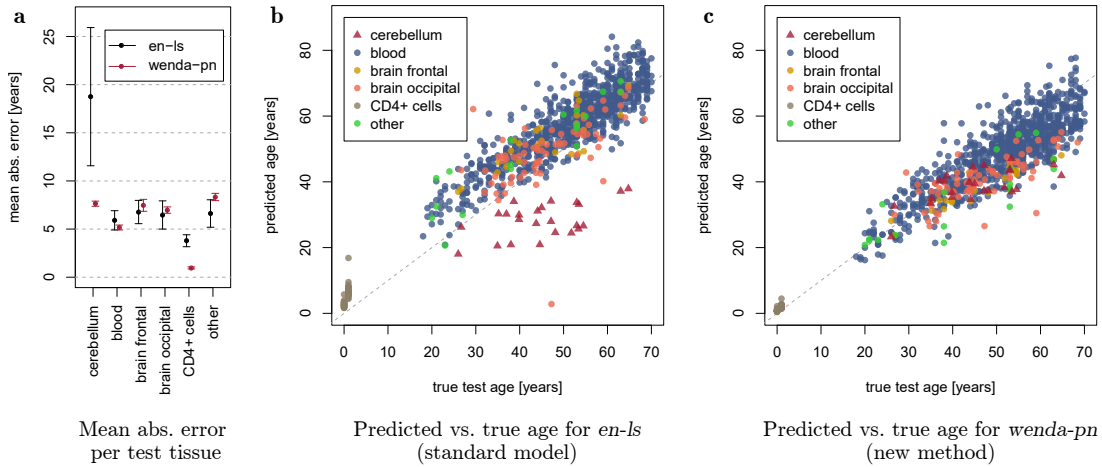


Figure 3.2: Prediction performance of *wenda-pn* on the DNA methylation dataset compared to a non-adaptive reference model. The reference model, *en-ls*, was a standard elastic net followed by a linear least-squares fit using only features that obtained non-zero coefficients in the elastic net. This model type had been used previously for age prediction from DNA methylation data [80]. **a**, Mean absolute error of *wenda-pn* and *en-ls* per test tissue, shown are the mean \pm standard deviation derived from cross-validation. **b–c**, Predicted versus true chronological age for typical runs of *en-ls* (**b**) and *wenda-pn* (**c**), samples are colored by tissue. **a–c**, Figure and caption adapted from Fig. 2 in Manuscript 2.

dimensional prediction tasks. Compared to a non-adaptive reference model, our methods both improved age prediction from DNA methylation data on samples with a distribution mismatch compared to the labeled training data. In Manuscript 1, where we used only information from a single target domain input for each prediction and performed strict feature selection, this improvement came at the cost of a larger error on test samples with no distribution mismatch. In contrast, *wenda* in Manuscript 2 performed well on test samples with and without distribution mismatch. This was achieved by combining information from multiple inputs from the same target domain and by using a feature weighting in the elastic-net regularization penalty instead of strict feature selection.

Both our methods are designed for the situation where the number of features exceeds the number of samples, which is common in molecular health data. They are complementary to previous methods performing unsupervised domain adaptation for regularized regression, which reweight samples rather than features [93, 94]. These were developed for the covariate shift case and while they have some proven theoretical guarantees in this scenario, sample reweighting may lead to high vari-

ance if the number of samples available for training is not large [56]. In addition, the covariate shift assumption states that the mismatch between source and target distribution arises only from sample selection bias and that the relationship between input and output is the same in both domains. In contrast, our methods allow some features to have a different impact on the output in source and target domain. While feature selection has previously been used successfully for supervised domain adaptation in an elastic net model [67], both the softer approach of using a feature weighting and the extension to the more challenging situation of unsupervised domain adaptation were new in *wenda* to the best of my knowledge.

By using feature selection or prioritization, our methods detect features which behave differently in source and target domain and reduce their impact on cross-domain prediction. Recent methods based on deep neural networks go even further and learn a new representation of the input data which is similar in both domains [59, 61, 95]. While this approach is more flexible and may avoid losing information by discarding some features, it also requires more unlabeled data from the target domain for training. One strength of *wenda* is that it requires only few unlabeled examples from the target domain. In simulations, as few as 100 target domain inputs were sufficient to learn suitable feature weights, and some target domains in the DNA methylation dataset were even smaller. Nevertheless, it would be an interesting and challenging next step to extend *wenda* to allow for correcting features with a distribution mismatch instead of penalizing them.

In *wenda*, regularization and domain adaptation are entangled. While this can make it difficult to control the strength of regularization and the level of domain adaptation independently, it allows for an implicit trade-off between how much information a feature contains on the output and how similar it behaves across domains. An alternative approach is transfer component analysis (TCA), which aims to extract a lower-dimensional feature representation that is transferable between domains [96]. This representation can subsequently be used in any standard machine learning model, e.g., in an elastic net, thus separating the two steps of identifying a representation suitable for domain adaptation and training a model to predict the output. However, TCA does not make use of the labels available for source domain data to guide the choice of the feature representation. As a consequence, features with a moderate domain mismatch but a strong influence on the output might be removed by TCA but retained by *wenda*.

Without labeled training data from the target domain, unsupervised domain adaptation requires assumptions on how the distributions in source and target domain differ. The core assumption of both our methods is that differences in the dependencies between inputs are informative of which features allow for a robust prediction of the output across domains. This assumption is motivated by the observation that different locations in the (epi)genome interact in a biologically meaningful way, and that the gene regulatory networks they form can differ between tissues [97]. Differences in these interactions could indicate parts of the DNA methylome which are related to tissue-specific biological processes and may therefore not share the same relationship with age across tissues. In Manuscripts 1 and 2, we only assessed this assumption indirectly, through changes in model performance, and found that it was appropriate to improve age prediction across tissues. It would be an interesting route for future work to analyze which features obtained particularly low confidences and whether this matches known functional differences between cell types.

Depending on the application scenario, other criteria could be suitable to define feature confidences. For instance, *wenda-mar* in Manuscript 2 explores the idea to measure differences in the marginal distributions of features rather than differences in the dependencies between them. Although this approach led to a smaller improvement than the dependency-based versions of *wenda* in our experiments, it may be useful for domain adaptation on datasets with no (informative) dependencies between features. An interesting next step would be to combine both approaches and to measure changes in marginal distributions and in dependencies simultaneously. Alternatively, extensions of *wenda* to classification tasks and to multi-omics settings could be promising directions for further research.

To capture the dependency structure between features, our methods model the conditional distribution of each feature given all other features. This step is computationally demanding, especially in a high-dimensional feature space. To accelerate model training, the initial implementation¹ of *wenda* made use of parallelization where possible, yet it could only utilize CPUs for model training. Ariel Hippen et al. later developed *wenda_gpu*, a *wenda* implementation based on GPyTorch that can utilize GPUs, reducing the training time and enabling applications to even

¹<https://github.com/PfeiferLabTue/wenda>

higher-dimensional settings [98].

An alternative approach to modeling all conditional distributions is to model the dependencies all at once by learning the joint distribution of all features. Here, structure learning for Bayesian networks could be applied to learn a network of dependencies between features [99, 100, 101]. Structure learning is NP hard and can only be solved approximately for large feature spaces. Nevertheless, a sparse Bayesian network would be a more concise representation of feature dependencies and could have additional advantages for *wenda*. If the relationship between two features differs in source and target domain, both features could receive low confidence in the current version of *wenda* and the confidence of other features which strongly depend on these two could also be influenced. A sparse dependency structure and directionality in feature relationships may help to reduce these effects and to pinpoint more clearly where the difference originated.

Finally, *wenda* could be applied to new prediction tasks based on heterogeneous molecular health data. An obvious example is age prediction from tissue-specific gene expression data such as transcriptomics [102] or proteomics [103], where similar heterogeneity effects and input dependencies as in DNA methylation data could be expected. Beyond age prediction, computational oncology offers further opportunities for applications of *wenda*. For instance, *wenda* would be promising to improve the prediction of cancer stage and prognosis across cancer types [104, 105, 106]. Hippen et al. also applied the GPU-accelerated version of *wenda* to identify loss-of-function mutations in a tumor suppressor gene across cancer types based on transcriptomic data [107]. Here, genetic differences between cancer types and subtypes and the genetic diversity of individual cancer cells are additional sources of heterogeneity. By improving prediction across cancer types, *wenda* could enable a more accurate prognosis prediction for rare cancers, where little or no labeled data is available. More generally, any prediction task across heterogeneous domains, where dependencies between features are informative of domain differences, are promising new application areas for *wenda*, especially if standard elastic net models previously proved successful within a domain.

3.2 Robust model-based analyses for HIV research

In Manuscripts 3 and 4, we developed robust models for the analysis of data from an early clinical trial evaluating the use of broadly-neutralizing antibodies (bNAbs) for the treatment of HIV-1.

To date, there is no treatment which can fully eradicate HIV-1 and patients require lifelong antiretroviral therapy (ART) to suppress the infection [108, 109]. While standard ART can effectively reduce the viral load below detectable levels, it can have side effects and needs to be taken every day. A promising candidate for a new class of therapies for HIV-1 are bNAbs, which neutralize a large number of virus variants by targeting conserved epitopes [110]. These are developed naturally by a small fraction of HIV-1 patients and can be isolated from their blood sera [111]. As a treatment for HIV-1, bNAbs could engage the patient's immune system and might require fewer doses than ART, owing to their long half-lives [112]. Since treatment with a single bNAb leads to the emergence of resistant HIV-1 variants, similar to treatment with a single antiretroviral agent, it is likely that a combination of multiple bNAbs would be needed to enable long-term control of HIV-1 [113, 114].

Both manuscripts report results of a phase 1b clinical trial (NCT02825797) studying the combination of two potent anti-HIV-1 bNAbs, 3BNC117 [115] and 10-1074 [116]. Although the main purpose of the trial was to evaluate the safety of this combination therapy in humans for the first time, its data also allowed to gain a first glimpse on treatment efficacy. The trial investigated the effect of combination therapy under two different initial conditions. In patients with suppressed HIV-1, who were previously on ART, it assessed if and for how long combination therapy could maintain viral suppression when ART was interrupted. In patients who were initially viremic and not on ART, it assessed if and for how long combination therapy could reduce the viral load. The trial participants were pre-screened for sensitivity to 3BNC117 and 10-1074 using an *in vitro* neutralization assay and received either one or three infusions of both bNAbs.

Data from early clinical trials is challenging to analyze because patient groups are small and control groups with well-matched characteristics are not always available. In our analyses, we compared current trial participants to patients from previous clinical trials who received only one of the two studied bNAbs or no intervention, and aimed to identify if there was a statistically significant difference between these

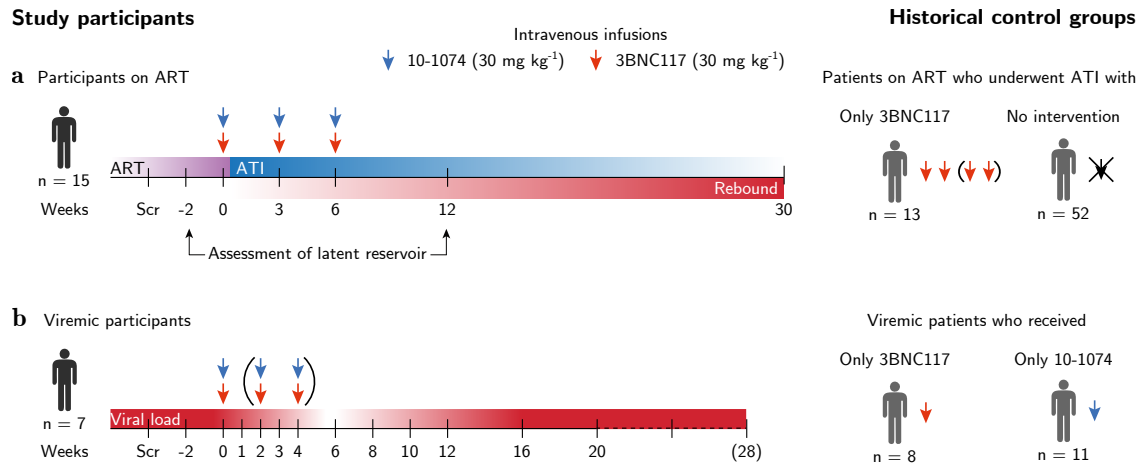


Figure 3.3: Schematic overview of the study design including current trial participants and historical control groups. **a**, Manuscript 3 reports on patients who were initially on antiretroviral therapy (ART) and received 3 infusions of 10-1074 and 3BNC117 while undergoing analytical treatment interruption (ATI). The time to viral rebound was compared to historical patients who underwent ATI and received only 3BNC117 [117] (2 doses 3 weeks apart or 4 doses at 2-week intervals) or no intervention [118]. **b**, Manuscript 4 reports on viremic patients off ART who received 1–3 infusions of combination therapy. The temporal development of the viral load was compared to viremic patients who received a single dose of monotherapy [113, 114]. **a–b**, Figure and caption adapted from Fig. 1 in Manuscript 3 and 4, respectively.

patient groups regarding the effect of the treatment on HIV-1. Doing so required robust models and statistical methods, which account for potential heterogeneity between patient groups and are appropriate for small sample sizes, depending on the precise application scenario. Figure 3.3 provides a schematic overview of the study design and control groups.

3.2.1 Combination therapy with anti-HIV-1 antibodies maintains viral suppression (Manuscript 3)

Manuscript 3 describes the results of the clinical trial for patients who were initially on ART and underwent analytical treatment interruption (ATI). Here, 15 patients each received three infusions of both 3BNC117 and 10-1074 at intervals of three weeks while ART was interrupted. The viral load was monitored regularly and ART was reinitiated at the time of viral rebound, which was defined as the first of two

subsequent viral load measurements detecting >200 copies per ml. We compared these patients to two control groups from previous clinical trials, who had received only 3BNC117 monotherapy [117] ($n = 13$) or no intervention [118] ($n = 52$) during ATI.

The main conclusion of Manuscript 3 was that combination therapy with 3BNC117 and 10-1074 can maintain viral suppression in patients sensitive to both bNAbs if ART is discontinued. Patients with complete viral suppression at the initiation of ATI had a median time to rebound of 21 weeks, compared to 6–10 weeks in the control group with 3BNC117 monotherapy and 2.3 weeks in the control group with no intervention during ATI. Two patients did not rebound in the entire follow-up period of 30 weeks. Figure 3.4a displays Kaplan-Meier plots of the time to viral rebound for current trial participants and both control groups. In patients with confirmed dual sensitivity to 10-1074 and 3BNC117, rebound only occurred after the serum concentration of at least one of the two bNAbs was very low and never earlier than 15 weeks after ATI initiation. Here, 10-1074 had a longer serum half-life than 3BNC117, leading to a period of effective 10-1074 monotherapy after 3BNC117 levels dropped. Although some of the patients consequently developed HIV-1 variants resistant to 10-1074, these variants remained sensitive to 3BNC117.

Building on previous work by Scheid *et al.* [117], we performed a statistical analysis to determine if the differences in time to rebound could be explained by any potential confounders. All statistical tests were performed at significance level $\alpha = 0.05$. We used parametric survival regression to describe the time-to-event data [37]. This model type assumes that the event density function, which describes the instantaneous rate at which events (in our case viral rebound) occur as a function of time, is the density of a parametric distribution and that covariates impact some parameter of this distribution via a linear function. Using the R package *flexsurv* [119], we fitted multiple parametric models for each available covariate (such as age, gender, or years on ART), comparing multiple parametric distributions (exponential, log-normal, Weibull, etc.) and selecting the best-fitting distribution for each covariate based on Akaike’s information criterion. We then compared each model including a covariate to a null model without covariates, using a likelihood ratio test to determine if the covariate had a significant impact on the time to viral rebound. In addition, we tested for differences in the distribution of relevant covariates between patient groups. The only confounder we identified

was years on ART, which had a significant impact on the time to rebound and was lower in the control group with no intervention than in current trial participants. All other available covariates did not differ significantly between patient groups, nor did they impact time to rebound.

We then tested whether the treatment group had a significant impact on the time to rebound while accounting for years on ART using two different methods, a likelihood ratio test including years on ART as an additional covariate and an adjusted log-rank test which reweights samples to account for differences between groups [120]. Both methods confirmed that the time to rebound differed significantly between patients treated with combination therapy and patients treated with monotherapy or no intervention, respectively, even when accounting for differences in the covariate years on ART. Although the differences in time to rebound were reported without statements of significance in Manuscript 3, this analysis corroborated their interpretation and the conclusion that combination therapy was more effective than monotherapy.

3.2.2 Safety and antiviral activity of combination HIV-1 broadly neutralizing antibodies in viremic individuals (Manuscript 4)

The results in viremic patients are detailed in Manuscript 4. For this part of the clinical trial, 7 viremic patients received either one infusion or three biweekly infusions of 3BNC117 and 10-1074. Afterwards, the viral load was monitored for 24 weeks. These patients were compared to a control group of 19 viremic patients who had received only 3BNC117 ($n = 8$) or only 10-1074 ($n = 11$) as monotherapy in previous clinical trials [113, 114].

Here, we performed a statistical analysis to compare the response to bNAb treatment between current trial participants and the control group. We modeled the development of the viral load over time using linear mixed-effects models, a popular model type that allows to account for dependencies between repeated measurements from the same subject [121]. More precisely, we included fixed effects for the treatment group and time, treating time as an ordered factor, and a subject-specific effect on the intercept to account for variability between individuals. The model output was defined as $Y_{ti} = \log_{10}(VL_{ti}) - \log_{10}(VL_{0i})$ for $t = 1, \dots, n_i$, where

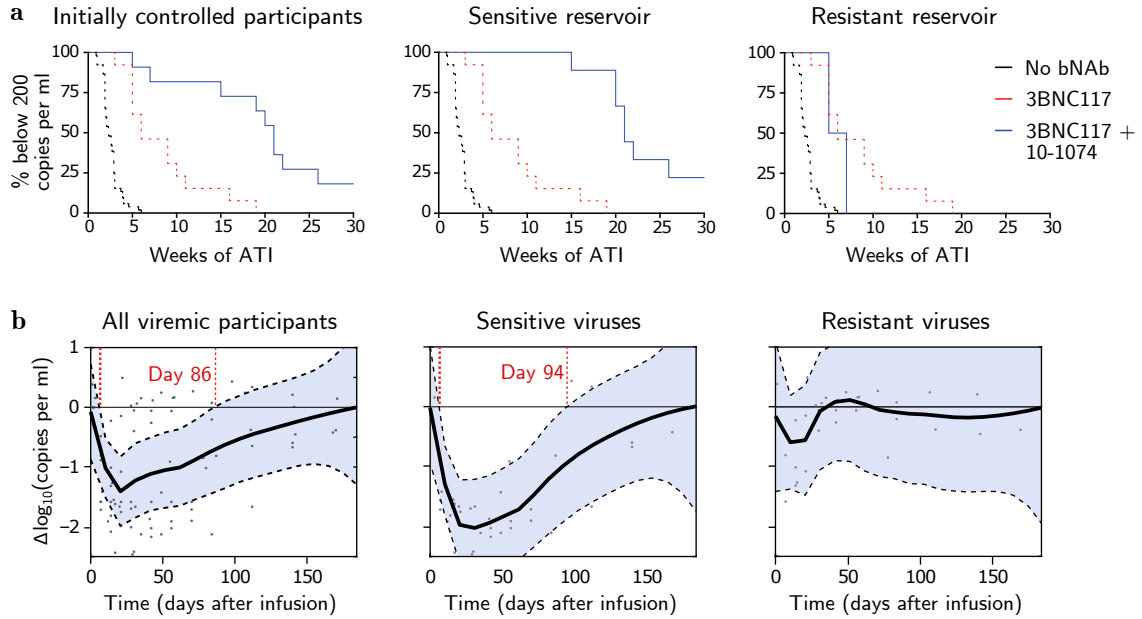


Figure 3.4: Plots summarizing the main results of the clinical trial. **a**, Kaplan–Meier plots depicting the time to viral rebound for the participants with <20 copies per ml two weeks before and at the start of ATI ($n = 11$, left), for the participants with confirmed sensitivity to both antibodies ($n = 9$, center), and for the participants that showed pre-existing resistance to one of the antibodies ($n = 2$, right). **b**, Simultaneous confidence bands for the development of the viral load in all viremic participants ($n = 7$, left), individuals harboring 3BNC117- and 10-1074-sensitive viruses ($n = 4$, center), and participants carrying viruses with partial or full bNAb resistance ($n = 3$, right). Each dot represents a viral load measurement. Solid and dashed lines represent the regression fit and simultaneous confidence bands at 95% certainty level, respectively. **a–b**, Figure and caption adapted from Fig. 1 in Manuscript 3 and Fig. 2 in Manuscript 4.

VL_{0i}, \dots, VL_{n_i} is the temporal sequence of viral load measurements of subject i . Thus, the model can be specified as follows.

$$\begin{pmatrix} Y_{1i} \\ \vdots \\ Y_{n_i i} \end{pmatrix} = X\beta + \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} u_i + \begin{pmatrix} \varepsilon_{1i} \\ \vdots \\ \varepsilon_{n_i i} \end{pmatrix} \quad (3.1)$$

$$u_i \sim N(0, \sigma_u^2) \quad (3.2)$$

$$\begin{pmatrix} \varepsilon_{1i} \\ \vdots \\ \varepsilon_{n_i i} \end{pmatrix} \sim N(\mathbf{0}, R_i) \quad (3.3)$$

Here, X and β are the design matrix and coefficient vector of a standard linear model including only the fixed effects, u_i is a random effect describing the subject-specific variation in the intercept, and R_i is the covariance matrix of the residual vector $(\varepsilon_{1i}, \dots, \varepsilon_{n_i i})^\top$. In contrast to a standard linear model, R_i is not a scaled identity matrix but has a structured parametric form that models the dependencies between repeated measurements of subject i . Measurements of different subjects as well as u_i and the residuals are assumed to be independent. We determined an appropriate correlation structure for R_i by evaluating multiple options and selecting the best model based on Akaike's information criterion, resulting in a first-order autoregressive correlation structure. Using a likelihood ratio test, we then compared this model to a model without the predictor treatment group. Our analysis confirmed a significant difference between combination therapy and monotherapy regarding how strongly and how long they reduced the viral load.

On average, combination therapy reduced the viral load by $1.65 \log_{10}$ copies per ml in viremic patients and it took 86 days for the viral load to return to its initial level. Figure 3.4b shows individual viral load measurements with estimated simultaneous confidence bands. Yet, the response varied considerably between individuals. Manuscript 4 includes a more detailed analysis of the sensitivity of each patient's circulating viruses to 3BNC117 and 10-1074 before and after combination therapy, based on single genome amplification (SGA). Although all patients were pre-screened for sensitivity, SGA revealed HIV-1 variants with resistance or reduced sensitivity to the studied bNABs in three patients. Consistent with this

finding, these patients showed a weaker or no response to combination therapy. Viral suppression below detectable levels was achieved in two of the four patients with confirmed dual sensitivity, who had the lowest initial viral loads, but not in the remaining participants. None of the four patients with confirmed dual sensitivity developed HIV-1 variants resistant to both studied bNAbs. We concluded that combination therapy with 3BNC117 and 10-1074 may be able to achieve and maintain viral suppression in sensitive patients with very low initial viral load, but can not be used to effectively treat viremic patients in general.

3.2.3 Discussion

Taken together, Manuscripts 3 and 4 demonstrate both the potential and the limitations of HIV-1 treatment with the combination of 3BNC117 and 10-1074. The long time to viral rebound in patients who were previously on ART suggests that this combination might be able to maintain longterm suppression of HIV-1 once it was established with ART. Replacing longterm ART with a bNAb treatment could benefit patients who experience severe side effects while on ART, and would allow for larger intervals between doses. On the other hand, the varying responses of viremic patients indicate that 3BNC117 and 10-1074 alone would not be sufficient to establish viral suppression. Achieving suppression of HIV-1 in viremic patients is harder than maintaining suppression after ART because the large number and diversity of circulating viruses in viremic patients facilitate the emergence of resistant HIV-1 variants. To potentially reach this second goal without ART, 3BNC117 and 10-1074 could potentially be combined with further bNAbs or drugs.

Our study is an important step towards bNAb treatments for HIV-1. While treatment with a single bNAb has consistently led to the development of resistant variants [113, 114], previous studies evaluating combinations of bNAbs during ATI were limited by the low potency of bNAbs available at the time and showed only a very short delay in viral rebound, if any [122, 123]. With 3BNC117 and 10-1074, we combined two newer-generation bNAbs with high potency, which target distinct epitopes, and demonstrated that longer control of HIV-1 using bNAbs is possible. In our study, no patient who was confirmed to be sensitive to both bNAbs developed dual resistance during the observation period, even among viremic patients. However, some pre-existing resistances were missed by the initial screening with an

in vitro neutralization assay and were only revealed by SGA. This highlights that more reliable methods for sensitivity screening would be needed to apply bNAbs in clinical practice safely and effectively.

From a methodological point of view, the main challenge of this project was to identify robust models and statistical tests which were appropriate to analyze the limited and heterogeneous data. For the survival analysis for Manuscript 3, we preferred parametric survival regression over the popular Cox proportional hazards model [124]. While the semi-parametric Cox model is generally more flexible than fully parametric models, it is restricted by the proportional hazards assumption and requires more samples. In contrast, parametric models make strong assumptions on the distribution of survival times but are highly efficient if these assumptions are justified [37]. To choose the most appropriate assumptions for our data, we compared multiple parametric models and selected the one with the best fit. In addition, we took special care to avoid confounding factors when drawing conclusions regarding the treatment effect. In Manuscript 4, we chose a linear mixed-effects model to analyze the repeated viral load measurements of only few subjects. Although this model is simplistic compared to true viral load dynamics, it can represent correlations between repeated measurements of the same subject and is suitable for small sample sizes. Since Manuscript 4 includes fewer patients than Manuscript 3, pooling of patients who received either one or three bNAb doses could not be avoided and some heterogeneity within groups remained. In both manuscripts, we relied on likelihood ratio tests to detect confounders or to verify our hypotheses regarding treatment effects. This test is well suited for the model types and hypotheses we considered and has a higher power than competing statistical tests [125].

It should be emphasized that our data and analyses are only a first glimpse at treatment efficacy. While we made every effort to use appropriate robust methods and to minimize the risk of confounding, we could only perform statistical tests for variables that were recorded in both the current and the previous clinical trials, and it is clear that not all external factors can be controlled in early clinical data. If combination therapy with 3BNC117 and 10-1074 is to advance to a novel treatment for HIV-1, larger-scale studies will have to confirm the observed effects and compare bNAb treatment to ART, using randomization and blinding where possible. Nevertheless, our results emphasize that combinations of potent bNAbs are highly promising for HIV treatment and that further clinical research is justified.

3.3 Robust models for transplantation medicine

In Manuscripts 5 and 6, we worked towards robust machine learning for allogeneic hematopoietic cell transplantation (HCT).

Allogeneic HCT is an effective and potentially curative treatment for hematological malignancies and other high-risk diseases [126, 127]. After conditioning therapy, typically with a combination of chemotherapy and total body irradiation, hematopoietic stem cells extracted from a healthy donor’s peripheral blood or bone marrow are injected into the patient’s blood stream. These transplanted stem cells can substitute the patient’s hematopoietic system and produce lymphocytes, which can eradicate remaining malignant cells [128, 129]. However, HCT comes with a high treatment-related mortality and can entail several severe complications. For instance, immunocompetent engrafted T lymphocytes may recognize antigens presented on healthy recipient cells, resulting in graft-versus-host disease (GVHD) [130]. Conversely, engraftment may be slow or unsuccessful and leave the patient prone to infections or reactivation of latent viruses such as cytomegalovirus (CMV) or Epstein-Barr virus [131]. To minimize complications, physicians need to weigh these risks and choose the right strategies for conditioning therapy, GVHD prophylaxis [132] and prophylactic or pre-emptive antiviral treatment [133].

In current clinical practice, HCT risk assessment is based on relatively simple risk categories and clinical scores like the Hematopoietic Cell Transplantation-specific Comorbidity Index (HCT-CI) [134] or the European Society for Blood and Marrow Transplantation (EBMT) risk score [135]. Such scores provide a risk assessment before HCT to guide the decisions of whether HCT is a suitable option and which conditioning regimen should be used. Machine learning models have been used to improve some of these scores or to predict more specific outcomes [136, 137, 138]. However, the generalization and performance of these models is not yet satisfactory [138, 139] and there remains a high need for more precise and robust models to predict outcomes after HCT.

This is in part owing to the limited availability and usability of detailed HCT data. International registries like the databases of the EBMT [140] or the Center for International Blood and Marrow Transplant Research (CIBMTR) [141] collect pre-HCT and outcome data from many HCT centers, but the information on each patient is limited to comparatively coarse categories. Individual HCT centers store

richer data on their patients, including the temporal course of laboratory values and virological tests, but here data formats are heterogeneous and differ between sites. In Europe, the integrated use of HCT data across multiple centers is further challenged by strict rules for the protection of personal health information [142].

Our work followed two different paths to meet these challenges. In Manuscript 5, we developed robust machine learning models for the accurate, time-dependent prediction of mortality and CMV reactivation after allogeneic HCT. The core idea of our models was to utilize longitudinal laboratory values and to update predictions whenever new measurements become available. In Manuscript 6, we contributed to the development of the XplOit platform, a software platform for the semantic integration of heterogeneous health data with a focus on HCT.

3.3.1 Time-dependent prediction of mortality and cytomegalovirus reactivation after allogeneic hematopoietic cell transplantation using machine learning (Manuscript 5)

The vast majority of existing models and scores for HCT-specific risk assessment rely solely on pre-HCT data as input and provide a prediction at a single point in time [134, 135, 136, 137, 143]. While these methods are valuable tools to guide the initial choice of the conditioning regimen and whether or not to perform HCT, they cannot adjust to new data becoming available after HCT. In Manuscript 5, we hypothesized that including time-dependent laboratory values and updating predictions whenever new data becomes available would allow for a more precise prediction of outcomes after HCT. Complementing previous approaches, such models could allow to adjust prophylactic and pre-emptive treatments to how individual risks develop after HCT.

We considered two endpoints, death and early CMV reactivation, and started from an extensive HCT dataset combining multiple data modalities of 1710 patients who received allogeneic HCT at the University Hospital Essen (UHE). The entire dataset contained several baseline characteristics of patient and donor, pre-HCT diagnosis, disease status and conditioning therapy, as well as unstructured medical letters and the results of routine laboratory tests and virological tests. Based on a combination of static pre-HCT data and time-dependent laboratory values, we aimed to predict at multiple time points after HCT whether these endpoints

would occur in a subsequent time window of 7 or 21 days. We performed standard data preprocessing but additionally included a time-dependent standardization of laboratory values to account for its heterogeneity over time. Building on previous work [17], we then trained a gradient-boosting machine (GBM) model for each task and calibrated the predicted probabilities as a postprocessing step. To evaluate the utility of time-dependent features, we compared these models to a baseline model that was trained for the same time-dependent prediction task but received only static features. Figure 3.5 depicts an overview of the entire process of data preparation, model development, and validation.

The final GBM models performed well on test data held out from our retrospective development cohort, which is shown in Figure 3.6 for prediction over a 21-day time window. Here, the GBM models predicting mortality and CMV reactivation achieved areas under the receiver operating characteristic of 0.92 and 0.83 and areas under the precision-recall curve of 0.58 and 0.62, respectively. For mortality prediction, time-dependent features proved highly valuable, indicating that this approach could improve current standards for HCT-specific risk assessment. Feature inspection using SHapley Additive exPlanations (SHAP values) [76] showed that laboratory values, mainly those related to inflammation or organ function, strongly influenced model predictions. In contrast, using time-dependent features led only to a modest improvement of CMV prediction. Here, feature inspection revealed that the CMV models relied mostly on static features like the patient’s CMV serostatus before HCT, in addition to the prediction day after HCT.

To assess the robustness of the developed models, we validated them in a prospective, non-interventional clinical trial (DRKS00026643) with 403 additional HCT patients from UHE. Overall, model validation was successful and the performance remained high on prospective data. While CMV prediction performance was unaltered, the performance of mortality prediction decreased slightly compared to retrospective data, more noticeably for the 7-day time window than for prediction over 21 days. This indicated a dataset shift over time, which was in part explained by changes in immunosuppression strategies in clinical practice. For 91 of the participants, we additionally performed a pilot comparison of the model predictions to the expectations of experienced HCT physicians. Here, we regularly asked multiple hematologists at UHE to prospectively estimate their patients’ overall performance and CMV status in 7 and 21 days, respectively. Except for 7-day mortality pre-

3.3 Robust models for transplantation medicine

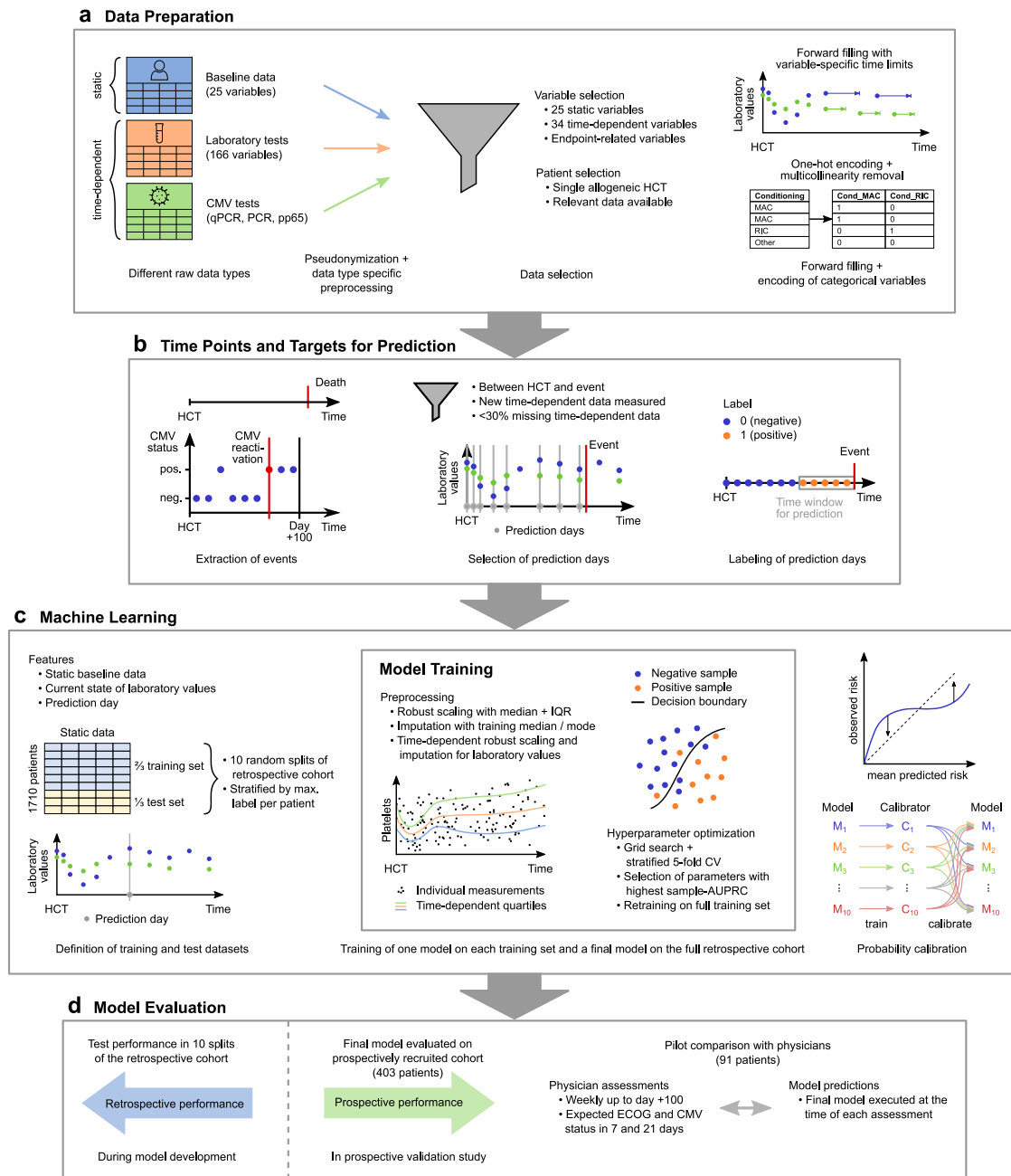


Figure 3.5: Overview of model development and evaluation. **a**, Data preparation, patient and variable selection. **b**, Time points and targets for prediction, selecting all days between HCT and an event or censoring with < 30% missing values as prediction days. **c**, Machine learning; models received static baseline data, current laboratory values and the prediction day after HCT as inputs. **d**, Model evaluation, using repeated splits into training and test data during model development and a prospective study including a pilot comparison with experienced HCT physicians for model validation. **a–d**, Figure and caption adapted from Figure 1 in Manuscript 5.

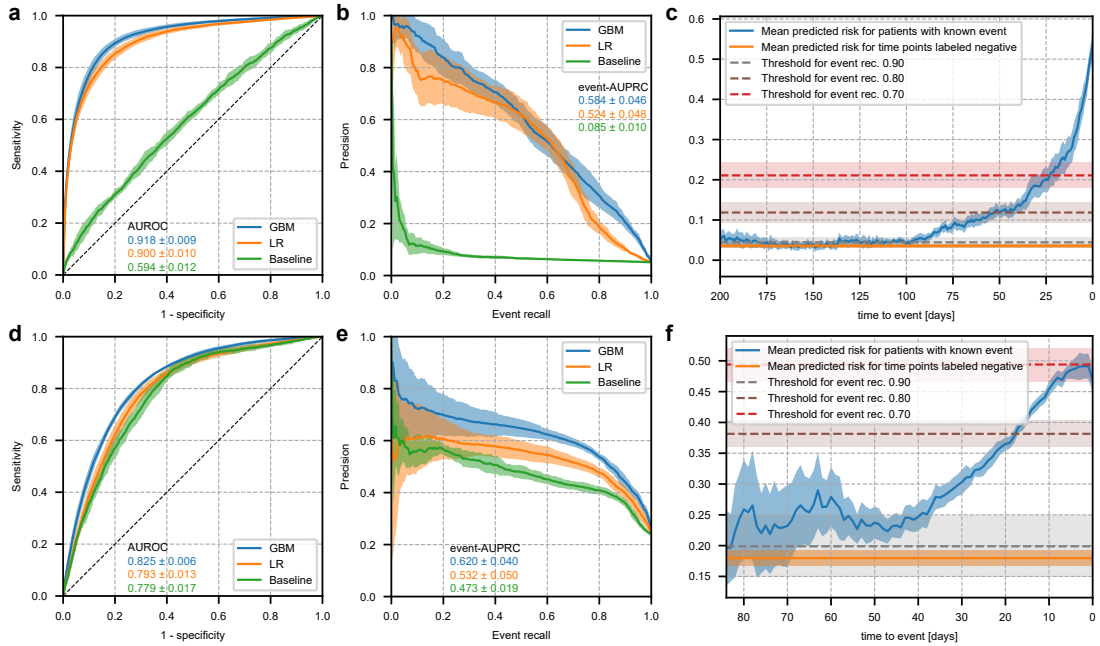


Figure 3.6: Performance of prediction of mortality (a–c) and CMV reactivation (d–f) in a 21-day time window on retrospective test data. Shown are the receiver-operating characteristic (a,d) and the precision-recall curve (b,e) of three models, respectively: a gradient-boosting machine (GBM) and regularized logistic regression (LR) model using time-dependent features, and a baseline LR model that received only static features. We used the definition of event recall from [17], i.e., the fraction of events that were predicted correctly on at least one of the preceding 21 days. c,f, Mean predicted risk of the GBM model as a function of time to event. Dashed horizontal lines indicate the thresholds required to achieve an event recall of 0.7, 0.8 and 0.9. a–f, Lines and shaded areas show the mean \pm standard deviation on the test set over 10 random splits into training and test data. Figure and caption adapted from Figures 2 and 3 in Manuscript 5.

diction, where only few time points with subsequent event were available for the comparison, all GBM models performed similar to the physicians. We concluded that time-dependent outcome prediction after HCT is possible with high accuracy and is more robust with the larger 21-day time window. An additional advantage of the larger time window is that it enables an earlier prediction of impending events.

3.3.2 XplOit: An ontology-based data integration platform supporting the development of predictive models for personalized medicine (Manuscript 6)

Manuscript 6 describes the XplOit platform, a software platform that supports various stages of model development for health applications and, in particular, for HCT. To date, a large proportion of a model developer's time is consumed by data preparation and integration tasks [144]. The key idea of the XplOit platform was to accelerate this process by providing efficient tools for the harmonization and detailed, semantic annotation of biomedical data. These annotations identify the precise meaning of each data point using a semantic ontology with clearly defined concepts. Thereby, they allow model developers to gain a deeper understanding of the data and to integrate datasets provided in different formats or by multiple institutions.

In the XplOit platform, each data point is annotated not with a single ontological concept but with an expressive path comprising multiple concepts and relationships between them. The semantic integration framework, a central platform component, provides semi-automatic tools for this task. It utilizes the Viral Disease Ontology Trunk (VDOT)², a modular domain ontology for biomedical data focusing on transplantation medicine and viral infections. To annotate a data point, it constructs and ranks all possible paths linking the patient to a matching concept in the ontology. Here, paths are preferred if they are similar to previously used annotations or if they contain concepts similar to the original label of the data point (e.g., the column name). VDOT reuses concepts from established ontologies and relates them with each other, which enables its flexible expansion if no matching concept is found.

²<https://bioportal.bioontology.org/ontologies/VDOT/>

The platform was designed to support heterogeneous data types and offers several additional functionalities. Data owners can upload pseudonymized raw data and manage data access via project communities. The platform allows to harmonize data formats using extract-transform-load (ETL) pipelines specific to each data type. After semantic annotation, all data is stored in a data warehouse as a graph of semantic triples. It can then be searched and visualized by model developers to assess correlations, data quality, or data differences between institutions.

3.3.3 Discussion

Manuscripts 5 and 6 advanced robust machine learning for transplantation medicine in two different ways. In Manuscript 5, we proposed a new model-based approach for HCT-specific risk assessment and demonstrated that longitudinal laboratory values allow for the accurate time-dependent prediction of HCT outcomes. In Manuscript 6, we introduced the XplOit platform, which facilitates the harmonization and integration of heterogeneous datasets from multiple HCT centers. Thus, it may accelerate the development of robust predictive models for HCT in the future. The work underlying both manuscripts was highly interconnected. While the XplOit platform simplified sharing and harmonizing the data we used to train the models in Manuscript 5, the experience gained from this first use case also enabled further improvements to the platform.

In both cases, our work goes beyond previously existing approaches. While similar time-dependent models have been developed to predict circulatory failure [17] or acute kidney injury [18] in patients requiring intensive care, this approach is new in HCT-specific risk assessment. Here, some existing methods utilize laboratory values [145, 143] or other longitudinal measurements [138], but these provide only a single risk assessment at a predefined point in time. Conversely, a recent web application offers personalized survival curves, i.e., a more detailed assessment than classical risk scores, but utilizes only static pre-HCT data as input [146]. In contrast, our models continuously monitor patient data and provide updated risk assessments, which are always based on the latest available information. The XplOit platform exceeds most existing data integration platforms [147, 148] in the expressivity of semantic annotations. While the p-medicine platform [149] is an exception that offers a similar level of detail, the semi-automatic tools provided by

the XplOit platform improve the usability of these expressive annotations.

In Manuscript 5, we considered the robustness of the developed models in several ways. For instance, we calibrated raw model predictions to ensure that the predicted risks agreed with observed event probabilities, capturing the uncertainty in model predictions. This is crucial in the medical application scenario, where physicians must be able to judge the relevance of predicted risks. In the prospective validation, we evaluated if the models can be applied robustly on new data from the same HCT center. Here, we discovered some dataset shift over time, highlighting the importance of monitoring model performance in medical applications. Occasional changes in clinical practice cannot be avoided and may erode model performance over time. To safely apply them in HCT care, models would have to be monitored continuously and retrained if necessary. That prediction performance on prospective data remained high overall is encouraging and indicates that the models may require retraining only at larger intervals. We additionally assessed the robustness of model performance. During model development, we analyzed multiple splits into training and test set to evaluate the sensitivity to changes in the data, and during model validation, we performed bootstrapping to quantify the uncertainty based on the limited prospective data.

We did not consider robustness across HCT centers in Manuscript 5 since we did not have access to comparable data from any center except UHE. Although Saarland University Medical Center (SUMC) also contributed to the XplOit platform, the provision of data from SUMC was delayed and complicated by technical obstacles. For instance, we did not receive access to the central information management system of SUMC to export HCT data automatically. While laboratory measurements and virological tests could be provided by individual departments, crucial pre-HCT data such as diagnosis or conditioning treatment would have had to be transcribed manually, a task which was not feasible in the limited time frame of the project. Consequently, we could use the data provided by SUMC to test data integration via the XplOit platform, but not to further validate the developed models.

Evaluating the robustness of the models across HCT centers would require a multi-center study, which could be conducted utilizing the XplOit platform. Although not described in Manuscript 6, we later extended the platform to allow for the execution of trained models on new data. Users can upload trained models

with a specification of the runtime environment and add semantic annotations for the model input and output parameters. These semantic annotations form a direct link between models and data, ensuring that the models can be executed robustly on any annotated dataset, irrespective of its origin and raw format. This makes the XplOit platform a helpful tool to support model validation and, potentially, model deployment.

An open question is how the risks predicted by our models would influence the decisions of HCT physicians and, ultimately, patient outcomes. Minimizing complications after HCT is a balancing act, where reducing one risk may increase another. For instance, reducing immunosuppression may counteract a CMV reactivation but increase the risk of GVHD [131]. Whether or not time-dependent risk prediction can improve overall survival will depend on the interplay between these complications, as well as graft failure and relapse. In addition, it will depend on the acceptance of the predictive models by HCT physicians. Recent studies have highlighted that medical experts tend to be skeptical of model-based recommendations, particularly if their decisions truly impact patients [150, 151]. Providing explainable predictions and efficiently integrating models into the clinical workflow may improve their acceptance [152, 153]. Ultimately, answering this question would require an interventional clinical trial, in which treatment with and without time-dependent risk prediction are compared directly.

The dependencies between HCT complications and relapse also pose a methodological challenge for model development. Currently, our models capture only part of these outcomes and are not interdependent. In the supplementary material of Manuscript 5, we evaluated whether including the result of the last CMV test or the diagnosis of post-HCT relapse as additional features improved survival prediction, which was not the case. However, CMV tests were encoded as binary categories, and the results may differ if quantitative viral load measurements or continuous predicted CMV risks are included as additional model features. Extending time-dependent risk prediction to other outcomes and capturing dependencies between them, ideally in more than one direction, are interesting steps for future work and may help physicians to consider multiple risks after HCT jointly.

Another open question and methodological challenge is how to make optimal use of heterogeneous, time-dependent laboratory measurements. After HCT, laboratory values are measured with varying frequency and follow a characteristic

nonlinear trend. In Manuscript 5, we performed a time-dependent normalization of laboratory values using smoothed window-based estimates of the median and quartiles of each laboratory value as a function of time. While our estimation method is pragmatic and allows for the adaptive selection of window sizes and bandwidths with varying measurement frequency, nonparametric quantile regression (e.g., using quantile regression forests [154]) may be a more elegant solution. In addition, our current models use only the most recent measurement of each laboratory value at the time of prediction. An obvious next step would be to include information on their history. We evaluated whether statistics computed in multiple time intervals before the prediction day (such as minimum, maximum, standard deviation, or slope of a least-squares regression line) could improve model performance. The results are included in the supplementary material of Manuscript 5 and showed no notable improvement. However, our approach was relatively simple, and more sophisticated methods to represent time-series data concisely despite missing values and varying measurement frequency may boost model performance.

Several challenges also remain for the XplOit platform. For instance, we currently represent the annotated data as a graph of semantic triples in the Resource Description Framework [155]. This data structure is well suited to represent knowledge in a machine-readable form and allows for automatic reasoning. However, searching it becomes inefficient when the dataset contains millions of individual measurements. Machine learning methods typically require tabular data, where the same parameters are available for many patients. Such data can be stored and searched efficiently in a relational database. A promising next step for the XplOit platform would be to combine the advantages of semantic annotations and relational databases, e.g., by representing only the meaning of columns in relation to a generic patient using semantic triples, instead of individual measurements.

To further accelerate model development, automatic tools to assess and improve data quality would be a valuable addition to the XplOit platform. In its current version, the platform already performs basic checks of the data types when raw data is loaded. For categorical parameters, it also allows annotating parameter values to define permitted values and their meaning. Yet, this proved difficult in practice in cases where the categories were obscure or inconsistent in the raw data. For numerical parameters, the semantic annotations could also be linked to a range of plausible values. Here, an unresolved challenge is the lack of clear definitions

of such ranges for clinical parameters. While laboratory tests, for instance, have a defined reference range of normal values in healthy individuals, there are typically no clear limits to distinguish abnormally high or low values from errors and artifacts. Building a database with expert-defined limits for this task would allow the well-founded automatic filtering of outliers and artifacts for machine learning applications.

4 Integrated discussion and conclusions

In my view, health applications remain one of the most promising and most challenging domains for machine learning. Making the most of available health data requires both generally robust machine learning methods and models that are carefully tailored to the task at hand. With the contributions I made throughout this thesis, I have advanced both aspects of robust machine learning for health applications. In Manuscripts 1 and 2, we developed robust methods that improve prediction performance across heterogeneous datasets and applied them to age prediction from DNA methylation data across tissues. In Manuscripts 3, 4, and 5, we developed robust models tailored to specific tasks either to ensure the appropriate interpretation of early clinical data on a potential novel treatment for HIV or to improve risk assessment after HCT. Here, heterogeneous and temporal data were challenges for both applications, and the HIV models additionally needed to be suitable for the limited sample size of early clinical trials. Finally, in Manuscript 6, we developed a software platform that facilitates the development and robust application of machine learning models across data formats.

There are several connection points across the application areas addressed in this thesis. For instance, Manuscripts 3, 4, and 5 all worked with temporal data, although in different ways. While the HIV model in Manuscript 4 modeled viral load measurements at multiple time points simultaneously, including correlation between them, the model for Manuscript 3 used classical parametric survival regression to fit right-censored time-to-event data. Both models were used for data analysis rather than prediction and made assumptions on the data distribution to cope with the small samples size. Manuscript 5 also used right-censored data, but instead of a statistical analysis of relevant factors, it aimed for the accurate time-dependent prediction of whether or not an event would occur in a subsequent time window.

In contrast to both HIV models, the large HCT dataset in Manuscript 5 allowed training flexible gradient boosting machines, which do not follow a predefined model shape.

Across all three application areas, we took the uncertainty of estimates into account, either to improve model robustness or to enable the robust interpretation of health data and model predictions based on it. For instance, in Manuscripts 1 and 2, we utilized uncertainty estimates to quantify differences between input dependencies in source and target domain. In Manuscripts 3 and 4, we performed statistical tests to distinguish meaningful differences from random fluctuations. And in Manuscript 5, we ensured calibrated risk predictions to include a measure of uncertainty in the predictions that would be communicated to physicians in a clinical application scenario.

The evaluation and potential application of the models for HCT-specific risk assessment in independent HCT centers would offer additional opportunities for connections between parts of my work. While the XplOit platform can provide technical support for an external validation by harmonizing data formats, it does not adjust distributional differences between centers, which might threaten model performance. If necessary, methods such as *wenda* could be applied to improve model generalization across HCT centers. We considered regression rather than classification in Manuscript 2, but the core idea of using a feature weighting in the regularization term could be applied directly to the regularized logistic regression models described in Manuscript 5. Although these had a lower performance than the more flexible gradient boosting machines, the benefit of domain adaptation may outweigh the cost of using a more restrictive model type on datasets with a distribution mismatch with respect to our training data. Since the HCT models use a combination of static and time-dependent features, defining feature weights based on the marginal distributions of features, as we did in *wenda-mar*, would be more straightforward than modeling all dependencies between them. An additional challenge would be the inclusion of categorical features in this approach.

Data heterogeneity, which is a common theme of all manuscripts in this thesis, is still a major obstacle to the successful and robust application of machine learning models on health data. Confounding factors within health datasets, heterogeneity between datasets, and dataset shift over time continue to cause models to fail prospective or multi-center validation [139, 156] and threaten their fairness [10].

Domain adaptation methods such as *wenda* have the potential to improve model generalization across heterogeneous datasets and over time. Yet, they may also be more challenging to apply in clinical practice since the prediction mechanism would change between target domains, which could affect both the trust into model predictions and their regulation as medical devices. In addition, domain adaptation does not address confounding factors and heterogeneity within the training dataset. Here, the best approach depends on the application scenario. In Manuscript 3 we used likelihood ratio tests and a sample weighting scheme to test and account for known potential confounders. On the larger dataset of Manuscript 5 we utilized SHAP values for model inspection and analyzed model performance in known heterogeneous subgroups. Generally, discovering unknown or unexpected sources of heterogeneity is more difficult than accounting for them once they are known. While inspecting the features that a trained model relies on can help to uncover artifacts and confounding factors, analyzing subgroups may reveal whether a model fits all parts of a heterogeneous dataset. If substantial heterogeneity between data subgroups is known or discovered, multi-task models could be used to learn separate models for subgroups while sharing some information.

Irrespective of how they were developed, machine learning models intended for use in clinical practice require extensive validation to assess their robustness, ideally on prospective data from independent medical centers. In Manuscript 5, we performed a prospective validation of our models for HCT-specific risk prediction, including a pilot comparison of model predictions to physicians' expectations. It showed that despite a minor dataset shift over time, the models remained applicable on prospective data from the same HCT center. Yet, as a single-center study, it could not assess the transferability of the models between centers. In current health research, there is a gap between many published machine learning models and clinical utility, partially due to a lack of thorough validation. The COVID-19 pandemic has especially highlighted this fact [157, 31]. Performing a large-scale model validation in an academic setting is challenging; it requires willing clinical partners, overcoming the legal challenge of obtaining access to protected health data from multiple sites and the technical challenge of integrating it. These steps are time-consuming and not always feasible in the limited funding period of a typical academic research project. Depending on the risks of a specific application, using a model in clinical practice may additionally require an interventional clini-

cal trial to evaluate its effect on patient outcomes and continuous monitoring and maintenance of model performance, entailing additional ethical and financial responsibilities. Advancing an existing model towards actual use in clinical practice may well be a research project on its own, which is no less demanding but potentially less attractive than developing new predictive models and comes with the risk of failure. The XplOit platform described in Manuscript 6 can support the technical part of this challenge by accelerating data integration. Nevertheless, additional incentives such as dedicated funding for model validation and clinical development or more scientific recognition of this task and of potential negative results may be needed to bridge the gap between model development and clinical application.

Making health data more accessible for machine learning use cases is perhaps the most direct way to promote robust machine learning for health. It includes simplifying data access while ensuring the protection of sensitive information and improving the usability of health data by developing standardized data structures and enhancing data quality. While the XplOit platform is a step towards this goal, a lesson learned during its development is that robust data management should start as early as possible, ideally as soon as the data is recorded in medical centers. If column names used within a hospital are unintuitive, no semantic annotation tool can recover its original meaning automatically. It is also easier to control data quality while collecting it than to find and correct mistakes in a dataset merged across institutions for large-scale machine learning applications. The German medical informatics initiative is currently addressing some of these issues in Germany by developing standards for storing and accessing health data across medical centers in a collaboration of computer scientists and university hospitals [158, 159]. Such efforts will reduce heterogeneity in data formats and allow to analyze larger multi-center datasets, reducing the risk of overfitting machine learning models. However, they can not resolve heterogeneity due to treatment protocols differing between medical centers or changing over time. Here, specialized methods for robust machine learning will remain vital and need to build on existing approaches for multi-task learning and domain adaptation.

In conclusion, unlocking the full potential of machine learning for health will require advancements on multiple levels. The standardization and accessibility of health data will need to improve to enable larger multi-center datasets for machine learning applications. More robust machine learning methods will be required to

learn generalizable patterns from this data and to reduce or avoid model deterioration over time. In addition, extensive model validation will need to assess the generalization of the developed models and their usefulness in clinical practice. And finally, successful models will need to be developed into medical products to impact patients' lives. Throughout this thesis, I have made contributions to some aspects of these advancements by developing robust machine learning methods for prediction across heterogeneous datasets, robust models tailored to specific tasks for HIV research and HCT-specific risk assessment, and contributing to a platform for semantically integrating health data. While much more remains to be done, I am convinced that the effort will ultimately be worth it.

Developing clinically useful models requires extensive interdisciplinary collaboration between machine learning scientists and health experts as well as genuine interest on both sides in achieving this goal together. Model developers need to understand the process their model aims to support, and health experts need to know the potential utility and limits of machine learning to guide model development. Close collaborations across fields as distinct as the computational sciences and medicine are demanding and intriguing at the same time. I have had the privilege to work on exciting interdisciplinary projects while preparing this thesis, and I hope to continue working at the intersection of machine learning and medicine in the future.

Bibliography

- [1] Hilbert, M. & López, P. The world's technological capacity to store, communicate, and compute information. *Science* **332**, 60–65 (2011).
- [2] Hastie, T., Tibshirani, R. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics (Springer, New York, 2017), 2nd edn.
- [3] Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-generation machine learning for biological networks. *Cell* **173**, 1581–1592 (2018).
- [4] Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
- [5] Carleo, G. *et al.* Machine learning and the physical sciences. *Reviews of Modern Physics* **91**, 045002 (2019).
- [6] Zheng, X., Zeng, Z., Chen, Z., Yu, Y. & Rong, C. Detecting spammers on social networks. *Neurocomputing* **159**, 27–34 (2015).
- [7] Stahlberg, F. Neural machine translation: A review. *Journal of Artificial Intelligence Research* **69**, 343–418 (2020).
- [8] Portugal, I., Alencar, P. & Cowan, D. The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications* **97**, 205–227 (2018).
- [9] Dressel, J. & Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* **4**, eaao5580 (2018).

- [10] Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
- [11] Floridi, L. *et al.* AI4People – an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines* **28**, 689–707 (2018).
- [12] Wright, M. C. *et al.* Toward designing information display to support critical care: A qualitative contextual evaluation and visioning effort. *Applied Clinical Informatics* **7**, 912–929 (2016).
- [13] Sutton, R. T. *et al.* An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine* **3**, 17 (2020).
- [14] Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
- [15] Arbabshirani, M. R. *et al.* Advanced machine learning in action: Identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *npj Digital Medicine* **1**, 9 (2018).
- [16] McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
- [17] Hyland, S. L. *et al.* Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine* **26**, 364–373 (2020).
- [18] Rank, N. *et al.* Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. *npj Digital Medicine* **3**, 139 (2020).
- [19] He, J. *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine* **25**, 30–36 (2019).
- [20] Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* **11**, 333–337 (2014).

- [21] Li, Q., Zhao, K., Bustamante, C. D., Ma, X. & Wong, W. H. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Genetics in Medicine* **21**, 2126–2134 (2019).
- [22] Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
- [23] Döring, M. *et al.* Geno2pheno[ngs-freq]: A genotypic interpretation system for identifying viral drug resistance using next-generation sequencing data. *Nucleic Acids Research* **46**, W271–W277 (2018).
- [24] Kouchaki, S. *et al.* Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics* **35**, 2276–2282 (2019).
- [25] Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* **18**, 463–477 (2019).
- [26] Zhang, P., Wang, F. & Hu, J. Towards drug repositioning: A unified computational framework for integrating multiple aspects of drug similarity and disease similarity. In *AMIA Annual Symposium Proceedings*, 1258–1267 (2014).
- [27] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
- [28] Sudlow, C. *et al.* UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine* **12**, e1001779 (2015).
- [29] Hutter, C. & Zenklusen, J. C. The Cancer Genome Atlas: Creating lasting value beyond its data. *Cell* **173**, 283–285 (2018).
- [30] The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
- [31] Roberts, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence* **3**, 199–217 (2021).

- [32] Maguolo, G. & Nanni, L. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *Information Fusion* **76**, 1–7 (2021).
- [33] Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N. D. (eds.) *Dataset Shift in Machine Learning*. Neural Information Processing Series (MIT Press, Cambridge, 2010).
- [34] Char, D. S., Shah, N. H. & Magnus, D. Implementing machine learning in health care – addressing ethical challenges. *The New England Journal of Medicine* **378**, 981–983 (2018).
- [35] McCradden, M. D., Joshi, S., Mazwi, M. & Anderson, J. A. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health* **2**, e221–e223 (2020).
- [36] Bellamy, D., Celi, L. & Beam, A. L. Evaluating progress on machine learning for longitudinal electronic healthcare data (2020). Preprint at <https://arxiv.org/pdf/2010.01149>.
- [37] Liu, X. *Survival Analysis: Models and Applications* (Wiley, Chichester, 2012).
- [38] Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
- [39] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**, 267–288 (1996).
- [40] Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320 (2005).
- [41] Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, 2006).
- [42] Schölkopf, B. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning Series (MIT Press, Cambridge, 2002).

- [43] Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).
- [44] Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**, 1189–1232 (2001).
- [45] Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, Cambridge, 2016).
- [46] Krogh, A. & Hertz, J. A. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems 4*, 950–957 (1991).
- [47] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014).
- [48] LeCun, Y. *et al.* Backpropagation applied to handwritten zip code recognition. *Neural Computation* **1**, 541–551 (1989).
- [49] Szegedy, C. *et al.* Intriguing properties of neural networks (2013). Preprint at <https://arxiv.org/pdf/1312.6199>.
- [50] Shafahi, A. *et al.* Poison frogs! Targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems 31*, 6103–6113 (2018).
- [51] Shafique, M. *et al.* Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead. *IEEE Design & Test* **37**, 30–57 (2020).
- [52] Qayyum, A., Qadir, J., Bilal, M. & Al-Fuqaha, A. Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering* **14**, 156–180 (2021).
- [53] Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* **22**, 1345–1359 (2010).
- [54] Patel, V. M., Gopalan, R., Li, R. & Chellappa, R. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine* **32**, 53–69 (2015).

- [55] Schweikert, G. B., Widmer, C., Schölkopf, B. & Rätsch, G. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *Advances in Neural Information Processing Systems 21*, 1433–1440 (2008).
- [56] Margolis, A. A literature review of domain adaptation with unlabeled data (2011). Technical Report, University of Washington.
- [57] Uzair, M. & Mian, A. Blind domain adaptation with augmented extreme learning machine features. *IEEE Transactions on Cybernetics* **47**, 651–660 (2016).
- [58] Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* **90**, 227–244 (2000).
- [59] Ganin, Y. *et al.* Domain-adversarial training of neural networks. *Journal of Machine Learning Research* **17**, 1–35 (2016).
- [60] Long, M., Zhu, H., Wang, J. & Jordan, M. I. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems 29*, 136–144 (2016).
- [61] Tzeng, E., Hoffman, J., Saenko, K. & Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7167–7176 (2017).
- [62] Blitzer, J., Dredze, M. & Pereira, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 440–447 (2007).
- [63] Glorot, X., Bordes, A. & Bengio, Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning*, 513–520 (2011).
- [64] Aljundi, R., Emonet, R., Muselet, D. & Sebban, M. Landmarks-based kernelized subspace alignment for unsupervised domain adaptation. In *Proceedings*

-
- of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, 56–63 (2015).
- [65] Gong, B., Shi, Y., Sha, F. & Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2066–2073 (2012).
- [66] Gong, B., Grauman, K. & Sha, F. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of the 30th International Conference on Machine Learning*, 222–230 (2013).
- [67] Li, Y., Vinzamuri, B. & Reddy, C. K. Constrained elastic net based knowledge transfer for healthcare information exchange. *Data Mining and Knowledge Discovery* **29**, 1094–1112 (2015).
- [68] Wachinger, C. & Reuter, M. Domain adaptation for Alzheimer’s disease diagnostics. *NeuroImage* **139**, 470–479 (2016).
- [69] Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series (MIT Press, Cambridge, 2012).
- [70] Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, 625–632 (2005).
- [71] Bruckert, S., Finzel, B. & Schmid, U. The next generation of medical decision support: A roadmap toward transparent expert companions. *Frontiers in Artificial Intelligence* **3**, 507973 (2020).
- [72] Markus, A. F., Kors, J. A. & Rijnbeek, P. R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics* **113**, 103655 (2021).
- [73] Fisher, A., Rudin, C. & Dominici, F. All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* **20**, 1–81 (2019).

- [74] Ribeiro, M. T., Singh, S. & Guestrin, C. “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144 (2016).
- [75] Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, 4765–4774 (2017).
- [76] Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56–67 (2020).
- [77] Babic, B., Gerke, S., Evgeniou, T. & Cohen, I. G. Beware explanations from AI in health care. *Science* **373**, 284–286 (2021).
- [78] Scherer, M. *Dissecting DNA Methylation in Human Aging*. Master thesis, Saarland University (2016).
- [79] Kirkwood, T. B. L. Understanding the odd science of aging. *Cell* **120**, 437–447 (2005).
- [80] Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology* **14**, R115 (2013).
- [81] Florath, I., Butterbach, K., Müller, H., Bewerunge-Hudler, M. & Brenner, H. Cross-sectional and longitudinal changes in DNA methylation with age: An epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Human Molecular Genetics* **23**, 1186–1201 (2014).
- [82] Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell* **49**, 359–367 (2013).
- [83] Levine, M. E. *et al.* DNA methylation age of blood predicts future onset of lung cancer in the women’s health initiative. *Aging* **7**, 690–700 (2015).
- [84] Levine, M. E., Lu, A. T., Bennett, D. A. & Horvath, S. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer’s disease related cognitive functioning. *Aging* **7**, 1198–1211 (2015).

- [85] Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biology* **16**, 25 (2015).
- [86] Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
- [87] Fraser, H. B., Khaitovich, P., Plotkin, J. B., Pääbo, S. & Eisen, M. B. Aging and gene expression in the primate brain. *PLOS Biology* **3**, e274 (2005).
- [88] Day, K. *et al.* Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biology* **14**, R102 (2013).
- [89] Chang, K. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **45**, 1113–1120 (2013).
- [90] Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**, 207–210 (2002).
- [91] Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
- [92] Jalali, A. & Pfeifer, N. Interpretable per case weighted ensemble method for cancer associations. *BMC Genomics* **17**, 501 (2016).
- [93] Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B. & Smola, A. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 19*, 601–608 (2007).
- [94] Cortes, C. & Mohri, M. Domain adaptation in regression. In *Proceedings of the 2011 International Conference on Algorithmic Learning Theory*, 308–323 (2011).
- [95] Kang, G., Jiang, L., Yang, Y. & Hauptmann, A. G. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4893–4902 (2019).

- [96] Pan, S. J., Tsang, I. W., Kwok, J. T. & Yang, Q. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* **22**, 199–210 (2011).
- [97] Thompson, D., Regev, A. & Roy, S. Comparative analysis of gene regulatory networks: From network reconstruction to evolution. *Annual Review of Cell and Developmental Biology* **31**, 399–428 (2015).
- [98] Hippen, A. A., Crawford, J., Gardner, J. R. & Greene, C. S. wenda_gpu: fast domain adaptation for genomic data. *Bioinformatics* **38**, 5129–5130 (2022).
- [99] Schmidt, M., Murphy, K., Fung, G. & Rosales, R. Structure learning in random fields for heart motion abnormality detection. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8 (2008).
- [100] Tsagris, M. Bayesian network learning with the PC algorithm: An improved and correct variation. *Applied Artificial Intelligence* **33**, 101–123 (2019).
- [101] Scanagatta, M., Salmerón, A. & Stella, F. A survey on Bayesian network structure learning from data. *Progress in Artificial Intelligence* **8**, 425–439 (2019).
- [102] Wang, F. *et al.* Improved human age prediction by using gene expression profiles from multiple tissues. *Frontiers in Genetics* **11**, 1025 (2020).
- [103] Moaddel, R. *et al.* Proteomics in aging research: A roadmap to clinical, translational research. *Aging Cell* **20**, e13325 (2021).
- [104] Das, J., Gayvert, K. M., Bunea, F., Wegkamp, M. H. & Yu, H. ENCAPP: Elastic-net-based prognosis prediction and biomarker discovery for human cancers. *BMC Genomics* **16**, 263 (2015).
- [105] Lin, Z. *et al.* Cancer progression prediction using gene interaction regularized elastic net. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics* *14*, 145–154 (2017).
- [106] Cheerla, A. & Gevaert, O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* **35**, i446–i454 (2019).

- [107] Hippen, A. A., Crawford, J., Gardner, J. R. & Greene, C. S. wenda_gpu: Fast domain adaptation for genomic data (2022). Preprint at <https://www.biorxiv.org/content/10.1101/2022.04.09.487671>.
- [108] Siliciano, J. D. *et al.* Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4⁺ T cells. *Nature Medicine* **9**, 727–728 (2003).
- [109] Finzi, D. *et al.* Latent infection of CD4⁺ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nature Medicine* **5**, 512–517 (1999).
- [110] Klein, F. *et al.* Antibodies in HIV-1 vaccine development and therapy. *Science* **341**, 1199–1204 (2013).
- [111] Simek, M. D. *et al.* Human immunodeficiency virus type 1 elite neutralizers: Individuals with broad and potent neutralizing activity identified by using a high-throughput neutralization assay together with an analytical selection algorithm. *Journal of Virology* **83**, 7337–7348 (2009).
- [112] Schoofs, T. *et al.* HIV-1 therapy with monoclonal antibody 3BNC117 elicits host immune responses against HIV-1. *Science* **352**, 997–1001 (2016).
- [113] Caskey, M. *et al.* Viraemia suppressed in HIV-1-infected humans by broadly neutralizing antibody 3BNC117. *Nature* **522**, 487–491 (2015).
- [114] Caskey, M. *et al.* Antibody 10-1074 suppresses viremia in HIV-1-infected individuals. *Nature Medicine* **23**, 185–191 (2017).
- [115] Scheid, J. F. *et al.* Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* **333**, 1633–1637 (2011).
- [116] Mouquet, H. *et al.* Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E3268–E3277 (2012).
- [117] Scheid, J. F. *et al.* HIV-1 antibody 3BNC117 suppresses viral rebound in humans during treatment interruption. *Nature* **535**, 556–560 (2016).

- [118] Li, J. Z. *et al.* The size of the expressed HIV reservoir predicts timing of viral rebound after treatment interruption. *AIDS* **30**, 343–353 (2016).
- [119] Jackson, C. H. flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software* **70**, 1–33 (2016).
- [120] Xie, J. & Liu, C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine* **24**, 3089–3110 (2005).
- [121] West, B. T., Welch, K. B. & Galecki, A. T. *Linear Mixed Models: A Practical Guide Using Statistical Software* (CRC Press, Boca Raton, 2007).
- [122] Trkola, A. *et al.* Delay of HIV-1 rebound after cessation of antiretroviral therapy through passive transfer of human neutralizing antibodies. *Nature Medicine* **11**, 615–622 (2005).
- [123] Mehandru, S. *et al.* Adjunctive passive immunotherapy in human immunodeficiency virus type 1-infected individuals treated with antiviral therapy during acute and early infection. *Journal of Virology* **81**, 11016–11031 (2007).
- [124] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **34**, 187–220 (1972).
- [125] Neyman, J. & Pearson, E. S. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A* **231**, 289–337 (1933).
- [126] Copelan, E. A. Hematopoietic stem-cell transplantation. *The New England Journal of Medicine* **354**, 1813–1826 (2006).
- [127] Gooley, T. A. *et al.* Reduced mortality after allogeneic hematopoietic-cell transplantation. *The New England Journal of Medicine* **363**, 2091–2101 (2010).
- [128] Weiden, P. L. *et al.* Antileukemic effect of graft-versus-host disease in human recipients of allogeneic-marrow grafts. *The New England Journal of Medicine* **300**, 1068–1073 (1979).

- [129] Bleakley, M. & Riddell, S. R. Molecules and mechanisms of the graft-versus-leukaemia effect. *Nature Reviews Cancer* **4**, 371–380 (2004).
- [130] Zeiser, R. & Blazar, B. R. Acute graft-versus-host disease – biologic process, prevention, and therapy. *The New England Journal of Medicine* **377**, 2167–2179 (2017).
- [131] Cho, S.-Y., Lee, D.-G. & Kim, H.-J. Cytomegalovirus infections after hematopoietic stem cell transplantation: Current status and future immunotherapy. *International Journal of Molecular Sciences* **20**, 2666 (2019).
- [132] Kröger, N. *et al.* Antilymphocyte globulin for prevention of chronic graft-versus-host disease. *The New England Journal of Medicine* **374**, 43–53 (2016).
- [133] Marty, F. M. *et al.* Letermovir prophylaxis for cytomegalovirus in hematopoietic-cell transplantation. *The New England Journal of Medicine* **377**, 2433–2444 (2017).
- [134] Sorrow, M. L. *et al.* Comorbidity and disease status based risk stratification of outcomes among patients with acute myeloid leukemia or myelodysplasia receiving allogeneic hematopoietic cell transplantation. *Journal of Clinical Oncology* **25**, 4246–4254 (2007).
- [135] Gratwohl, A. *et al.* Risk score for outcome after allogeneic hematopoietic stem cell transplantation: A retrospective analysis. *Cancer* **115**, 4715–4726 (2009).
- [136] Shouval, R. *et al.* Prediction of allogeneic hematopoietic stem-cell transplantation mortality 100 days after transplantation using a machine learning algorithm: A European Group for Blood and Marrow Transplantation Acute Leukemia Working Party retrospective data mining study. *Journal of Clinical Oncology* **33**, 3144–3152 (2015).
- [137] Arai, Y. *et al.* Using a machine learning algorithm to predict acute graft-versus-host disease following allogeneic transplantation. *Blood Advances* **3**, 3626–3634 (2019).

- [138] Tang, S. *et al.* Predicting acute graft-versus-host disease using machine learning and longitudinal vital sign data from electronic health records. *JCO Clinical Cancer Informatics* **4**, 128–135 (2020).
- [139] Buturovic, L. *et al.* Evaluation of a machine learning-based prognostic model for unrelated hematopoietic cell transplantation donor selection. *Biology of Blood and Marrow Transplantation* **24**, 1299–1306 (2018).
- [140] Gratwohl, A., Mohty, M. & Apperley, J. The EBMT: History, present, and future. In Carreras, E., Dufour, C., Mohty, M. & Kröger, N. (eds.) *The EBMT Handbook*, 11–17 (Springer, Cham, 2019).
- [141] Phelan, R., Arora, M. & Chen, M. Current use and outcome of hematopoietic stem cell transplantation: CIBMTR US summary slides (2020).
- [142] Council of European Union. Regulation (EU) no 2016/679 (2016).
- [143] Luft, T. *et al.* EASIX and mortality after allogeneic stem cell transplantation. *Bone Marrow Transplantation* **55**, 553–561 (2020).
- [144] Anaconda, Inc. The state of data science 2020: Moving from hype toward maturity (2020). <https://www.anaconda.com/state-of-data-science-2020> (accessed on 28.12.2021).
- [145] Lu, C.-C. *et al.* A BLSTM with attention network for predicting acute myeloid leukemia patient’s prognosis using comprehensive clinical parameters. In *Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2455–2458 (2019).
- [146] Okamura, H. *et al.* Interactive web application for plotting personalized prognosis prediction curves in allogeneic hematopoietic cell transplantation using machine learning. *Transplantation* **105**, 1090–1096 (2021).
- [147] Murphy, S. N. *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association* **17**, 124–130 (2010).

-
- [148] Scheufele, E. *et al.* tranSMART: An open source knowledge management and high content data analytics platform. *AMIA Joint Summits on Translational Science Proceedings* **2014**, 96–101 (2014).
- [149] Marés, J. *et al.* p-medicine: A medical informatics platform for integrated large scale heterogeneous patient data. *AMIA Annual Symposium Proceedings* **2014**, 872–881 (2014).
- [150] McIntosh, C. *et al.* Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nature Medicine* **27**, 999–1005 (2021).
- [151] Gaube, S. *et al.* Do as AI say: Susceptibility in deployment of clinical decision-aids. *npj Digital Medicine* **4**, 31 (2021).
- [152] Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **9**, e1312 (2019).
- [153] Diprose, W. K. *et al.* Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association* **27**, 592–600 (2020).
- [154] Meinshausen, N. Quantile regression forests. *Journal of Machine Learning Research* **7**, 983–999 (2006).
- [155] Cyganiak, R., Wood, D. & Lanthaler, M. RDF 1.1 concepts and abstract syntax: W3C recommendation (2014). <http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/> (accessed on 16.01.2022).
- [156] Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine* **15**, e1002683 (2018).
- [157] Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *The BMJ* **369**, m1328 (2020).
- [158] Semler, S. C., Wissing, F. & Heyder, R. German medical informatics initiative. *Methods of Information in Medicine* **57**, e50–e56 (2018).

Bibliography

- [159] Bild, R. *et al.* Towards a comprehensive and interoperable representation of consent-based data usage permissions in the German medical informatics initiative. *BMC Medical Informatics and Decision Making* **20**, 103 (2020).

Appendix A

Publications

A.1 Manuscript 1

Title:

Partially blind domain adaptation for age prediction from DNA methylation data

Authors:

Lisa Handl, Adrin Jalali, Michael Scherer & Nico Pfeifer

Published in:

arXiv:1612.06650 [q-bio.QM]

<https://arxiv.org/abs/1612.06650>

Presented as a poster on the workshop “Machine Learning for Health” at the Thirtieth Conference on Neural Information Processing Systems (NeurIPS), which took place December 5–10, 2016 in Barcelona, Spain.

Time of Publication:

December 2016

License information:

The publisher arXiv.org was granted a perpetual, non-exclusive license to distribute the article (<https://arxiv.org/licenses/nonexclusive-distrib/1.0/license.html>). As one of the authors, I retain the right to reuse it as part of this thesis.

Partially blind domain adaptation for age prediction from DNA methylation data

Lisa Handl

Max Planck Institute for Informatics
Saarland Informatics Campus
66123 Saarbrücken
lisa.handl@mpi-inf.mpg.de

Adrin Jalali

Max Planck Institute for Informatics
Saarland Informatics Campus
66123 Saarbrücken
ajalali@mpi-inf.mpg.de

Michael Scherer

Max Planck Institute for Informatics
Saarland Informatics Campus
66123 Saarbrücken
mscherer@mpi-inf.mpg.de

Nico Pfeifer

Max Planck Institute for Informatics
Saarland Informatics Campus
66123 Saarbrücken
nico.pfeifer@mpi-inf.mpg.de

Abstract

Over the last years, huge resources of biological and medical data have become available for research. This data offers great chances for machine learning applications in health care, e.g. for precision medicine, but is also challenging to analyze. Typical challenges include a large number of possibly correlated features and heterogeneity in the data. One flourishing field of biological research in which this is relevant is epigenetics. Here, especially large amounts of DNA methylation data have emerged. This epigenetic mark has been used to predict a donor's "epigenetic age" and increased epigenetic aging has been linked to lifestyle and disease history. In this paper we propose an adaptive model which performs feature selection for each test sample individually based on the distribution of the input data. The method can be seen as partially blind domain adaptation. We apply the model to the problem of age prediction based on DNA methylation data from a variety of tissues, and compare it to a standard model, which does not take heterogeneity into account. The standard approach has particularly bad performance on one tissue type on which we show substantial improvement with our new adaptive approach even though no samples of that tissue were part of the training data.

1 Introduction

Epigenetics, the heritable modification of phenotypes that is not encoded by DNA, has become an important field in biological research. The best-studied epigenetic mark is DNA methylation, which was detected to play a role in long-term repression of genes through promoter methylation, X-chromosomal inactivation and genomic imprinting [1]. It refers to the covalent addition of methyl groups to the C5 position of cytosines, predominately found in CpG dinucleotides. Due to the growing number of datasets in this field, a connection between the methylation pattern of genomic DNA and its donor's chronological age was reported [2, 3, 4]. On this basis, several studies created models to predict chronological age from DNA methylation data [5, 6, 7]. They defined the outcome of the prediction as the "epigenetic age" of the person and linked increased epigenetic aging to lifestyle factors and disease history. As a concept of biological age, the epigenetic age is more informative about the individual's health status than chronological age and can be useful to optimize disease treatment.

Due to the large number of sometimes strongly correlated features, DNA methylation data at the CpG level is challenging to model. Ordinary least squares regression leads to predictors with large variance because a large positive coefficient of one variable can be compensated by a large negative coefficient of a correlated variable. One way to prevent this is to use feature selection, e.g., by penalizing the L_1 norm of the coefficient vector in the loss function (LASSO). This type of regularization will set many coefficients to zero, leading to sparse and more robust models. An alternative approach is ridge regression, which penalizes the L_2 norm instead. Ridge regression forces coefficients to be small, but does not strictly set them to zero. In the presence of correlated features, ridge regression averages the coefficients while LASSO tends to pick one of the correlated variables. The elastic net penalizes a linear combination of the L_1 and L_2 norm of the coefficients and has been proposed to combine the advantages of LASSO and ridge regression [8]. It still performs feature selection, but tends to average the coefficients of included correlated features in a similar way as ridge regression.

Another difficulty, which is present in many biological and medical datasets, is the heterogeneity of the data. Small differences in data acquisition and processing (e.g., different protocols in laboratories or standards in clinics) may lead to biases and make it hard to compare data from different sources. Domain adaptation attempts to correct for mismatches between distributions in scenarios where large amounts of data from a source domain and small amounts of data from a target domain are available [9]. An even harder problem is blind domain adaptation, where data from the target domain is not available at training time [10].

In this paper, we present an approach which performs feature selection for each test sample individually to reduce effects of data heterogeneity. We build on ideas from [11] to find features that behave similarly in training and test data, but do not use a predefined set of weak learners. Instead, we train a full model for each test sample. Since the models are still trained only on the training data, but information from the test samples is used to select appropriate features, our setting can be seen as partially blind domain adaptation. We apply the method to the problem of age prediction based on a large DNA methylation dataset. The main source of heterogeneity in this data comes from the use of different tissues, some of which are not present in our training data. We show that our approach leads to improved test errors for samples from the cerebellum of the human brain, which is the tissue in our data that leads to the largest errors with standard models that do not account for the bias.

2 Methods

The core idea of our approach is to train test sample-specific models, considering only features in which we have high confidence for the test sample at hand. In a large heterogeneous dataset, it is possible that only some features cause the heterogeneity while others behave similarly in training and test data. Obviously, features that behave very differently should not be used in a predictive model. Excluding them and relying only on similarly behaving features can thus lead to a more robust model.

This can be expressed more formally in the framework of domain adaptation. Assume that the training and test samples are drawn independently from two joint probability distributions $P_S(X, Y) = P_S(Y | X) \cdot P_S(X)$ and $P_T(X, Y) = P_T(Y | X) \cdot P_T(X)$, respectively. Here S stands for source domain and T for target domain. A classical assumption in domain adaptation is that the conditional distributions, $P_S(Y | X) = P_T(Y | X)$, are the same in source and target domain while the distributions of input features may be different, i.e., $P_S(X) \neq P_T(X)$. This setting is called the covariate shift case. We weaken the covariate shift assumption by requiring equal conditional distributions only for part of the available features. More precisely, we assume that there is a subset $M \subset \{1, \dots, m\}$ of all features on which the same model can accurately predict the outcome from training and test inputs. This means that $P_S(Y | X_M) = P_T(Y | X_M)$, where X_M denotes the subvector of the random vector X containing only features in the reduced feature set M . The distribution of input features as well as the relationship between Y and the remaining features may be different in source and target domain, i.e., $P_S(X) \neq P_T(X)$ and $P_S(Y | X_N) \neq P_T(Y | X_N)$ for $N = \{1, \dots, m\} \setminus M$. In addition, we allow that M , the set of features that behave similarly in predicting Y , may be different for different test samples. Thus, a good choice of M has to be determined for each test sample separately.

For this purpose, we propose a model-based approach to estimate a confidence of each feature for a given test sample. We then train a full model for each test sample, learning from the training data

and using only high-confidence features. Since we do not know the response variable Y for the test samples, we explore the dependency structure within X to determine confidences. The underlying assumption is that if there is a subset of features, M , whose dependency structure is very similar in training and test data, then the relationship between Y and these features will also be similar in training and test data. More formally, writing X_f for the value of feature f and X_{-f} for the values of all other features, we assume that if $P_S(X_f|X_{-f}) \approx P_T(X_f|X_{-f})$ holds for all features $f \in M$, then $P_S(Y|X_M) \approx P_T(Y|X_M)$.

Model types We apply two main model types in this paper: elastic net and Gaussian process models. The elastic net is a form of regularized linear regression, which penalizes a combination of the L_1 and L_2 norm of the coefficient vector [8]. More precisely, it finds

$$\hat{\beta} = \arg \min_{\beta} \left(\frac{1}{2n} \|y - \mathbf{X}\beta\|_2^2 + \lambda \left(\alpha \|\beta\|_1 + \frac{1-\alpha}{2} \|\beta\|_2^2 \right) \right),$$

where \mathbf{X}, y is the training data and n is the number of samples that it contains. While $\alpha \in [0, 1]$ determines the mixing ratio of L_1 and L_2 penalty and is often set to a fixed value, $\lambda \geq 0$ controls the strength of regularization and is usually determined using cross-validation. Gaussian process models are a type of non-parametric Bayesian regression, where the prior distribution over regression functions is a Gaussian process with mean zero and a covariance function which is typically specified in the form of a kernel [12]. Bayesian models have the advantage that they provide not only a predicted value, but a distribution of possible output values for any new input. In the setting applied in this paper this distribution is Gaussian and known explicitly.

Datasets We collected 26 datasets from the Gene Expression Omnibus (GEO, ncbi.nlm.nih.gov/geo) and the Cancer Genome Atlas (TCGA, cancergenome.nih.gov), which analyzed DNA methylation by the Illumina Infinium HumanMethylation450 BeadChip. Then, we combined these datasets using RnBeads [13] and split it into a training and test set consisting of 1866 and 1007 samples, respectively. All samples included were obtained only from healthy tissues. The training set contains 16 and the test set 6 different tissues, with a focus on blood samples for both sets. For the training set, samples from donors with chronological ages between 0 and 103 years were used. The age range for the test set is 0-70 years, accordingly. SNP-removal, removal of gonosomal CpGs and data normalization with the BMIQ method [14] were performed by RnBeads. We reduced the initial number of features from 466,094 to 12,980 features using an elastic net model with strong regularization ($\lambda = 1.1 \cdot 10^{-4}$). This is necessary for computational reasons since we train a very large number of models.

Reference model We used a similar type of model as baseline as presented in [5], namely, an elastic net model with $\alpha = 0.8$, followed by least squares linear regression based on the selected features. This model has been trained on our training dataset and the regularization parameter λ has been selected via 10-fold cross-validation.

Adaptive model To estimate confidences of the features of test samples, we first trained a Gaussian process model for each feature, based on all other features. We chose a linear kernel and additive Gaussian noise, and determined the kernel parameter and noise variance of each model using marginal likelihood maximization. For a given test sample, X_i , these models can be used to predict a posterior distribution of $X_{i,f}$ (the value of X_i for some feature f), given the values of all other features, which we denote by $X_{i,-f}$. In our setting, we obtain a Gaussian posterior distribution, $N(\mu_{g_f}(X_{i,-f}), \sigma_{g_f}^2(X_{i,-f}))$. By comparing the observed value, $X_{i,f}$, to the predicted distribution, we can quantify how well $X_{i,f}$ fits to what is expected according to the training data. We quantify the confidence of feature f for X_i as proposed in [11] by

$$c_f(X_i) = 2 \cdot \Phi \left(- \left| \frac{X_{i,f} - \mu_{g_f}(X_{i,-f})}{\sigma_{g_f}(X_{i,-f})} \right| \right), \quad (1)$$

where Φ denotes the cumulative distribution function of the standard normal distribution. This can be interpreted as the probability that a value like $X_{i,f}$ or more extreme occurs according to its predicted distribution. After estimating confidences for all test samples and features, we use this information to train an age predictor for each test sample individually, based on only its high-confidence features. Here we used the same model type as for the reference model described in the previous paragraph,

Table 1: Mean and median absolute test errors of the reference model for the full test dataset and for cerebellum (CRBM) samples.

Type of test error		Test error
Full test dataset	mean	4.82
	median	3.45
CRBM samples	mean	16.95
	median	16.57

Table 2: Mean and median absolute test errors of the adaptive model for the full test dataset and for cerebellum (CRBM) samples.

Type of test error		Percentage of high-confidence features			
		Top 10%	Top 20%	Top 30%	Top 40%
Full test dataset	mean	7.96	6.61	6.16	5.78
	median	6.82	5.69	4.87	4.30
CRBM samples	mean	12.78	12.96	13.36	14.11
	median	10.19	12.63	13.78	14.94

but only 3-fold cross-validation. We tried multiple thresholds for defining high-confidence features, choosing the top 10%, 20%, 30% or 40% for each test sample. Note that the confidence estimation (and feature selection) is specific to the test sample, but each model is trained on the same training data. Moreover, no information on the output of test samples is used.

The adaptive model is computationally expensive since it involves fitting a large number of models. If m is the number of features and k is the number of test samples, then $m + k$ models are fitted in total. However, each of the main steps (i.e., fitting m models for confidence estimation and fitting k final models) can easily be parallelized to speed up computations.

3 Results and discussion

Reference model We trained the reference model on the training dataset with 12,980 features. The optimal regularization parameter determined by cross-validation is $\lambda = 0.01$, which corresponds to 436 features with nonzero coefficients. Table 1 shows the mean and median absolute test errors for the full test dataset and for cerebellum samples separately. We obtained a mean absolute error of 4.82 on the full test dataset. Given the wide range of ages and tissues considered, an error of this size seems reasonable. For cerebellum samples, however, we obtained a mean absolute error of 16.95, which is more than three times larger. This is not surprising as cerebellum samples are not present in our training data, but much larger than desirable. Both for the full test dataset and for cerebellum samples, the median absolute error is slightly lower than the mean.

Adaptive model In addition, we trained the adaptive model described in Section 2 for different thresholds defining high-confidence features. The resulting mean and median absolute test errors are presented in Table 2. For cerebellum samples, each of the adaptive models gave lower errors than the reference model. The performance on cerebellum samples is best when only features with the top 10% of confidences are used, leading to a mean absolute error of 12.78 and an even lower median of 10.19. When increasing the threshold, the errors on cerebellum samples slowly become larger, but still stay well below the corresponding errors of the reference model. These results demonstrate that restricting the model to high-confidence features can reduce the error on samples for which a distribution mismatch with the training data is present. A stronger restriction, which corresponds to a stronger focus on high confidences, leads to a larger improvement. At the same time, the errors on the full test dataset are larger for the adaptive models than for the reference model. Here we observe the opposite development. Errors decrease continuously with increasing threshold, from 7.96 for a threshold of 10% to 5.78 for a threshold of 40% in the case of mean absolute error. This can be explained by the fact that if all features behave the same way for training and test data, selecting only the “best” of them will not lead to an improvement. Thus, if no distribution mismatch is present,

restricting the model to far less features than the reference model is expected to lead to increased errors. Despite this, all errors on the full test dataset are still below the errors on cerebellum samples.

4 Conclusions and outlook

Heterogeneous data is ubiquitous in applications of machine learning in biology and medicine. In this paper we analyzed a large dataset of DNA methylation, which is heterogeneous because it was derived from multiple tissues. We proposed an adaptive model for predicting the donor's chronological age from this data. For each test sample the model selects features according to which the test sample behaves in a similar way as the training data. Then, it uses only these reliable features for prediction. Our model performs better than a non-adaptive reference model on samples from the cerebellum of the human brain. This tissue was not represented in the training data and led to the largest errors in the reference model. Thus, we demonstrated that our approach to partially blind domain adaptation can be a powerful way to reduce test errors on samples that are different from the training data. This improvement has a price when applying the model to test samples with the same or a very similar distribution as the training data. The main reason is that strictly excluding features restricts the model, which is not beneficial if no distribution mismatch is present. Of course, these findings need to be verified on additional datasets.

One possibility for improvement of the proposed model might be to weight features according to their confidences instead of including or excluding them strictly. This might improve the performance on samples without a distribution mismatch and will be subject of future work.

Acknowledgments

This work was prepared within the project *Xploit* of the initiative "i:DSem – Integrative Datensemantik in der Systemmedizin", which is funded by the German Federal Ministry of Education and Research (BMBF).

References

- [1] D. Schübeler, Function and information content of DNA methylation, *Nature* 517 (2015) 321–326.
- [2] H. Heyn, N. Li, H. Ferreira, J. Wang, M. Esteller, Distinct DNA methylomes of newborns and centenarians, *PNAS* 109 (2012) 10522–10527.
- [3] J. Bell, P.-C. Tsai, T.-P. Yang, R. Pidsley, P. Deloukas, Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population, *PLoS Genetics* 8 (2012) e1002629.
- [4] A. Teschendorff, J. West, S. Beck, Age-associated epigenetic drift: implications, and a case of epigenetic thrift?, *Human Molecular Genetics* 15 (2013) R7–R15.
- [5] S. Horvath, DNA methylation age of human tissues and cell types, *Genome Biology* 14 (2013) 3156.
- [6] G. Hannum, J. Guinney, L. Zhao, L. Zhang, K. Zhang, Genome-wide methylation profiles reveal quantitative views of human aging rates, *Molecular Cell* 49 (2013) 359–367.
- [7] I. Florath, K. Butterbach, H. Müller, M. Bewerunge-Hudler, H. Brenner, Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites, *Human Molecular Genetics* 23 (2013) 1186–1201.
- [8] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2005) 301–320.
- [9] G. Schweikert, G. Rätsch, C. Widmer, B. Schölkopf, An empirical analysis of domain adaptation algorithms for genomic sequence analysis, in: *Advances in Neural Information Processing Systems*, 2009, pp. 1433–1440.
- [10] M. Uzair, A. Mian, Blind domain adaptation with augmented extreme learning machine features, *IEEE Transactions on Cybernetics* PP (2016) 1–10.
- [11] A. Jalali, N. Pfeifer, Interpretable per case weighted ensemble method for cancer associations, *BMC Genomics* 17 (2016) 501.

- [12] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, Cambridge, 2006.
- [13] Y. Assenov, F. Müller, P. Lutsik, J. Walter, T. Lengauer, C. Bock, Comprehensive analysis of DNA methylation data with RnBeads, *Nature Methods* 11 (2014) 1138–1140.
- [14] A. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, S. Beck, A beta-mixture quantile normalization method for correcting probe bias in Illumina Infinium 450k DNA methylation data, *Bioinformatics* 29 (2013) 189–196.

A.2 Manuscript 2

Title:

Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data

Authors:

Lisa Handl, Adrin Jalali, Michael Scherer, Ralf Eggeling & Nico Pfeifer

Published in:

Bioinformatics, Volume 35, Pages i154–i163

<https://doi.org/10.1093/bioinformatics/btz338>

Presented as a proceedings talk in the section “Machine Learning for Computational and Systems Biology” at ISMB/ECCB 2019, which took place July 21–25, 2019 in Basel, Switzerland.

Time of Publication:

July 2019

License information:

This article was published under a Creative Commons CC-BY-NC license (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial reuse and reproduction in any medium, provided the original work is properly cited.

Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data

Lisa Handl^{1,2,3,*}, Adrin Jalali¹, Michael Scherer¹, Ralf Eggeling^{2,3} and Nico Pfeifer^{1,2,3,*}

¹Department for Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, 66123 Saarbrücken, Germany and ²Department of Computer Science and ³Institute for Biomedical Informatics, University of Tübingen, 72076 Tübingen, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: Predictive models are a powerful tool for solving complex problems in computational biology. They are typically designed to predict or classify data coming from the same unknown distribution as the training data. In many real-world settings, however, uncontrolled biological or technical factors can lead to a distribution mismatch between datasets acquired at different times, causing model performance to deteriorate on new data. A common additional obstacle in computational biology is scarce data with many more features than samples. To address these problems, we propose a method for unsupervised domain adaptation that is based on a weighted elastic net. The key idea of our approach is to compare dependencies between inputs in training and test data and to increase the cost of differently behaving features in the elastic net regularization term. In doing so, we encourage the model to assign a higher importance to features that are robust and behave similarly across domains.

Results: We evaluate our method both on simulated data with varying degrees of distribution mismatch and on real data, considering the problem of age prediction based on DNA methylation data across multiple tissues. Compared with a non-adaptive standard model, our approach substantially reduces errors on samples with a mismatched distribution. On real data, we achieve far lower errors on cerebellum samples, a tissue which is not part of the training data and poorly predicted by standard models. Our results demonstrate that unsupervised domain adaptation is possible for applications in computational biology, even with many more features than samples.

Availability and implementation: Source code is available at <https://github.com/PfeiferLabTue/wenda>.

Contact: lisa.handl@uni-tuebingen.de or pfeifer@informatik.uni-tuebingen.de

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Machine learning has gained wide popularity in recent years and has proved its potential to solve important problems in computational biology on many occasions (Almagro Armenteros *et al.*, 2017; Angermueller *et al.*, 2017; Farh *et al.*, 2015; Jansen *et al.*, 2003; Krogan *et al.*, 2006). Enabled by the increasing amounts of available data, predictive models have the potential to uncover new relationships, e.g. between genotypes and phenotypes (Leffler *et al.*, 2017; Stranger *et al.*, 2011), and to improve health care by offering treatment decision support systems to predict critical events (Hoiles and

van der Schaar, 2016) or a patient's response to treatment (Lengauer and Sing, 2006).

Traditionally, machine learning assumes that the training data originates from the same distribution as the data on which the learned model is later applied. While this assumption forms the statistical basis of all standard models, it is often violated in real-world settings. If new data does not have exactly the same distribution as the training data, learned relationships may no longer be valid, causing model performance to deteriorate.

For example, a model may be developed in a highly controlled setting, but when it is later put to use in the real world, the

conditions are less ideal. New data might be measured in different institutions with different devices or protocols, or batch effects might lead to differences in the distributions of data acquired at different times (Akey *et al.*, 2007; Leek *et al.*, 2010). Biological variability can also lead to a distribution mismatch, e.g. when cell composition or other confounders cannot be precisely controlled (Saito and Sætrum, 2012). A distribution mismatch may even arise intentionally, if training data for the problem of interest are not directly available and different but related data are used as a replacement, e.g. for knowledge transfer between species.

Building predictive models that perform well even on data with a certain distribution mismatch with respect to the training data is known as domain adaptation (Pan and Yang, 2010; Patel *et al.*, 2015). The general setting considers data from two domains with different but related underlying distributions: a source domain, from which a sufficient amount of labeled data is available, and a target domain, from which little or no labeled data are available. The goal is to predict well on the target domain while training (mostly) on source domain data. There are multiple flavors of domain adaptation, differing in how much information from the target domain is known.

A particularly challenging variant is unsupervised domain adaptation (Margolis, 2011), where only unlabeled examples from the target domain are available for training. In this setting, there is no direct way to measure a model’s predictive performance on the target domain during training. It is necessary to make assumptions on the structure of the distribution mismatch, which can vary with the data type or application of interest. Otherwise, the source and target distributions could be arbitrarily far apart, eliminating any chance of successful prediction. For some applications, e.g. in computer vision for object recognition from digital images, unsupervised domain adaptation has been studied extensively with promising results (Aljundi *et al.*, 2015; Gong *et al.*, 2012, 2013) and especially domain adaptation methods based on (deep) neural networks have proven successful (Ganin *et al.*, 2016; Long *et al.*, 2016).

Despite the recent success of deep learning methods, applications in computational biology often demand other approaches since models are required to be interpretable and data are less abundant. A popular example are regularized regression models like the elastic net (Zou and Hastie, 2005), which limit the complexity of a model by penalizing large coefficients. Such models are well suited for prediction problems with a much larger number of possibly correlated features than samples, and are thus frequently used in computational biology (Garnett *et al.*, 2012; Hughey and Butte, 2015; Schmidt *et al.*, 2017). Specifically, the elastic net uses a convex combination of L_1 and L_2 penalty, combining advantages of LASSO (Tibshirani, 1996) and ridge regression (Hoerl and Kennard, 1970) regarding sparsity and the handling of correlated features.

In this article we propose *wenda* (weighted elastic net for unsupervised domain adaptation). Our method compares the dependency structure between inputs in source and target domain to measure how similar features behave. It then encourages the use of similarly behaving features using a target domain-specific feature weighting. We build on ideas from Jalali and Pfeifer (2016) to measure the similarity of features in source and target domain, but do not use strict feature selection or a predefined set of weak learners. Instead, we learn a full weighted model for each considered target domain. *Wenda* retains all advantages of the standard elastic net regarding interpretability and the effects of regularization, but prioritizes features according to how well they agree in both domains.

As a concrete application example, we consider the problem of age prediction from DNA methylation data across tissues. DNA

methylation is a well-studied epigenetic mark, which has been shown to play a role in important gene regulatory processes like the long-term repression of genes, genomic imprinting and X-chromosome inactivation (Schübeler, 2015). In addition, DNA methylation patterns of genomic DNA have been found to be associated with its donor’s chronological age (Bell *et al.*, 2012; Heyn *et al.*, 2012; Teschendorff *et al.*, 2013a). Several studies used DNA methylation data to predict donor age and elastic net models turned out to be particularly useful for this task (Florath *et al.*, 2014; Hannum *et al.*, 2013; Horvath, 2013). While these models were trained on the DNA methylation and chronological age of healthy donors, their predictions are interpreted as a biological epigenetic age. Increased epigenetic aging could be linked to lifestyle factors and disease history, suggesting that the epigenetic age contains useful information on an individual’s health status.

DNA methylation patterns are known to be highly tissue specific (Varley *et al.*, 2013; Ziller *et al.*, 2013). While some age-associated changes in DNA methylation are similar across tissues (Christensen *et al.*, 2009; Zhu *et al.*, 2018), this does not hold for all of them (Day *et al.*, 2013; Fraser *et al.*, 2005). Predicting age on different tissues than the ones that are available for training can therefore be seen as an unsupervised domain adaptation problem. As more tissue-specific data have recently become available (Aguet *et al.*, 2017), predicting age on data from multiple tissues can serve as an example for many future prediction scenarios, making this problem an ideal candidate for evaluating *wenda* on real biological data.

We consider DNA methylation data from multiple tissues and explicitly unmatched tissue compositions in training and test set. Compared with a non-adaptive standard model, we show that our method strongly improves performance on samples from the cerebellum of the human brain, which were not part of the training data and very poorly predicted by a non-adaptive standard model. In addition, we study the performance of *wenda* in simulation experiments, where it is possible to vary the severity of the distribution mismatch between domains in a controlled setting. We show that our method reduces test error compared with a simple elastic net without domain adaptation also in this scenario, suggesting a wide applicability in computational biology.

2 The *wenda* method

We assume to have n labeled examples, $(x_1, y_1), \dots, (x_n, y_n)$, from the source domain and m labeled examples, $(\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_m, \tilde{y}_m)$, from the target domain. In both domains, the inputs, $\{x_i\}_{i=1}^n$ and $\{\tilde{x}_i\}_{i=1}^m$, are p -dimensional vectors with $p \in \mathbb{N}$, and the outputs, $\{y_i\}_{i=1}^n$ and $\{\tilde{y}_i\}_{i=1}^m$, are scalars. The goal of our method is to use the source domain examples and the target domain inputs to come up with a good prediction of target domain output. The data in source and target domain follow two different joint probability distributions $P_S(X, Y) = P_S(Y|X) \cdot P_S(X)$ and $P_T(X, Y) = P_T(Y|X) \cdot P_T(X)$, respectively. A classical assumption in domain adaptation, called the covariate shift assumption, is that the difference between these distributions arises only from the inputs, i.e. $P_S(X) \neq P_T(X)$, while the conditional distributions, $P_S(Y|X) = P_T(Y|X)$, are identical. We weaken this assumption by allowing some features to have a different influence on the output in source and target domain. More precisely, we assume that a subset M of all p features, $M \subset \{1, \dots, p\}$, that shares the same dependency structure in source and target domain will also have the same influence on Y in both domains. Features which are not in M might influence Y differently in source and target domain. More formally, the core assumption is

$$\begin{aligned} P_S(X_f|X_{-f}) &\approx P_T(X_f|X_{-f}) \text{ for all } f \in M \\ \Rightarrow P_S(Y|X_M) &\approx P_T(Y|X_M), \end{aligned} \quad (1)$$

where X_f and X_{-f} denote feature f and all features except f in X , respectively, and X_M is the subvector of X containing only features in M . We propose a model-based approach to quantify how well $P_S(X_f|X_{-f})$ and $P_T(X_f|X_{-f})$ agree for different features. Instead of strictly including or excluding features, we enforce stronger regularization on features for which larger differences exist. This allows for a tradeoff between a feature’s suitability for adaptation and its importance for prediction. If $P_S(Y|X_{\{1,\dots,p\}\setminus M})$ and $P_T(Y|X_{\{1,\dots,p\}\setminus M})$ differ noticeably, reducing the influence of features outside M on the model should improve its robustness and capability to transfer between domains.

Wenda consists of the following three main components, which we describe in detail in the following sections:

1. *Feature models*: We estimate the dependency structure between inputs in the source domain using Bayesian models.
2. *Confidence scores*: We evaluate the estimated input dependency structure on the target domain to quantify the confidence into each feature for domain adaptation.
3. *Final adaptive model*: We train the final model on source domain data while adjusting the strength of regularization for each feature depending on its confidence.

For simplicity, we explain this method considering only one target domain even though it can easily be applied to multiple target domains as we do in Sections 3 and 4.

2.1 Feature models

We capture the dependency structure between inputs in the source domain using Bayesian models. For each feature f , we train a model g_f which predicts f based on all other features using the source domain inputs, x_1, \dots, x_n , as training data. These feature models estimate all conditional distributions $P_S(X_f|X_{-f})$. Since we consider high-dimensional feature spaces, we use Gaussian process models (Rasmussen and Williams, 2006) with a simple linear kernel and additive noise. This model has two hyper parameters, the variance of the prior on the coefficients σ_p^2 , and the variance of the noise σ_n^2 , which we determine by maximum marginal likelihood for each feature. More precisely, we write $x_{:,f} = (x_{1,f}, \dots, x_{n,f})^\top$ for the vector containing feature f , and $x_{:-f}$ for the $(n \times (p-1))$ -matrix containing all remaining features of the training samples, and maximize

$$\begin{aligned} \log p(x_{:,f}|x_{:-f}) &= -\frac{1}{2} x_{:,f}^\top (K + \sigma_n^2 I_n)^{-1} x_{:,f} \\ &\quad - \frac{1}{2} \log |K + \sigma_n^2 I_n| - \frac{n}{2} \log(2\pi). \end{aligned} \quad (2)$$

Here $K = \sigma_p^2 x_{:-f} x_{:-f}^\top$ is the linear kernel matrix, I_n is the n -dimensional identity matrix and $|\cdot|$ denotes the determinant. Given σ_p^2 and σ_n^2 , the posterior distribution of the coefficients, ω , of the linear model is Gaussian and has the closed-form solution

$$p(\omega|x_{:,f}, x_{:-f}) \sim \mathcal{N}(\sigma_n^{-2} A^{-1} x_{:-f}^\top x_{:,f}, A^{-1}), \quad (3)$$

where $A = \sigma_n^{-2} x_{:-f}^\top x_{:-f} + \sigma_p^{-2} I_{p-1}$. The advantage of using Bayesian models in this step is that they offer not only a single prediction, but a posterior distribution including uncertainty information.

2.2 Confidence scores

This uncertainty information can be used to define a score that quantifies how closely each feature in the target domain follows the source-domain dependency structure. Consider a given test input, \tilde{x}_i , and feature, f . We denote the value of f in \tilde{x}_i by $\tilde{x}_{i,f}$, and the values of all features except f in \tilde{x}_i by $\tilde{x}_{i,-f}$. Given $\tilde{x}_{i,-f}$, the feature model g_f outputs a posterior distribution, describing which values of $\tilde{x}_{i,f}$ would be expected according to the source-domain dependency structure. For Gaussian processes this is a normal distribution, $\mathcal{N}(\mu_{g_f}(\tilde{x}_{i,-f}), \sigma_{g_f}(\tilde{x}_{i,-f}))$. We quantify how well the observed value, $\tilde{x}_{i,f}$, fits to this predicted distribution using the confidence proposed by Jalali and Pfeifer (2016),

$$c_f(\tilde{x}_i) = 2 \cdot \Phi \left(- \frac{|\tilde{x}_{i,f} - \mu_{g_f}(\tilde{x}_{i,-f})|}{\sigma_{g_f}(\tilde{x}_{i,-f})} \right), \quad (4)$$

where Φ denotes the cumulative distribution function of a standard normal distribution. This confidence is the probability that a value as far from $\mu_{g_f}(\tilde{x}_{i,-f})$ as $\tilde{x}_{i,f}$ or further occurs in the posterior distribution predicted by g_f . We define the confidence of feature f for prediction on the target domain as the average of $c_f(\tilde{x}_i)$ over all target inputs,

$$c_f = \frac{1}{m} \sum_{i=1}^m c_f(\tilde{x}_i). \quad (5)$$

For each feature, c_f describes how well the source-domain dependencies of feature f fit in the target domain and, according to the core assumption stated in Equation (1), how suitable f is for the considered domain adaptation task.

2.3 Final adaptive model

To predict the output, $\tilde{y}_1, \dots, \tilde{y}_m$, in the target domain, we train a final model on the source domain data using the confidences defined in Equation (5) to prioritize features. Here we use a weighted version of the elastic net, which scales the contributions of features to the regularization term according to predefined feature weights. The weighted elastic net solves the problem

$$\hat{\beta} = \arg \min_{\beta} (\text{RSS}(\beta) + \lambda J(\beta)) \quad (6)$$

$$J(\beta) = \alpha \sum_{f=1}^p w_f |\beta_f| + \frac{1}{2} (1 - \alpha) \sum_{f=1}^p w_f \beta_f^2, \quad (7)$$

where $\text{RSS}(\beta)$ denotes the residual sum of squares on the training data, w_f are the feature weights, $\lambda > 0$ is the regularization parameter and $\alpha \in [0, 1]$ determines the proportion of L_1 and L_2 penalty. If $w_f = 1$ for all features, Equation (7) reduces to the standard elastic net penalty. We choose these feature weights based on the confidences defined in Equation (5) to encourage the use of features which were estimated to be useful for domain adaptation. More precisely, we set

$$w_f = (1 - c_f)^k, \quad (8)$$

where $k > 0$ is a user-specified model parameter. This means that coefficients of features with a low confidence are penalized more severely than coefficients of high-confidence features. The parameter k controls how exactly confidences are translated into weights. For $k = 1$, the feature weight increases linearly with decreasing confidence, for higher values of k the model puts an increasingly high penalty on very low confidences while penalizing medium to high

confidences less severely. The resulting model still attempts to predict well on the training data by achieving a small $RSS(\beta)$, but is encouraged to prefer features with high confidence. It takes into account both a feature’s importance for predicting the output according to the source domain data and its confidence, i.e. its estimated suitability for domain adaptation.

2.4 The challenge of parameter selection

Wenda has three external parameters: the weighting parameter k , the proportion of L_1 and L_2 penalty α and the regularization parameter λ . Parameters α and λ are inherited from the standard elastic net and usually optimized via cross-validation on the training data. Alternatively, α is sometimes treated as a design choice (Horvath, 2013; Hughey and Butte, 2015), as its effect, i.e. the interpolation between ridge regression and LASSO, is fairly straightforward to interpret.

Cross-validation approximates the error on unseen samples drawn from the same distribution as the training data. The goal of unsupervised domain adaptation, however, is to achieve low error on samples from the target domain, which follow a different distribution. The absence of labeled output examples from the target domain for training is an obstacle for model selection. While parameters can be optimized with respect to the source-domain distribution, it is uncertain whether they generalize to the target domain. Furthermore, simultaneously optimizing multiple parameters constitutes a non-negligible computational burden.

Considering these aspects, we treat α as a design choice and keep it fixed at $\alpha = 0.8$. Parameter λ determines the strength of regularization and can thus not be globally set to one value that performs well across different datasets. Since data-dependent tuning of λ is inevitable, we evaluate and compare two approaches, which are described in Sections 2.5 and 2.6. The parameter k is introduced by our method, so we evaluate its sensitivity in the empirical studies (Sections 3 and 4).

2.5 *Wenda-pn*: prior knowledge on size of mismatch

In *wenda*, λ does not only affect the strength of regularization but also how strongly the feature weights are taken into account. For very small λ , e.g. all features are weakly penalized and differences among feature weights have only a minor influence. For large λ , redistributing coefficients between features with different weights can strongly change the value of the objective function, giving feature weights a large influence on the final result. Hence, for any target domain T , the optimal value, λ_{opt}^T , depends on how much adaptation is needed for transfer between the source and target domain.

If the size or severity of the distribution mismatch between domains has a major influence on which λ is optimal, prior knowledge on the similarity between the domains could help to choose λ . Note that prior knowledge here refers to information known from other sources, but not to a prior distribution in the Bayesian sense. This approach requires:

1. A quantitative measure of similarity or dissimilarity between source domain and target domain(s).
2. A mapping from domain (dis)similarity to a good choice of λ .

If and how prior knowledge on domain similarity is available depends on the application and will be described in Sections 3.3 and 4.2 for the datasets used in this work.

The mapping usually has to be estimated from data, which is possible if multiple target domains, T_1, \dots, T_ℓ , are considered and

labeled examples are available for some of them. We model $\log(\lambda_{\text{opt}}^T)$ as a linear function of domain similarity since λ is non-negative and typically chosen from a grid of equidistant points on a logarithmic scale (Friedman *et al.*, 2010).

We call the version of *wenda* using prior knowledge *wenda-pn* and evaluate it using the following cross-validation scheme. We first partition the indexes $\{1, \dots, \ell\}$ of all available target domains into two subsets, I_1 and I_2 . For all $i \in I_1$ we determine $\lambda_{\text{opt}}^{T_i}$ by varying λ on a grid and choosing the value which leads to the lowest mean absolute error (MAE) on the target domain T_i , disclosing the corresponding labels. Next, we fit the model for the relationship between domain similarity and λ_{opt}^T via least squares, using $\{\lambda_{\text{opt}}^{T_i}\}_{i \in I_1}$ and the corresponding domain similarities as training data. With this model we predict $\lambda_{\text{opt}}^{T_i}$ for all $i \in I_2$ and measure the resulting performance of *wenda-pn*. This process is repeated for multiple splits of the target domains into subsets I_1 and I_2 . The exact number and ratio of splits is problem dependent and will be described in Sections 3.3 and 4.2.

2.6 *Wenda-cv*: cross-validation on training data

If no knowledge on domain similarity is available, an alternative option is to still use cross-validation on the training data to determine λ . Cross-validation will choose a regularization strength which is optimal on the source domain for the given feature weights, rather than the target domain. Including the feature weighting can still lead to an improvement compared with a standard elastic net, but choosing λ with cross-validation on source domain data may not fully exploit its potential. We call this version of our method *wenda-cv*.

2.7 Implementation

We implemented all models in python 3.5.4., the source code is available on GitHub (<https://github.com/PfeiferLabTue/wenda>). For computing the regularization paths of (weighted or unweighted) elastic net models, we used python-glmnet (Civis Analytics, 2016), a python wrapper around the original Fortran code which is also the basis of the R package glmnet (Friedman *et al.*, 2010). For optimizing the Gaussian process models needed for the feature models described in Section 2.1, we used the python package GPy (GPy, 2012).

3 Experiments on simulated data

To evaluate how *wenda* performs on datasets with varying degrees of domain mismatch in a controlled setting, we simulate multiple datasets with dependent inputs and a defined distribution mismatch between source and target domain. In each simulated dataset we use 1000 inputs, 3000 training samples from the source domain and 1000 test samples from the target domain. To account for variability, we run 10 fully independent simulations.

3.1 Source domain model

We model the complex dependency structure between inputs using Bayesian networks (Pearl, 1988) with Gaussian marginal distributions. For each simulation, we first randomly generate 20 directed acyclic graphs (DAGs) with 50 nodes each and a maximum degree of 5 (indegree + outdegree) using BNGenerator (Ide and Cozman, 2002). These graphs model 20 groups of input variables with dependencies within but not between groups. BNGenerator uses a Markov chain Monte Carlo approach to sample uniformly from all possible DAGs which satisfy the specified constraints. It additionally outputs categorical distributions and conditional distributions for the nodes, which we ignore for this application. Instead of

categorical distributions, we assign independent standard normal distributions to all root nodes and define the distributions of all child nodes as linear combinations of their parent nodes plus a fixed amount of Gaussian noise. To control the variance of child nodes, we move through each graph according to its topological ordering, draw random weights for parent edges from a standard normal distribution, and scale them to achieve a total variance of 1 (including noise). We set the noise variance for input dependencies to $\sigma_e^2 = 0.1$, i.e. 10% of the marginal variance of each node.

For the output, we use a sparse linear model with Gaussian noise. We randomly choose 20 out of 1000 coefficients to be non-zero, one in each of the 20 graphs. As for the relationships between inputs, we set the noise variance to $\sigma_{\text{out}}^2 = 0.1$, draw the nonzero coefficients from a standard normal distribution and scale them to achieve variance 1.

3.2 Target domain model

To model target domain data with a distribution mismatch, we start from the source domain model, but make changes to some of the variables and their influence on the output. The Bayesian networks allow us to directly change dependencies between inputs in the model, instead of just distorting simulated data. Depending on the degree of domain mismatch we wish to introduce, we randomly pick a certain number of the 20 graphs representing the inputs and multiply the weights of all their edges with -1 , thus inverting the dependencies they have in the source domain. This is an attractive choice because it specifically changes the dependencies of inputs while not strongly distorting their marginal distributions. In addition, we change the influence of these altered variables on the output by setting the corresponding coefficients in the output model to zero. In each simulation, we consider four different target domains with varying size of distribution mismatch: no mismatch, 10%, 20% and 30% altered variables. When training the weighted models, we average confidences only over groups of 100 samples at a time, to account for the variability in feature weights caused by smaller target domain sample sizes.

3.3 Prior knowledge on domain mismatch

Incorporating knowledge on the size of the domain mismatch is simple for simulated data since the ground truth of how many variables were altered is known. We define domain similarity as the fraction of unchanged variables and use leave-one-out cross-validation on the four sizes of distribution mismatch to evaluate the performance of *wenda-pn* (Section 2.5). When predicting with *wenda-pn* for the target domains with a certain size of distribution mismatch, we use the remaining target domains (from all simulations) to learn the relationship between domain similarity and λ_{opt}^T .

3.4 Baseline models

We compare the results of *wenda-pn* and *wenda-cv* on the simulated datasets to two baseline models. The first is a simple elastic net without feature weights (*en*), which is the natural baseline for our adaptive model. Here we choose $\alpha = 0.8$ in agreement with *wenda*, and determine λ via 10-fold cross-validation on the training data.

The second baseline is a weighted elastic net with a simpler feature weighting, for which we use the abbreviation *wenda-mar*. This model has the same structure as proposed in Section 2, but feature weights are computed based on the marginal distributions of features instead of the dependency structure between them, eliminating the need to train feature models as described in Section 2.1. It still detects differences between the distributions of inputs in source and

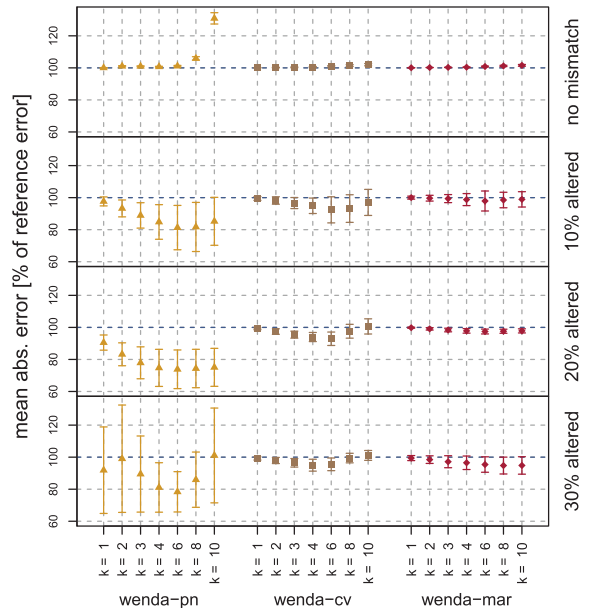


Fig. 1. Mean absolute error (MAE) of *wenda-pn*, *wenda-cv* and *wenda-mar* on simulated test data. Each row shows results on one target domain (no mismatch, 10–30% altered variables). We report all errors relative to the MAE of *en* showing the mean \pm standard deviation over 10 simulations

target domain, but does not utilize dependencies between features to do so. More precisely, the confidence defined in Equation (4) is replaced by the simplified version

$$c_f^s(\tilde{x}_i) = 2 \cdot \min\{\hat{F}_f(\tilde{x}_{i,f}), 1 - \hat{F}_f(\tilde{x}_{i,f})\}, \quad (9)$$

where \hat{F}_f denotes the empirical cumulative distribution function of feature f in the training data. As in *wenda-pn* and *wenda-cv*, we average these confidences over all target-domain inputs and translate them to feature weights in analogy to Equations (5) and (8). Consistently with *wenda-pn* and *wenda-cv*, we keep $\alpha = 0.8$ fixed and report results for multiple values of k . To determine the regularization parameter λ , we use 10-fold cross-validation on the training data.

The score $c_f^s(\tilde{x}_i)$ is chosen to be very similar to Equation (4). A comparison of *wenda-mar* to an alternative score based on KL divergence can be found in Supplementary Figures S1 and S2.

3.5 Results on simulated data

Figure 1 summarizes the MAE of *wenda-pn*, *wenda-cv* and *wenda-mar* on the simulated test data. We report all errors relative to the MAE of the standard (unweighted) elastic net (*en*), the error bars indicate mean and standard deviation over 10 simulations. A similar plot of the correlation between true and predicted output is shown in Supplementary Figure S3.

With *wenda-pn* we obtain considerable improvements for the intermediate target domains with 10% and 20% altered variables, reducing the MAE of *en* by up to 18.7% and 26.2%, respectively. For the more extreme target domains the results are mixed. With 30% altered variables we still observe an improvement for some values of k , but the variability is very high (both within one choice and between choices of k). For the target domain without mismatch, the MAE even increases compared with *en* for high values of k . This can be explained by the cross-validation scheme we employ to learn the

relationship between λ_{opt}^T and domain similarity (Section 3.3). For each size of distribution mismatch, the model describing this relationship has been trained on the remaining target domains. This is an interpolation for the intermediate target domains (10% and 20% altered variables), but an extrapolation for the target domains with 30% altered variables and no mismatch. Extrapolation is a harder problem and can lead to a less accurate estimate of λ_{opt}^T and increased variability.

It should be noted that using domain adaptation even though prior knowledge suggests that there is no distribution mismatch between domains is not a realistic scenario. We include the results of *wenda-pn* on data without distribution mismatch for the sake of completeness.

The other two weighted models, *wenda-cv* and *wenda-mar* show no or only very little improvement over *en*. On target domains with mismatch, *wenda-cv* consistently receives a slightly lower MAE than *en*, but the improvement is only 7.6% at best. It uses the same feature weights as *wenda-pn*, but obviously chooses a less suitable value for λ . The simpler confidences used by *wenda-mar* can only pick up changes in the marginal distributions of features, not in their dependency structure, leading to almost the same results as *en*. Only for 30% altered variables a slight improvement can be noted. Since marginal distributions are only altered very subtly in the target domain model, we expected a weak performance of *wenda-mar* in this simulation study.

4 Age prediction from DNA methylation data

Now we consider our primary application on real data, i.e. the problem of age prediction from DNA methylation data across multiple tissues.

4.1 DNA methylation dataset and preprocessing

We use DNA methylation data and donor age from two sources, the Cancer Genome Atlas (TCGA; Chang *et al.*, 2013) and the Gene Expression Omnibus (GEO; Edgar *et al.*, 2002). We include only DNA methylation data which were measured with the Illumina Infinium HumanMethylation450 BeadChip and only samples from healthy tissue. Using RnBeads (Assenov *et al.*, 2014), we perform several preprocessing steps on the DNA methylation data. In particular, we remove SNPs and gonosomal CpGs, and normalize the data with the BMIQ method (Teschendorff *et al.*, 2013b). In addition, we impute missing values (<0.5% of all measurements) using 10-nearest-neighbor imputation in the R package *impute* (Hastie *et al.*, 2017). Finally, we split the dataset into a training and test set with 1866 and 1001 samples, respectively.

The final training set contains data from 19 different tissues, with a focus on blood, and from donors with a chronological age ranging from 0 to 103 years. The test set consists of data from 13 different tissues initially, including blood as well as tissues which are not present in the training data, e.g. samples from the cerebellum of the human brain. We slightly aggregate them, combining ‘blood’, ‘whole blood’ and ‘menstrual blood’, as well as ‘Brain MedialFrontalCortex’ and ‘Brain FrontalCortex’ to increase sample sizes per tissue. The range of ages represented in the test set is 0–70 years. When applying *wenda*, we keep the training set fixed and consider each tissue in the test set as a separate target domain.

To limit the computational burden of training feature models, we reduce the initial number of 466 094 features to 12 980 using a standard elastic net model with $\alpha = 0.8$ and fixed regularization parameter, $\lambda = 1.1 \times 10^{-4}$. Furthermore, we use the following

transformation for the chronological ages, which was proposed by Horvath (2013). We transform all training ages with the function

$$F(y) = \begin{cases} \log(y+1) - \log(y_{\text{adult}}+1), & \text{if } y \leq y_{\text{adult}} \\ (y - y_{\text{adult}})/(y_{\text{adult}} + 1), & \text{otherwise} \end{cases}$$

with adult age $y_{\text{adult}} = 20$ prior to training, and later re-transform the model’s predictions with the inverse function, F^{-1} . This transformation is logarithmic for ages below and linear for ages above y_{adult} , which is motivated by the fact that the methylation landscape changes more quickly and dramatically in childhood and adolescence than later in life. Subsequently, we standardize all data to zero mean and unit variance.

4.2 Prior knowledge on domain mismatch

As prior knowledge for *wenda-pn* (Section 2.5), we make use of published data on similarities between human tissues. The GTEx consortium published an analysis of a large dataset of (among others) genotype and gene expression data across 42 human tissues (Aguet *et al.*, 2017). In this article, Aguet *et al.* (2017) identified tissue-specific expression quantitative trait loci (eQTLs), i.e. locations in the genome where genetic variants have a significant effect on gene expression levels. Furthermore, the authors estimated tissue-specific effect sizes for each eQTL using a linear mixed model, and reported the correlation (Spearman’s ρ) of effect sizes between all pairs of tissues (see Figure 2a in Aguet *et al.*, 2017), providing a comprehensive measure of tissue similarity. Here we focus on the correlations reported for cis-eQTLs, where the location of the genetic variation is within 1 Mb of the target gene’s transcription start site, since these were identified in larger numbers and with a lower false discovery rate than trans-eQTLs.

We map each tissue in our data to the corresponding tissue(s) contained in the GTEx study, allowing multiple matches if the GTEx classification is more detailed than the one available for our data (Supplementary Table S1). Next, we compute similarities between tissues in our data by looking up (and potentially averaging) the similarities between matched GTEx tissues. Finally, we define the similarity between each target domain and the source domain as the average over all pairwise similarities between samples from the two sets. Our data contains several samples from tissues for which no close match is available in the GTEx data (240 samples in the training set, 56 in the test set). For these we impute the similarity to other tissues with the mean of all pairwise tissue similarities.

When evaluating the performance of *wenda-pn*, we repeatedly split the test tissues into one part for fitting the relationship between domain similarity and λ_{opt}^T and one part for evaluation (Section 2.5). Here, we iterate over all combinations of 3 tissues with at least 20 samples each for training and evaluate the performance on the remaining tissues.

4.3 Baseline models

We compare *wenda-pn* and *wenda-cv* to the two baseline models described in Section 3.4 with the following minor modification: instead of using a simple elastic net directly, we use *en* followed by a linear least-squares fit based only on features which received non-zero coefficients in *en*. We refer to this baseline as *en-ls*. This model type was suggested by Horvath (2013) for age prediction from DNA methylation data, who reported that the subsequent least-squares fit reduced test errors on his dataset. We observe a similar effect on our data, where *en-ls* produces lower test errors than *en* on cerebellum samples while making almost no difference on the remaining samples.

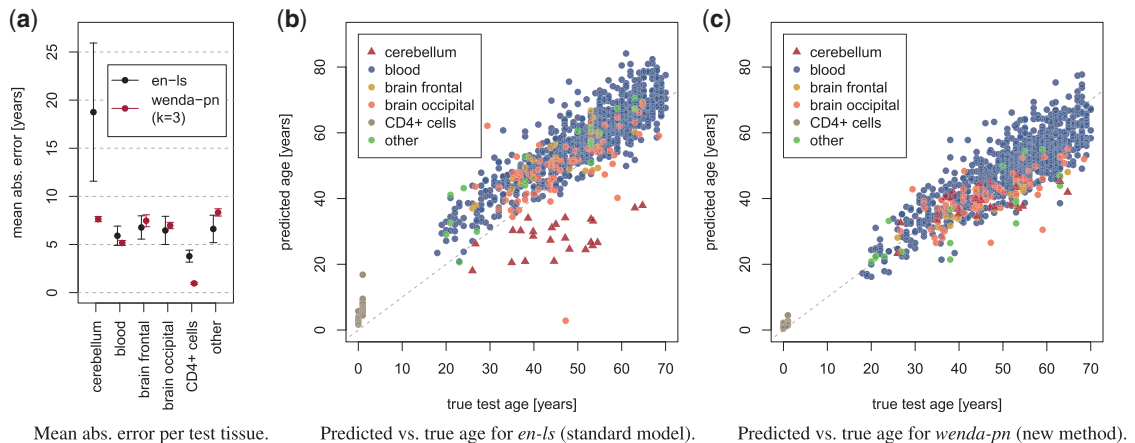


Fig. 2. (a) Mean absolute error of *en-ls* and *wenda-pn* with $k=3$ per test tissue. We show the mean \pm standard deviation over 10 runs of 10-fold cross-validation for *en-ls*, and over all splits of the test tissues where the tissue of interest was in the evaluation set for *wenda-pn*. Predicted versus true chronological age for typical runs of *en-ls* (b) and *wenda-pn* with $k=3$ (c). In each plot, we show samples colored by tissue. As a typical run for *en-ls* we show the one with closest to median performance on cerebellum samples and full test set. For *wenda-pn*, we choose a typical run for each tissue: among all models with this tissue in the holdout set, we plot predictions of the one with closest to median performance

4.4 Results on DNA methylation data

We compare the results of *wenda-pn*, *wenda-cv* and the two baseline models on the dataset described in Section 4.1 and measure performance by MAE on the test set (Supplementary Figure S4 for correlation instead of MAE). Due to the heterogeneous nature of the data, the random split of the training data used for 10-fold cross-validation has a large influence on the results, especially for *en-ls*. Hence, we report the mean and standard deviation over 10 runs. For *wenda-pn*, we do not perform cross-validation on the training data but iterate over multiple splits of the test tissues to learn the relationship between domain similarity and λ_{opt}^T . Here, we measure MAE only on samples which were not used for the similarity-lambda fit, and report mean and standard deviation over all splits.

When training the weighted models, we regard each tissue in the test dataset as a separate target domain. To be precise, we average the confidences defined in Equation (5) only over samples of the same tissue and train a separate model for each tissue, using always the same training data but tissue-specific feature weights.

With *en-ls* we obtain an MAE of 6.19 ± 0.90 years on the full test set. Figure 2a illustrates the MAE of *en-ls* and a representative example of a weighted model (*wenda-pn*, $k=3$) on each test tissue. It shows that *en-ls* yields a considerably higher MAE on cerebellum samples than on other tissues. Figure 2b shows the predicted versus true ages for the test set in a typical cross-validation run, colored by tissue, and reveals that the predicted age is consistently far below the true chronological age. Both plots demonstrate that *en-ls* predicts age well on all test tissues except cerebellum. In fact, on cerebellum samples *en-ls* produces an MAE of 18.75 ± 7.18 years.

Cerebellum samples are especially hard to predict for two reasons: they are not represented in the training data and they are known to be biologically very different even from other brain tissues regarding function and gene expression patterns (Aguet et al., 2017; Fraser et al., 2005). Therefore, the focus of our evaluation is whether domain adaptation as implemented by *wenda* can improve performance on these samples.

The predictions of *wenda-pn* with $k=3$ versus the true ages are shown in Figure 2c. Here, we plot the predictions of a typical run for each tissue by choosing the model with closest to median performance among all models with this tissue in the holdout set. The

ages predicted by *wenda-pn* for cerebellum samples are far closer to the corresponding true ages than they were for *en-ls* (Fig. 2b), and predictions of *wenda-pn* on the remaining test tissues are of a similar quality as those of *en-ls*. This observation is confirmed by the quantitative comparison in Figure 2a, where *wenda-pn* has far lower errors than *en-ls* on cerebellum samples, and similar or better performance than *en-ls* on the remaining test tissues.

While *en-ls* predicts age far worse on cerebellum samples than on other tissues, *wenda-pn* shows no major difference in prediction quality between cerebellum samples and the remaining test tissues. Consequently, *wenda-pn* demonstrates to be considerably more robust to the distribution mismatch between cerebellum samples and the training data than *en-ls*.

Figure 3 shows the MAE of all models on cerebellum samples. Here, all weighted models strongly improve upon *en-ls*. The lowest errors on cerebellum samples are achieved by *wenda-cv*, reaching as low as 6.07 ± 0.10 years for $k=4$. This is closely followed by *wenda-pn*, which achieves an MAE between 7.60 and 8.70 years on average on cerebellum samples for $k \leq 4$. Even *wenda-mar*, which uses only marginal distributions to weight features, improves upon *en-ls* with an MAE of 9.42 ± 0.69 years at best. All weighted models achieve their best result for k between 2 and 4 with not too much variation in this range. However, even when k is far from optimal for cerebellum samples, they still perform better than *en-ls*.

A comparison of the MAE of all models on the full test set is shown in Figure 4 and indicates an overall similar performance of *wenda* and the two baselines. For $k \leq 4$, *wenda-cv* and *wenda-mar* yield a slightly lower MAE than *en-ls*, and for large k , *wenda-cv* and *wenda-pn* yield a slightly higher MAE than *en-ls*. Given that *en-ls* already shows acceptable performance on all tissues except cerebellum, we did not expect a big improvement here. The results show, however, that the improvement on cerebellum samples is not bought by a loss of performance on other tissues.

5 Discussion

Predictive models are widely used in computational biology, but differences between the distribution of their training data and new data to which they are later applied can severely threaten their

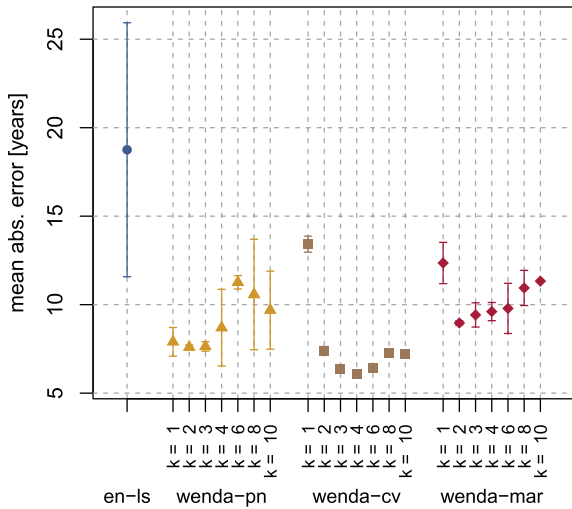


Fig. 3. Mean absolute error of all models on cerebellum samples. We show the mean and standard deviation over 10 runs of 10-fold cross-validation or, in case of *wenda-pn*, over all splits where cerebellum samples were in the evaluation set

performance. In this article we propose *wenda*, a method for unsupervised domain adaptation based on the elastic net. It detects differences in the dependency structure between inputs in source and target domain and enforces stronger regularization on features that behave differently. Our method is different from previous studies on the combination of the elastic net and domain adaptation techniques (Li *et al.*, 2015; Wachinger and Reuter, 2016). Both consider only the easier problem of supervised domain adaptation, i.e. the situation where some labeled examples from the target domain are available for training, and are not applicable in the setting we consider. Our method is also different from the approach proposed by Cortes and Mohri (2011), which uses a sample weighting rather than a feature weighting and is thus better suited for situations with $n > p$ than for the ones we consider.

The key idea of our approach, which separates it from many other domain adaptation methods, is to learn the dependency structure between inputs for calculating feature weights. This property is of particular relevance to applications within computational biology where, in contrast to, e.g. image analysis, the dependency structure is irregular and not known a priori. For example, even distant locations in the (epi)genome can interact and form complex gene regulatory networks, which vary with cell type and differentiation state (Thompson *et al.*, 2015). While we used Gaussian process models with linear kernels as feature models, any other Bayesian model type would be applicable in principle, subject only to the data and computational resources.

Like any domain adaptation method, *wenda* makes the assumption that source and target distribution are not too far apart, so that some features are useful for predicting the output and behave similarly in source and target domain. Another central assumption of our method is that the dependency structure between inputs is informative of which features are useful for domain adaptation. There are certain extreme cases, where this is clearly violated. For example, when features are entirely independent, the distribution predicted by each feature model g_f would be approximately the feature’s marginal distribution, and *wenda-pn* and *wenda-cv* would behave similarly to *wenda-mar*. Another such case is the presence of

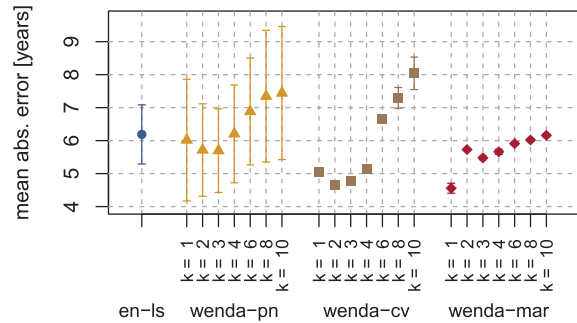


Fig. 4. Mean absolute error (MAE) of all models on the full test set of DNA methylation data. We show the mean and standard deviation over 10 runs of 10-fold cross-validation. In case of *wenda-pn*, we compute the MAE only based on samples in the evaluation set, and plot the mean and standard deviation over all considered splits of the test tissues

duplicates or extremely strong correlations between variables. These could arise, e.g. in sequencing-based methylation assays, where the DNA methylation of consecutive CpG sites is highly correlated in all tissues. Thus, each feature would always be well predicted by its neighbor, regardless of changes on a larger scale. In situations like this, we suggest to aggregate extremely correlated features before training, which is also advisable for a standard elastic net.

Our method is computationally demanding since it requires to train one Bayesian model per feature (for confidence estimation) and one weighted elastic net per target domain (for prediction). While both of these steps can be parallelized to speed up calculations, fitting the feature models remains challenging for large datasets. For example, training 12 980 feature models for the DNA methylation data on 10 CPUs of the type Intel Xeon CPU E7-4850 with 2.30 GHz takes about 51 h.

However, the structure of *wenda* allows additional speed-ups, as feature models have to be trained only once (as long as the training data remain fixed) and can be reused to predict on multiple target domains or with different parameter settings. If the confidence scores for a given test dataset are precomputed as well, the final model for one target domain is only a weighted elastic net trained on the training data, whose regularization path can be computed quickly, e.g. with glmnet. With the same computational setup as before and with precomputed feature models and confidence scores, training all models required for *wenda-pn* with $k = 3$ (Fig. 2c) takes about 43 s.

Wenda allows to incorporate prior knowledge on the size of the domain mismatch (*wenda-pn*), but a simplified version can also be applied without it (*wenda-cv*). *Wenda-cv* uses cross-validation on the training data to determine λ , which is not ideal in a domain adaptation setting. Nevertheless, our results on the DNA methylation data demonstrate that it can still lead to a surprisingly large improvement over a non-adaptive model. This makes it a valuable alternative to *wenda-pn*, especially if no prior knowledge on the size of domain mismatch is available.

Wenda introduces a new parameter k , which controls how confidences are translated into feature weights. We empirically studied the impact of choosing k on the MAE and observed satisfying performance in the interval $k \in [2, 4]$. Hence, $k = 3$ might constitute a relatively robust choice for future applications, albeit it is unlikely that any single parameter choice is optimal for each and every target domain. We note that *wenda* never performs substantially worse than the non-adaptive reference. Hence, the precise value of k determines only the magnitude of improvement obtained and a

suboptimal choice poses relatively little risk. Nevertheless, without labeled training examples from the target domain, parameter selection remains a non-trivial problem. Finding a data-driven way to determine an optimal choice for k , or evaluating whether α can be optimized additionally, are challenging themes for future research.

6 Conclusions

In this article we propose *wenda*, a method for unsupervised domain adaptation which is based on the elastic net and utilizes dependencies between inputs to detect differences between source and target domain. Using a weighted elastic net penalty, *wenda* enforces stronger regularization on features that behave differently in the two domains, reducing the effects of a distribution mismatch.

We compare two variants of our method, *wenda-pn* and *wenda-cv*, on simulated datasets and on real data, where we considered the problem of age prediction from DNA methylation data across tissues. Our experimental results demonstrate that both variants can reduce test errors on samples with a distribution mismatch. While *wenda-cv* outperforms the non-adaptive reference only on real data, *wenda-pn* strongly reduces errors on test samples with a distribution mismatch both on real and simulated data, which makes it the more promising variant for future applications.

From a wider perspective, this article demonstrates that the ambitious goal of unsupervised domain adaptation is indeed feasible not only for big data analysis with deep learning methods, but also for traditional machine learning methods that are useful for analyzing relatively small datasets as they frequently occur in computational biology and medicine.

Acknowledgements

We would like to thank Dr Alexis Battle and Ben Strober for kindly providing the matrix of similarities plotted in Figure 2a in [Aguet et al. \(2017\)](#). We additionally thank Martina Feierabend for reviewing the mapping of tissues between their data and ours.

Funding

This work was prepared within the project *Xploit* of the initiative ‘i: DSEM—Integrative Datensemantik in der Systemmedizin’, funded by the German Federal Ministry of Education and Research (BMBF).

Conflict of Interest: none declared.

References

- Aguet, F. et al. (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.
- Akey, J.M. et al. (2007) On the design and analysis of gene expression studies in human populations. *Nat. Genet.*, **39**, 807–808.
- Aljundi, R. et al. (2015) Landmarks-based kernelized subspace alignment for unsupervised domain adaptation. In: *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015: 8–10 June, 2015, pp. 56–63. Boston, Massachusetts, USA.
- Almagro Armenteros, J.J. et al. (2017) DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, **33**, 3387–3395.
- Angermueller, C. et al. (2017) DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.*, **18**, 67.
- Assenov, Y. et al. (2014) Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods*, **11**, 1138–1140.
- Bell, J.T. et al. (2012) Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet.*, **8**, e1002629.
- Chang, K. et al. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
- Christensen, B.C. et al. (2009) Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet.*, **5**, e1000602.
- Civis Analytics (since 2016) python-glmnet: A Python Port of the glmnet Package for Fitting Generalized Linear Models via Penalized Maximum Likelihood. Python Package Version 2.0.0. <http://github.com/civisanalytics/python-glmnet> (10 May 2019, date last accessed).
- Cortes, C. and Mohri, M. (2011) Domain adaptation in regression. In: *Proceedings of the 2011 International Conference on Algorithmic Learning Theory (ALT)*, 5–7 October, 2011, pp. 308–323. Espoo, Finland.
- Day, K. et al. (2013) Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biol.*, **14**, R102.
- Edgar, R. et al. (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
- Farh, K.K.-H. et al. (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**, 337–343.
- Florath, I. et al. (2014) Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Hum. Mol. Genet.*, **23**, 1186–1201.
- Fraser, H.B. et al. (2005) Aging and gene expression in the primate brain. *PLoS Biol.*, **3**, e274.
- Friedman, J. et al. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.
- Ganin, Y. et al. (2016) Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, **17**, 1–35.
- Garnett, M.J. et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.
- Gong, B. et al. (2012) Geodesic flow kernel for unsupervised domain adaptation. In: *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 16–21 June, 2012, pp. 2066–2073. Rhode Island, USA.
- Gong, B. et al. (2013) Connecting the dots with landmarks: discriminatively learning domain-invariant features for unsupervised domain adaptation. In: *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 16–21 June, 2013, pp. 222–230. Atlanta, Georgia, USA.
- GPy (since 2012) GPy: A Gaussian Process Framework in Python. Python Package Version 1.5.3. <http://github.com/SheffieldML/GPy> (10 May 2019, date last accessed).
- Hannum, G. et al. (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell*, **49**, 359–367.
- Hastie, T. et al. (2017) *impute: Imputation for Microarray Data*. R Package Version 1.52.0. <http://www.biocconductor.org/packages/release/bioc/html/impute.html> (10 May 2019, date last accessed).
- Heyn, H. et al. (2012) Distinct DNA methylomes of newborns and centenarians. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 10522–10527.
- Hoerl, A.E. and Kennard, R.W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.
- Hoiles, W. and van der Schaar, M. (2016) A non-parametric learning method for confidently estimating patient’s clinical state and dynamics. *Adv. Neural Inform. Process. Syst.*, **29**, 2020–2028.
- Horvath, S. (2013) DNA methylation age of human tissues and cell types. *Genome Biol.*, **14**, R115.
- Hughey, J.J. and Butte, A.J. (2015) Robust meta-analysis of gene expression using the elastic net. *Nucleic Acids Res.*, **43**, e79.
- Ide, J.S. and Cozman, F.G. (2002) Random generation of Bayesian networks. In: *Advances in Artificial Intelligence*, Lecture Notes in Computer Science, Springer, Berlin.
- Jalali, A. and Pfeifer, N. (2016) Interpretable per case weighted ensemble method for cancer associations. *BMC Genom.*, **17**, 501.
- Jansen, R. et al. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
- Krogan, N.J. et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, **440**, 637–643.
- Leek, J.T. et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.

- Leffler, E.M. *et al.* (2017) Resistance to malaria through structural variation of red blood cell invasion receptors. *Science*, **356**, eaam6393.
- Lengauer, T. and Sing, T. (2006) Bioinformatics-assisted anti-HIV therapy. *Nat. Rev. Microbiol.*, **4**, 790–797.
- Li, Y. *et al.* (2015) Constrained elastic net based knowledge transfer for health-care information exchange. *Data Min. Knowl. Discov.*, **29**, 1094–1112.
- Long, M. *et al.* (2016) Unsupervised domain adaptation with residual transfer networks. *Adv. Neural Inform. Process. Syst.*, **29**, 136–144.
- Margolis, A. (2011) A Literature Review of Domain Adaptation with Unlabeled Data. Technical Report, University of Washington.
- Pan, S.J. and Yang, Q. (2010) A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, **22**, 1345–1359.
- Patel, V.M. *et al.* (2015) Visual domain adaptation: a survey of recent advances. *IEEE Signal Process. Mag.*, **32**, 53–69.
- Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco.
- Rasmussen, C.E. and Williams, C.K.I. (2006) *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.
- Saito, T. and Sætrom, P. (2012) Target gene expression levels and competition between transfected and endogenous microRNAs are strong confounding factors in microRNA high-throughput experiments. *Silence*, **3**, 3.
- Schmidt, F. *et al.* (2017) Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.*, **45**, 54–66.
- Schübeler, D. (2015) Function and information content of DNA methylation. *Nature*, **517**, 321–326.
- Stranger, B.E. *et al.* (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, **187**, 367–383.
- Teschendorff, A.E. *et al.* (2013) Age-associated epigenetic drift: implications, and a case of epigenetic thrift? *Hum. Mol. Genet.*, **22**, R7–R15.
- Teschendorff, A.E. *et al.* (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, **29**, 189–196.
- Thompson, D. *et al.* (2015) Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annu. Rev. Cell Dev. Biol.*, **31**, 399–428.
- Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Ser. B (Stat. Methodol.)*, **58**, 267–288.
- Varley, K.E. *et al.* (2013) Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.*, **23**, 555–567.
- Wachinger, C. and Reuter, M. (2016) Domain adaptation for Alzheimer's disease diagnostics. *NeuroImage*, **139**, 470–479.
- Zhu, T. *et al.* (2018) Cell and tissue type independent age-associated DNA methylation changes are not rare but common. *Aging*, **10**, 3541–3557.
- Ziller, M.J. *et al.* (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**, 477–481.
- Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. Royal Stat. Soc. Ser. B (Stat. Methodol.)*, **67**, 301–320.

A.3 Manuscript 3

Title:

Combination therapy with anti-HIV-1 antibodies maintains viral suppression

Authors:

Pilar Mendoza, Henning Gruell, Lilian Nogueira, Joy A. Pai, Allison L. Butler, Katrina Millard, Clara Lehmann, Isabelle Suárez, Thiago Y. Oliveira, Julio C. C. Lorenzi, Yehuda Z. Cohen, Christoph Wyen, Tim Kümmerle, Theodora Karagounis, Ching-Lan Lu, Lisa Handl, Cecilia Unson-O'Brien, Roshni Patel, Carola Ruping, Maike Schlotz, Maggi Witmer-Pack, Irina Shimeliovich, Gisela Kremer, Eleonore Thomas, Kelly E. Seaton, Jill Horowitz, Anthony P. West Jr, Pamela J. Bjorkman, Georgia D. Tomaras, Roy M. Gulick, Nico Pfeifer, Gerd Fätkenheuer, Michael S. Seaman, Florian Klein, Marina Caskey & Michel C. Nussenzweig

Published in:

Nature, Volume 561, Pages 479–484

<https://doi.org/10.1038/s41586-018-0531-2>

Time of Publication:

September 2018

License information:

As part of their editorial policies, Springer Nature allows authors to reuse the version of record of their article published in any Nature portfolio journal in their own dissertation without obtaining an additional, explicit written permission (<https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish>).

Combination therapy with anti-HIV-1 antibodies maintains viral suppression

Pilar Mendoza^{1,19}, Henning Gruell^{2,3,4,19}, Lilian Nogueira¹, Joy A. Pai¹, Allison L. Butler¹, Katrina Millard¹, Clara Lehmann^{3,4,5}, Isabelle Suárez^{3,4,5}, Thiago Y. Oliveira¹, Julio C. C. Lorenzi¹, Yehuda Z. Cohen¹, Christoph Wyen^{3,6}, Tim Kümmerle^{3,6}, Theodora Karagounis¹, Ching-Lan Lu¹, Lisa Handl⁷, Cecilia Unson-O'Brien¹, Roshni Patel¹, Carola Ruping², Maike Schlotz², Maggi Witmer-Pack¹, Irina Shimeliovich¹, Gisela Kremer³, Eleonore Thomas³, Kelly E. Seaton⁸, Jill Horowitz¹, Anthony P. West Jr⁹, Pamela J. Bjorkman⁹, Georgia D. Tomaras^{8,10,11,12}, Roy M. Gulick¹³, Nico Pfeifer^{7,14,15,16}, Gerd Fätkenheuer^{3,4}, Michael S. Seaman¹⁷, Florian Klein^{2,4,5,20*}, Marina Caskey^{1,20*} & Michel C. Nussenzweig^{1,18,20*}

Individuals infected with HIV-1 require lifelong antiretroviral therapy, because interruption of treatment leads to rapid rebound viraemia. Here we report on a phase 1b clinical trial in which a combination of 3BNC117 and 10-1074, two potent monoclonal anti-HIV-1 broadly neutralizing antibodies that target independent sites on the HIV-1 envelope spike, was administered during analytical treatment interruption. Participants received three infusions of 30 mg kg⁻¹ of each antibody at 0, 3 and 6 weeks. Infusions of the two antibodies were generally well-tolerated. The nine enrolled individuals with antibody-sensitive latent viral reservoirs maintained suppression for between 15 and more than 30 weeks (median of 21 weeks), and none developed viruses that were resistant to both antibodies. We conclude that the combination of the anti-HIV-1 monoclonal antibodies 3BNC117 and 10-1074 can maintain long-term suppression in the absence of antiretroviral therapy in individuals with antibody-sensitive viral reservoirs.

During infection, HIV-1 is reverse transcribed and integrated as a provirus into the host genome. Although the vast majority of infected cells die by apoptosis or pyroptosis¹, a small percentage survive and harbour transcriptionally silent, integrated proviruses that comprise a reservoir that can be reactivated. Once established, the latent reservoir has an estimated half-life of 44 months, resulting in the lifelong requirement for antiretroviral therapy (ART)². Passive administration of potent broadly neutralizing monoclonal anti-HIV-1 antibodies (bNAbs) represents a potential alternative to antiretroviral drugs because, in addition to neutralizing the virus, antibodies engage the host immune system and have long half-lives³⁻⁵.

In human clinical trials, viraemic individuals who received 3BNC117 or VRC01, two related bNAbs that target the CD4 binding site on the HIV-1 envelope spike, or 10-1074, a bNAb that targets the base of the V3 loop and surrounding glycans, showed significant reductions in viremia⁶⁻⁸. Moreover, in HIV-1-infected individuals undergoing analytical treatment interruption (ATI) of antiretroviral therapy, four infusions of 3BNC117 maintained virus suppression for a median of 10 weeks compared to 2.3 weeks in historical controls^{9,10}. By contrast, six infusions of VRC01 maintained suppression for 5.6 weeks¹¹. The difference in activity between VRC01 and 3BNC117 in preclinical experiments^{12,13} and clinical trials^{6,7,9,11} is consistent with the lower relative neutralization potency of VRC01.

Across all bNAb clinical trials to date, and similar to monotherapy with antiretroviral drugs, treatment with any single bNAb was associated with the emergence of antibody-resistant viral variants^{6-9,11}. Like

antiretroviral drugs, combinations of bNAbs are more effective than individual antibodies in HIV-1 infected humanized mice and simian/human immunodeficiency virus (SHIV)-infected macaques¹⁴⁻¹⁶. By contrast, antibody combinations showed little if any efficacy in suppressing viraemia during ATI in humans^{17,18}. However, these earlier studies were performed using bNAbs that were less potent than 3BNC117 and 10-1074. Here we investigate whether the bNAb combination of 3BNC117 and 10-1074 can maintain viral suppression during ATI in HIV-1-infected humans.

Combination bNAb infusion is well-tolerated

To evaluate the effects of the combination of 3BNC117 and 10-1074 on maintaining HIV-1 suppression during ATI, we conducted a phase 1b clinical trial (Fig. 1a). HIV-1-infected individuals on ART were pre-screened for 3BNC117 and 10-1074 sensitivity of bulk outgrowth culture-derived viruses in an in vitro neutralization assay using TZM-bl cells¹⁹. Consistent with previous results, 64% and 71% of the outgrowth viruses were sensitive to 3BNC117 and 10-1074, respectively, and 48% were sensitive to both^{8,9,20} (half-maximum inhibitory concentration (IC₅₀) ≤ 2 μg ml⁻¹; Extended Data Fig. 1a and Supplementary Table 1).

Study eligibility criteria included ongoing ART for at least 24 months with plasma HIV-1 RNA levels of <50 copies per ml for at least 18 months (one blip <500 copies per ml was allowed) and <20 copies per ml at screening, as well as CD4⁺ T cell counts >500 cells per μl (Extended Data Figs. 1b, 2a). Enrolled participants received three infusions of 30 mg kg⁻¹ of 3BNC117 and 10-1074 each at

¹Laboratory of Molecular Immunology, The Rockefeller University, New York, NY, USA. ²Laboratory of Experimental Immunology, Institute of Virology, University Hospital Cologne, Cologne, Germany. ³Department I of Internal Medicine, University Hospital Cologne, Cologne, Germany. ⁴German Center for Infection Research, partner site Bonn-Cologne, Cologne, Germany. ⁵Center for Molecular Medicine Cologne (CMCC), University of Cologne, Cologne, Germany. ⁶Praxis am Ebertplatz, Cologne, Germany. ⁷Methods in Medical Informatics, Department of Computer Science, University of Tübingen, Tübingen, Germany. ⁸Duke Human Vaccine Institute, Duke University, Durham, NC, USA. ⁹Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. ¹⁰Department of Surgery, Duke University, Durham, NC, USA. ¹¹Department of Immunology, Duke University, Durham, NC, USA. ¹²Department of Molecular Genetics and Microbiology, Duke University, Durham, NC, USA. ¹³Division of Infectious Diseases, Weill Cornell Medicine, New York, NY, USA. ¹⁴Medical Faculty, University of Tübingen, Tübingen, Germany. ¹⁵German Center for Infection Research, partner site Tübingen, Tübingen, Germany. ¹⁶Max Planck Institute for Informatics, Saarbrücken, Germany. ¹⁷Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. ¹⁸Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA. ¹⁹These authors contributed equally: Pilar Mendoza, Henning Gruell. ²⁰These authors jointly supervised this work: Florian Klein, Marina Caskey, Michel C. Nussenzweig. *e-mail: florian.klein@uk-koeln.de; mcaskey@rockefeller.edu; nussen@rockefeller.edu

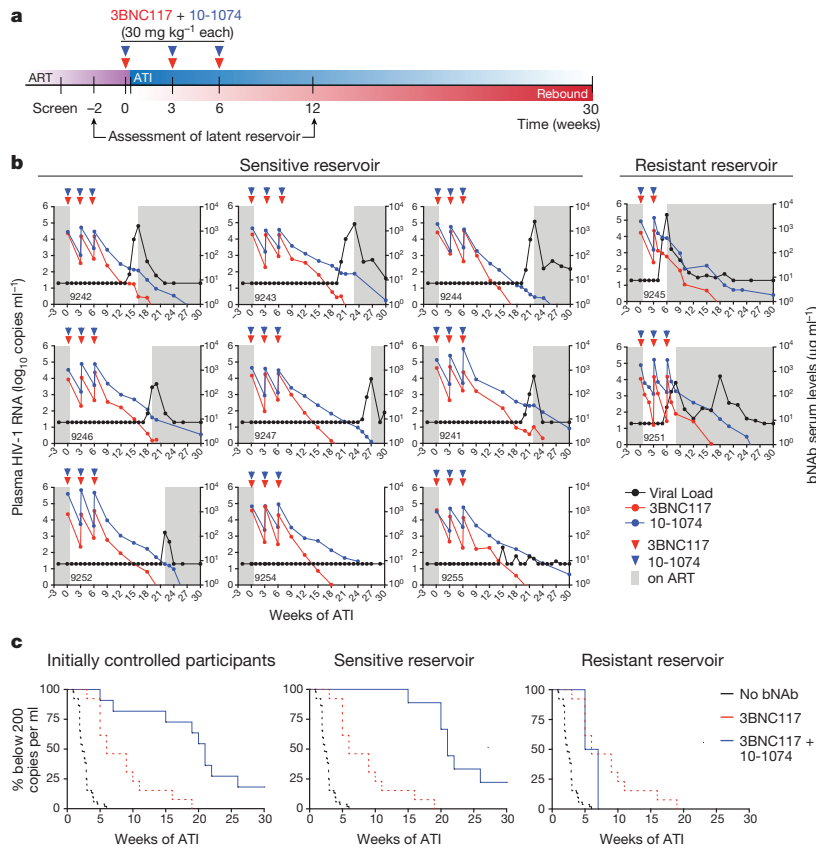


Fig. 1 | Delayed viral rebound with 3BNC117 and 10-1074 combination therapy during ATI. **a**, Study design. Red and blue triangles represent 3BNC117 and 10-1074 infusions, respectively. **b**, Plasma HIV-1 RNA levels (black; left y axis) and bNAb serum concentrations (3BNC117, red; 10-1074, blue; right y axis) in the nine bNAb-sensitive participants (left) and the two participants with pre-existing resistance against one of the antibodies (right). Red and blue triangles indicate 3BNC117 and 10-1074 infusions, respectively. Serum antibody concentrations were determined by TZM-bl assay. Grey shaded areas indicate time on ART. Lower limit of detection of HIV-1 RNA was 20 copies per ml. **c**, Kaplan-Meier plots summarizing time to viral rebound for the participants with HIV-1

RNA <20 copies per ml two weeks before and at the start of ATI ($n = 11$, left), for the participants sensitive to both antibodies ($n = 9$, centre), and for the participants that showed pre-existing resistance to one of the antibodies ($n = 2$, right). The y axis shows the percentage of participants that maintain viral suppression. The x axis shows the time in weeks after start of ATI. Participants receiving the combination of 3BNC117 and 10-1074 are indicated by the blue line. Dotted red lines indicate a cohort of individuals receiving 3BNC117 alone during ATI⁹ ($n = 13$) and dotted black lines indicate a cohort of participants who underwent ATI without intervention¹⁰ ($n = 52$).

three-week intervals beginning two days before treatment interruption (Fig. 1a). Individuals whose regimens contained non-nucleoside reverse transcriptase inhibitors were switched to an integrase inhibitor-based regimen four weeks before discontinuing ART (Extended Data Figs. 1b, 2a). Viral load and CD4⁺ T cell counts were monitored every 1–2 weeks (Supplementary Table 2). ART was reinitiated and antibody infusions were discontinued if viraemia of >200 copies per ml was confirmed. Time of viral rebound was defined as the first of two consecutive viral loads of >200 copies per ml. Fifteen individuals were enrolled, but four of them showed viral loads of >200 copies per ml two weeks before or at the time of the first bNAb infusion and they were excluded from efficacy analyses (Extended Data Fig. 1b and Supplementary Table 2).

Antibody infusions were generally safe and well-tolerated with no reported serious adverse events or antibody-related adverse events, except for mild fatigue in two participants (Supplementary Table 3). The mean CD4⁺ T cell count was 685 and 559 cells per μ l at the time of first antibody infusion and at rebound, respectively (Extended Data Fig. 2b and Supplementary Table 2). Reinitiation of ART after viral rebound resulted in resuppression of viraemia (Supplementary Table 2). We conclude that combination therapy with 3BNC117 and 10-1074 is generally safe and well-tolerated.

The serum half-life of each antibody was measured independently by TZM-bl assay and anti-idiotype enzyme-linked immunosorbent assay (ELISA, Extended Data Fig. 2c, d and Supplementary Table 2). 3BNC117 had a half-life of 12.5 and 17.6 days as measured using TZM-bl and ELISA, respectively (Extended Data Fig. 2c, d). The half-life of 10-1074 was 19.1 and 23.2 days as measured by TZM-bl and ELISA, respectively; significantly longer than 3BNC117 in both assays ($P = 0.0002$ and $P = 0.02$, Extended Data Fig. 2e, f). These measurements are similar to those observed when each antibody was administered alone in ART-treated HIV-1-infected individuals^{6,8,9}. We conclude that the pharmacokinetic profiles of 3BNC117 and 10-1074 are not altered when they are used in combination.

The combination of bNAbs maintains viral suppression

For the 11 individuals who had complete viral suppression (HIV-1 RNA <20 copies per ml) during the screening period and at day 0, combination antibody therapy was associated with maintenance of viral suppression for between 5 and more than 30 weeks (Fig. 1b, c and Supplementary Table 2). The median time to rebound was 21 weeks compared to 2.3 weeks for historical controls who participated in previous non-interventional ATI studies¹⁰ and 6–10 weeks for monotherapy with 3BNC117⁹ (Fig. 1c). Together, 9 of the 11 participants maintained viral

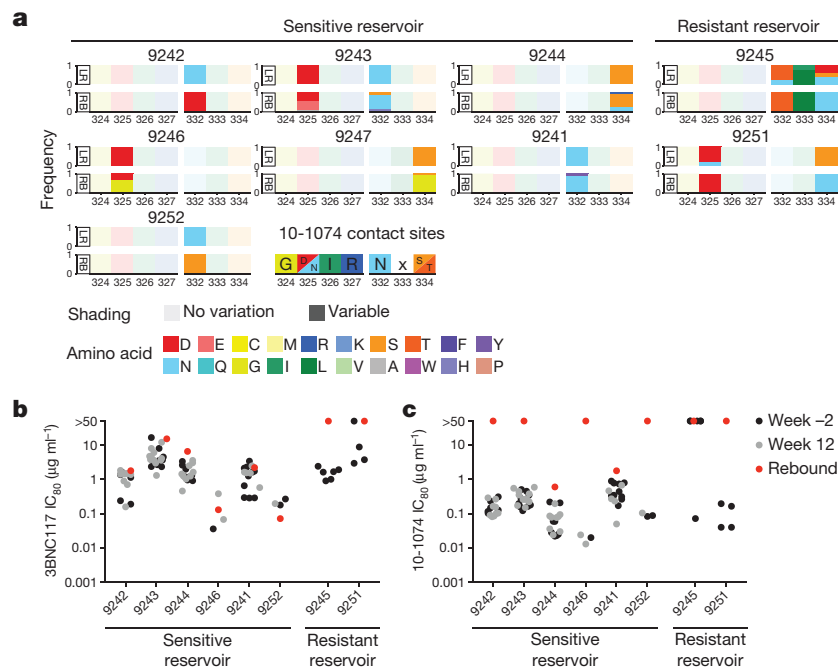


Fig. 2 | Amino acid variants at 10-1074 contact sites and bNAb sensitivity of reactivated latent and rebound viruses. **a**, Colour charts show Env contact sites of 10-1074 at the G(D/N)IR motif (positions 324–327, according to HXB2 numbering) and the glycan at the potential N-linked glycosylation site at position 332 (NxS/T motif at positions 332–334). Diagram shows the seven bNAb-sensitive participants that rebounded before week 30 (left) and the two individuals with pre-existing resistance to one of the two antibodies (right). LR, latent reservoir viruses isolated by Q²VOA; RB, rebound viruses isolated by SGA (plasma) or viral outgrowth (PBMCs). Each amino acid is represented by a colour and the frequency of each amino acid is indicated by the height of

rectangle. Shaded rectangles indicate the lack of variation between latent reservoir virus and rebound virus at the indicated position. Full-colour rectangles represent amino acid residues with changes in distribution between reservoir and rebound viruses. **b**, **c**, Dot plots indicating IC₈₀ (µg ml⁻¹) of 3BNC117 (**b**) and 10-1074 (**c**) against latent and rebound viruses determined by TZM-bl neutralization assay. Q²VOA-derived latent viruses from week -2 and week 12 are shown as black and grey circles, respectively. For outgrowth culture-derived rebound viruses, the highest IC₈₀ is shown as red circle. For 9246, 9252, 9245 and 9251 viruses could not be obtained from rebound outgrowth cultures and pseudoviruses were made from *env* sequences from Q²VOA and plasma SGA.

suppression for at least 15 weeks, although two rebounded at weeks 5 and 7 (Fig. 1b, c).

Quantitative and qualitative viral outgrowth assays (Q²VOA) were used to retrospectively analyse the replication-competent latent viral reservoir in all individuals. Phylogenetic analysis showed that the trial participants were infected with epidemiologically distinct clade B viruses (Extended Data Fig. 3). Q²VOA analysis revealed that the pre-infusion latent reservoir in the two individuals who rebounded early, 9245 and 9251, harboured 10-1074- or 3BNC117-resistant viruses, respectively (Fig. 2 and Supplementary Table 4). Therefore, these two individuals were effectively subjected to antibody monotherapy, because there was pre-existing resistance in the reservoir of these individuals to one of the two bNAbs. Consistent with this idea, the delay in rebound in these two participants was within the anticipated range of antibody monotherapy^{9,11} (Fig. 1c). In addition, all four of the individuals excluded from the analysis due to incomplete viral suppression showed pre-existing resistance or viruses that were not fully neutralized by one or both of the antibodies and these individuals rebounded before week 12 (Extended Data Figs. 4, 5 and Supplementary Table 4).

To examine the viruses that arose in the early rebounding individuals, we performed single genome analysis (SGA) of plasma viruses obtained at the time of rebound. In addition to the pre-existing sequences associated with resistance in the 10-1074 target site (N332T and S334N, Fig. 2a), rebound viruses in 9245 also carried an extended V5 loop and potential N-linked glycosylation sites that could interfere with 3BNC117 binding (Extended Data Fig. 6). Conversely, genetic features associated with resistance to 3BNC117 were found in the pre-infusion reservoir of 9251 and were accompanied by mutations in the 10-1074 target site in the rebounding viruses (S334N, Fig. 2a and Extended

Data Fig. 6). For both individuals, resistance of rebound viruses to both antibodies was confirmed by the TZM-bl neutralization assay (Fig. 2b, c and Supplementary Table 4). Thus, bulk outgrowth cultures used for screening failed to detect pre-existing resistance in the reservoir of 2 of the 11 studied individuals. This result is not surprising given that bulk cultures are dominated by a limited number of rapidly growing viral species that may not be representative of the diversity of the latent reservoir.

The median time to rebound in the seven individuals that had no detectable resistant viruses in the pre-infusion latent reservoir, and rebounded during the study period, was also 21 weeks and different from the 6–10 weeks found for monotherapy with 3BNC117⁹ (Fig. 1c). In these participants, viral suppression was maintained for 15–26 weeks after ART discontinuation (Supplementary Table 2). The two remaining participants (9254 and 9255) completed study follow-up at 30 weeks without experiencing rebound (Supplementary Table 2). Notably, viral rebound never occurred when the concentration of both administered antibodies was above 10 µg ml⁻¹. The average serum concentration of 3BNC117 (determined by TZM-bl assay) at the time of rebound in sensitive individuals that rebounded during study follow-up was 1.9 µg ml⁻¹ (Fig. 1b and Supplementary Table 2). By contrast, the average serum concentration of 10-1074 at rebound was 14.8 µg ml⁻¹ (Fig. 1b and Supplementary Table 2). The difference in the antibody concentrations at the time of rebound is consistent with the longer half-life of 10-1074, which resulted in a period of 10-1074 monotherapy (Fig. 1b, Extended Data Fig. 2c–f and Supplementary Table 2). Finally, these nine individuals showed little or no pre-existing neutralizing antibodies against a diagnostic panel of viruses before bNAb infusion (Supplementary Table 5).

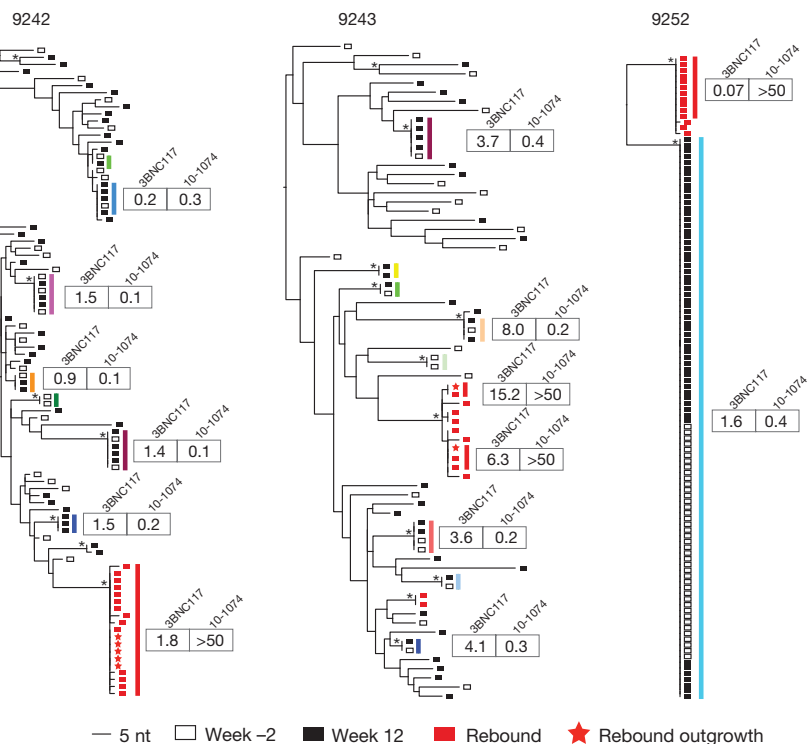


Fig. 3 | Comparison of the circulating latent reservoir and rebound viruses. Maximum likelihood phylogenetic trees of full-length *env* sequences of viruses isolated from Q²VOA, rebound plasma SGA and rebound PBMC outgrowth cultures from three out of seven participants (9242, 9243 and 9252) that rebounded before week 30 (9241, 9244, 9247 and 9246 are depicted in Extended Data Fig. 7). Open and closed black rectangles indicate Q²VOA-derived viruses from week -2 and week 12, respectively. Viruses obtained at the time of rebound are indicated by

red rectangles (plasma SGA) and red stars (rebound PBMC outgrowth cultures). Asterisks indicate nodes with significant bootstrap values (bootstrap support $\geq 70\%$). Clones are denoted by coloured lines mirroring the colours of slices in Extended Data Fig. 10a. Boxes indicate IC₈₀ values ($\mu\text{g ml}^{-1}$) of 3BNC117 and 10-1074 against representative viruses throughout the phylogenetic tree and clones, when possible (Supplementary Table 4). nt, nucleotide.

Rebound and latent viruses

To examine the relationship between rebound viruses and the circulating latent reservoir, we compared *env* sequences obtained from plasma rebound viruses by SGA with sequences obtained by Q²VOA from both pre-infusion and week 12 samples. In addition, we measured the sensitivity of rebound outgrowth viruses and/or pseudoviruses to 3BNC117 and 10-1074 using the TZM-bl neutralization assay (Fig. 2b, c, 3, Extended Data Fig. 7 and Supplementary Table 4). A total of 154 viral *env* sequences obtained by plasma SGA were analysed and compared to 408 sequences obtained from the latent reservoir by Q²VOA. Although rebound and reservoir viruses clustered together for each individual (Extended Data Fig. 3), we found no identical sequences between the two compartments in any of the individuals studied (Figs. 3, 4a and Extended Data Fig. 7). The difference could be accounted for by distinct requirements for HIV-1 reactivation in vitro and in vivo, compartmentalization of reservoir viruses, HIV-1 mutation during the course of the trial and/or by viral recombination in some individuals^{20,21} (Extended Data Fig. 8). Whether or not bNAb therapy influences selection for recombination events remains to be determined.

Similar to 3BNC117 monotherapy, the vast majority of rebounding viruses clustered within low-diversity lineages consistent with expansion of 1–2 recrudescence viruses⁹ (Fig. 3, Extended Data Figs. 7, 9). By contrast, rebound viruses are consistently polyclonal during ATI in the absence of antibody therapy^{22,23}. Thus, the antibodies restrict the outgrowth of latent viruses in vivo.

The emerging viruses in 6 of the 7 individuals who rebounded when the mean 3BNC117 and 10-1074 serum concentrations were 1.9 and 14.8 $\mu\text{g ml}^{-1}$, respectively, carried resistance-associated mutations in the 10-1074 target site (Figs. 1b, 2a). Consistent with the sequence data,

these rebound viruses were generally resistant to 10-1074, as shown by the TZM-bl neutralization assay, but remained sensitive to 3BNC117 (Fig. 2b, c and Supplementary Table 4). The level of sensitivity to 3BNC117 in these emerging viruses was similar to that found in the reservoir viruses in each of the individuals (Fig. 2b and Supplementary Table 4). One individual, 9244, showed rebound viruses that remained sensitive to both antibodies in TZM-bl neutralization assays. Rebound occurred when 3BNC117 and 10-1074 concentrations in serum of this individual were undetectable and 11.6 $\mu\text{g ml}^{-1}$, respectively (Fig. 1b and Supplementary Table 2). The sensitivity of the plasma rebound viruses was similar to that of latent pre-infusion and week 12 viruses obtained in viral outgrowth cultures (Fig. 2b, c and Supplementary Table 4). Therefore, this individual did not develop resistance to either of the antibodies despite prolonged exposure to both. In conclusion, none of the nine individuals with pre-infusion reservoirs containing viruses that were sensitive to both antibodies developed double resistance during the observation period.

The latent reservoir

To determine whether there were changes in the circulating reservoir during the observation period, we compared the results of Q²VOA assays performed at entry and 12 weeks after the start of ATI for 8 of the 9 individuals that remained suppressed for at least 12 weeks (Fig. 4 and Extended Data Fig. 10). Similar to previous reports, 63% of all viruses obtained by Q²VOA belonged to expanded clones^{20,24–26} (Extended Data Fig. 10a, b). Comparison of the *env* sequences of the viruses that emerged in outgrowth cultures revealed that 60% of the sequences could be found at both time points. However, there were numerous examples of clones that appeared or disappeared between

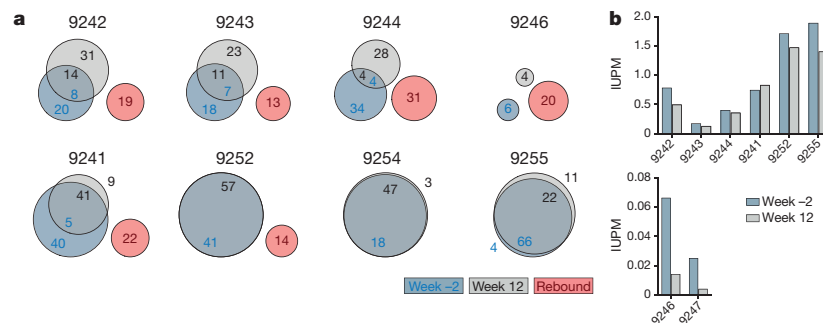


Fig. 4 | Distribution of the circulating latent reservoir and rebound viruses. **a**, Venn diagrams showing sequence identity between *env* sequences obtained from Q²VOA at week -2 (blue) and week 12 (grey), and plasma SGA or rebound PBMC outgrowth culture at the time of viral rebound (red). Area of overlap is proportional to the number of identical sequences. The number of obtained sequences is indicated. **b**, IUPM

CD4⁺ T cells at weeks -2 and 12 as determined by Q²VOA. Participants with IUPMs that were higher and lower than 0.1 are shown at the top and bottom, respectively. Participant 9254 is not shown owing to lack of sample availability. The two time points were not statistically different ($P = 0.078$ (paired Student's *t*-test)).

the time points and some of the changes were significant (Extended Data Fig. 10a). To determine the number of infectious units per million (IUPM, <http://silicianolab.johnshopkins.edu/>), 6.0×10^7 – 6.2×10^8 CD4⁺ T cells were assayed by Q²VOA for each time point for each individual (Fig. 4b). The difference between the two time points was never greater than 6.5-fold for any individual, and the IUPM values at the two time points were not statistically different ($P = 0.078$). Moreover, time to rebound was not directly correlated with IUPM (Extended Data Fig. 10c). Additional time points would be required to calculate the half-life of the reservoir in individuals who received immunotherapy²⁷.

Discussion

First-generation anti-HIV-1 bNAbs were generally ineffective in suppressing viraemia in animal models and humans leading to the conclusion that this approach should not be pursued^{17,18,28}. The advent of new methods for anti-HIV-1 antibody cloning²⁹ and subsequent discovery of a new, more potent generation of bNAbs revitalized this area of research^{30,31}.

bNAb monotherapy with 3BNC117 or VRC01 is not enough to maintain control during ATI in HIV-1-infected humans^{9,11}. Similar results were obtained in participant 9251 who effectively received 10-1074 monotherapy due to pre-existing resistance to 3BNC117. By contrast the combination of 3BNC117 and 10-1074 is sufficient to maintain viral suppression in sensitive individuals when the concentration of both antibodies remains above of $10 \mu\text{g ml}^{-1}$ in serum. Rebound occurred when 3BNC117 levels dropped below $10 \mu\text{g ml}^{-1}$ effectively leading to 10-1074 monotherapy, from which viruses in nearly all individuals rapidly escaped by mutations in the 10-1074 contact site. The observation that nine individuals infected with distinct viruses were unable to develop viruses who were resistant to both antibodies over a median period of 21 weeks suggests that viral replication was severely limited by this combination of antibodies.

In human studies, monotherapy with 3BNC117 is associated with enhanced humoral immunity and accelerated clearance of HIV-1-infected cells^{5,32}. In addition, when administered early to macaques infected with the chimeric simian/human immunodeficiency virus SHIV_{AD8}, combined 3BNC117 and 10-1074 immunotherapy induced host CD8⁺ T cell responses that contributed to the control of viraemia in nearly 50% of the animals³. However, virus-specific CD8⁺ T cells that were responsible for control of viraemia in these macaques were not detected in the circulation, and their contribution to viral suppression was only documented after CD8⁺ T cell depletion³. In most macaques that maintained viral control, complete viral suppression was only established after rebound viraemia that followed antibody clearance³.

Two individuals in this study remained suppressed for over 30 weeks after ATI, 9254 and 9255. Neither participant had detectable levels of

ART in the blood or carried the B*27 and B*57 HLA alleles that are most frequently associated with elite control³³. The first, 9254, reports starting ART within 4–5 months after probable exposure to the virus with an initial viral load of 860,000 copies per ml. Despite relatively early therapy, and excellent virological control for 21 years on therapy, this individual had an IUPM of 0.68 by Q²VOA at the 12-week time point (Extended Data Fig. 10b). The second individual, 9255, showed several viral blips that were spontaneously controlled beginning 15 weeks after ATI when antibody levels were waning. This individual was infected for at least 7 months before starting ART with an initial viral load of 85,800 copies per ml and had an IUPM of 1.4 at the 12-week time point. A small fraction of individuals on ART¹⁰ show spontaneous prolonged virologic control after ART is discontinued, and this number appears to increase when ART treatment is initiated during the acute phase of infection^{34–38}. Whether antibody-enhanced CD8⁺ T cell responses contribute to the prolonged control in the two out of nine individuals who received combination immunotherapy and whether this effect can be enhanced by latency reactivating agents or immune checkpoint inhibitors remains to be determined.

A substantial fraction of the circulating latent reservoir is composed of expanded clones of infected T cells^{24,26,39–42}. These T cell clones appear to be dynamic in that the specific contribution of individual clones of circulating latently infected CD4⁺ T cells to the reservoir of individuals receiving ART fluctuates over time^{24,25}. Individuals that maintain viral suppression by antibody therapy appear to show similar fluctuations in reservoir clones that do not appear to be associated with antibody sensitivity. Whether the apparent differences observed in the reservoir during immunotherapy lead to changes in the reservoir half-life cannot be determined from the available data and will require reservoir assessments in additional individuals at multiple time points over an extended observation period.

Individuals harbouring viruses sensitive to 3BNC117 and 10-1074 maintained viral suppression during ATI for a median of almost four months after the final antibody administration. However, HIV-1 is a highly diverse virus with varying levels of sensitivity to specific bNAbs. As a result, maintenance therapy with just the combination of 3BNC117 and 10-1074 would only be possible for the approximately 50% of clade B-infected individuals that are sensitive to both antibodies. This problem may be overcome by addition of or substitution with other antibodies^{14,15,43}, or long-acting small-molecule antiretroviral drugs.

In macaques, the therapeutic efficacy of anti-HIV-1 antibodies is directly related to their half-life^{4,12,13}, which can be extended by mutations that enhance Fc domain interactions with the neonatal Fc receptor^{4,13,44}. These mutations also increase the half-life of antibodies in humans by 2–4-fold⁴⁵. Our data suggest that a single administration of combinations of bNAbs with extended half-lives could maintain suppression for 6–12 months in individuals harbouring sensitive viruses.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0531-2>.

Received: 6 June 2018; Accepted: 30 July 2018;

Published online 26 September 2018.

- Doitsh, G. & Greene, W. C. Dissecting how CD4 T cells are lost during HIV infection. *Cell Host Microbe* **19**, 280–291 (2016).
- Churchill, M. J., Deeks, S. G., Margolis, D. M., Siliciano, R. F. & Swanstrom, R. HIV reservoirs: what, where and how to target them. *Nat. Rev. Microbiol.* **14**, 55–60 (2016).
- Nishimura, Y. et al. Early antibody therapy can induce long-lasting immunity to SHIV. *Nature* **543**, 559–563 (2017).
- Gautam, R. et al. A single injection of crystallizable fragment domain-modified antibodies elicits durable protection from SHIV infection. *Nat. Med.* **24**, 610–616 (2018).
- Schoofs, T. et al. HIV-1 therapy with monoclonal antibody 3BNC117 elicits host immune responses against HIV-1. *Science* **352**, 997–1001 (2016).
- Caskey, M. et al. Viraemia suppressed in HIV-1-infected humans by broadly neutralizing antibody 3BNC117. *Nature* **522**, 487–491 (2015).
- Lynch, R. M. et al. Virologic effects of broadly neutralizing antibody VRC01 administration during chronic HIV-1 infection. *Sci. Transl. Med.* **7**, 319ra206 (2015).
- Caskey, M. et al. Antibody 10-1074 suppresses viremia in HIV-1-infected individuals. *Nat. Med.* **23**, 185–191 (2017).
- Scheid, J. F. et al. HIV-1 antibody 3BNC117 suppresses viral rebound in humans during treatment interruption. *Nature* **535**, 556–560 (2016).
- Li, J. Z. et al. The size of the expressed HIV reservoir predicts timing of viral rebound after treatment interruption. *AIDS* **30**, 343–353 (2016).
- Bar, K. J. et al. Effect of HIV antibody VRC01 on viral rebound after treatment interruption. *N. Engl. J. Med.* **375**, 2037–2050 (2016).
- Shingai, M. et al. Passive transfer of modest titers of potent and broadly neutralizing anti-HIV monoclonal antibodies block SHIV infection in macaques. *J. Exp. Med.* **211**, 2061–2074 (2014).
- Gautam, R. et al. A single injection of anti-HIV-1 antibodies protects against repeated SHIV challenges. *Nature* **533**, 105–109 (2016).
- Klein, F. et al. HIV therapy by a combination of broadly neutralizing antibodies in humanized mice. *Nature* **492**, 118–122 (2012).
- Horwitz, J. A. et al. HIV-1 suppression and durable control by combining single broadly neutralizing antibodies and antiretroviral drugs in humanized mice. *Proc. Natl Acad. Sci. USA* **110**, 16538–16543 (2013).
- Shingai, M. et al. Antibody-mediated immunotherapy of macaques chronically infected with SHIV suppresses viraemia. *Nature* **503**, 277–280 (2013).
- Trkola, A. et al. Delay of HIV-1 rebound after cessation of antiretroviral therapy through passive transfer of human neutralizing antibodies. *Nat. Med.* **11**, 615–622 (2005).
- Mehandru, S. et al. Adjunctive passive immunotherapy in human immunodeficiency virus type 1-infected individuals treated with antiviral therapy during acute and early infection. *J. Virol.* **81**, 11016–11031 (2007).
- Sarzotti-Kelsoe, M. et al. Optimization and validation of the T2M-bl assay for standardized assessments of neutralizing antibodies against HIV-1. *J. Immunol. Methods* **409**, 131–146 (2014).
- Cohen, Y. Z. et al. Relationship between latent and rebound viruses in a clinical trial of anti-HIV-1 antibody 3BNC117. *J. Exp. Med.* **215**, <https://doi.org/10.1084/jem.20180936> (2018).
- Robertson, D. L., Sharp, P. M., McCutchan, F. E. & Hahn, B. H. Recombination in HIV-1. *Nature* **374**, 124–126 (1995).
- Rothenberger, M. K. et al. Large number of rebounding/founder HIV variants emerge from multifocal infection in lymphatic tissues after treatment interruption. *Proc. Natl Acad. Sci. USA* **112**, E1126–E1134 (2015).
- Kearney, M. F. et al. Lack of detectable HIV-1 molecular evolution during suppressive antiretroviral therapy. *PLoS Pathog.* **10**, e1004010 (2014).
- Lorenzi, J. C. et al. Paired quantitative and qualitative assessment of the replication-competent HIV-1 reservoir and comparison with integrated proviral DNA. *Proc. Natl Acad. Sci. USA* **113**, E7908–E7916 (2016).
- Wang, Z. et al. Expanded cellular clones carrying replication-competent HIV-1 persist, wax, and wane. *Proc. Natl Acad. Sci. USA* **115**, E2575–E2584 (2018).
- Hosmane, N. N. et al. Proliferation of latently infected CD4⁺ T cells carrying replication-competent HIV-1: potential role in latent reservoir dynamics. *J. Exp. Med.* **214**, 959–972 (2017).
- Crooks, A. M. et al. Precise quantitation of the latent HIV-1 reservoir: implications for eradication strategies. *J. Infect. Dis.* **212**, 1361–1365 (2015).
- Poignard, P. et al. Neutralizing antibodies have limited effects on the control of established HIV-1 infection in vivo. *Immunity* **10**, 431–438 (1999).
- Scheid, J. F. et al. A method for identification of HIV gp140 binding memory B cells in human blood. *J. Immunol. Methods* **343**, 65–67 (2009).
- Escolano, A., Dosenovic, P. & Nussenzweig, M. C. Progress toward active or passive HIV-1 vaccination. *J. Exp. Med.* **214**, 3–16 (2017).
- Kwong, P. D. & Mascola, J. R. HIV-1 vaccines based on antibody identification, B cell ontogeny, and epitope structure. *Immunity* **48**, 855–871 (2018).
- Lu, C. L. et al. Enhanced clearance of HIV-1-infected cells by broadly neutralizing antibodies against HIV-1 in vivo. *Science* **352**, 1001–1004 (2016).
- Walker, B. D. & Yu, X. G. Unravelling the mechanisms of durable control of HIV-1. *Nat. Rev. Immunol.* **13**, 487–498 (2013).
- Colby, D. J. et al. Rapid HIV RNA rebound after antiretroviral treatment interruption in persons durably suppressed in Fiebig I acute HIV infection. *Nat. Med.* **24**, 923–926 (2018).
- Sáez-Cirión, A. et al. Post-treatment HIV-1 controllers with a long-term virological remission after the interruption of early initiated antiretroviral therapy ANRS VISCONTI Study. *PLoS Pathog.* **9**, e1003211 (2013).
- Sneller, M. C. et al. A randomized controlled safety/efficacy trial of therapeutic vaccination in HIV-infected individuals who initiated antiretroviral therapy early in infection. *Sci. Transl. Med.* **9**, eaan8848 (2017).
- Fidler, S. et al. Virological blips and predictors of post treatment viral control after stopping ART started in primary HIV infection. *J. Acquir. Immune Defic. Syndr.* **74**, 126–133 (2017).
- Martin, G. E. et al. Post-treatment control or treated controllers? Viral remission in treated and untreated primary HIV infection. *AIDS* **31**, 477–484 (2017).
- Cohn, L. B. et al. Clonal CD4⁺ T cells in the HIV-1 latent reservoir display a distinct gene profile upon reactivation. *Nat. Med.* **24**, 604–609 (2018).
- Maldarelli, F. et al. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345**, 179–183 (2014).
- Wagner, T. A. et al. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345**, 570–573 (2014).
- Cohn, L. B. et al. HIV-1 integration landscape during latent and active infection. *Cell* **160**, 420–432 (2015).
- Halper-Stromberg, A. et al. Broadly neutralizing antibodies and viral inducers decrease rebound from HIV-1 latent reservoirs in humanized mice. *Cell* **158**, 989–999 (2014).
- Ko, S. Y. et al. Enhanced neonatal Fc receptor function improves protection against primate SHIV infection. *Nature* **514**, 642–645 (2014).
- Gaudinski, M. R. et al. Safety and pharmacokinetics of the Fc-modified HIV-1 human monoclonal antibody VRC01LS: a phase 1 open-label clinical trial in healthy adults. *PLoS Med.* **15**, e1002493 (2018).

Acknowledgements We thank all study participants who devoted time to our research; members of the Klein and Nussenzweig laboratories for helpful discussions, especially Y. Bar-On, L. Cohn and M. Jankovic; R. Levin for study coordination and the Rockefeller University Hospital Clinical Research Support Office and nursing staff as well as K. Fiddike, C. Golder, S. Margane, M. Platten, E. Voigt and D. Weiland for help with recruitment and study implementation; K. Jain for help with sample processing; S. Kiss for ophthalmologic assessments; T. Keler and the CellDex Therapeutics team for 3BNC117 and 10-1074 manufacturing and regulatory support; C. Conrad for regulatory support; U. Kerkweg, R. Macarthur and A. Johnson for pharmaceutical services; H. Janicki, M. Ercanoglu, P. Schommers and R. Kaiser for help with virus cultures; C. Scheid and U. Holtick for leukaphereses; S. McMillan, S. Mosher, S. Sawant, D. Beaumont, M. Sarzotti-Kelsoe, K. Greene, H. Gao and D. Montefiori for help with PK assay development, validation, reporting, and/or project management; P. Fast and H. Park for clinical monitoring; and S. Schlesinger for input on study design. This work was supported by the Bill and Melinda Gates Foundation Collaboration for AIDS Vaccine Discovery (CAVD) grants OPP1092074, OPP1124068 (M.C.N.), CAVIMC OPP1146996 (G.D.T., M.S.S.); the Heisenberg-Program of the DFG (KL 2389/2-1), the European Research Council (ERC-StG639961), and the German Center for Infection Research (DZIF) (F.K.); the NIH grants 1UM1 AI100663 and R01AI-129795 (M.C.N.); the Einstein-Rockefeller-CUNY Center for AIDS Research (1P30AI124414-01A1); BEAT-HIV Delaney grant UM1 AI126620 (M.C.); and the Robertson fund of the Rockefeller University. M.C.N. is a Howard Hughes Medical Institute Investigator.

Reviewer information *Nature* thanks G. Silvestri and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.C. is the principal investigator for the work in the United States and F.K. is the principal investigator for Germany; M.C., F.K. and M.C.N. designed the trial; P.M., H.G., F.K., M.C. and M.C.N. analysed the data and wrote the manuscript; P.M., L.N. and T.Ka. performed Q²VOA, rebound cultures and SGA; H.G., M.W.-P., G.K., E.T., J.H., M.C. and F.K. implemented the study; A.L.B., K.M., Y.Z.C., C.L., I.Su., C.W., T.Kü. and C.S. contributed to recruitment and clinical assessments; P.M., H.G. and L.N. performed bulk viral cultures; J.A.P. and T.Y.O. performed bioinformatics processing; K.E.S. and G.D.T. conducted anti-idiotypic ELISA; M.S.S. conducted T2M-bl assays; C.U.-O., R.P., C.R., M.S. and I.Sh. coordinated and performed sample processing; L.H., A.P.W., P.J.B. and N.P. contributed to data analysis; J.C.C.L., C.L.L., R.M.G. and G.F. contributed to study design and implementation.

Competing interests : There are patents on 3BNC117 (PTC/US2012/038400) and 10-1074 (PTC/US2013/065696) that list M.C.N. as an inventor.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0531-2>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0531-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to F.K., M.C. or M.C.N.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Study design. An open-label phase 1b study was conducted in HIV-1-infected participants who were virologically suppressed on ART (<http://www.clinicaltrials.gov>; NCT02825797; EudraCT: 2016-002803-25). Study participants were enrolled sequentially according to eligibility criteria. Participants received 3BNC117 and 10-1074 intravenously at a dose of 30 mg kg⁻¹ body weight of each antibody, at weeks 0, 3 and 6, unless viral rebound occurred. ART was discontinued 2 days after the first infusion of antibodies (day 2). Plasma HIV-1 viral RNA levels were monitored weekly and ART was resumed if the viral load increased to ≥ 200 copies per ml or CD4⁺ T cell counts decreased to < 350 cells per μ l in two consecutive measurements. Time to viral rebound was determined by the first of two consecutive viral loads of > 200 copies per ml. Study participants were followed for 30 weeks after the first infusion. Safety data are reported until the end of study follow-up. All participants provided written informed consent before participation in the study and the study was conducted in accordance with Good Clinical Practice. The protocol was approved by the Federal Drug Administration in the USA, the Paul-Ehrlich-Institute in Germany, and the Institutional Review Boards (IRBs) at the Rockefeller University and the University of Cologne.

Study participants. Study participants were recruited at the Rockefeller University Hospital, New York, USA, and the University Hospital Cologne, Cologne, Germany. Eligible participants were adults aged 18–65 years, HIV-1-infected, on ART for a minimum of 24 months, with plasma HIV-1 RNA levels of < 50 copies per ml for at least 18 months (one viral blip of > 50 but < 500 copies per ml during this 18-month period was allowed), plasma HIV-1 RNA levels < 20 copies per ml at the screening visit, and a current CD4⁺ T cell count > 500 cells per μ l. In addition, participants were prescreened for sensitivity of latent proviruses against 3BNC117 and 10-1074 by bulk PBMC viral outgrowth culture as described in 'Prescreening bulk PBMC cultures'. Sensitivity was defined as an IC₅₀ $< 2 \mu$ g ml⁻¹ for both 3BNC117 and 10-1074 against outgrowth virus. Participants on an ART regimen that included a non-nucleoside reverse transcriptase inhibitor (NNRTI) were switched to an integrase inhibitor-based regimen (dolutegravir plus tenofovir disoproxil fumarate and emtricitabine) four weeks before treatment interruption due to the prolonged half-life of NNRTIs. Exclusion criteria included reported CD4⁺ T cell nadir of < 200 cells μ l⁻¹, concomitant hepatitis B or C infection, previous receipt of monoclonal antibodies of any kind, clinically relevant physical findings, medical conditions or laboratory abnormalities, and pregnancy or lactation.

Study procedures. 3BNC117 and 10-1074 were administered intravenously at a dose of 30 mg kg⁻¹. The appropriate stock volume of 3BNC117 and 10-1074 was calculated according to body weight and diluted in sterile normal saline to a total volume of 250 ml per antibody. Monoclonal antibody infusions were administered sequentially and intravenously over 60 min. Study participants were observed at the Rockefeller University Hospital or the University Hospital Cologne for 1 h after the last antibody infusion. Participants returned for weekly follow-up visits during the ATI period for safety assessments, which included physical examination and measurements of clinical laboratory parameters such as haematology, chemistries, urinalysis and pregnancy tests (for women). Plasma HIV-1 RNA levels were monitored weekly during the ATI period and CD4⁺ T cell counts were measured every 1–2 weeks. After ART was re-initiated, participants returned for follow-up every two weeks until viral re-suppression was achieved, and every eight weeks thereafter. Study investigators evaluated and graded adverse events according to the Division of AIDS (DAIDS) Table for Grading the Severity of Adult and Pediatric Adverse Events (version 2.0, November 2014) and determined causality. Leukapheresis was performed at the Rockefeller University Hospital or at the University Hospital Cologne at week -2 and week 12. Blood samples were collected before and at multiple times after 3BNC117 and 10-1074 infusions. Samples were processed within 4 h of collection, and serum and plasma samples were stored at -80°C . PBMCs were isolated by density gradient centrifugation. The absolute number of PBMCs was determined using an automated cell counter (Vi-Cell XR; Beckman Coulter) or manually, and cells were cryopreserved in fetal bovine serum plus 10% DMSO.

Plasma HIV-1 RNA Levels. HIV-1 RNA levels in plasma were measured at the time of screening, at week -2, day 0 (before infusion), weekly during ATI, and every two weeks to every eight weeks after viral rebound had occurred. HIV-1 RNA levels were determined using the Roche COBAS AmpliPrep/COBAS TaqMan HIV-1 Assay (version 2.0) or the Roche COBAS HIV-1 quantitative nucleic acid test (COBAS 6800), which quantify HIV-1 RNA over a range of 2×10^1 to 1×10^7 copies per ml. These assays were performed at LabCorp or at the University Hospital Cologne.

CD4⁺ T cells. CD4⁺ T cell counts were determined by clinical flow cytometry assay, performed at LabCorp or at the University Hospital Cologne, at screening, week 0 (before infusion), weeks 2, 3, 5, 6, 8, 10, and weekly thereafter, while participants remained off ART.

Determination of baseline neutralizing antibody activity. Purified IgG (Protein G Sepharose 4 Fast Flow, GE Life Sciences) obtained before antibody infusions was tested against a panel of 12 HIV-1 pseudoviruses as described previously⁵.

Measurement of 3BNC117 and 10-1074 serum levels. Blood samples were collected before, at the end of each 3BNC117 infusion and at the end of each 10-1074 infusion at weeks 0, 3 and 6, and weekly during the ATI period, up to week 30. Serum levels of 3BNC117 and 10-1074 were determined by a TZM-bl assay and by ELISA from samples obtained before and after each antibody infusion, and approximately every three weeks during follow-up as well as at the time of viral rebound.

Serum concentrations of 3BNC117 and 10-1074 were measured by a validated sandwich ELISA. High-bind polystyrene plates were coated with 4μ g ml⁻¹ of an anti-idiotypic antibody that specifically recognizes 3BNC117 (anti-ID 1F1-2E3 monoclonal antibody) or 2μ g ml⁻¹ of an anti-idiotypic antibody that specifically recognizes 10-1074 (anti-ID 3A1-4E11 monoclonal antibody), and incubated overnight at 2–8 °C. After washing, plates were blocked with 5% Milk Blotto (w/v), 5% NGS (v/v) and 0.05% Tween 20 (v/v) in PBS. Serum samples, quality controls and standards were added (1:50 minimum dilution in 5% Milk Blotto (w/v), 5% NGS (v/v) and 0.05% Tween 20 (v/v) in PBS) and incubated at room temperature. 3BNC117 or 10-1074 were detected using a horseradish peroxidase (HRP)-conjugated mouse anti-human IgG kappa-chain-specific antibody (Abcam) for 3BNC117 or an HRP-conjugated goat anti-human IgG Fc-specific antibody for 10-1074 (Jackson ImmunoResearch) and the HRP substrate tetra-methylbenzidine. 3BNC117 and 10-1074 concentrations were then calculated from the standard curves of 3BNC117 or 10-1074 that were run on the same plate using a 5-PL curve-fitting algorithm (Softmax Pro, v.5.4.5). Standard curves and positive controls were created from the drug product lots of 3BNC117 and 10-1074 used in the clinical study. The capture anti-idiotypic monoclonal antibodies were produced using a stable hybridoma cell line (Duke Protein Production Facility⁶). The lower limit of quantification for the 3BNC117 ELISA is 0.78μ g ml⁻¹ and for the 10-1074 ELISA is 0.41μ g ml⁻¹. The lower limit of detection was determined to be 0.51μ g ml⁻¹ and 0.14μ g ml⁻¹ in HIV-1 seropositive serum for the 3BNC117 and 10-1074 ELISA, respectively. For values that were detectable (that is, positive for the monoclonal antibodies) but were below the lower limit of quantification, values are reported as $< 0.78 \mu$ g ml⁻¹ and $< 0.41 \mu$ g ml⁻¹ for 3BNC117 and 10-1074 ELISA, respectively. If day 0 baseline samples had measurable levels of antibody by the respective assays, the background measured antibody level was subtracted from subsequent results. In addition, samples with antibody levels measured to be within threefold from background were excluded from the analysis of pharmacokinetic parameters.

Serum concentrations of active 3BNC117 and 10-1074 were also measured using a validated luciferase-based neutralization assay in TZM-bl cells as previously described¹⁹. In brief, serum samples were tested using a primary 1:20 dilution with a fivefold titration series against HIV-1 Env pseudoviruses Q769.d22 and X2088_c9, which are highly sensitive to neutralization by 3BNC117 and 10-1074, respectively, while fully resistant against the other administered antibody. In the case of the post-infusion time points of 10-1074, instances for which the serum 50% inhibitory dilution (ID₅₀) titres against X2088_c9 were $> 100,000$, serum samples were also tested against a less sensitive strain, Du422 (Supplementary Table 2). To generate standard curves, clinical drug products of 3BNC117 and 10-1074 were included in every assay set-up using a primary concentration of 10μ g ml⁻¹ with a fivefold titration series. Serum concentrations of 3BNC117 and 10-1074 for each sample were calculated as follows: serum ID₅₀ titre (dilution) \times 3BNC117 IC₅₀ or 10-1074 IC₅₀ titre (μ g ml⁻¹) = serum concentration of 3BNC117 or 10-1074 (μ g ml⁻¹). Env pseudoviruses were produced using an ART-resistant backbone vector that reduces background inhibitory activity of antiretroviral drugs if present in the serum sample (SG3 Δ Env/K101P.Q148H.Y181C; M.S.S., unpublished data). Virus that was pseudotyped with the envelope protein of murine leukaemia virus (MuLV) was used as a negative control. Antibody concentrations were calculated using the serum ID₈₀ titre and monoclonal antibody IC₈₀ if non-specific activity against MuLV was detected (ID₅₀ > 20 ; 9246, week 30; 9248, baseline, day 0, week 18). All assays were performed in a laboratory that meets Good Clinical Laboratory Practice standards.

Prescreening bulk PBMC cultures. To test HIV-1 viral strains for sensitivity to 3BNC117 and 10-1074, we performed bulk viral outgrowth cultures by coculturing isolated CD4⁺ T cells with the MOLT-4/CCR-5 cell line (NIH AIDS Reagent Program, Ca. No. 4984) or CD8⁺ T cell-depleted healthy donor lymphoblasts. PBMCs for prescreening were obtained up to 72 weeks (range 54–505 days) before enrollment under separate protocols approved by the IRBs of the Rockefeller University and the University of Cologne. Sensitivity was determined by TZM-bl neutralization assay as described below. Culture supernatants with IC₅₀ $< 2 \mu$ g ml⁻¹ were deemed sensitive.

Quantitative and qualitative viral outgrowth assay. The Q²VOA was performed using isolated PBMCs from leukapheresis at week -2 and week 12 as previously described²⁴. In brief, isolated CD4⁺ T cells were activated with 1μ g ml⁻¹ phytohaemagglutinin (PHA; Life Technologies) and 100 U ml^{-1} IL-2 (Peprotech) and cocultured with 1×10^6 irradiated PBMCs from a healthy donor in 24-well plates.

A total of 6×10^7 – 6.2×10^8 cells were assayed for each individual at each of the two time points. After 24 h, PHA was removed and 0.1×10^6 MOLT-4/CCR5 cells were added to each well. Cultures were maintained for two weeks, splitting the MOLT-4/CCR5 cells in half seven days after the initiation of the culture and every other day after that. Positive wells were detected by measuring p24 by ELISA. The frequency of latently infected cells was calculated through the IUPM algorithm developed by the Siliciano laboratory (<http://silicianolab.johnshopkins.edu>).

Rebound outgrowth cultures. CD4⁺ T cells isolated from PBMCs from the rebound time points were cultured at limiting dilution exactly as described for Q²VOA. CD4⁺ T cells were activated with T cell activation beads (Miltenyi) at a concentration of 0.5×10^6 beads per 10^6 CD4⁺ T cells and 20 U ml^{-1} of IL-2. Rebound outgrowth cultures were performed using PBMCs from the highest viral load sample (usually the repeat measurement ≥ 200 copies per ml). Viruses for which the sequences matched the SGA *env* sequences, and therefore were identical to those present in plasma, as opposed to potentially reactivated PBMC-derived latent reservoir viruses, were selected to test for neutralization.

Viral sensitivity testing. Supernatants from p24-positive bulk PBMC cultures, rebound PBMC outgrowth cultures and Q²VOA wells were tested for sensitivity to 3BNC117 and 10-1074 by TZM-bl neutralization assay as previously described¹⁹.

Sequencing. HIV-1 RNA extraction and single-genome amplification was performed as previously described⁴⁶. In brief, HIV-1 RNA was extracted from plasma samples or Q²VOA-derived virus supernatants using the MinElute Virus Spin kit (Qiagen) followed by first-strand cDNA synthesis using SuperScript III reverse transcriptase (Invitrogen). cDNA synthesis for plasma-derived HIV-1 RNA was performed using the antisense primer *envB3out* 5'-TTGCTACTTGTGATTGCTCCATGT-3'. *gp160* was amplified using *envB5out* 5'-TAGAGCCCTGGAAGCATCCAGGAAG-3' and *envB3out* 5'-TTGCTACTTGTGATTGCTCCATGT-3' in the first round and in the second round with nested primers *envB5in* 5'-CACCTTAGGCATCTCCATGGCAGGAAGAAG-3' and *envB3in* 5'-GTCTCGAGATACTGCTCCACCC-3'. PCRs were performed using High Fidelity Platinum Taq (Invitrogen) and run at 94°C for 2 min; 35 cycles of 94°C for 15 s, 55°C for 30 s and 68°C for 4 min; and 68°C for 15 min. Second-round PCR was performed with 1 µl of the PCR product from the first round as template and High Fidelity Platinum Taq at 94°C for 2 min; 45 cycles of 94°C for 15 s, 55°C for 30 s and 68°C for 4 min; and 68°C for 15 min. cDNA synthesis for Q²VOA-derived HIV-1 RNA was performed using the antisense primer R3B6R 5'-TGAAGCACTCAAGGCAAGCTTTATTGAGGC-3'. The *env* 3' half-genome was amplified in a single PCR using B3F3 primer 5'-TGGAAAGGTGAAGGGCAGTAGTAATAC-3' and R3B6R primer 5'-TGAAGCACTCAAGGCAAGCTTTATTGAGGC-3'. PCR was performed using High Fidelity Platinum Taq and run at 94°C for 2 min; 45 cycles of 94°C for 15 s, 55°C for 30 s and 68°C for 5 min; and 68°C for 15 min.

Pseudovirus generation. Selected single genome sequences from outgrowth culture supernatants and plasma were used to generate pseudoviruses that were tested for sensitivity to bNAbs in a TZM-bl neutralization assay. To produce the pseudoviruses, plasmid DNA containing the cytomegalovirus (CMV) promoter was amplified by PCR using forward primer 5'-AGTAATCAATTACGGGGTCATTAGTTCAT-3' and reverse primer 5'-CATAGGAGATGCCTAAGCCGGTGGAGCTCTGCTTATATAGACCTC-3'. Individual *env* amplicons were amplified using forward primer 5'-CACC GGCTTAGGCATCTCCTATGGCAGGAAGAA-3' and reverse primer 5'-GTCTCGAGATACTGCTCCACCC-3'. The CMV promoter amplicon was fused to individual purified *env* amplicons by PCR using forward primer 5'-AGTAATCAATTACGGGGTCATTAGTTCAT-3' and reverse primer 5'-ACTTTTGTACCACCTTGCCACCCAT-3'. Overlapping PCR was carried out using the High Fidelity Platinum Taq (Invitrogen) in a 50-µl reaction consisting of 1 ng purified CMV promoter amplicon, 0.125 µl purified *env* SGA amplicon, 400 nM each forward and reverse primers, 200 µM dNTP mix, 1 × Buffer HiFi and

1 µl DNA polymerase mix. PCR was run at 94°C for 2 min; 25 cycles of 94°C for 12 s, 55°C for 30 s and 68°C for 4 min; and 72°C for 10 min. Resulting amplicons were analysed by gel electrophoresis, purified by gel extraction, and cotransfected with pSG3Δ*env* into HEK293T cells to produce pseudoviruses as described previously⁴⁷.

Sequence and phylogenetic analysis. Nucleotide alignments of intact *env* sequences were translation-aligned using ClustalW v.2.1⁴⁸ under the BLOSUM cost matrix. Sequences with premature stop codons and frameshift mutations that fell in the gp120 surface glycoprotein region were excluded from all analyses. Maximum likelihood phylogenetic trees were then generated from these alignments with PhyML v.3.1⁴⁹ using the GTR model with 1,000 bootstraps. For the combined analysis of sequences from all participants, *env* sequences were aligned using MAFFT v.7.309 and clustered using RAxML v.8.2.9⁵⁰ under the GTRGAMMA model with 1,000 bootstraps. To analyse changes between reservoir and rebound viruses, *env* sequences were aligned at the amino acid level to a HXB2 reference using ClustalW v.2.1.

Statistical analyses. For sample size considerations, one-sided Clopper-Pearson confidence intervals were calculated for varying number of observed rebounds. A sample size of 15 HIV-1 infected individuals was determined to allow for the rejection of the null hypothesis (rate = 0.85) with 80% power for an effect size equal to or higher than 0.33, if at least 6 out of 15 enrolled participants did not experience viral rebound by week 8 (2 weeks after the last antibody infusions). Pharmacokinetic parameters were estimated by performing a non-compartmental analysis using Phoenix WinNonlin Build 8 (Certara), using all available PK data starting with the time point after the last infusion of 10-1074 from either TZM-bl assay (using the X2088_c9 pseudovirus to determine 10-1074 levels) or ELISA, and compared by a two-tailed unpaired Student's *t*-test. CD4⁺ T cell counts on day 0 and at the time of viral rebound were compared by two-tailed paired Student's *t*-test. IUPMs determined at week -2 and week 12 were compared using a two-tailed paired Student's *t*-test. Time to rebound in current trial participants (combination therapy with 3BNC117 and 10-1074), participants receiving 3BNC117 monotherapy⁹ and participants in previous non-interventional ATI studies conducted by ACTG¹⁰ were plotted using Kaplan–Meier survival curves. Potential correlation between IUPM and time to rebound was analysed by two-tailed Pearson's correlations.

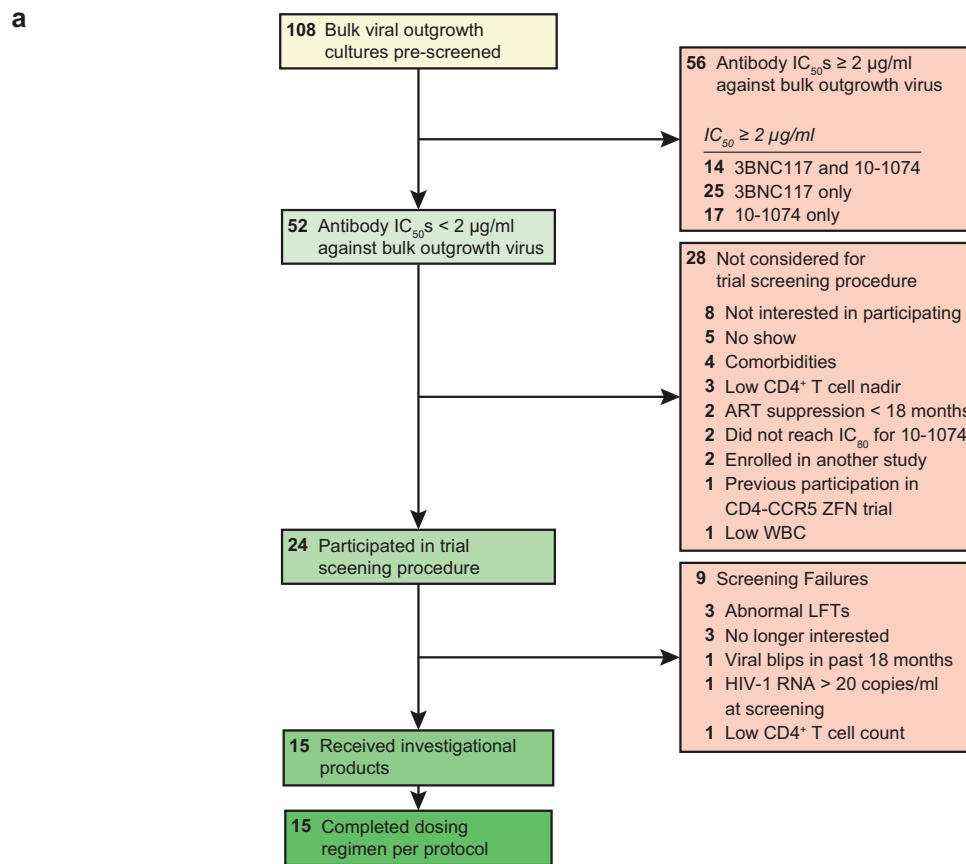
Recombination analysis of *env* sequences. Multiple alignment of nucleotide sequences guided by amino acid translations of *env* sequences was performed by TranslatorX (<http://translatorx.co.uk/>). Latent and rebound sequences were analysed for the presence of recombination using the 3SEQ recombination algorithm (<http://mol.ax/software/3seq/>). Sequences that showed statistical evidence of recombination (rejection of the null hypothesis of clonal evolution) in which 'parent' sequences were derived from the latent reservoir and the 'child' sequence was a rebound sequence are represented in a circos plot (<http://circos.ca/>).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The sequences from all isolated viruses are available in GenBank, accession numbers MH575375–MH576416.

- Salazar-Gonzalez, J. F. et al. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J. Virol.* **82**, 3952–3970 (2008).
- Kirchherr, J. L. et al. High throughput functional analysis of HIV-1 *env* genes without cloning. *J. Virol. Methods* **143**, 104–111 (2007).
- Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
- Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
- Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).



b

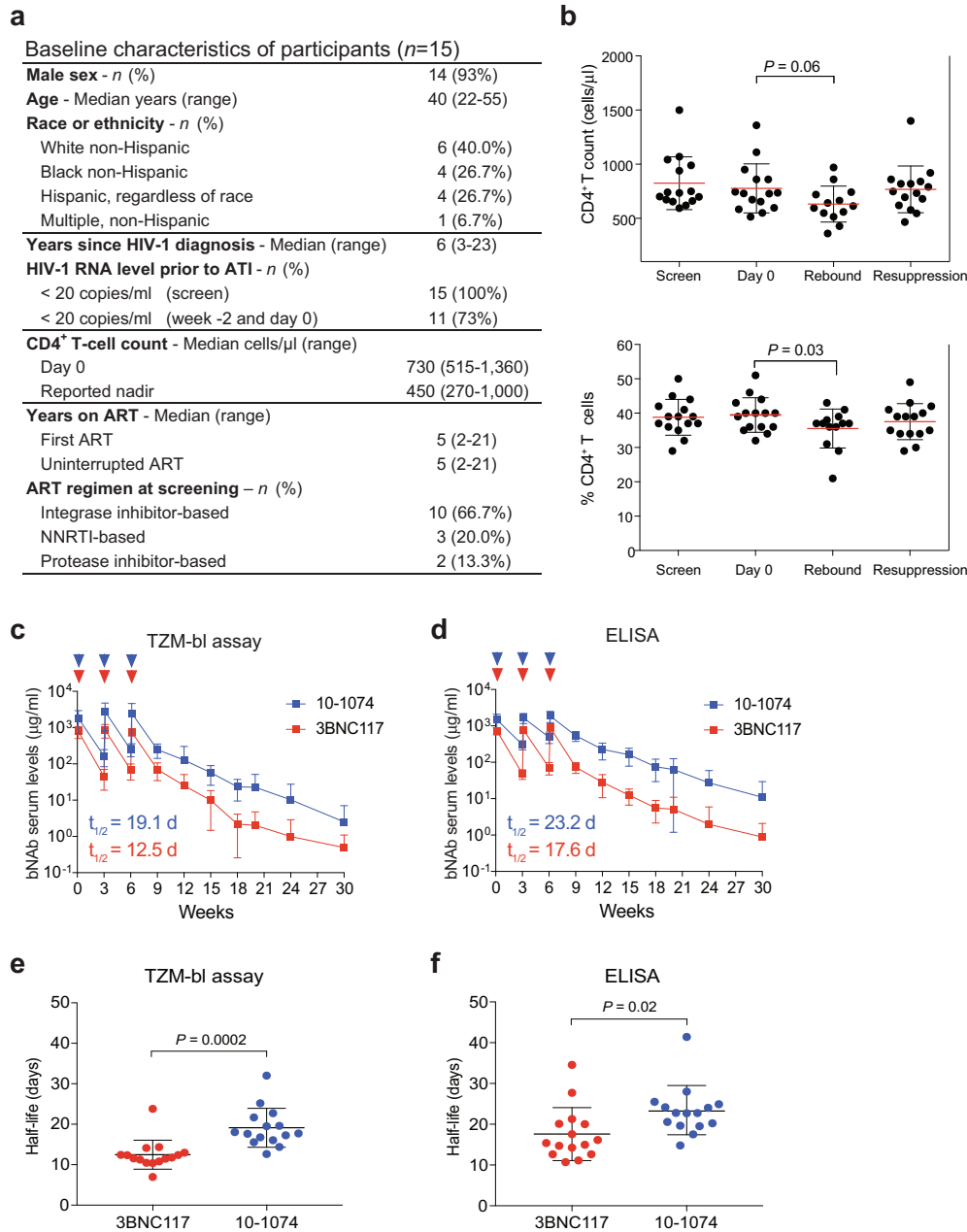
ID	Age	Gender	Race	Years since			Uninterr. ART before ATI (yrs)	ART at Screening*	Switched ART**	Reported CD4 nadir	HLA alleles	Pre-Screen Sensitivity (µg/ml)***				CD4 count (d0)	HIV-1 RNA (cp/ml)#			Weeks to viral rebound
				HIV-1 dx	first ART	ATI						3BNC117	10-1074	IC ₅₀	IC ₈₀		IC ₉₀	IC ₉₅	Scr	
9241	40	M	White/Hisp	6	5	5	EVG/cobi/TDF/FTC	-	500	n.d.	0.809	2.212	0.090	0.243	515	<20	<20	<20	21	
9242	43	M	White/Hisp	3	3	2	EVG/cobi/TDF/FTC	-	450	n.d.	0.160	0.433	0.144	0.389	654	<20	<20	<20	15	
9243	29	M	Amer Indian/Hisp	5	5	5	RPV/TDF/FTC	DTG/TDF/FTC	350	n.d.	0.641	2.913	0.072	0.241	583	<20	<20 D	<20 D	20	
9244	36	M	Amer Indian/not Hisp	9	5	5	EFV/TDF/FTC	DTG/TDF/FTC	730	n.d.	0.277	0.966	0.025	0.068	1,110	<20	<20	<20	21	
9245	22	M	White/Hisp	5	5	5	EVG/cobi/TAF/FTC	-	360	n.d.	0.417	1.423	0.038	0.089	736	<20	<20	<20	5	
9246	30	M	Black	5	5	5	EVG/cobi/TAF/FTC	-	500	n.d.	0.144	0.387	0.040	0.105	745	<20	<20	<20 D	19	
9247	31	M	Black	6	6	6	EVG/cobi/TAF/FTC	-	600	n.d.	0.556	1.930	0.072	0.326	728	<20	<20	<20	26	
9248	52	M	White	11	11	11	DRV/RTV/TAF/FTC	-	310	n.d.	1.863	9.738	0.676	2.252	730	<20 D	<20 D	58	12	
9249	49	M	White	23	20	6	DRV/RTV/ABC/3TC	-	426	n.d.	0.562	2.095	0.260	0.983	860	<20 D	<20 D	32	3	
9250	55	M	White	7	5	5	EVG/cobi/TAF/FTC	-	350	n.d.	0.558	3.174	0.447	2.644	550	<20	40	50	6	
9251	40	M	Black	6	2	2	EVG/cobi/TDF/FTC	-	1,000	n.d.	1.200	3.125	0.073	0.153	672	<20	<20	<20	7	
9252	51	F	Black	11	11	11	EFV/TDF/FTC	DTG/TDF/FTC	270	n.d.	0.630	3.074	0.243	0.640	598	<20	<20	<20	22	
9253	41	M	White	5	2	2	DTG/TAF/FTC	-	387	n.d.	0.558	2.644	0.020	0.317	950	<20	41	<20 D	5	
9254	48	M	White	21	21	21	EVG/cobi/TAF/FTC	-	590	A1,29 B38,44	0.142	0.386	0.085	0.240	860	<20	<20	<20	>30	
9255	30	M	White	5	4	4	EVG/cobi/TAF/FTC	-	779	A3,25 B18,44	0.324	0.833	0.006	0.015	1,360	<20	<20 D	<20	>30	

Extended Data Fig. 1 | Study participant selection and demographics.

a, Flow diagram indicating the selection of study participants.

b, Individual participant demographics and baseline clinical characteristics. Grey-shaded rows indicate participants who were found to have detectable viraemia (HIV-1 viral load of >20 copies ml⁻¹) at week -2 or day 0. These participants were not included in the efficacy analyses given the lack of viral suppression at baseline. Amer Indian, American Indian; Hisp, Hispanic. *3TC, lamivudine; ABC, abacavir; cobi, cobicitast; DRV, darunavir; DTG, dolutegravir; EFV, efavirenz; EVG, elvitegravir;

FTC, emtricitabine; RPV, rilpivirine; RTV, ritonavir; TAF, tenofovir alafenamide fumarate; TDF, tenofovir disoproxil fumarate. **NNRTI-based regimens were switched four weeks before ART interruption due to longer half-lives of NNRTIs. ***Pre-screening of bulk outgrowth virus obtained from PBMC cultures by TZM-bl assay. #All participants harboured clade B viruses. Viral load <20 D, plasma HIV-1 RNA detected but not quantifiable by clinical assay. d0, day 0; Dx, diagnosis; Scr, screening; Wk -2, week -2.



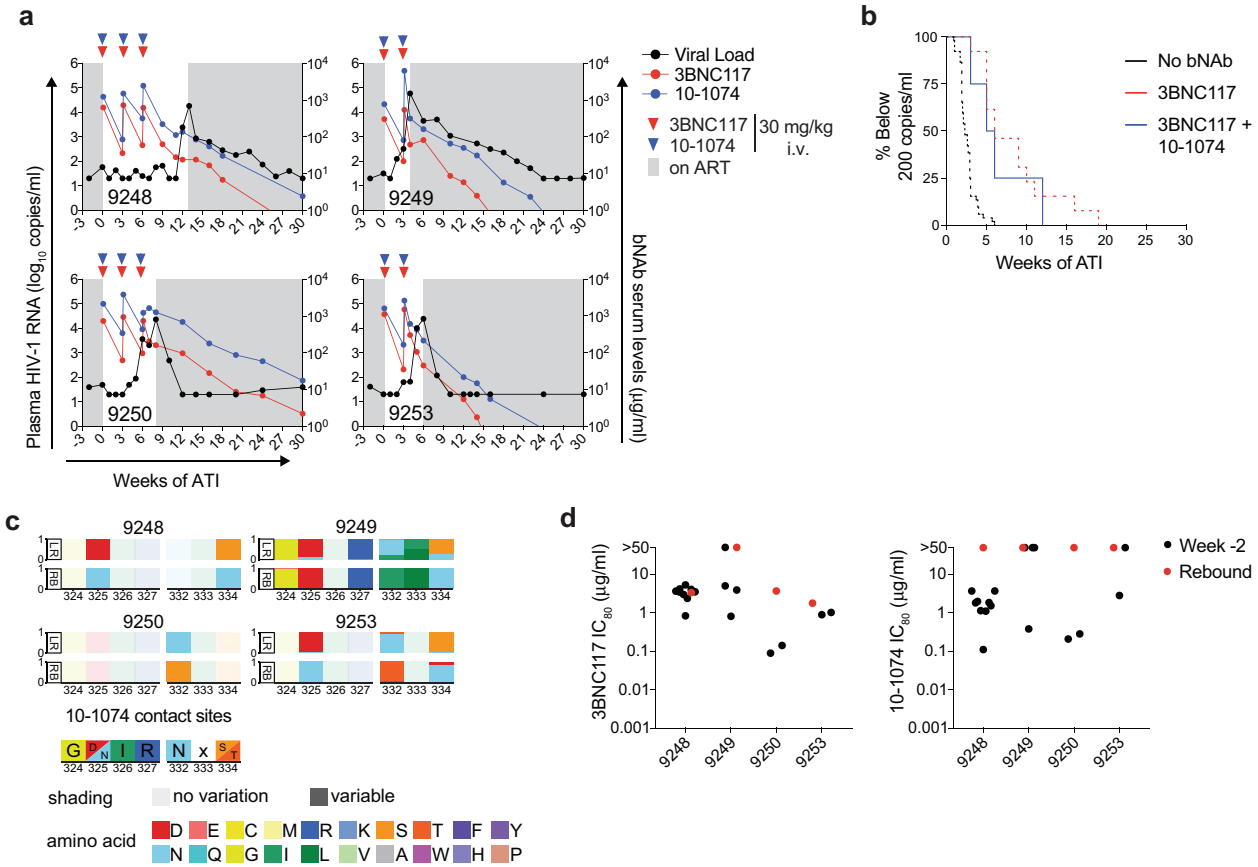
Extended Data Fig. 2 | Demographics, CD4⁺ T cells during study period in participants and pharmacokinetics of 3BNC117 and 10-1074.

a, Baseline participant demographics. **b**, Absolute CD4⁺ T cell counts and percentage of CD4⁺ T cells among CD3⁺ T cells at screening ($n = 15$), day 0 ($n = 15$), at the time of viral rebound ($n = 13$) and at the end of the study are shown ($n = 15$) (see also Supplementary Table 2). The last available time point after resuppression was used as end of the study time point for the participants that reinitiated ART. Red lines indicate mean, error bars indicate standard deviation and individual participants are shown as dots. P values were obtained using a two-tailed paired Student's t -test comparing CD4⁺ T cell counts between day 0 and the time of viral rebound. **c**, **d**, 3BNC117 (red) and 10-1074 (blue) levels in serum ($n = 15$) as determined by TZM-bl assay (**c**) and ELISA (**d**). In cases in which participants only received 2 infusions due to early viral rebound (9245,

9249 and 9253), only antibody concentrations up to the second infusion were included. Half-life of each bNAbs is indicated in days. Curves indicate mean serum antibody concentrations and error bars represent standard deviation. Red and blue triangles indicate 3BNC117 and 10-1074 infusions, respectively. **c**, In the TZM-bl assay, lower limits of quantification were $0.46 \mu\text{g ml}^{-1}$ and $0.10 \mu\text{g ml}^{-1}$ for 3BNC117 and 10-1074, respectively. **d**, In the ELISA, lower limits of detection were $0.78 \mu\text{g ml}^{-1}$ and $0.41 \mu\text{g ml}^{-1}$, respectively. **e**, **f**, Half-lives of both antibodies as measured by TZM-bl assay (**e**) and ELISA (**f**). Each dot represents a single participant. The half-lives of both antibodies from the 15 participants enrolled in the study are represented. Black lines indicate the mean value and standard deviation ($n = 15$). P values were obtained using a two-tailed unpaired Student's t -test comparing the two antibodies.

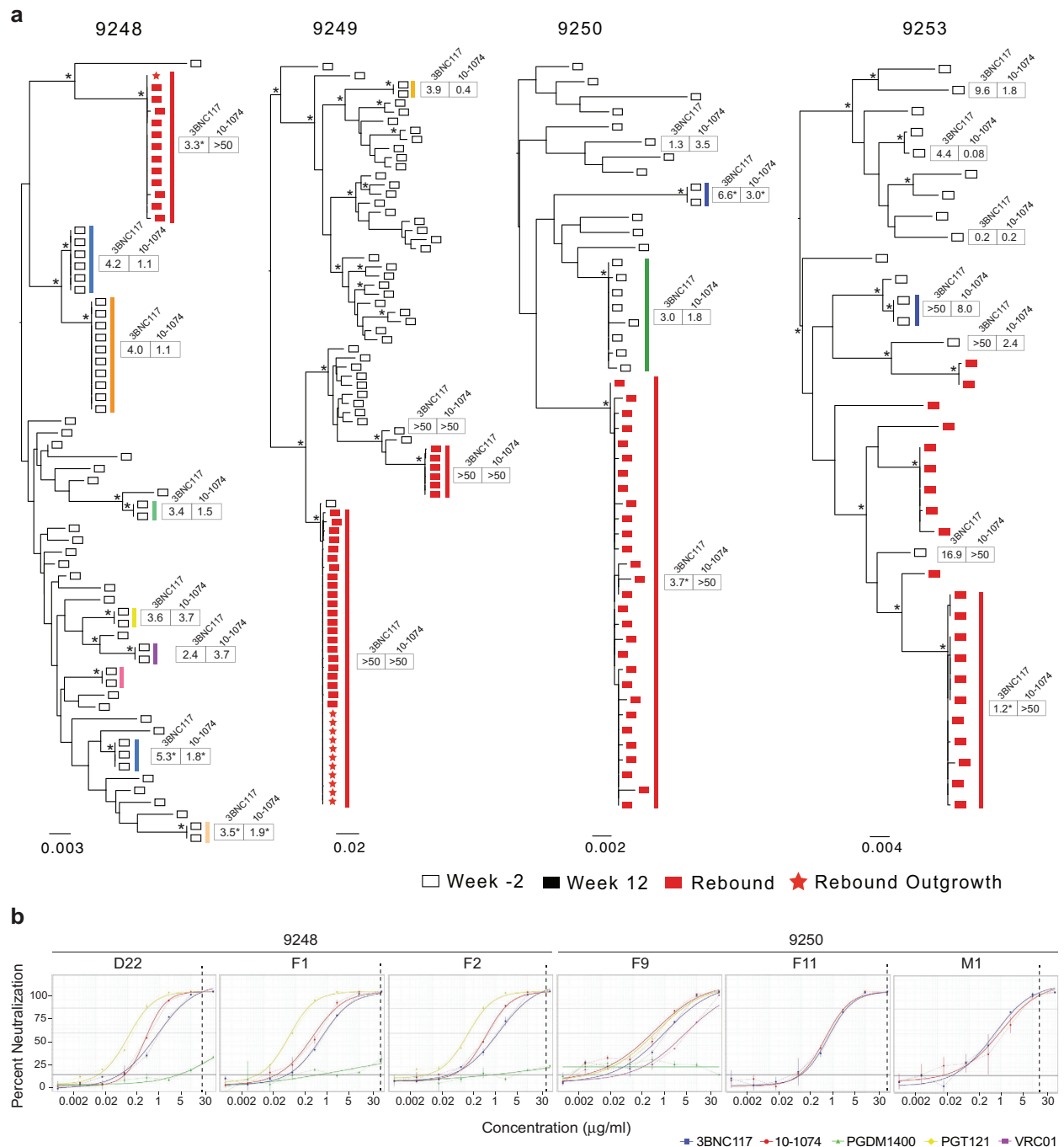


Extended Data Fig. 3 | Phylogenetic tree of viruses from all enrolled participants. Maximum likelihood phylogenetic trees of full-length *env* sequences containing all sequences obtained from Q²VOA cultures and rebound viruses from SGA or rebound outgrowth of the 15 participants enrolled in the study. Participants are indicated by individual colours.



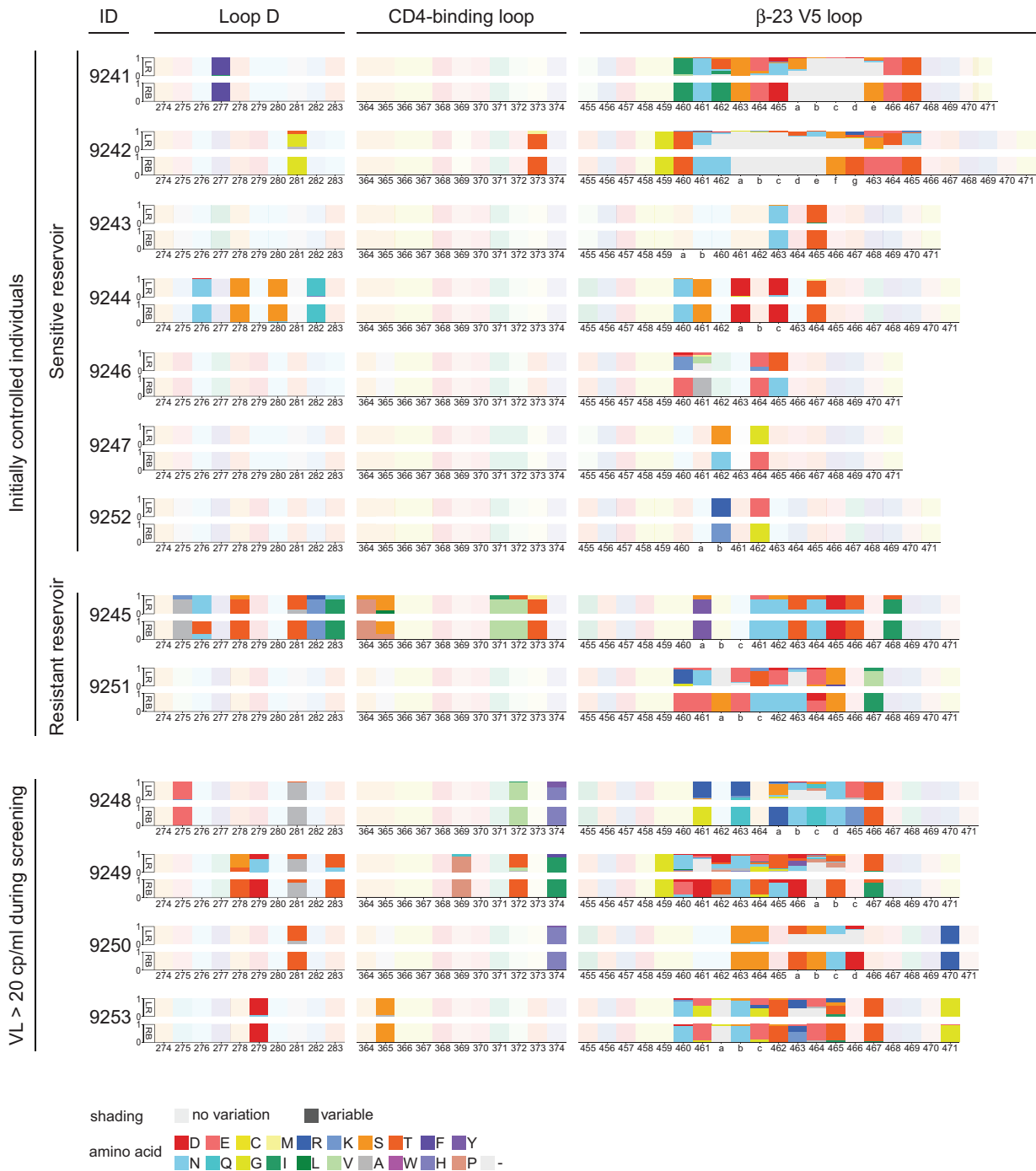
Extended Data Fig. 4 | Viral rebound, amino acid variants at 10-1074 contact sites and sensitivities of latent and rebound viruses in the participants with detectable viraemia (>20 copies per ml) two weeks before or at the start of ATI. a, Plasma HIV-1 RNA levels (black; left y axis) and bNAb serum concentrations (3BNC117, red; 10-1074, blue; right y axis). Red and blue triangles indicate 3BNC117 and 10-1074 infusions, respectively. Serum antibody concentrations were determined by TZM-bl assay. Grey-shaded areas indicate time on ART. Lower limit of detection of HIV-1 RNA was 20 copies per ml. **b**, Kaplan-Meier plots summarizing time to viral rebound. The y axis shows the percentage of participants that maintained viral suppression. The x axis shows the time in weeks after the start of ATI. Participants receiving the combination of 3BNC117 and 10-1074 are indicated by the blue line ($n = 4$). The dotted red line indicates a cohort of individuals receiving 3BNC117 alone during ATI⁹ ($n = 13$) and the dotted black line indicates a cohort of participants who underwent ATI without any intervention¹⁰ ($n = 52$). **c**, Colour charts show Env contact sites of 10-1074 at the G(D/N)IR motif (positions 324–327, according to HXB2 numbering) and the glycan at the potential N-linked glycosylation site at position 332 (NxS/T motif at positions 332–334). LR, latent reservoir viruses isolated by Q²VOA

(week -2); RB, rebound viruses isolated by SGA (plasma) or viral outgrowth (PBMCs). Each amino acid is represented by a colour and the frequency of each amino acid is indicated by the height of the rectangle. Shaded rectangles indicate the lack of variation between latent reservoir and rebound viruses at the indicated position. Full-colour rectangles represent amino acid residues with changes in distribution between reservoir and rebound viruses. **d**, Dot plots showing the IC_{80} ($\mu\text{g ml}^{-1}$) of 3BNC117 (left) and 10-1074 (right) against latent and rebound viruses determined by TZM-bl neutralization assay. Q²VOA-derived latent viruses from week -2 are shown as black circles. For outgrowth culture-derived rebound viruses, the highest IC_{80} determined is shown as red circle. For 9250 and 9253, no viruses could be obtained from rebound outgrowth cultures and pseudoviruses were made from *env* sequences of the latent reservoir (Q²VOA) and rebound viruses (plasma SGA). Note that 9249 and 9253 had pre-existing resistant viruses in the reservoir ($\text{IC}_{50} > 2 \mu\text{g ml}^{-1}$). 9248 and 9250 had pre-existing viruses that failed to reach an IC_{100} when tested up to $50 \mu\text{g ml}^{-1}$ for 3BNC117 (Extended Data Fig. 5). Rebound viruses of all four participants had an IC_{80} or IC_{100} of $> 50 \mu\text{g ml}^{-1}$ for both 3BNC117 and 10-1074.



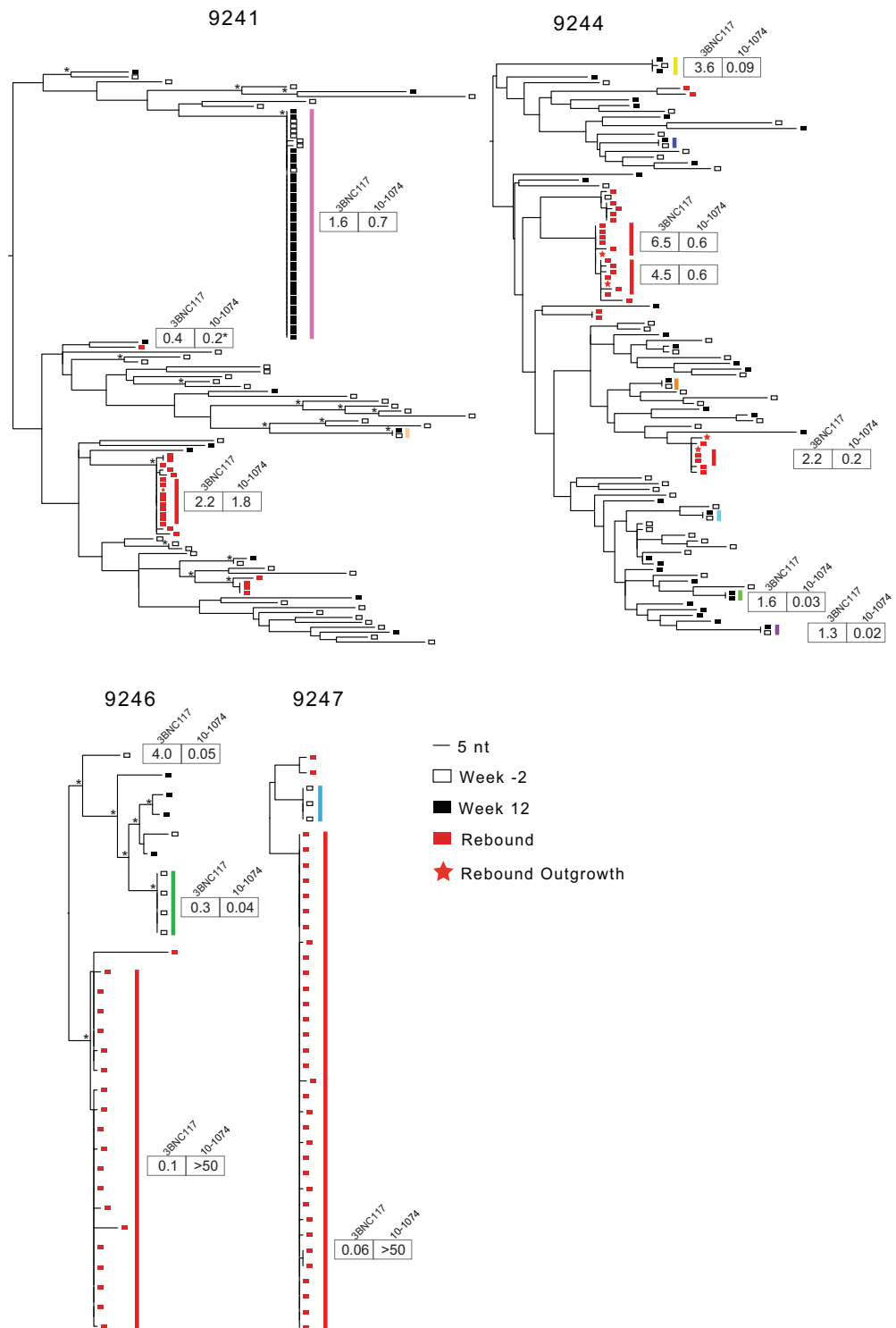
Extended Data Fig. 5 | Phylogenetic *env* trees and Tz-m-bl neutralization curves for individuals with viral blips. a, Circulating reservoir and viral rebound in study participants with detectable viraemia at week -2 or day 0. Maximum likelihood phylogenetic trees of full-length *env* sequences of viruses isolated from week -2 Q²VOA cultures, rebound plasma SGA and rebound outgrowth from the four participants with viral blips. Open black rectangles indicate Q²VOA-derived viruses from week -2. Viruses obtained at the time of rebound are indicated by red rectangles (plasma SGA) and red stars (rebound PBMC outgrowth cultures), respectively. Asterisks indicate nodes with significant bootstrap values (bootstrap support $\geq 70\%$). Clones are denoted by coloured lines. Boxes indicate IC₈₀ values ($\mu\text{g ml}^{-1}$) of 3BNC117 and 10-1074 against

individual clones, with asterisks indicating $\text{IC}_{100} > 50 \mu\text{g ml}^{-1}$. **b**, Latent reservoir virus Tz-m-bl neutralization curves for two participants that had a viral load of >20 copies per ml at day 0 (9248 and 9250). Curves show neutralization titres by 3BNC117 (blue), 10-1074 (red) and other bNAbs, when available, for week -2 Q²VOA-derived viruses present in the circulating reservoir. Three representative viruses from 9248 (left) and 9250 (right) are shown. Although these viruses had low 3BNC117 and 10-1074 IC₅₀ or IC₈₀ titres, the IC₁₀₀ (black dotted line) is reached only at a high concentration or not reached at all. The neutralization titre was measured by Tz-m-bl neutralization assay using a five-parameter curve fit method.



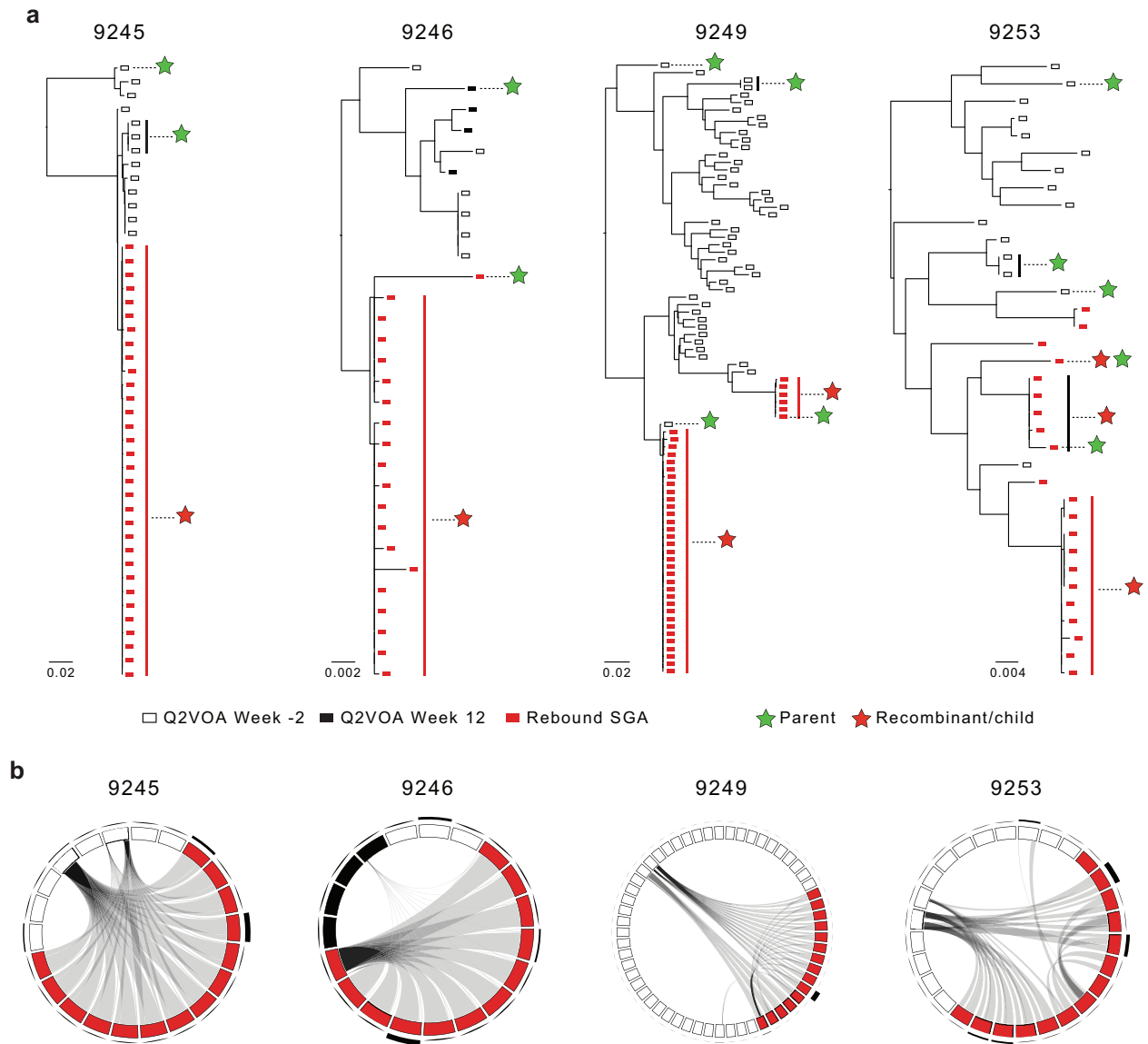
Extended Data Fig. 6 | Amino acid variants at 3BNC117 contact sites of reactivated latent and rebound viruses. Colour charts show 3BNC117 contact sites in Env according to HXB2 numbering. Diagram shows the 13 participants that experienced viral rebound before week 30. LR, latent reservoir viruses isolated by Q²VOA (on weeks -2 and 12 when available); RB, rebound viruses isolated by SGA (plasma) and viral

outgrowth (PBMCs). Each amino acid is represented by a colour and the frequency of each amino acid is indicated by the height of the rectangle. Shaded rectangles indicate the lack of variation and full-colour rectangles represent amino acid residues with changes in the distribution between the reservoir and rebound.



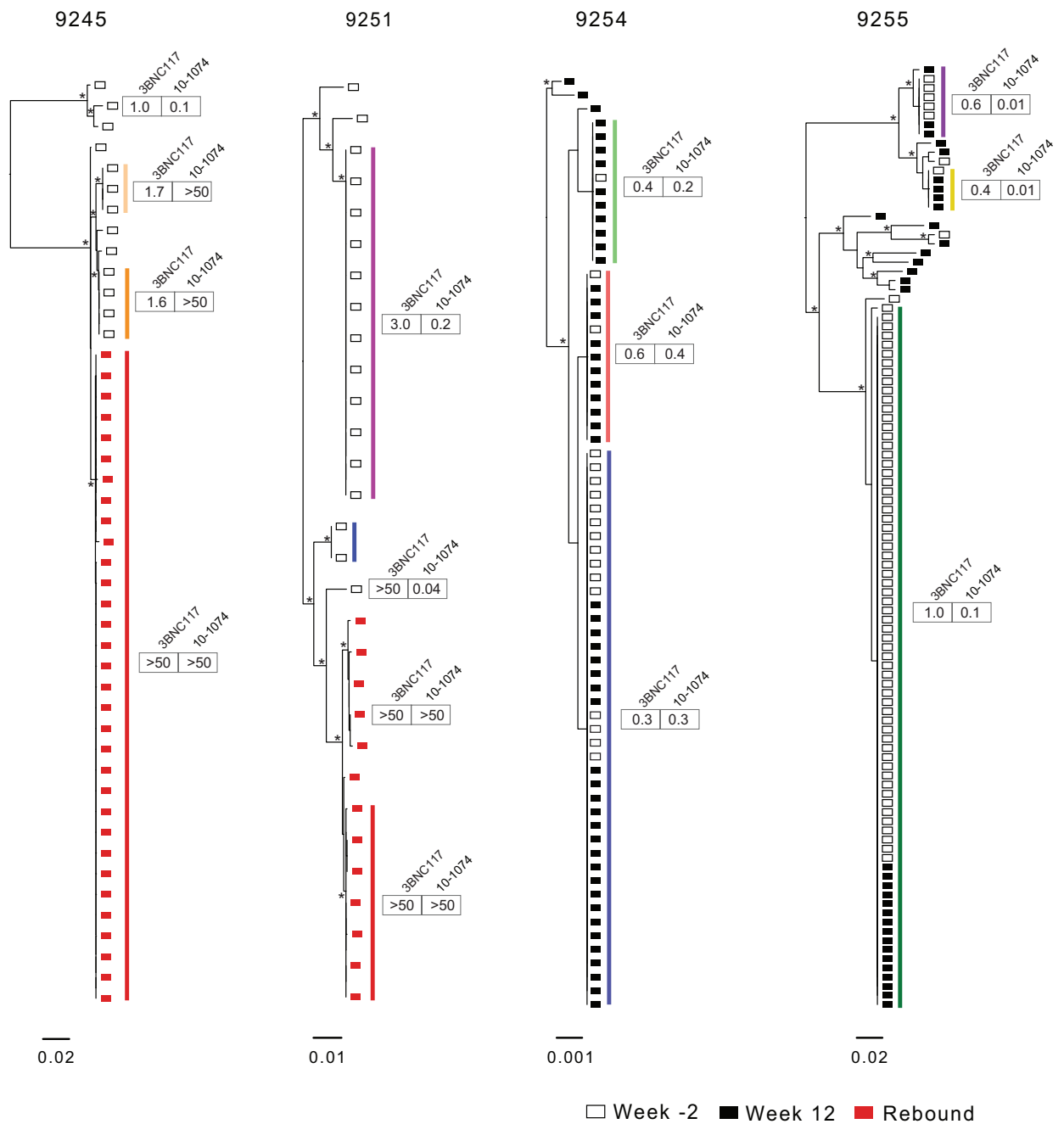
Extended Data Fig. 7 | Comparison of the circulating latent reservoir and rebound viruses. Maximum likelihood phylogenetic trees of full-length *env* sequences of viruses isolated from Q²VOA, rebound plasma SGA and rebound PBMC outgrowth cultures from participants 9241, 9244, 9246 and 9247, who rebounded before week 30. Open and closed black rectangles indicate Q²VOA-derived viruses from week -2 and week 12, respectively. Viruses obtained at the time of rebound are

indicated by red rectangles (plasma SGA) and red stars (rebound PBMC outgrowth cultures). Asterisks indicate nodes with significant bootstrap values (bootstrap support $\geq 70\%$). Clones are denoted by coloured lines mirroring the colours of slices in Extended Data Fig. 10a. Boxes indicate IC₈₀ values ($\mu\text{g ml}^{-1}$) of 3BNC117 and 10-1074 against representative viruses throughout the phylogenetic tree and clones, when possible (Supplementary Table 4). Asterisks in boxes indicate IC₁₀₀ $> 50 \mu\text{g ml}^{-1}$.



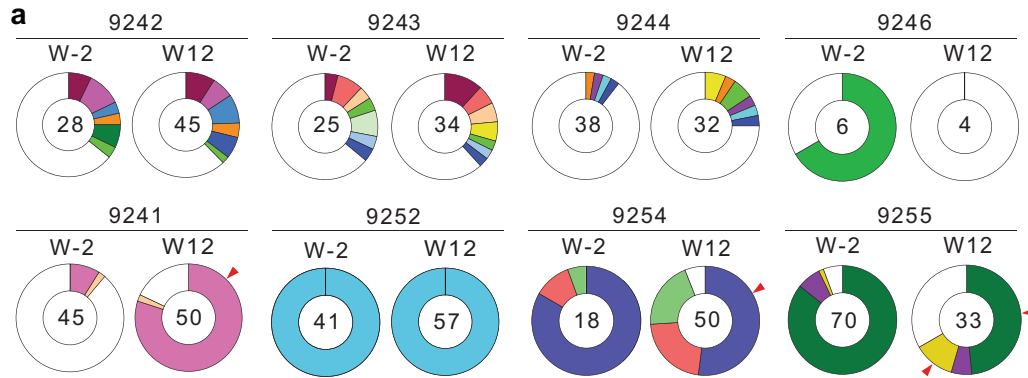
Extended Data Fig. 8 | Recombination events in rebound viruses.
a, Maximum likelihood phylogenetic trees of full-length *env* sequences of viruses isolated from Q²VOA cultures and rebound SGA in the four participants for whom rebound viruses showed recombination events. Open and closed black rectangles indicate Q²VOA-derived viruses from week -2 and week 12, respectively. Rebound plasma SGA- or outgrowth-derived viruses are indicated by closed red rectangles. Green stars

represent parent sequences that underwent recombination to produce the child sequences (red stars). **b**, Circos plots indicating the relationship between the parent sequences and the recombinants. Open and closed black rectangles indicate Q²VOA-derived sequences from week -2 and week 12, respectively. Rebound virus sequences are indicated by red rectangles. The thickness of the black outer bars represents the number of sequences obtained from that particular clone.



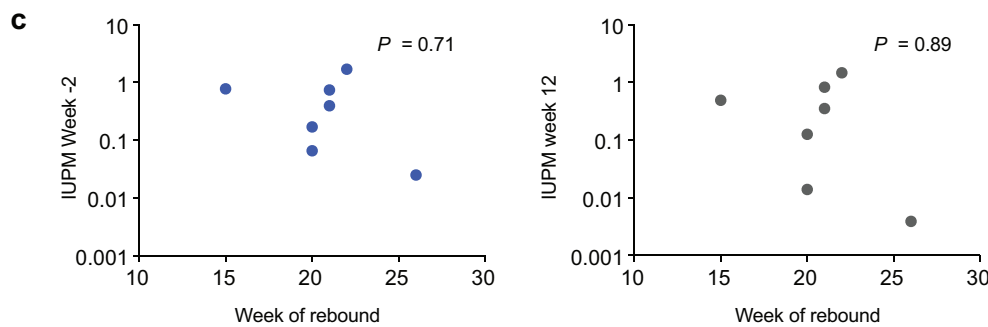
Extended Data Fig. 9 | Phylogenetic trees of participants 9245, 9251, 9254 and 9255. Maximum likelihood phylogenetic trees of full-length *env* sequences of viruses isolated from Q²VOA cultures and rebound plasma SGA and rebound outgrowth from the two participants (9245 and 9251) with pre-existing resistance to one of the two antibodies and the two sensitive participants (9254 and 9255) who maintained viral suppression for >30 weeks (end of the study). Open and closed black

rectangles indicate Q²VOA-derived viruses from week -2 and week 12, respectively. Rebound plasma SGA viruses are indicated by closed red rectangles. Asterisks indicate nodes with significant bootstrap values (bootstrap support ≥ 70%). Clones are denoted by coloured lines beside the phylogenetic tree. Numbers correspond to 3BNC117 and 10-1074 IC₅₀ neutralization titres.



b

Study ID	Total <i>env</i> seqs by Q ² VOA (no.)	Clonal <i>env</i> sequences		IUPM	
		(no.)	(%)	week -2	week 12
9242	73	27	37.0	0.781	0.493
9243	59	22	37.3	0.170	0.126
9244	70	12	17.1	0.397	0.354
9246	10	4	40.0	0.066	0.014
9247	3	3	100.0	0.025	0.004
9241	95	46	48.4	0.743	0.828
9252	98	98	100.0	1.709	1.470
9254	68	65	95.6	N/A	0.680
9255	103	88	85.4	1.890	1.400
Total	579	365	63.0		



Extended Data Fig. 10 | Clonal distribution of the circulating latent reservoir and IUPM changes. **a**, Pie charts depicting the distribution of Q²VOA-derived *env* sequences obtained at weeks -2 (W-2) and week 12 (W12). Number in the inner circle indicates the total number of analysed *env* sequences. White represents sequences isolated only once across both time points and coloured slices represent identical sequences that appear more than once (clones). The size of each pie slice is proportional to the

size of the clone. Red arrows denote clones that significantly change in size ($P \leq 0.05$ (two-sided Fisher's exact test)) between the two time points. **b**, Summary of clonal *env* sequences and IUPM in the nine individuals with an antibody-sensitive reservoir. **c**, IUPM versus time of viral rebound in the antibody-sensitive individuals ($n = 7$) who rebounded within the study observation period (30 weeks). P values were obtained using a two-tailed Pearson correlation test comparing the two variables.

A.4 Manuscript 4

Title:

Safety and antiviral activity of combination HIV-1 broadly neutralizing antibodies in viremic individuals

Authors:

Yotam Bar-On, Henning Gruell, Till Schoofs, Joy A. Pai, Lilian Nogueira, Allison L. Butler, Katrina Millard, Clara Lehmann, Isabelle Suárez, Thiago Y. Oliveira, Theodora Karagounis, Yehuda Z. Cohen, Christoph Wyen, Stefan Scholten, Lisa Handl, Shiraz Belblidia, Juan P. Dizon, Jörg J. Vehreschild, Maggi Witmer-Pack, Irina Shimeliovich, Kanika Jain, Kerstin Fiddike, Kelly E. Seaton, Nicole L. Yates, Jill Horowitz, Roy M. Gulick, Nico Pfeifer, Georgia D. Tomaras, Michael S. Seaman, Gerd Fätkenheuer, Marina Caskey, Florian Klein & Michel C. Nussenzweig

Published in:

Nature Medicine, Volume 24, Pages 1701–1707
<https://doi.org/10.1038/s41591-018-0186-4>

Time of Publication:

September 2018

License information:

As part of their editorial policies, Springer Nature allows authors to reuse the version of record of their article published in any Nature portfolio journal in their own dissertation without obtaining an additional, explicit written permission (<https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish>).

Safety and antiviral activity of combination HIV-1 broadly neutralizing antibodies in viremic individuals

Yotam Bar-On^{1,17}, Henning Gruell^{2,3,4,17}, Till Schoofs^{1,2}, Joy A. Pai¹, Lilian Nogueira¹, Allison L. Butler¹, Katrina Millard¹, Clara Lehmann^{3,4,5}, Isabelle Suárez^{3,4,5}, Thiago Y. Oliveira¹, Theodora Karagounis¹, Yehuda Z. Cohen⁶, Christoph Wyen^{3,6}, Stefan Scholten⁷, Lisa Handl⁸, Shiraz Belblidia¹, Juan P. Dizon¹, Jörg J. Vehreschild^{3,4}, Maggi Witmer-Pack¹, Irina Shimeliovich¹, Kanika Jain², Kerstin Fiddike³, Kelly E. Seaton⁹, Nicole L. Yates⁹, Jill Horowitz¹, Roy M. Gulick¹⁰, Nico Pfeifer^{8,11,12,13}, Georgia D. Tomaras^{9,14}, Michael S. Seaman¹⁵, Gerd Fätkenheuer^{3,4}, Marina Caskey^{1,18*}, Florian Klein^{2,4,5,18*} and Michel C. Nussenzweig^{1,16,18*}

Monotherapy of HIV-1 infection with single antiretroviral agents is ineffective because error-prone HIV-1 replication leads to the production of drug-resistant viral variants^{1,2}. Combinations of drugs can establish long-term control, however, antiretroviral therapy (ART) requires daily dosing, can cause side effects and does not eradicate the infection^{3,4}. Although anti-HIV-1 antibodies constitute a potential alternative to ART^{5,6}, treatment of viremic individuals with a single antibody also results in emergence of resistant viral variants⁷⁻⁹. Moreover, combinations of first-generation anti-HIV-1 broadly neutralizing antibodies (bNAbs) had little measurable effect on the infection¹⁰⁻¹². Here we report on a phase 1b clinical trial (NCT02825797) in which two potent bNAbs, 3BNC117¹³ and 10-1074¹⁴, were administered in combination to seven HIV-1 viremic individuals. Infusions of 30 mg kg⁻¹ of each of the antibodies were well-tolerated. In the four individuals with dual antibody-sensitive viruses, immunotherapy resulted in an average reduction in HIV-1 viral load of 2.05 log₁₀ copies per ml that remained significantly reduced for three months following the first of up to three infusions. In addition, none of these individuals developed resistance to both antibodies. Larger studies will be necessary to confirm the efficacy of antibody combinations in reducing HIV-1 viremia and limiting the emergence of resistant viral variants.

3BNC117 and 10-1074 are potent bNAbs that target the CD4 binding site and the base of the V3 loop on the HIV-1 envelope spike, respectively^{13,14}. Infusion of the combination of 3BNC117 and

10-1074 during ART interruption maintains suppression of viremia and prevents the emergence of resistant variants¹⁵.

Controlling infection in viremic individuals represents a much more difficult problem than maintaining suppression in ART-treated individuals undergoing treatment interruption simply because of the large diversity of circulating HIV-1 variants that are present during active infection. Thus, although monotherapy with any one of three different bNAbs reduced viremia by 1.1–1.5 log₁₀ copies per ml, these effects were transient and superseded by the emergence of antibody-resistant viral variants⁷⁻⁹. To determine whether the combination of 3BNC117 and 10-1074 is safe and results in improved antiviral activity against HIV-1 compared to monotherapy, we conducted a phase 1b trial in viremic individuals (Fig. 1a).

Viremic participants were selected from a cohort that was screened for sensitivity to 3BNC117 and 10-1074 by TZM-bl cell neutralization assays performed on viruses derived from bulk CD4⁺ T cell outgrowth cultures¹⁶ (Supplementary Fig. 1). In agreement with previous reports, 67 and 58% of the individuals tested showed half-maximum inhibitory concentration (IC₅₀) values of <2 µg ml⁻¹ to 3BNC117 and 10-1074, respectively, and 40% were sensitive to both^{8,17,18} (Supplementary Table 1). The seven viremic participants had been diagnosed with HIV-1 infection for a median of five years and had a geometric mean viral load of 11,494 copies per ml on the day of the first infusion (Fig. 1b and Supplementary Tables 2, 3). In addition, eight individuals on ART with viral loads below the limit of detection were included for safety and pharmacokinetic assessments (Fig. 1a, Supplementary Fig. 1 and Supplementary Tables 2, 3).

¹Laboratory of Molecular Immunology, The Rockefeller University, New York, NY, USA. ²Laboratory of Experimental Immunology, Institute of Virology, University Hospital Cologne, Cologne, Germany. ³Department I of Internal Medicine, University Hospital Cologne, Cologne, Germany. ⁴German Center for Infection Research, Partner Site Bonn-Cologne, Cologne, Germany. ⁵Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany. ⁶Praxis am Ebertplatz, Cologne, Germany. ⁷Praxis Hohenstaufenring, Cologne, Germany. ⁸Methods in Medical Informatics, Department of Computer Science, University of Tübingen, Tübingen, Germany. ⁹Duke Human Vaccine Institute, Duke University, Durham, NC, USA. ¹⁰Division of Infectious Diseases, Weill Cornell Medicine, New York, NY, USA. ¹¹Medical Faculty, University of Tübingen, Tübingen, Germany. ¹²German Center for Infection Research, Partner Site Tübingen, Tübingen, Germany. ¹³Max Planck Institute for Informatics, Saarbrücken, Germany. ¹⁴Departments of Surgery, Immunology and Molecular Genetics and Microbiology, Duke University, Durham, NC, USA. ¹⁵Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. ¹⁶Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA. ¹⁷These authors contributed equally: Yotam Bar-On, Henning Gruell. ¹⁸These authors jointly supervised this work: Marina Caskey, Florian Klein, Michel C. Nussenzweig. *e-mail: mcaskey@rockefeller.edu; florian.klein@uk-koeln.de; nussen@rockefeller.edu

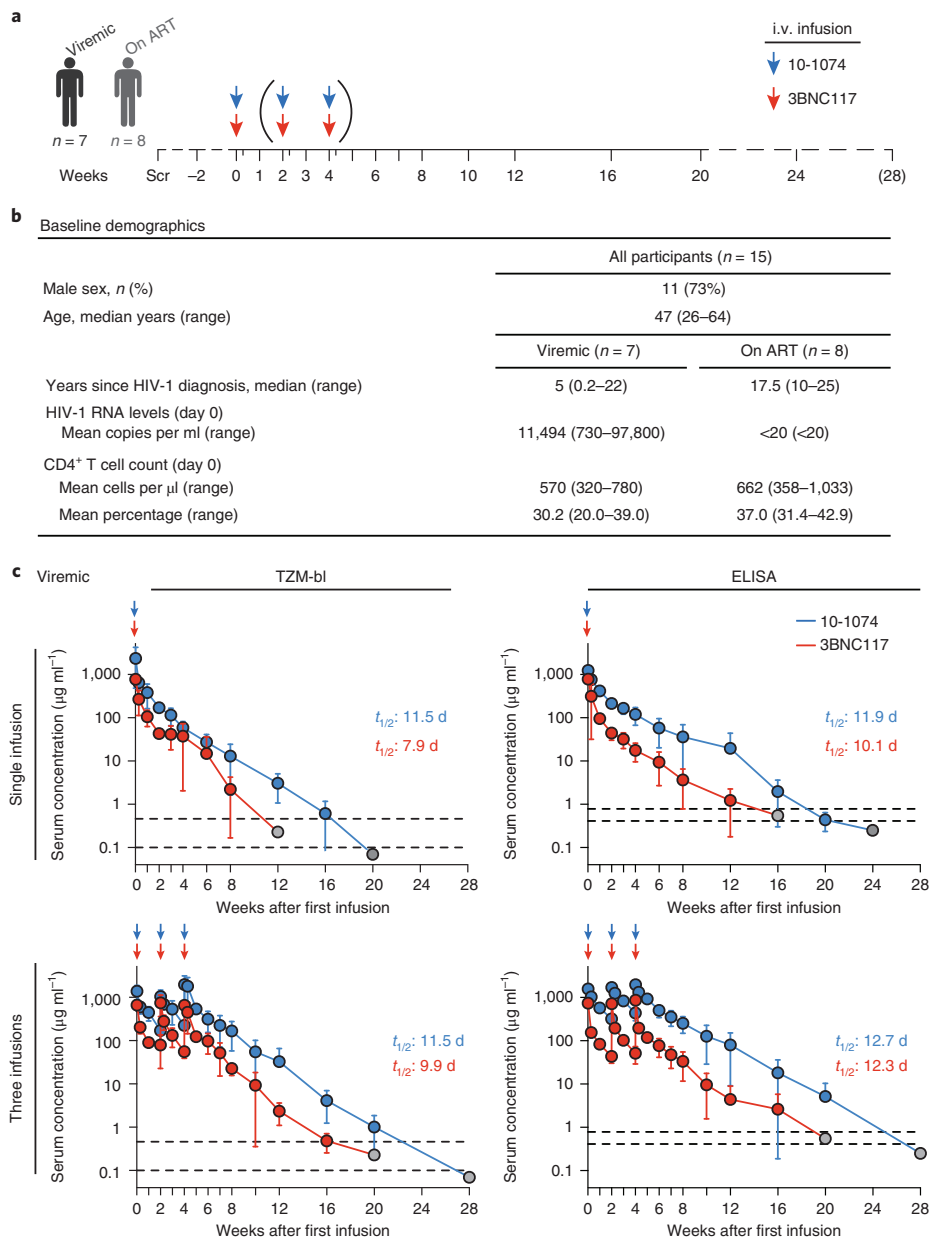


Fig. 1 | Study design and pharmacokinetics of 3BNC117 and 10-1074 in HIV-1-infected individuals. **a**, Schematic representation of the study design. i.v., intravenous. **b**, Baseline demographics of study participants. **c**, Serum concentrations ($\mu\text{g ml}^{-1}$) of 3BNC117 (red) and 10-1074 (blue) in viremic individuals after a single infusion (top) and three infusions given every two weeks (bottom) of 3BNC117 and 10-1074 (30 mg kg^{-1} of each antibody). bNAB concentrations were determined by TZM-bl assay (left) and ELISA (right). Lines indicate arithmetic mean concentration and standard deviation. Dotted grey lines indicate lower limits of quantitation (TZM-bl, $0.46 \mu\text{g ml}^{-1}$ and $0.1 \mu\text{g ml}^{-1}$ for 3BNC117 and 10-1074, respectively; ELISA, $0.78 \mu\text{g ml}^{-1}$ and $0.41 \mu\text{g ml}^{-1}$ for 3BNC117 and 10-1074, respectively). Grey circles indicate antibody levels below the limit of quantification. $t_{1/2}$, average half-life.

Participants received either a single intravenous infusion of 3BNC117 and 10-1074 at a dose of 30 mg kg^{-1} per antibody, or three infusions of 30 mg kg^{-1} per antibody every two weeks (Fig. 1a). Viral loads, antibody serum levels, CD4⁺ T cell counts and clinical parameters were monitored for 24 weeks after the last antibody infusion (Fig. 1 and Supplementary Tables 3, 4).

Administration of both antibodies was well-tolerated. No serious adverse events or treatment-related adverse events graded as moderate or severe were observed (Supplementary Table 4). CD4⁺ T cell

counts did not change significantly during the observation period (Supplementary Fig. 2 and Supplementary Table 3). We conclude that the combination of 3BNC117 and 10-1074 is generally safe and well-tolerated.

3BNC117 and 10-1074 antibody levels in serum were determined via enzyme-linked immunosorbent assays (ELISAs) using anti-idiotypic antibodies and the TZM-bl assay, which measures the neutralizing activities of the antibodies in serum. In viremic individuals, the half-lives of 3BNC117 and 10-1074 were 11.1 and 12.2 days

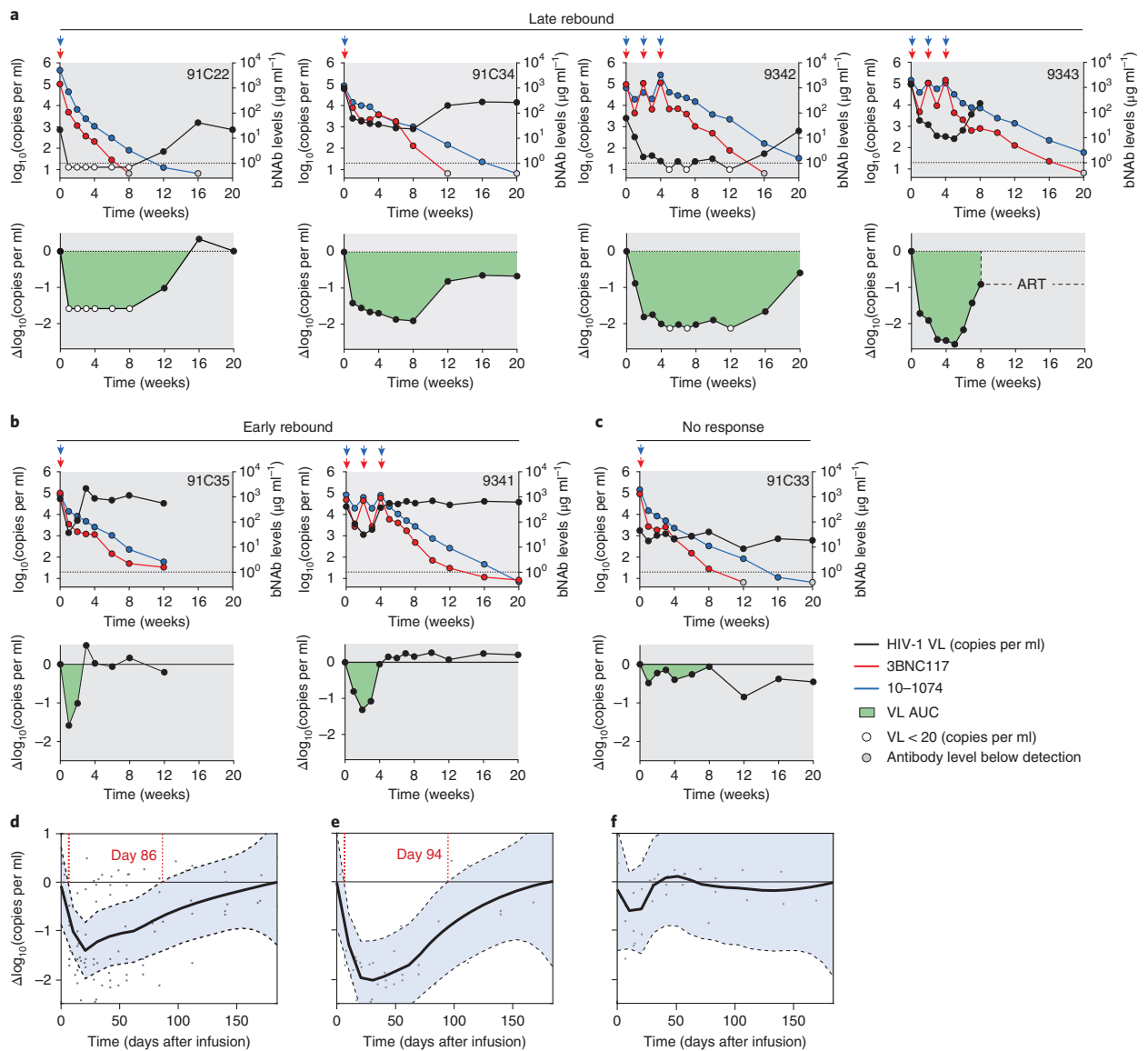


Fig. 2 | Viral load following 3BNC117/10-1074 infusions in HIV-1-infected participants. a–c, Changes in viremia and bNAb serum concentrations in HIV-1-infected participants showing late rebound (**a**), early rebound (**b**) or no response (**c**) after 3BNC117 and 10-1074 combination therapy. Top, HIV-1 RNA in \log_{10} copies per ml (black, left y axis), and 3BNC117 (red) and 10-1074 (blue) serum levels (right y axis, determined by TZM-bl). The x axis shows the time in weeks after the first antibody infusion. The dashed line indicates the lower limit of detection of HIV-1 RNA (20 copies per ml). Arrows indicate antibody infusions. Bottom, \log_{10} changes per ml in HIV-1 RNA copies compared to day 0. Green shading depicts viral suppression compared to day 0. VL, viral load. **d–f,** Simultaneous confidence band estimation to determine time of significant suppression (red dotted lines) of HIV-1 viremia in all viremic participants (**d**; $n = 7$, participants shown in **a–c**), individuals harboring 3BNC117- and 10-1074-sensitive viruses (**e**; $n = 4$, participants shown in **a**), and participants carrying viruses with partial or full bNAb resistance (**f**; $n = 3$, participants shown in **b,c**). Each dot represents a viral load measurement. Solid and dashed lines represent the regression fit and simultaneous confidence bands at 95% certainty level, respectively, and were computed using the Gaussian family for the local likelihood function using R package locfit (version 1.5–9.1).

when measured using ELISA, and 8.5 and 11.5 days when determined by the TZM-bl assay, respectively (Fig. 1, Supplementary Figs. 3, 4 and Supplementary Tables 3, 5). In ART-treated individuals, half-lives of 3BNC117 and 10-1074 were 14.5 and 19.0 days as measured using ELISA, and 11.5 and 18.4 days in the TZM-bl assay, respectively (Supplementary Figs. 3, 4 and Supplementary Table 5). Viremic individuals generally showed lower antibody half-lives than individuals on ART with suppressed viral loads, possibly owing to

an antigen sink effect^{7,8,19}. Overall, these values are consistent with the results obtained when both antibodies were administered individually^{7,8,17} (Supplementary Fig. 3). Thus, pharmacokinetics of 3BNC117 and 10-1074 do not appear to be altered when the antibodies are administered in combination.

Plasma HIV-1 RNA levels were measured on a weekly basis for four weeks after antibody infusions and every 2–4 weeks thereafter (Fig. 2a–c, Supplementary Fig. 4 and Supplementary Table 3). The

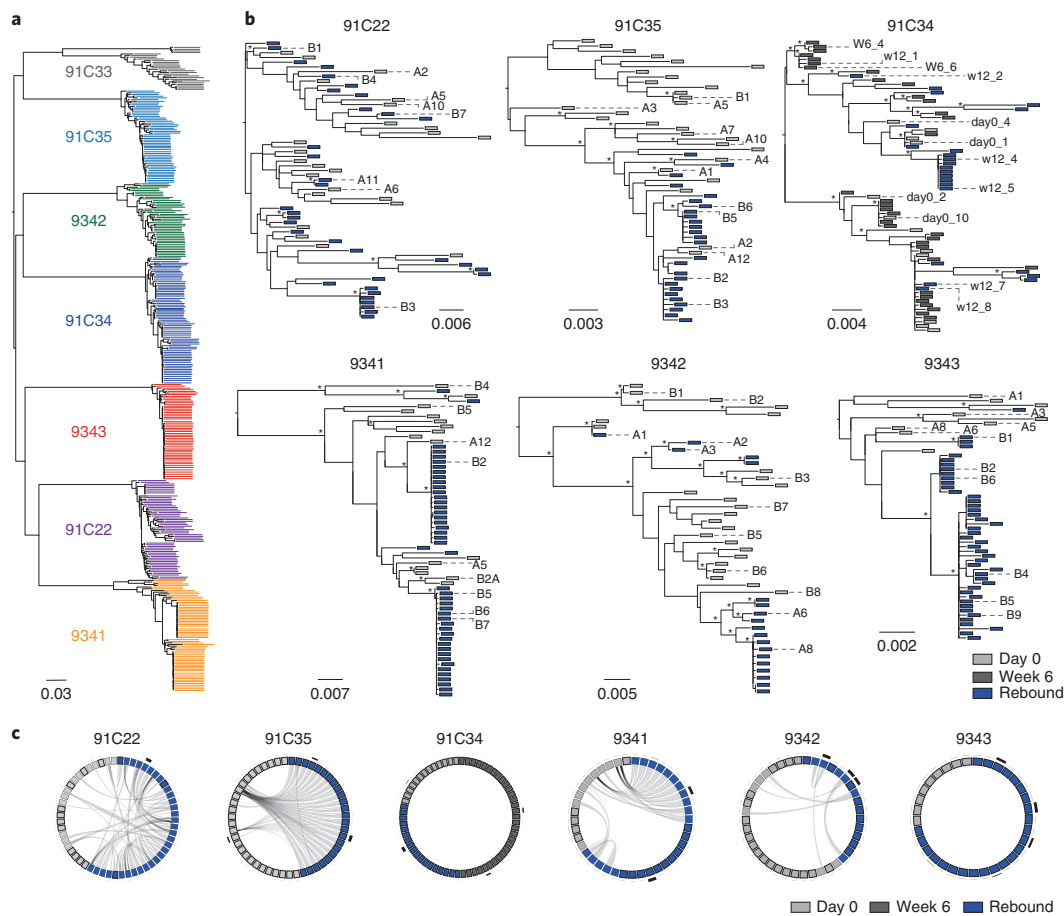


Fig. 3 | Phylogenetic sequence analysis of viremic participants. **a**, Maximum likelihood phylogenetic tree of all SGA-derived *env* gene sequences ($n=382$) obtained from plasma of viremic study participants ($n=7$). **b**, Maximum likelihood phylogenetic trees of *env* sequences ($n=356$, 91C33 not shown) obtained from plasma of single participants before antibody therapy (light grey) and at viral rebound (blue). Dark grey indicates sequences amplified at week 6 after antibody infusion (participant 91C34). *env* sequences that were used to produce pseudoviruses for neutralization testing are indicated. Asterisks indicate nodes with significant bootstrap values (bootstrap support of $\geq 70\%$). **c**, Circos plots indicating the relationship between parent sequences and recombinant sequences in single participants ($n=6$). SGA sequences are depicted by light-grey (day 0), dark-grey (week 6) and blue (rebound) rectangles. Grey lines indicate recombination events between different viruses. Thickness of the black outer bars represents the number of sequences obtained from that particular clone.

average drop in viral load for all viremic individuals was $1.65 \log_{10}$ copies per ml and viremia remained significantly reduced until day 86 (Fig. 2d). The four individuals with sensitive viruses (see below) showed a more pronounced drop in viral load compared to the other individuals (average of $2.05 \log_{10}$ copies per ml) and were significantly suppressed until day 94 (Fig. 2a,e,f). In comparison to a single infusion of either 3BNC117⁷ or 10-1074⁸, viremic individuals receiving one or three infusions of the combination of both antibodies showed significantly prolonged viral suppression ($P=0.00018$) (Fig. 2d and Supplementary Fig. 5). We conclude that the combination of 3BNC117 and 10-1074 is more effective in suppressing viremia than either antibody alone.

Despite the pronounced difference in the duration of viremia reduction between monotherapy and combination therapy, there was considerable variation in the response of individual participants receiving 3BNC117 and 10-1074 combination treatment (Fig. 2 and Supplementary Table 3). To define the relationship between individual responses to antibody therapy and circulating virus sensitivity to the antibodies, we performed single genome amplification (SGA)

of plasma viruses. Initially, 382 intact full-length *env* sequences were analyzed from the seven viremic participants (Supplementary Fig. 6). All of these individuals were infected with epidemiologically distinct clade B virus (Fig. 3a). In addition, sequences of circulating viruses at the time of viral rebound were polyclonal, and as expected for viremic individuals, recombination events were detected between circulating viruses in most individuals (Fig. 3b,c).

Pseudoviruses constructed from plasma SGA were tested for bNAb sensitivity in the TZM-bl assay (Fig. 4a and Supplementary Table 6). Participant 91C33, who failed to respond to antibody infusions, had preexisting circulating viruses that were resistant to both antibodies (Fig. 4a and Supplementary Table 6). These viruses carried mutations in 3BNC117 contact sites (N280S and A281H) and in 10-1074 contact sites (N332T and S334N, Supplementary Fig. 6). Two individuals, 91C35 and 9341, responded to antibody therapy with a decrease in viremia of -1.58 and $-1.32 \log_{10}$ copies per ml but HIV-1 RNA levels returned to baseline within 3 and 4 weeks, respectively (Fig. 2b). 91C35 was found to have pre-infusion circulating viruses with reduced sensitivity to 3BNC117, and carried

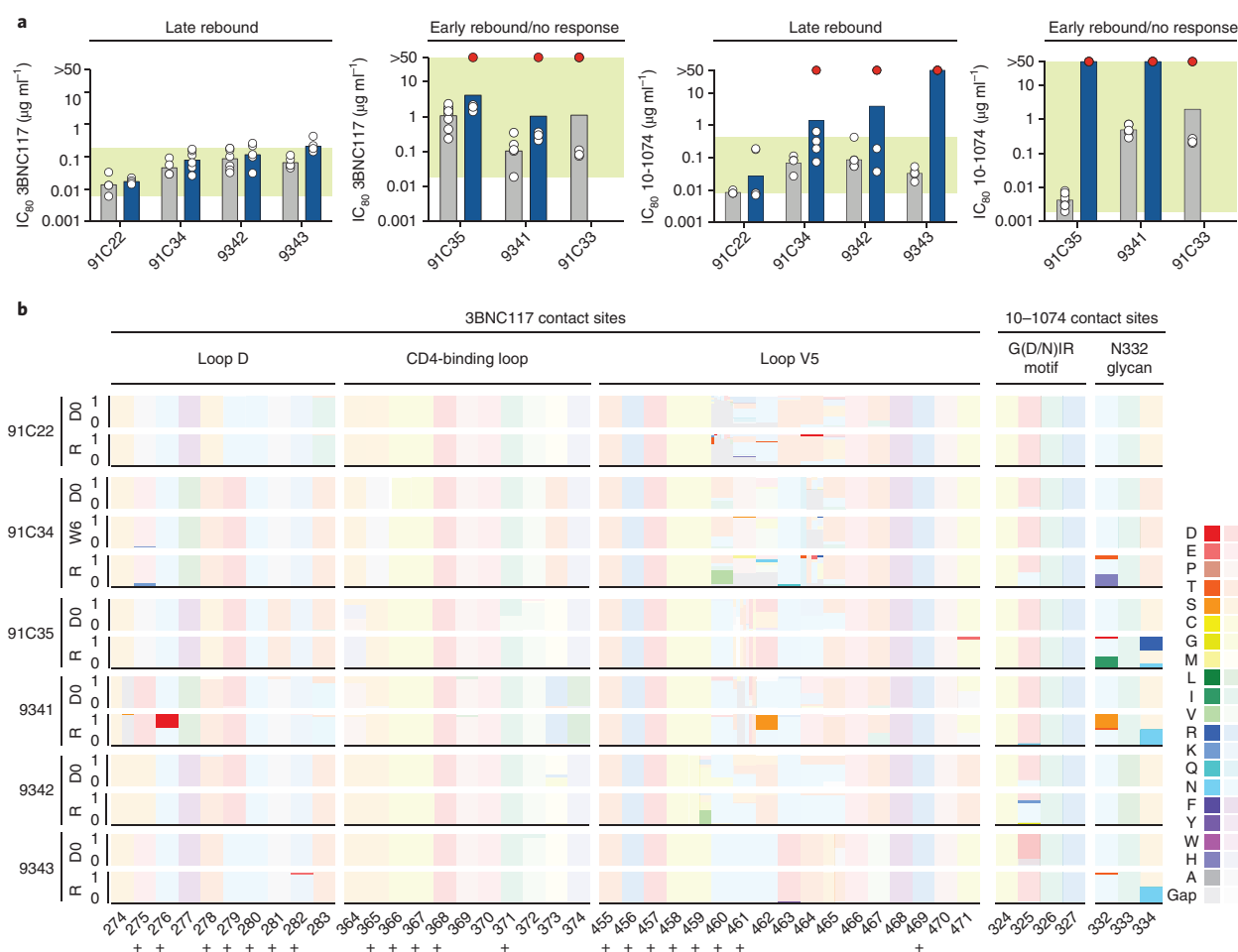


Fig. 4 | HIV-1 escape analysis of individuals receiving 3BNC117 and 10-1074 therapy. a, Viral sensitivities (IC_{80} , $\mu\text{g ml}^{-1}$) of pseudoviruses constructed from SGA-derived *env* sequences obtained on day 0 (grey) and at the time of rebound (blue). Columns reflect geometric mean IC_{80} values of viruses tested against 3BNC117 (left) and 10-1074 (right). Each circle represents one viral isolate. Fully resistant viruses ($\text{IC}_{80} > 50 \mu\text{g ml}^{-1}$) are depicted by red circles. Green shading indicates the range of IC_{80} values. **b**, Frequency of amino acids in and around known 3BNC117 and 10-1074 contact residues in Env (3BNC117, amino acids 274–283, 364–374 and 455–471; 10-1074, amino acids 324–327 and 332–334). Amino acids are numbered according to HXB2. 3BNC117 contact sites are indicated by ‘+’³². D0 indicates viruses isolated from plasma by SGA before antibody infusions (day 0); R indicates rebound viruses isolated by SGA; W6 indicates viruses isolated at week 6. Each amino acid is represented by a color and the frequency of each amino acid is indicated by the height of the rectangle. Shaded rectangles represent instances in which amino acids that were found in rebound viruses were also found in day 0 viruses at the indicated position. Full-color rectangles represent instances in which an amino acid was found in rebound sequences but not in day 0 sequences.

a CD4 contact residue mutation (A281T) that was associated with viral escape from 3BNC117²⁰ (Fig. 4a, Supplementary Figs. 6 and 7 and Supplementary Table 6). Pre-infusion viruses derived from bulk CD4⁺ T cell outgrowth cultures of 9341 showed a 10-1074 IC_{80} that was 1.3 \log_{10} higher than the geometric mean IC_{80} of all other enrolled viremic individuals (Supplementary Table 1). In both of these cases, rebounding viruses were resistant to both antibodies and carried mutations resulting in the loss of the potential N-linked glycosylation site at position 332 that is critical for 10-1074 binding (Fig. 4a,b, Supplementary Figs. 6, 7 and Supplementary Table 6). In addition, rebound viruses from 91C35 and 9341 contained G471E and N276D mutations, respectively, that are associated with increased resistance to 3BNC117 (Supplementary Fig. 6)^{7,17,21,22}. These mutations were not found in the pre-infusion circulating viruses described above or in the additional 113 pre-infusion *env* sequences that were analyzed from these two participants

(Supplementary Fig. 8). Thus, 91C35 and 9341 were infected with viruses with reduced sensitivity to one of the two antibodies and resemble individuals that received antibody monotherapy, both in the magnitude of the drop in viremia and time required to return to baseline viremia^{7–9}. We conclude that the bulk outgrowth cultures used for initial screening failed to detect partial or complete pre-existing resistance against one or both of the antibodies in three of the seven individuals studied.

The four remaining individuals showed no detectable pre-existing resistant viruses in circulation and experienced significantly suppressed viremia until day 94 after the first antibody infusion with an average maximum drop in viral load of $-2.05 \log_{10}$ copies per ml (Figs. 2a,e, 4a and Supplementary Table 6). The individual in this group with the highest initial viral load (97,800 copies per ml; patient 9343) was the first to rebound at eight weeks (Fig. 2a and Supplementary Table 3). The two individuals with the lowest

initial viral loads, 91C22 and 9342 (750 and 2,550 copies per ml, respectively), demonstrated suppression to near or below the limit of detection for 12 and 16 weeks, respectively (Fig. 2a and Supplementary Table 3). Finally, viremia in participant 91C34 was reduced for a period of 12 weeks, however it never dropped below 810 copies per ml. Despite the persistent viremia, no resistance against both antibodies developed in this individual for as long as bNAb serum levels were above $10 \mu\text{g ml}^{-1}$ (Supplementary Figs. 7, 9 and Supplementary Table 3).

In three of the four initially sensitive individuals, rebound viremia was associated with the appearance of viruses that were resistant to 10-1074, but these individuals remained sensitive to 3BNC117 (Fig. 4a and Supplementary Table 6). This is consistent with the relatively shorter half-life of 3BNC117, which means that participants were effectively exposed to 10-1074 monotherapy at the end of the observation period. In accordance with the increased resistance to 10-1074, rebound viruses carried mutations in 10-1074 contact sites (Fig. 4b and Supplementary Figs. 6, 7). By contrast, there was no accumulation of de novo mutations in 3BNC117 contact sites (Fig. 4b and Supplementary Figs. 6, 7). 91C22, the participant with the lowest initial viral load, only returned to baseline viremia after both antibodies were below the limit of detection, and rebound viruses remained sensitive to both antibodies (Fig. 2a, Supplementary Fig. 4 and Supplementary Tables 3, 6). Overall none of the four participants that were initially sensitive to the two antibodies developed de novo resistance to 3BNC117 over a cumulative observation period of over one year (56 weeks), despite the residual viremia observed in three of these participants and frequent recombination events between circulating viruses (Fig. 3c).

Combination bNAb therapy for HIV-1 in humans showed a number of similarities with bNAb therapy for macaques infected with chimeric simian/human immunodeficiency virus AD8. For example, suppression was incomplete in macaques with higher initial viral loads; however, despite persistent low-level viremia, there was no emergence of 3BNC117 and 10-1074 double-resistant variants²³. In contrast to the macaque infection with a clonal virus, each of the four antibody-sensitive individuals in this study was infected with a uniquely diverse swarm of viruses. Thus, the relative difficulty of HIV-1 to develop resistance to the combination of 3BNC117 and 10-1074 is not limited to any particular strain of HIV-1. Macaque CD8⁺ T cell responses can control viremia and this type of cellular immunity can be enhanced by bNAb therapy²⁴. CD8⁺ T cells have also been implicated in HIV-1 control in humans²⁵. Whether such responses can also be enhanced by immunotherapy in humans remains to be determined.

3BNC117 and 10-1074 target distinct epitopes on the Env trimer. 3BNC117 interacts with the CD4 binding site, which is critical for HIV-1 binding to its cellular receptor CD4. Thus, escape mutations from 3BNC117 are limited by the requirement of continued affinity to CD4 and are associated with a reduction in viral fitness²⁶. Combinations of just two antibodies that synergize to further restrict viral escape may be even more effective than 3BNC117 and 10-1074²⁷.

Should antibodies enter clinical practice for HIV-1, adequate safeguards will be required to minimize the emergence of resistant variants. Reliable screening methods that identify viral resistance against individual drugs facilitate the selection of antiretroviral drug combinations with full activity. By contrast, the culture-based method used to screen for resistance in this study failed to detect partial or complete pre-existing antibody resistance in three of the seven viremic participants. This is likely due to outgrowth of a limited set of viruses in vitro that fails to represent the entire population that is circulating or archived in vivo^{8,15,17}. Sequence-based screening methods that encompass a much larger group of viruses are currently being developed and should be far more effective than the bulk cultures.

This study highlights some of the limitations of immunotherapy with the combination of 3BNC117 and 10-1074 in viremic individuals. 3BNC117 and 10-1074 infusions failed to suppress viremia to undetectable levels in the two dual antibody-sensitive individuals with the highest pre-infusion viral load despite persistent reductions for up to 12 weeks. Sustained suppression of plasma HIV-1 RNA levels to below 20 copies per ml was only achieved in individual 91C22, who had the lowest pre-infusion viral load (730 copies per ml). Thus, whereas two antibodies may be sufficient to achieve and/or maintain suppression in sensitive individuals with very low levels of viremia or ART-suppressed individuals undergoing analytical treatment interruption¹⁵, additional antibodies or combinations of small molecule drugs and antibodies would be required if this type of therapy is to be considered for viremic individuals.

This trial was limited to three bNAb infusions. However, despite the small number of infusions, sensitive individuals maintained reductions in viral load for up to three months after the last infusion. In the case of anti-RSV antibodies and the anti-HIV-1 antibody VRC01, antibody half-life can be increased by up to more than a factor of 4 by mutations that alter binding to the neonatal Fc receptor^{28–30}. In macaques, the same half-life extension mutations lead to a significant increase in the half-life and protective efficacy of 3BNC117 and 10-1074³¹. Should they also do so in humans, intermittent infusions of combinations of antibodies or antibodies plus long-acting antiretroviral drugs every 3–6 months might be an alternative to daily ART.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41591-018-0186-4>.

Received: 19 June 2018; Accepted: 16 August 2018;

Published online: 26 September 2018

References

1. Ndung'u, T. & Weiss, R. A. On HIV diversity. *AIDS* **26**, 1255–1260 (2012).
2. Bailey, J., Blankson, J. N., Wind-Rotolo, M. & Siliciano, R. F. Mechanisms of HIV-1 escape from immune responses and antiretroviral drugs. *Curr. Opin. Immunol.* **16**, 470–476 (2004).
3. Siliciano, J. D. et al. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4⁺ T cells. *Nat. Med.* **9**, 727–728 (2003).
4. Finzi, D. et al. Latent infection of CD4⁺ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat. Med.* **5**, 512–517 (1999).
5. Walker, L. M. & Burton, D. R. Passive immunotherapy of viral infections: 'super-antibodies' enter the fray. *Nat. Rev. Immunol.* **18**, 297–308 (2018).
6. Klein, F. et al. Antibodies in HIV-1 vaccine development and therapy. *Science* **341**, 1199–1204 (2013).
7. Caskey, M. et al. Viraemia suppressed in HIV-1-infected humans by broadly neutralizing antibody 3BNC117. *Nature* **522**, 487–491 (2015).
8. Caskey, M. et al. Antibody 10-1074 suppresses viremia in HIV-1-infected individuals. *Nat. Med.* **23**, 185–191 (2017).
9. Lynch, R. M. et al. Virologic effects of broadly neutralizing antibody VRC01 administration during chronic HIV-1 infection. *Sci. Transl. Med.* **7**, 319ra206 (2015).
10. Armbuster, C. et al. Passive immunization with the anti-HIV-1 human monoclonal antibody (hMAb) 4E10 and the hMAb combination 4E10/2F5/2G12. *J. Antimicrob. Chemother.* **54**, 915–920 (2004).
11. Mehandru, S. et al. Adjunctive passive immunotherapy in human immunodeficiency virus type 1-infected individuals treated with antiviral therapy during acute and early infection. *J. Virol.* **81**, 11016–11031 (2007).
12. Trkola, A. et al. Delay of HIV-1 rebound after cessation of antiretroviral therapy through passive transfer of human neutralizing antibodies. *Nat. Med.* **11**, 615–622 (2005).
13. Scheid, J. F. et al. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* **333**, 1633–1637 (2011).
14. Mouquet, H. et al. Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. *Proc. Natl Acad. Sci. USA* **109**, E3268–E3277 (2012).

15. Mendoza, P. et al. Combination therapy with anti-HIV-1 antibodies maintains viral suppression. *Nature* <https://doi.org/10.1038/s41586-018-0531-2> (2018).
16. Sarzotti-Kelsoe, M. et al. Optimization and validation of the TZM-bl assay for standardized assessments of neutralizing antibodies against HIV-1. *J. Immunol. Methods* **409**, 131–146 (2014).
17. Scheid, J. F. et al. HIV-1 antibody 3BNC117 suppresses viral rebound in humans during treatment interruption. *Nature* **535**, 556–560 (2016).
18. Cohen, Y. Z. et al. Neutralizing activity of broadly neutralizing anti-HIV-1 antibodies against clade B clinical isolates produced in peripheral blood mononuclear cells. *J. Virol.* **92**, e01883–17 (2018).
19. Keizer, R. J., Huitema, A. D., Schellens, J. H. & Beijnen, J. H. Clinical pharmacokinetics of therapeutic monoclonal antibodies. *Clin. Pharmacokinet.* **49**, 493–507 (2010).
20. Horwitz, J. A. et al. HIV-1 suppression and durable control by combining single broadly neutralizing antibodies and antiretroviral drugs in humanized mice. *Proc. Natl Acad. Sci. USA* **110**, 16538–16543 (2013).
21. Klein, F. et al. HIV therapy by a combination of broadly neutralizing antibodies in humanized mice. *Nature* **492**, 118–122 (2012).
22. Klein, F. et al. Enhanced HIV-1 immunotherapy by commonly arising antibodies that target virus escape variants. *J. Exp. Med.* **211**, 2361–2372 (2014).
23. Shingai, M. et al. Antibody-mediated immunotherapy of macaques chronically infected with SHIV suppresses viraemia. *Nature* **503**, 277–280 (2013).
24. Nishimura, Y. et al. Early antibody therapy can induce long-lasting immunity to SHIV. *Nature* **543**, 559–563 (2017).
25. Walker, B. D. & Yu, X. G. Unravelling the mechanisms of durable control of HIV-1. *Nat. Rev. Immunol.* **13**, 487–498 (2013).
26. Lynch, R. M. et al. HIV-1 fitness cost associated with escape from the VRC01 class of CD4 binding site neutralizing antibodies. *J. Virol.* **89**, 4201–4213 (2015).
27. Diskin, R. et al. Restricting HIV-1 pathways for escape using rationally designed anti-HIV-1 antibodies. *J. Exp. Med.* **210**, 1235–1249 (2013).
28. Gaudinski, M. R. et al. Safety and pharmacokinetics of the Fc-modified HIV-1 human monoclonal antibody VRC01LS: a phase 1 open-label clinical trial in healthy adults. *PLoS Med.* **15**, e1002493 (2018).
29. Ko, S. Y. et al. Enhanced neonatal Fc receptor function improves protection against primate SHIV infection. *Nature* **514**, 642–645 (2014).
30. Robbie, G. J. et al. A novel investigational Fc-modified humanized monoclonal antibody, motavizumab-YTE, has an extended half-life in healthy adults. *Antimicrob. Agents Chemother.* **57**, 6147–6153 (2013).
31. Gautam, R. et al. A single injection of crystallizable fragment domain-modified antibodies elicits durable protection from SHIV infection. *Nat. Med.* **24**, 610–616 (2018).
32. Zhou, T. et al. Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* **39**, 245–258 (2013).

Acknowledgements

We thank all study participants who devoted time to our research; members of the Klein and Nussenzweig laboratories for helpful discussions, especially P. Mendoza, C.-L. Lu, J. C. C. Lorenzi, L. Cohn and M. Jankovic; R. Levin, G. Kremer and D. Weiland for study coordination; the Rockefeller University Hospital Clinical Research Support Office and nursing staff as well as C. Golder, E. Thomas, M. Platten, S. Margane and T. Kümmerle for help with recruitment and study implementation; C. Ruping and M. Schlotz for help with sample processing; S. Kiss for ophthalmologic assessments; T. Keler and the Celldex Therapeutics team for 3BNC117 and 10-1074 manufacturing and regulatory support; C. Conrad for regulatory support; U. Kerkweg for pharmaceutical services; H. Janicki, M. Ercanoglu, P. Schommers and R. Kaiser for help with virus cultures; P. Fast and H. Park for clinical monitoring; S. McMillan, S. Mosher, S. Sawant, D. Beaumont, M. Sarzotti-Kelsoe, K. Greene, H. Gao and D. Montefiori for help with PK assay development, validation, reporting and/or project management; and S. Schlesinger for input on study design. This work was supported by The Bill and Melinda Gates Foundation Collaboration for AIDS Vaccine Discovery (CAVD) grants OPP1092074, OPP1124068 (M.C.N.), CAVIMC OPP1146996 (G.D.T., M.S.S.); the NIH grants 1UM1 AI100663 and R01AI-129795 (M.C.N.); the Heisenberg-Program of the DFG (KL 2389/2-1), the European Research Council (ERC-StG639961) and the German Center for Infection Research (DZIF) (FK.); the Einstein-Rockefeller-CUNY Center for AIDS Research (1P30AI124414-01A1); BEAT-HIV Delaney grant UM1 AI126620 (M.C.); and the Robertson fund. M.C.N. is a Howard Hughes Medical Institute Investigator.

Author contributions

M.C. (principal investigator in the United States), F.K. (principal investigator in Germany) and M.C.N. designed the trial; Y.B.-O., H.G., M.C., F.K. and M.C.N. analyzed the data and wrote the manuscript; Y.B.-O., T.S. and T.K. performed single-genome sequencing; H.G., A.L.B., K.M., M.W.-P., K.F., J.H., M.C. and F.K. implemented the study; Y.Z.C., R.M.G. and G.F. contributed to study design and implementation; C.L., I.S., C.W. and S.S. contributed to participant recruitment and clinical assessments; J.A.P. and T.Y.O. performed bioinformatics processing; H.G., L.N. and T.K. performed viral cultures; L.H. and N.P. contributed to statistical analyses; S.B., J.P.D., J.J.V., I.Sh. and K.J. performed, coordinated or contributed to sample processing; K.E.S., N.L.Y. and G.D.T. performed anti-idiotypic ELISAs; and M.S.S. performed neutralization assays.

Competing interests

There are patents on 3BNC117 and 10-1074 on which M.C.N. is an inventor.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41591-018-0186-4>.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.C. or F.K. or M.C.N.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Study design. We conducted a dose-escalation phase 1b study in HIV-1-infected individuals to evaluate the safety, pharmacokinetics and antiretroviral activity of the combination of the antibodies 3BNC117 and 10-1074 (<http://www.clinicaltrials.gov>; NCT02825797; EudraCT: 2016-002803-25). Study participants were enrolled sequentially into groups 1A, 1B, 1C and 3 according to eligibility criteria (Supplementary Fig. 1). Participants in groups 1A and 1B were virologically suppressed on ART and were randomized in a 2:1 ratio (six participants per group) to receive one intravenous infusion of 3BNC117 and 10-1074 (group 1A, 10 mg kg⁻¹ per antibody; group 1B, 30 mg kg⁻¹ per antibody) or placebo (sterile saline). Study participants and investigators were blinded to the assignment in groups 1A and 1B. Placebo recipients were not included in the data analysis. Viremic individuals off ART were enrolled in group 1C (four participants) or group 3 (three participants), and received one intravenous infusion (group 1C) or three intravenous infusions (group 3, every two weeks) of 3BNC117 and 10-1074 at a dose of 30 mg kg⁻¹. Participation in groups 1C and 3 was open-label. All study participants were followed for 24 weeks after the last administration of the antibodies or placebo. Participants off ART were encouraged to initiate ART six weeks after the last antibody infusion. Safety data are reported until the end of study follow-up. All participants provided written informed consent before participation in the study and the trial was conducted in accordance with Good Clinical Practice. The study protocol was approved by the Food and Drug Administration in the USA, the Paul-Ehrlich-Institute in Germany, and the Institutional Review Boards at the Rockefeller University and the University of Cologne.

Study participants. Study participants were recruited at the Rockefeller University Hospital, New York, USA, and at the University Hospital Cologne, Cologne, Germany. Eligible participants were HIV-1-infected adults aged 18–65 years with a current CD4⁺ T cell count >300 cells per μ l. Individuals on ART were eligible for participation and enrollment in groups 1A and 1B if HIV-1 RNA levels were <20 copies per ml at screening. Viremic individuals were eligible for enrollment in groups 1C and 3, if they were off ART with detectable HIV-1 RNA plasma levels of <100,000 copies per ml. Exclusion criteria included concomitant hepatitis B or C infection, previous receipt of monoclonal antibodies of any kind, clinically relevant physical findings, medical conditions or laboratory abnormalities, and pregnancy or lactation. Viremic participants were prescreened for the sensitivity of bulk CD4⁺ T cell outgrowth culture-derived viruses to 3BNC117 and 10-1074 as described below. Antibody sensitivity was defined as an IC₅₀ <2 μ g ml⁻¹ for both 3BNC117 and 10-1074 measured in a TZM-bl neutralization assay.

Study procedures. The required stock volume of 3BNC117 or 10-1074 was calculated according to body weight and diluted in sterile normal saline to a total volume of 250 ml. Each monoclonal antibody was administered intravenously over 60 min. Both antibodies were administered individually and sequentially. Placebo recipients received equivalent volumes of sterile normal saline. Study participants were observed at the Rockefeller University Hospital or the University Hospital Cologne for 4 h (groups 1A–C) or 1 h (group 3) after the last antibody infusion. Participants returned for scheduled follow-up visits for safety assessments, which included physical examination as indicated and measurements of clinical laboratory parameters, such as hematology, CD4⁺ T cell counts, chemistries, urinalysis and pregnancy tests. Plasma HIV-1 RNA levels were monitored at each visit. Study investigators evaluated and graded adverse events according to the Division of AIDS (DAIDS) Table for Grading the Severity of Adult and Pediatric Adverse Events (version 2.0, November 2014) and determined the causality of events. Blood samples were collected before and at multiple times after the infusions of 3BNC117 and 10-1074 or placebo. Samples were processed within 4 h of collection. Serum and plasma samples were stored at –80 °C. Peripheral blood mononuclear cells (PBMCs) were isolated by density gradient centrifugation and the absolute number of PBMCs was determined using an automated cell counter (Vi-Cell XR; Beckman Coulter) or manually. Isolated cells were cryopreserved in fetal bovine serum and 10% DMSO.

Plasma HIV-1 RNA levels. Plasma HIV-1 RNA levels were determined at every study visit, including the screening (day –49 to –7) and pre-infusion (day –42 to –2) visits, as well as before the first infusion on day 0 and two days after each infusion. Following the last infusion, HIV-1 RNA levels were monitored weekly for four weeks, and continued to be monitored with two- to four-week intervals and at the final study visit. HIV-1 RNA levels were determined using the Roche COBAS AmpliPrep/COBAS TaqMan HIV-1 Assay (version 2.0) or the Roche COBAS HIV-1 quantitative nucleic acid test (COBAS 6800). These assays have a linear quantification range between 2×10^3 and 1×10^7 viral copies per ml and were performed at LabCorp or at the University Hospital Cologne.

CD4⁺ and CD8⁺ T cell counts. CD4⁺ and CD8⁺ T cell counts were determined using a clinical flow cytometry assay performed at LabCorp or at the University Hospital Cologne every 2–4 weeks and at the final study visit.

TZM-bl neutralization assay to measure 3BNC117 and 10-1074 serum levels. This assay was performed as previously described¹⁶. In brief, serum samples were

heat-inactivated for 1 h at 56 °C and tested using a primary 1:20 dilution and a fivefold titration series against HIV-1 Env pseudoviruses Q769.d22 and X2088_c9. These pseudoviruses are highly sensitive to neutralization by 3BNC117 and 10-1074, respectively, and fully resistant against the other administered antibody. If serum half-maximum inhibitory dilution (ID₅₀) titers exceeded 100,000 against X2088_c9, immediate post-infusion levels of 10-1074 were also determined using the less sensitive Du422 strain. 3BNC117 and 10-1074 clinical drug products were tested in parallel at a starting concentration of 10 μ g ml⁻¹ with a fivefold titration series. Pseudoviruses were produced with an ART-resistant backbone vector that reduces the inhibitory activity of antiretroviral drugs (SG3ΔEnv/K101P.Q148H.Y181C, M.S.S., unpublished data). In viremic individuals, serum concentrations of 3BNC117 and 10-1074 were calculated by multiplying the determined ID₅₀ titer of the respective serum sample and the determined IC₅₀ concentration of each monoclonal standard antibody. In individuals on ART, serum tNAb concentrations were calculated using the ID₈₀ serum titers and IC₈₀ values of the monoclonal antibodies as described above to minimize the influence of nonspecific ART-mediated background activity. Viruses pseudotyped with the envelope protein murine leukemia virus (MuLV) were used as negative control and measurements were excluded if nonspecific serum activity against MuLV-pseudotyped viruses was observed (ID₅₀ or ID₈₀ >20 in viremic individuals or individuals on ART, respectively). All assays were performed in a laboratory that met Good Clinical Laboratory Practice standards. The lower limit of detection was determined to be 0.24 μ g ml⁻¹ and 0.10 μ g ml⁻¹ for the 3BNC117 and 10-1074 TZM-bl assay, respectively. The lower limit of quantification was 0.46 mcg/ml for 3BNC117 and 0.1 mcg/ml for 10-1074.

ELISA-based measurement of 3BNC117 and 10-1074 serum levels. Serum concentrations of 3BNC117 and 10-1074 were measured by a validated sandwich ELISA. High bind polystyrene plates were coated overnight at 2–8 °C with 4 μ g ml⁻¹ of an anti-idiotypic antibody that specifically recognizes 3BNC117 (anti-ID 1F1-2E3 monoclonal antibody) or 2 μ g ml⁻¹ of an anti-idiotypic antibody that specifically recognizes 10-1074 (anti-ID 3A1-4E11 monoclonal antibody). After washing, plates were blocked with 5% Milk Blotto (w/v), 5% normal goat serum (v/v), and 0.05% Tween 20 (v/v) in PBS. Serum samples, quality controls and standards were added (1:50 minimum dilution in 5% Milk Blotto (w/v), 5% normal goat serum (v/v) and 0.05% Tween 20 (v/v) in PBS) and incubated at room temperature. A horseradish peroxidase (HRP)-conjugated mouse anti-human IgG kappa-chain-specific antibody (Abcam) was used to detect 3BNC117 and an HRP-conjugated goat anti-human IgG Fc-specific antibody (Jackson ImmunoResearch) to detect 10-1074. For detection, the HRP substrate tetra-methylbenzidine was added. A 5-PL curve-fitting algorithm (Softmax Pro, v.5.4.5, Molecular Devices) was used to calculate serum concentrations of 3BNC117 and 10-1074 from respective standard curves run on the same plate. Standards and positive controls were created from the drug product lots of 3BNC117 and 10-1074 that were used in the clinical study. The capture anti-idiotypic monoclonal antibodies were produced in a stable hybridoma cell line (Duke Protein Production Facility⁷). If day 0 samples had measurable levels of antibody by the respective assays, the measured background antibody level was subtracted from subsequent results. In addition, samples with measured antibody levels within threefold of background values were excluded from the analysis of pharmacokinetic (PK) parameters. The lower limit of detection was determined to be 0.51 μ g ml⁻¹ and 0.14 μ g ml⁻¹ in HIV-1 seropositive serum for the 3BNC117 and 10-1074 ELISA, respectively. For values that were detectable (that is, positive for monoclonal antibodies) but below the lower limit of quantification, values are reported as <0.78 μ g ml⁻¹ and <0.41 μ g ml⁻¹ for 3BNC117 and 10-1074 ELISA.

SGA of viral env genes. SGA and sequencing of HIV-1 *env* genes was performed for plasma samples as described previously^{17,33}. All *env* sequences were translated to amino acids and aligned using ClustalW³⁴. Sequences containing premature stop codons or large internal deletions that would compromise Env functionality were removed from downstream analysis. Frequency plots were produced to analyze changes in 3BNC117 and 10-1074 binding sites between day 0 and rebound viruses. Amino acids were numbered according to the HXB2 *env* sequence (GenBank accession number K03455). Logo plots were generated using the 'longitudinal antigenic sequences and sites from intra-host evolution' tool (LASSIE)³⁵. Maximum likelihood phylogenetic trees were generated from the alignments with PhyML v.3.1³⁶ using the GTR model³⁷ with 1,000 bootstraps. For the combined analysis of sequences from all participants, *env* sequences were aligned using MAFFT v.7.309³⁸ and clustered using RAxML v.8.2.9 using the GTRGAMMA model³⁷ with 1,000 bootstraps.

Pseudovirus production. Selected viral sequences that were isolated from the plasma of each participant by SGA were used to generate CMV-promoter-based pseudoviruses as previously described^{33,39}. The CMV promoter was amplified using the forward primer 5'-AGTAATCAATTACGGGGTCATTAGTTCAT-3' and the reverse primer 5'-CATAGGAGATGCCAAGCCGGTGGAGCTCTGCTTATA TAGACCTC-3'. Individual *env* amplicons were amplified using the forward primer 5'-CACCGGCTTAGGCATCTCTATGGCAGGAAGAA-3' and the reverse primer 5'-GTCTCGAGATACTGCTCCACCC-3'. To fuse the individual

purified *env* amplicons to the CMV promoter, overlapping PCR was performed using the forward primer 5'-AGTAATCAATTACGGGGTCATTAGTTCAT-3' and the reverse primer 5'-ACTTTTTTGACCACTTGGCACCCAT-3'. Pseudoviruses were generated by transfecting HEK293T cells as previously described⁹.

Prescreening bulk PBMC culture. Candidate viremic individuals were prescreened for sensitivity of bulk culture-derived outgrowth viruses against 3BNC117 and 10-1074 as described previously^{7,8,15,17}. PBMCs for prescreening were obtained a median of 27 weeks (range 4.9–38 weeks) before enrollment under separate protocols approved by the Institutional Review Boards of the Rockefeller University and the University of Cologne. In brief, isolated CD4⁺ T cells were cocultured with the MOLT-4/CCR-5 cell line (NIH AIDS Reagent Program, cat. no. 4984) or CD8⁺ T cell-depleted donor lymphoblasts and culture supernatants were regularly monitored for p24 levels. Viral supernatants from p24-positive cultures were tested for sensitivity against 3BNC117 and 10-1074 by the TZM-bl neutralization assay as described below. Cultures were deemed sensitive if the determined individual IC₅₀ values for 3BNC117 and 10-1074 were <2 µg ml⁻¹.

Virus neutralization assays. Supernatants from p24-positive bulk CD4⁺ T cell cultures and pseudoviruses were tested for sensitivity to antibodies as previously described¹⁶.

Pharmacokinetic analyses. PK parameters were estimated by performing a non-compartmental analysis using Phoenix WinNonlin Build 8 (Certara), using all PK data available starting with the time point after the infusion of 3BNC117 from either TZM-bl assay or ELISA.

Viral *env* recombination analysis. Multiple sequence alignment of *env* genes guided by amino acid translations of *env* sequences was done by TranslatorX (<http://translatorx.co.uk/>). The 3SEQ recombination algorithm (<http://mol.ax/software/3seq/>) was used to detect recombination between day 0 viruses and rebound viruses or between different rebound viruses. Instances in which statistical evidence of recombination was found (rejection of the null hypothesis of clonal evolution) are shown in a circos plot (<http://circos.ca/>).

Statistical analyses. The sample size to detect a decline in viremia of >0.9 log₁₀ copies per ml with 80% power at 5% significance level with *P* of 0.05 was determined to be six viremic HIV-1-infected individuals, assuming that the standard deviation would be similar to 3BNC117 or 10-1074 monotherapy in humans (s.d. of 0.75 and 0.6, respectively)^{7,8}. To measure the effect of the combination treatment on viral load, we estimated simultaneous confidence bands for the Δlog₁₀ viral loads. The viral load was considered significantly suppressed whenever the two dashed lines representing the simultaneous confidence bands at 95% certainty level excluded zero (Fig. 2d–f). We computed simultaneous confidence bands with the R package locfit (version 1.5–9.1) using the Gaussian family for the local likelihood function (Fig. 2d–f). To estimate whether there is a significant difference between the 3BNC117 and 10-1074 combination therapy and 3BNC117 or 10-1074 monotherapy in viremic individuals off antiretroviral therapy, we fit a linear mixed-effects model to the data, using time and treatment as fixed effects and a random intercept for each participant. Data for 3BNC117 and 10-1074 monotherapy have been published previously and only time points

from viral load measurements off antiretroviral therapy and subjects responding to antibody infusions by a drop in viremia were included^{7,8}. We compared it to a model without treatment as predictor using a likelihood ratio test. The time point of viral load measurement was modeled as an ordered factor and the correlation structure between measurements from the same individual was modeled based on the order of measurements using different options available in nlme (exponential, linear, rational quadratic and spherical correlation structure, as well as different combinations of autocorrelation and moving average). The models were fitted maximizing the log-likelihood with the lme function of the R package nlme (version 3.1–131). We decided on the best model using Akaike information criterion (see Supplementary Fig. 5). Time points were restricted to day 0, week 1, week 2, week 3, week 4, week 6, week 8, week 12, week 16, week 20 and week 24 to have a sufficient number of measurements per time point. Marginal means (also known as least-squares means) are shown in Supplementary Fig. 5. CD4⁺ T cell counts before and after 3BNC117 plus 10-1074 infusions were compared by one-way ANOVA using GraphPad Prism (version 7.0).

Reporting Summary. Further information on research design can be found in the Nature Research Reporting Summary linked to this article.

Data availability

All requests for raw and analyzed data and materials are promptly reviewed by the Rockefeller University to verify whether the request is subject to any intellectual property or confidentiality obligations. Patient-related data not included in the paper were generated as part of clinical trials and may be subject to patient confidentiality. Any data and materials that can be shared will be released via a Material Transfer Agreement. HIV-1 envelope SGA data are available in GenBank, accession numbers MH632763–MH633255.

References

- Schoofs, T. et al. HIV-1 therapy with monoclonal antibody 3BNC117 elicits host immune responses against HIV-1. *Science* **352**, 997–1001 (2016).
- Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
- Hraber, P. et al. Longitudinal antigenic sequences and sites from intra-host evolution (LASSIE) identifies immune-selected HIV variants. *Viruses* **7**, 5443–5475 (2015).
- Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
- Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
- Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
- Kirchherr, J. L. et al. High throughput functional analysis of HIV-1 *env* genes without cloning. *J. Virol. Methods* **143**, 104–111 (2007).

A.5 Manuscript 5

Title:

Time-dependent prediction of mortality and cytomegalovirus reactivation after allogeneic hematopoietic cell transplantation using machine learning

Authors:

Lisa Eisenberg, the Xploit consortium, Christian Brossette, Jochen Rauch, Andrea Grandjean, Hellmut Ottinger, Jürgen Rissland, Ulf Schwarz, Norbert Graf, Dietrich W. Beelen, Stephan Kiefer, Nico Pfeifer & Amin T. Turki

Published in:

American Journal of Hematology, Volume 97, Pages 1309–1323
<https://doi.org/10.1002/ajh.26671>

Time of Publication:






August 2022

License information:

This article was published under a Creative Commons CC-BY-NC license (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial reuse and reproduction in any medium, provided the original work is properly cited.

RESEARCH ARTICLE

Time-dependent prediction of mortality and cytomegalovirus reactivation after allogeneic hematopoietic cell transplantation using machine learning

Lisa Eisenberg^{1,2}  | the XpIOit consortium | Christian Brossette³ |
Jochen Rauch⁴ | Andrea Grandjean⁵ | Hellmut Ottinger⁶ | Jürgen Rissland⁷ |
Ulf Schwarz⁸ | Norbert Graf³  | Dietrich W. Beelen⁶  | Stephan Kiefer⁴ |
Nico Pfeifer^{1,2}  | Amin T. Turki⁶ 

¹Department of Computer Science, University of Tübingen, Tübingen, Germany

²Institute of Bioinformatics and Medical Informatics (IBMI), University of Tübingen, Tübingen, Germany

³Department of Pediatric Oncology and Hematology, Saarland University, Homburg, Germany

⁴Department of Biomedical Data & Bioethics, Fraunhofer Institute for Biomedical Engineering (IBMT), Sulzbach, Germany

⁵Averbis GmbH, Freiburg, Germany

⁶Department of Hematology and Stem Cell Transplantation, University Hospital Essen, Essen, Germany

⁷Institute of Virology, Saarland University Medical Center, Homburg, Germany

⁸Institute for Formal Ontology and Medical Information Science (IFOMIS), Saarland University, Saarbrücken, Germany

Correspondence

Amin T. Turki, Department of Hematology and Stem Cell Transplantation & Computational Hematology Lab, University Hospital Essen, Essen 45147, Germany. Nico Pfeifer, Department of Computer Science, University of Tübingen, Tübingen 72076, Germany. Email: amin.turki@uk-essen.de; nico.pfeifer@uni-tuebingen.de

Funding information

Bundesministerium für Bildung und Forschung, Grant/Award Number: 031L0027A-F; Deutsche Forschungsgemeinschaft, Grant/

Abstract

Allogeneic hematopoietic cell transplantation (HCT) effectively treats high-risk hematologic diseases but can entail HCT-specific complications, which may be minimized by appropriate patient management, supported by accurate, individual risk estimation. However, almost all HCT risk scores are limited to a single risk assessment before HCT without incorporation of additional data. We developed machine learning models that integrate both baseline patient data and time-dependent laboratory measurements to individually predict mortality and cytomegalovirus (CMV) reactivation after HCT at multiple time points per patient. These gradient boosting machine models provide well-calibrated, time-dependent risk predictions and achieved areas under the receiver-operating characteristic of 0.92 and 0.83 and areas under the precision–recall curve of 0.58 and 0.62 for prediction of mortality and CMV reactivation, respectively, in a 21-day time window. Both models were successfully validated in a prospective, non-interventional study and performed on par with expert hematologists in a pilot comparison.

Nico Pfeifer and Amin T. Turki contributed equally to the manuscript.

A list of all members of the XpIOit consortium is provided at the end of this article.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *American Journal of Hematology* published by Wiley Periodicals LLC.

1 | INTRODUCTION

Allogeneic hematopoietic cell transplantation (HCT) is an effective and potentially curative treatment for patients suffering from high-risk hematological malignancies and other non-malignant and congenital disorders.¹ Despite its success and continuous improvement over the past decades,^{2,3} the treatment-related non-relapse mortality (NRM) after HCT remains high. HCT recipients are at risk for multiple potentially life-threatening complications, such as graft-versus-host disease (GVHD) or cytomegalovirus (CMV) reactivation. Accurate risk assessment and an appropriate choice of prophylactic and pre-emptive treatments are crucial to minimize these risks.^{4,5} Registries such as the databases of the European Society for Blood and Marrow Transplantation (EBMT) or of the Center for International Blood and Marrow Transplant Research (CIBMTR) collect individual patients' pre-HCT and outcome data from centers via standardized reporting forms.^{6,7} Using these databases, the prevalence and risk factors of HCT complications can be analyzed on a large scale. Due to the data collection process, registry data per patient is limited to a set of categorical variables. While time-dependent endpoint data are available regarding the time of relapse or death, continuously measured laboratory values from electronic health records (EHR), or unstructured data from reports cannot yet be integrated into these registries. Since the 2000s, a number of relevant predictive risk scores have been developed utilizing static registry data to improve outcome assessment before HCT and to adjust the toxicity of the intervention by reducing the conditioning intensity. The hematopoietic cell transplantation-specific comorbidity index (HCT-CI)⁸ is to date the most relevant and utilized score to predict NRM. Other Cox-regression models based on categorical, pre-HCT variables, such as the EBMT risk score⁹ or the disease risk index,¹⁰ have additionally improved pre-HCT and relapse risk assessment for different hematologic malignancies. However, the overwhelming majority of existing methods for assessing such HCT-specific risks offer only a single risk assessment before HCT.

Across medical areas, machine learning (ML) techniques have proven their value as powerful tools for diagnosis¹¹⁻¹⁴ or risk assessment.¹⁵⁻¹⁷ ML models are ideally suited to discover associations in large datasets and can automatically identify important parameters and relationships between them without the need for a predefined model shape. In recent years, several ML models have been proposed for HCT-specific risk assessment at a single point in time.¹⁸⁻²⁰ For instance, an alternating decision tree model produced more accurate predictions of 100-day mortality after HCT than the EBMT score for acute leukemia patients,¹⁸ demonstrating that ML can improve standard scores for pre-HCT risk assessment.

The endothelial activation and stress index (EASIX) measured before conditioning therapy is associated with overall survival after HCT, highlighting the potential of including laboratory parameters in pre-HCT risk assessment.²¹ Additionally, EASIX measured at the onset

of acute GVHD predicts overall survival after GVHD onset.²² Despite its added value, EASIX is calculated from a limited set of three parameters (creatinine, platelets, LDH) using a predefined formula, and each study only evaluated its prognostic value at a single point in time.

Integrating time-dependent measurements into ML models can not only improve predictive performance but also allows to update risk assessments whenever new data become available. For instance, early-warning systems developed for intensive care units (ICU) continuously monitor patient data and predict critical events such as acute kidney injury¹⁶ or circulatory failure,¹⁷ which may help physicians to react earlier to critical events or to prevent them. Given the high variability of individual outcomes after HCT and the importance of optimal patient management, we hypothesized that ML-based models for precise, time-dependent risk prediction after HCT may provide a valuable tool to support treatment decisions. Compared with the large, annotated, public EHR datasets of ICU patients,^{17,23} time-dependent HCT data are scarce. Their use in ML models is further challenged by high variability in laboratory measurement frequencies and a characteristic nonlinear development of laboratory values after HCT, which requires context-dependent evaluation of identical numerical results. In addition, longer observation times may entail missing values and censored data. Major national and international efforts are currently directed toward digitizing medicine,^{24,25} developing unified standards for data management, and facilitating the increasingly widespread use of EHR systems. As a consequence, we expect the accessibility and usability of health data to improve, with impacts on different fields of medicine including HCT care.

In this article, we describe the development and prospective validation of ML models, which accurately predict death and early CMV reactivation at multiple time points after HCT. These are the first models for continuous time-dependent risk assessment of these outcomes after HCT.

2 | METHODS

2.1 | Patients

Between January 2005 and June 2020, 2191 patients with hematologic malignancies, inherited stem cell disorders, or acquired bone marrow failure underwent allogeneic HCT in the Department of Hematology and Stem Cell Transplantation of the West-German Cancer Center at University Hospital Essen (UHE). Patients with HCT before September 1, 2017 were included in the retrospective cohort. Patients with HCT between September 2017 and June 2020 were prospectively recruited into the non-interventional XploIt validation study (Figure S1). We excluded patients with multiple allogeneic HCTs, with hemoglobinopathies or without data on relevant laboratory tests, resulting in retrospective and prospective cohort sizes of

1710 and 403 patients, respectively. For models and analyses related to CMV reactivation, we additionally excluded patients without CMV data before day +30 after HCT. Donors were HLA-matched related donors (MRD, 23.0%), haploidentical related donors (haplo, 3.8%), 10/10 HLA-A-, -B, -C, -DRB1, -DQB1 matched unrelated donors (MUD, 53.6%), or mismatched unrelated donors (MMUD, 19.6%; Table S1). HLA-DPB1 was not considered for donor–recipient matching. Typically, patients were followed up for 60 months after transplantation. Long-term surviving patients were censored. Early supportive and follow-up care was identical for all patients. In the retrospective cohort, the predominant calcineurin inhibitor based GVHD prophylaxis consisted of Ciclosporin A plus Methotrexate. Patients with higher GVHD risk were assigned to additional in vivo T cell depletion using anti-T-Lymphocyte globulin (ATG) based on standardized clinical treatment protocols.

2.2 | Ethics

This study was conducted in accordance with German legislation and the revised Helsinki Declaration. Study design and data acquisition were evaluated by the institutional review board (IRB) of the University Duisburg-Essen (Protocol No. 17-7576-BO) and by the IRB of the medical association of the Saarland (Protocol No. 33/17). All patients included in the prospective, non-interventional XplOit study (registered in the German Clinical Trials Register (DRKS), registration No. DRKS00026643) have given written consent to collection, electronic storage, and scientific analysis of pseudonymized HCT-specific patient data.

2.3 | Data preparation, endpoint assessment, and statistical analysis

The sections on data preparation, endpoint assessment, and statistical analysis are detailed in the supplementary material.

2.4 | Preprocessing

We selected 60 features for model development, including all static features available in structured format, the prediction day, and 34 of the most frequently performed laboratory tests (Table S3). For time-dependent laboratory tests, we only used the most recent value of each parameter at the time point of prediction. Static and time-dependent features were preprocessed separately and concatenated directly before model training. Preprocessing is detailed in the supplement.

2.5 | Prediction times and classification target

We aimed for the application scenario where models are executed once per day whenever new time-dependent data become available.

Therefore, we considered all days between HCT and the event of interest (or censoring) where any laboratory measurements were reported as potential prediction days.

For each event (death or CMV reactivation), we defined binary classification targets based on two different window sizes d of 7 and 21 days, respectively. Each time point was labeled with 1 (positive) if the event occurred within the following d days and 0 (negative) otherwise. We excluded time points where patients were censored in this time window or where more than 30% of time-dependent features were missing after forward filling. For prediction of CMV reactivation, we considered only events in the first 100 days and excluded prediction days after day $100 - d$. The final number of time points is listed in Table S5 for each prediction task.

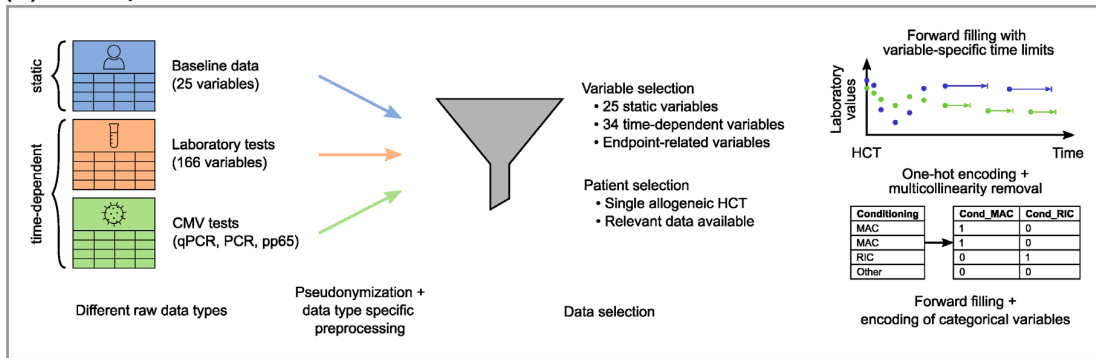
2.6 | Machine learning models and training

We trained gradient boosting machine (GBM) models using LightGBM, which provides an efficient implementation of a gradient boosted ensemble of decision trees.²⁶ For the comparative L2-regularized logistic regression (LR) model and baseline we used the LogisticRegressionCV class in scikit-learn.²⁷

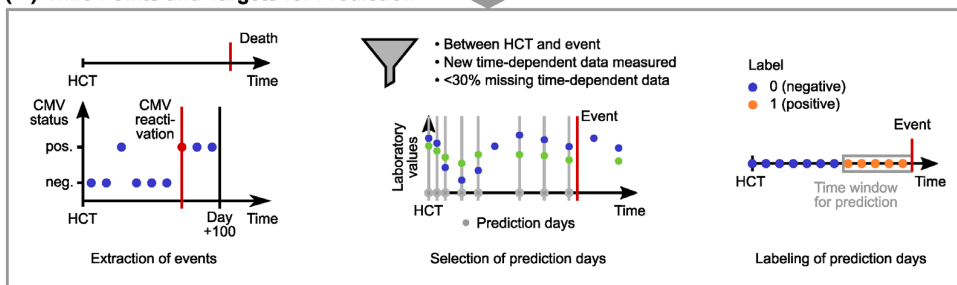
For both model types, we optimized hyperparameters with grid search and five-fold cross-validation (CV). CV folds were defined on patient level to ensure their independence and were stratified by the maximum label per patient. We selected the parameters producing the highest mean sample-AUPRC and retrained on the full training set with these parameters. For GBM models, we used early stopping during CV to determine the number of boosted trees in the ensemble. For each combination of hyperparameters, model training was stopped early when the mean logistic loss over CV folds did not improve for 50 iterations. When retraining on the full training set, we used the number of boosted trees, which produced the lowest logistic loss during CV. The exact parameter choices, grids, and optimal values are provided in Table S6.

To evaluate model performance and variability on retrospective data, we repeatedly split the patients of the retrospective cohort into two-thirds training and one-third test set (stratified by the maximum label per patient). We ran the entire training process, including imputation, normalization, and hyperparameter search, on each training set independently and evaluated model performance on the corresponding test set using AUROC, sample-AUPRC, and event-AUPRC. Here, sample-AUPRC is the standard area under the precision–recall curve, where recall is defined as the fraction of correctly predicted samples (i.e., time points) with positive label (sample recall). In contrast, event-AUPRC defines recall as the fraction of events, which were correctly predicted on at least one of the positive labeled time points (event recall) and was previously introduced for time-dependent event prediction.¹⁷ Unless specified otherwise, model performance on retrospective data is reported as mean and SD over 10 random splits into training and test set. Using the same methodology, we additionally trained a final model on the entire retrospective cohort for prospective validation.

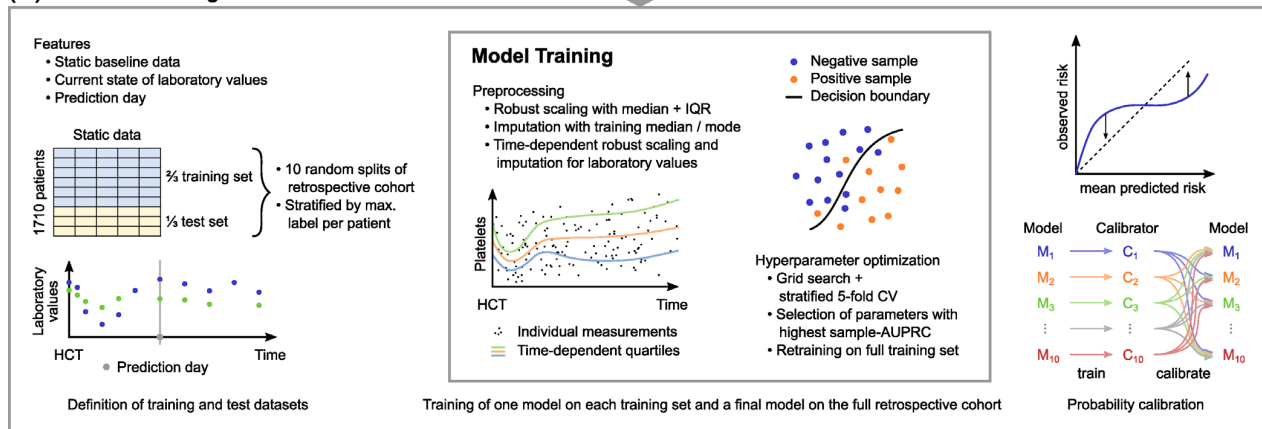
(A) Data Preparation



(B) Time Points and Targets for Prediction



(C) Machine Learning



(D) Model Evaluation

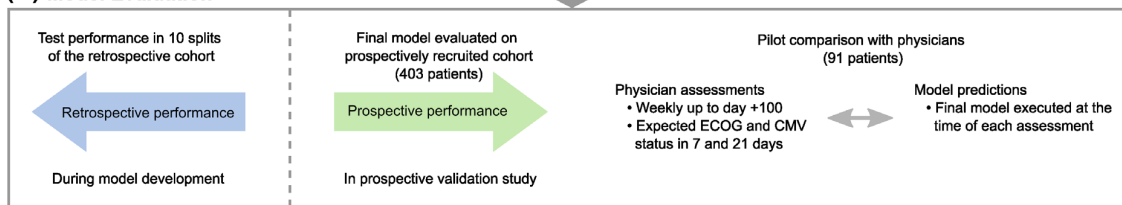


FIGURE 1 Legend on next page.

2.7 | Models with additional features

We evaluated whether additional information from unstructured medical letters or information on the history of laboratory values improved

the performance of survival and CMV prediction. For this purpose, we trained two further GBM models per task, which received additional input features (Table S8). Since the added features led to little or no performance improvement on the retrospective data (Figure S15), we

selected only the simpler models with the initial feature set for prospective validation. An overview of all developed models and the included features is provided in Table S9.

2.8 | Model calibration

We calibrated all trained models as a postprocessing step using isotonic regression. For this purpose, we trained a separate calibrator for each split of the retrospective cohort into training and test set, using the raw model predictions on the test set. To apply calibration to any of the models trained for these splits, we averaged the output of the nine calibrators trained on the remaining splits (Figure 1C). The predicted probabilities of the final model trained on the entire retrospective cohort were calibrated using the average over all 10 calibrators.

2.9 | Prospective validation

In order to prospectively validate the developed models on an independent cohort, we recruited 408 patients to the prospective non-interventional XplOit study (inclusion criteria: first allogeneic HCT, ≥ 18 years, written informed consent) from September 2017 to June 2020. We applied the final GBM and LR models to generate predictions on the prospective cohort, selecting prediction times with the same methodology described for the retrospective cohort. Throughout the prospective study, both physicians and patients were blinded for the model predictions.

We compared model predictions to the observed outcome and measured performance with the same metrics used on retrospective data. To assess variability in performance measures, we applied bootstrapping with 10 000 bootstrap samples on the prospective dataset. During bootstrapping, we kept the total number of positive labeled samples fixed at its original value and adjusted the number of negative labeled samples to obtain the same positive fraction as observed in the retrospective dataset to enable a direct performance comparison between retrospective and prospective cohort.

2.10 | Head-to-head comparison to physicians' expectations

Within the last quarter of prospectively recruited patients, we performed a pilot study to compare the performance of the developed ML models to the expectations of experienced physicians regarding early complications after HCT. For 91 patients in the prospective cohort, we prospectively assessed the expectations of the treating physicians regarding overall survival and CMV reactivation between day 0 and day +100 after HCT. Physicians were requested to estimate each patient's performance status (ECOG, 0–5) and risk to have a CMV reactivation (low, moderate, high) in 7 and 21 days after the assessment date. Assessment was performed weekly between day –7 and day +100 after HCT by physicians of the Department of Hematology and Stem Cell Transplantation at UHE. Whenever an assessment was made (starting at HCT), the GBM and LR models were executed on the most recent available data to allow for a head-to-head comparison of the predictions. Treating and risk assessing physicians were blinded for the model predictions.

To enable model predictions on the day of each assessment, we used indefinite forward filling on laboratory measurements for this analysis. Since the physicians' assessments were recorded as categories rather than probabilities, we binarized their answers and the model predictions, and compared performance measures on these binary predictions. Specifically, we compared Matthews correlation coefficients (MCC) and F1 scores, choosing the optimal binarization threshold for models and physicians, respectively. To assess variability, we repeated this evaluation on 10 000 bootstrap samples drawn from the dataset for this pilot comparison. Here, we kept the positive fraction fixed by drawing the same number of samples with positive and negative labels, respectively, as were originally in the dataset.

2.11 | Implementation

Preprocessing was in part performed within the XplOit platform (version 20201130_1700) using extract–transform–load pipelines specific to each data type. All remaining steps of preprocessing, model

FIGURE 1 Overview of model development and evaluation. (A) Data preparation. Raw data tables were pseudonymized and combined into one coherent dataset. After patient and variable selection, sparsity in laboratory values was reduced by forward filling with variable-specific time limits and categorical features were converted into a binary representation. (B) Time points and targets for prediction. Of the two considered events, death was directly documented and CMV reactivation was extracted from virological tests as the first positive CMV test, which was not an isolated positive. We selected all days between HCT and an event or censoring as prediction days where new laboratory values were measured and <30% of them were missing. Each prediction day was labeled positive if the event occurred in a fixed subsequent time window, and negative otherwise. (C) Machine learning. Models received static baseline data, current laboratory values, and the prediction day after HCT as inputs. We randomly split the retrospective cohort into training and test sets 10 times, and trained a separate model on the training set of each split and a final model on the full retrospective cohort. We defined the splits on patient level and stratified the proportion of patients with at least one positive labeled time point. Preprocessing included a time-dependent normalization and imputation of laboratory values. We trained one calibrator for each split into training and test set. To calibrate each model, we averaged over the calibrators trained on the remaining splits or over all calibrators in case of the final model. (D) Model evaluation. During model development, performance was evaluated on the test set of the 10 splits of the retrospective cohort. In a prospective validation study, we additionally evaluated the performance of the final model on 403 prospectively recruited patients and, in a subset of 91 patients, performed a pilot comparison with experienced HCT physicians.

building, and analysis were implemented in python (version 3.8.2) using scikit-learn (version 0.22.1),²⁷ numpy (version 1.18.1),²⁸ and pandas (version 1.0.3).²⁹ GBM models were trained with LightGBM (version 2.3.1)²⁶ and SHAP values for these models were computed using the TreeExplainer implemented in shap (version 0.37.0).³⁰

3 | RESULTS

Using ML, we developed GBM and LR models to predict at multiple time points after HCT whether an event, that is, death or CMV reactivation, would occur in a subsequent time window of 21 or 7 days (Figure 1A–C). Each model received a combination of routinely collected static and time-dependent HCT data as input and was trained to predict a continuous risk score for one specific event. We then validated these ML models in the prospective non-interventional XpIOit study, which also included a pilot comparison between ML model predictions and prospectively collected outcome expectations of experienced HCT physicians (Figure 1D).

3.1 | Assembling an extensive longitudinal HCT dataset

Utilizing the XpIOit data integration platform for medical research,³¹ we assembled an extensive, well-annotated retrospective dataset incorporating static and time-dependent data of 1710 HCT patients to form the basis of model development. Based on their relevance, we selected 60 parameters as input features for the ML models (Figure 1), including static pre-HCT constellations, such as diagnosis, conditioning regimen, and donor information, as well as the day of the prediction and current laboratory values (Table S3). During the non-interventional XpIOit validation study, we additionally recruited 403 patients for prospective model validation.

Relevant baseline characteristics were balanced between the development and validation cohort and are detailed in Table S1. As expected, the largest fraction of patients presented with acute myeloid leukemia for HCT. Cyclosporin A (CSA) was the predominant calcineurin inhibitor for baseline immunosuppression. Following changes in HCT practices, such as the introduction of post-transplant cyclophosphamide, the prospective cohort had a higher proportion of patients with tacrolimus-based immunosuppression. Time-dependent laboratory values were available at 163 425 and 31 889 time points in the retrospective and prospective cohort, respectively, comprising more than 5.4 million individual measurements in total. In accordance with international best-practice HCT guidelines, the measurement intervals were shortest during the inpatient care of 35–40 days and were extended for outpatients (Figure S5).

The endpoints of this study were adequately covered by the analyzed data. The time of death was known for 1134 patients (53.7%), and 925 patients (43.8%) developed an early CMV reactivation (within 100 days after HCT), with the median first episode of CMV

reactivation at day +34. After 24 months, the overall survival (OS) rate was 55% in the retrospective cohort (Figure S2a), which is representative of HCT outcomes across different risk groups in real-world data. After a median follow-up of 14.4 months, the median overall survival was not reached in the prospective XpIOit study (Figure S2b). While the cumulative incidence of NRM was comparable between the retrospective and prospective cohort, overall survival differed significantly consistent with reduced relapse rates in recent HCT (Figure S2c). The GBM model predicts 21-day mortality with an AUROC of 0.92 and an event-AUPRC of 0.58.

We evaluated model performance using the standard area under the receiver-operating characteristic (AUROC) and two versions of the area under the precision–recall curve (AUPRC), event-AUPRC and sample-AUPRC. While sample-AUPRC is based on the standard recall on individual samples, event-AUPRC defines recall as the fraction of correctly predicted events and specifically addresses time-dependent event prediction.¹⁷ Following data preprocessing, as detailed in the Methods section, the retrospective dataset for the development of 21-day mortality models contained 143 669 time points of 1695 patients, 7354 of these time points (5.14%) were labeled positive (death occurred within 21 days).

The developed GBM model for 21-day mortality prediction achieved a very high AUROC of 0.918 and good event AUPRC of 0.584 (Figure 2A,B). It outperformed the LR model, which had an AUROC of 0.900 and an event-AUPRC of 0.524. To assess the value of including time-dependent data for outcome prediction, we compared these models with a baseline LR model receiving only static input data. The time-dependent GBM and LR models both vastly outperformed the static LR baseline, which achieved an AUROC of only 0.594 and event-AUPRC of 0.085. The same trend was observed in sample-AUPRC (Figure S6). After calibration, we obtained very close agreement between predicted and observed risk, with areas of 0.04 and 0.06 between the line representing ideal calibration and the calibration curve of the GBM and the LR model, respectively (Figure S7).

We then analyzed the performance of the GBM model for 21-day mortality prediction over time in more detail. As expected, the fraction of correctly predicted events increased with shorter time to the event (Figure S3a). This finding was independent of the exact threshold chosen to convert continuous risks into binary event predictions. With a threshold chosen to obtain an overall event recall of 0.8, the majority of events was predicted at least 2 weeks in advance. The predicted continuous risks evolved similarly with a steady increase as patients approached an event (Figure 2C), which supports the plausibility of the model. Compared with the average risk predicted for negatively labeled time points, that is, without any event in the subsequent 21 days, this increase was detectable as early as 85 days before the event. Although the GBM model recognized initial signs of an impending event much earlier than 21 days before, these were not yet sufficient for a confident event prediction. Analyzing GBM model performance as a function of the prediction day after HCT, we found that AUROC increased slightly over time (Figure 2D). Sample-AUPRC varied more noticeably; it was lower early after HCT and highest

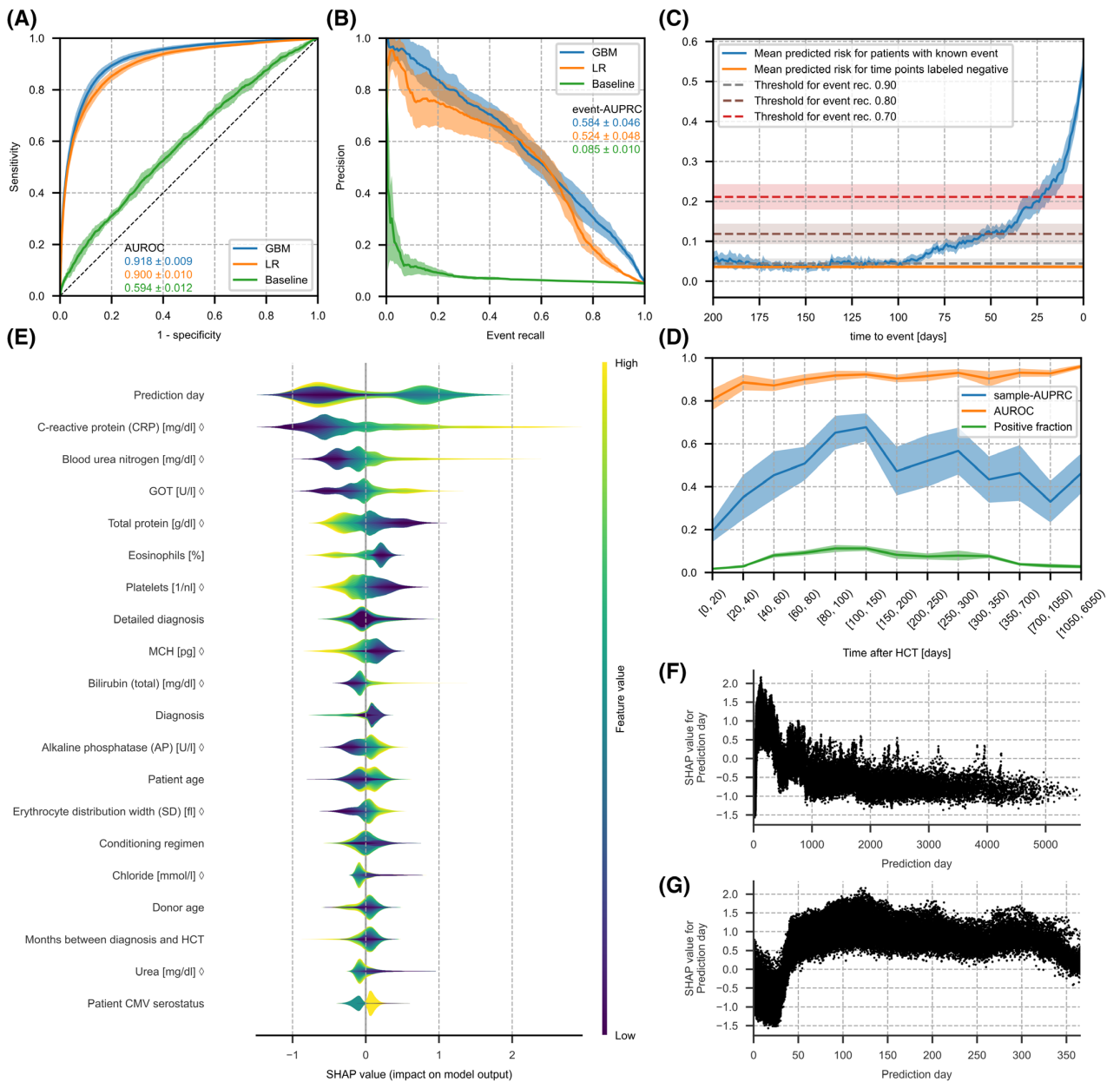


FIGURE 2 Performance and feature importance of the GBM model for 21-day mortality prediction. (A) Receiver-operating characteristic of GBM and LR model, which received a combination of static and time-dependent input features, and a baseline model which received only static features. (B) Precision-recall curve for the same models as shown in (A), based on event recall, that is, the fraction of events that were correctly predicted on any of the previous 21 days. (C) Mean predicted risk of the GBM model as a function of time to event. For reference, the orange horizontal line indicates the mean predicted risk over all time points labeled negative. Dashed horizontal lines indicate the thresholds to reach the target event recall stated in the figure legend. (D) AUROC and sample-AUPRC of the GBM model and fraction of samples with positive label as functions of time after HCT. Bin size increases because fewer samples were available late after HCT. (A–D) Lines and shaded areas show the mean \pm SD on the test set over 10 random splits into training and test data. (E) Layered violin plot of SHAP values of the GBM model for the 20 features with highest mean absolute SHAP value. The thickness of the violins corresponds to the estimated density of each feature's SHAP values, colors show the magnitude of feature values (percentiles). For features marked with \diamond , the feature value is the time-normalized score that the model received as input, not the raw value in its original unit. For categorical features, the colors are based on an integer representation and should not be interpreted as ordered. All SHAP values were computed based on raw model output in log-odds space. (F–G) Scatter plots of individual SHAP values over feature values. Shown are plots for the feature prediction day after HCT on the entire range of feature values (F) and zoomed in on the first year after HCT (G).

between days 80 and 150. This correlated with the fraction of positive labeled samples at different times after HCT since a small positive fraction makes it difficult to achieve a high precision score.

3.2 | Prediction day, CRP, and urea nitrogen had the highest impact on mortality predictions

Using SHapley Additive exPlanations (SHAP values),³⁰ we analyzed the impact of individual features on GBM model predictions (Figure 2E). SHAP values indicate how much the value of a feature has contributed to the prediction generated for a specific sample. High values (>0) indicate that the feature value increased the predicted risk, while low values (<0) indicate that it reduced the predicted risk. For the GBM model predicting 21-day mortality, the most important features were the day of the prediction (in days after HCT), C-reactive protein (CRP), blood urea nitrogen, glutamate oxaloacetate transaminase (GOT), and protein levels (Figure 2E). Especially high blood levels of CRP, urea nitrogen, and GOT as compared with other patients at the same time after HCT led the model to predict an increased mortality risk. In contrast, high values of total protein led to a lower predicted risk. These features are markers of inflammation or infection, or reflect liver or kidney function. For the prediction day after HCT, the relationship between feature value and SHAP value was more complex. Within the first year after HCT, the prediction day appeared to increase the predicted risk, while after 1 year the SHAP values continuously decreased, falling below zero about 3 years after HCT (Figure 2F). A closer inspection of the first year after HCT revealed that prediction days up to day +40 decreased the predicted risk, while all later prediction days of the first year had constantly high SHAP values (Figure 2G).

For 7-day mortality prediction, the GBM and LR models both had a higher AUROC and lower event- and sample AUPRC than the corresponding 21-day models (Figure S8). As a consequence of the narrower time window, fewer samples were labeled positive (1.88% for 7-day prediction), which can partially explain the lower event and sample-AUPRC. Detailed results for the 7-day prediction models are provided in the supplementary material (Figures S8 and S9). While these models focused on all-cause mortality to enable prediction for all HCT patients, independently of their relapse status, we also tested if our modeling approach would result in comparable prediction performance for NRM, which was indeed confirmed (Table S10b).

3.3 | The performance of 21-day mortality prediction models remained high on prospective data

In the second step, we validated the developed ML models on an independent prospectively recruited cohort ($n = 403$) from the same HCT center (Table 1A). Depending on the time window for prediction, we observed specific differences in the performance of mortality prediction on prospective data. The models for 21-day mortality

prediction remained relatively stable; AUROC and event-AUPRC of the GBM model faded only slightly from 0.918 to 0.895 and from 0.584 to 0.522, respectively. Responding to changes in HCT practices, we additionally compared subgroups of the two main distinct immunosuppressive regimens (CSA and TAC) within the prospective cohort (Table S2), and found no major differences between these subgroups. However, for 7-day prediction, we observed a quite pronounced decrease in model performance on prospective data, with AUROC and event-AUPRC of the GBM model dropping from 0.951 to 0.931 and from 0.525 to 0.372, respectively. Here, model performance was noticeably higher for patients with CSA instead of TAC immunosuppression, which were better represented in the retrospective cohort. Model calibration remained appropriate on prospective data (Figure S10).

Despite some differences between retrospective and prospective patient outcomes and model performance, the AUROC of both GBM and LR models remained high on prospective data. Event- and sample-AUPRC were also acceptable given the low fraction of positive labeled time points. Next, we tested if the models trained to predict all-cause mortality could also be leveraged to predict NRM. The validation of the GBM model for 21-day mortality on the subgroup of 361 prospectively recruited patients without relapse resulted in a comparably high AUROC of 0.900, an event-AUPRC of 0.536, and a sample-AUPRC of 0.428 (Table S10a). Thus, the developed ML models were successfully validated on the prospective dataset for both all-cause mortality and NRM.

3.4 | For 21-day mortality prediction the GBM models performed similar to HCT physicians

In a pilot study, which was part of the prospective validation, we additionally compared the predictive performance of the final GBM and LR models during the first 100 days after HCT to the outcome expectations of experienced HCT physicians. Within the last year of the prospective Xploit study, each treating physician was requested once per week to estimate their patients' expected Eastern Cooperative Oncology Group (ECOG) performance score and risk of CMV reactivation (low, medium, high) in 7 and 21 days. In total, we collected 649 forms containing post-HCT assessments for 91 patients. In parallel, we executed GBM and LR models at the time of each assessment with the latest available time-dependent data. All physicians were blinded to the model predictions.

The results of this comparison are displayed in Table 1B. For 21-day mortality prediction, GBM model and physicians showed a similar performance, as measured by MCC values of 0.461 ± 0.086 and 0.488 ± 0.089 , respectively. Although the differences were small compared with the SD derived from bootstrapping, trends showed a slight advantage of the physicians' expectations over the GBM model predictions and of the GBM model over the LR model. For 7-day prediction, the physicians achieved a very high MCC and F1 score of 0.796 ± 0.180 and 0.767 ± 0.214 , respectively,

TABLE 1 Model performance on prospective data and comparison of the prediction performance of ML models and treating physicians

A. Comparison of model performance on retrospective and prospective cohort ^a				
Prediction task	Model	Performance metric	Retrospective cohort	Prospective cohort
Mortality 21 days	GBM	AUROC	0.918 ± 0.009	0.895 ± 0.005
		Event-AUPRC	0.584 ± 0.046	0.522 ± 0.023
		Sample-AUPRC	0.488 ± 0.042	0.414 ± 0.015
	LR	AUROC	0.900 ± 0.010	0.866 ± 0.006
		Event-AUPRC	0.524 ± 0.048	0.549 ± 0.021
		Sample-AUPRC	0.445 ± 0.043	0.413 ± 0.015
Mortality 7 days	GBM	AUROC	0.951 ± 0.006	0.931 ± 0.006
		Event-AUPRC	0.525 ± 0.038	0.372 ± 0.029
		Sample-AUPRC	0.410 ± 0.034	0.303 ± 0.021
	LR	AUROC	0.940 ± 0.008	0.894 ± 0.009
		Event-AUPRC	0.464 ± 0.038	0.348 ± 0.026
		Sample-AUPRC	0.375 ± 0.023	0.269 ± 0.020
CMV 21 days	GBM	AUROC	0.825 ± 0.006	0.846 ± 0.004
		Event-AUPRC	0.620 ± 0.040	0.574 ± 0.011
		Sample-AUPRC	0.565 ± 0.025	0.549 ± 0.009
	LR	AUROC	0.793 ± 0.013	0.818 ± 0.004
		Event-AUPRC	0.532 ± 0.050	0.515 ± 0.012
		Sample-AUPRC	0.502 ± 0.033	0.496 ± 0.009
CMV 7 days	GBM	AUROC	0.846 ± 0.010	0.875 ± 0.005
		Event-AUPRC	0.335 ± 0.023	0.323 ± 0.015
		Sample-AUPRC	0.295 ± 0.017	0.302 ± 0.012
	LR	AUROC	0.777 ± 0.014	0.802 ± 0.006
		Event-AUPRC	0.192 ± 0.017	0.176 ± 0.007
		Sample-AUPRC	0.188 ± 0.014	0.181 ± 0.006
B. Comparison of the prediction performance of ML models and treating physicians ^b				
Prediction task	Performance metric	Physicians	GBM	LR
Mortality 21 days	MCC	0.488 ± 0.089	0.461 ± 0.086	0.417 ± 0.087
	F1 score	0.453 ± 0.086	0.427 ± 0.085	0.360 ± 0.084
Mortality 7 days	MCC	0.796 ± 0.180	0.377 ± 0.064	0.304 ± 0.069
	F1 score	0.767 ± 0.214	0.272 ± 0.077	0.204 ± 0.069
CMV 21 days	MCC	0.234 ± 0.051	0.329 ± 0.062	0.266 ± 0.023
	F1 score	0.289 ± 0.055	0.322 ± 0.049	0.281 ± 0.026
CMV 7 days	MCC	0.170 ± 0.067	0.147 ± 0.033	0.143 ± 0.042
	F1 score	0.168 ± 0.063	0.110 ± 0.025	0.117 ± 0.030

^aFor the retrospective cohort, the table displays mean ± SD on the test set over 10 random splits into training and test data. For the prospective cohort, it shows the performance of the final models, trained on the entire retrospective cohort, as mean ± SD over 10 000 bootstrap samples.

^bPerformance of models and physicians was measured using Matthews correlation coefficient (MCC) and F1 score after binarization with the respective optimal threshold. Displayed is the mean ± SD over 10 000 bootstrap samples.

outperforming both ML models. However, the dataset for comparing predictive performance over a 7-day window in this pilot sub-study was limited due to a low number of fatalities preceded by prospective assessments. In addition, these deceased patients were less representative of the training cohort since they received TAC immunosuppression.

3.5 | The GBM models for 21-day CMV prediction had AUROC 0.83 and event-AUPRC 0.62

For two reasons, the dataset for the development of models predicting early CMV reactivation was smaller than for mortality prediction: first, we focused on the first 100 days after HCT, where the

earliest episode of CMV reactivation almost exclusively occurs in the absence of prophylaxis. Second, we excluded patients without CMV testing during the first 30 days after HCT since the earliest CMV episode could have been missed without regular tests. For CMV prediction over 21 days, the dataset contained 52 008 time points from 1561 patients, of which 12 413 (23.87%) were labeled positive.

Here, the GBM model also had the best performance with an AUROC of 0.825 compared with 0.793 and 0.779 for LR and baseline, respectively (Figure 3A). The same trend was observed in event-AUPRC (Figure 3B), which was 0.620, 0.532, and 0.473 for GBM, LR, and baseline models, respectively, and in sample-AUPRC (Figure S6). For CMV prediction, the gap between models using time-dependent data (GBM and LR) and the static baseline was much smaller than for mortality prediction. The primary reason is that even the CMV models with access to time-dependent data relied on static features for their predictions, while time-dependent laboratory values had only a minor impact (Figure 3E). Calibrated predictions agreed closely with the observed risk; GBM and LR models both had an area of 0.05 between the calibration curve and the line representing perfect calibration (Figure S11).

We performed the same analysis of GBM model performance over time for 21-day CMV prediction as described for 21-day mortality prediction. Again, the fraction of correctly predicted events increased while approaching the event, and this trend was independent of the exact decision threshold chosen (Figure S3b). With a threshold offering an event recall of 0.8, the GBM model predicted 60% of events at least 2 weeks before they occurred. For patients approaching a CMV event, the mean predicted risk rose almost linearly, starting about 40 days beforehand (Figure 3C). While AUROC remained nearly constant over time after HCT, sample-AUPRC dropped after day +40 post-HCT as fewer events occurred (Figure 3D).

3.6 | The CMV predictions were mainly based on prediction day and static features

SHAP value analysis of the GBM model for 21-day CMV prediction revealed that patient CMV serostatus had the highest impact on model predictions, followed by prediction day after HCT and underlying hematologic disorder (Figure 3E). Conditioning regimen, anti-thymocyte globulin as GVHD prophylaxis, donor CMV serostatus, and patient age were also relevant. Interestingly, the time-dependent laboratory values had only a minor role in the predictions of this CMV model, with the exception of the percentage of lymphocytes, which ranked among the top 10 features. Consequently, the CMV model relied predominantly on static data. The joint analysis of feature values and SHAP values confirmed that a positive patient CMV serostatus led to a strongly increased risk prediction, while a negative serostatus reduced the predicted risk (Figure 3F). This dichotomy was even more pronounced among patients who received additional T cell depletion with anti-thymocyte globulin as GVHD prophylaxis. The

SHAP values for the prediction day after HCT peaked between days +20 and +50, indicating a typical timing for early CMV reactivation events (Figure 3G). This peak was most pronounced for patients with recipient-positive CMV serostatus. Interestingly, donor age did not have a differential impact on the risk of CMV reactivation predicted by the GBM model, except for very young donors (<17 years) (Figure S4b). However, these samples were limited in our dataset and were also associated with young patient age.

For prediction of CMV reactivation over 7 days, the GBM and LR models both had a similar AUROC but considerably lower event- and sample-AUPRC than the corresponding models for prediction over 21 days. Again, this may be influenced by the lower positive fraction of 7.50% with the narrower 7-day time window. An analysis of model performance over time and of the impact of individual features on predictions of the 7-day GBM are included in the supplementary material (Figures S12 and S13).

3.7 | CMV models were successfully validated and performed similar to HCT physicians

In the prospective validation cohort ($n = 398$), the performance of all CMV models remained very close to their performance on retrospective data (Table 1A). Compared with the retrospective cohort, the AUROC of the GBM model for 21-day CMV prediction increased slightly from 0.825 to 0.846, while its event-AUPRC decreased slightly from 0.620 to 0.574. This performance remained stable across patient subgroups with distinct immunosuppressive regimens (Table S2). In contrast, the 21-day LR model had a higher performance for patients who received CSA instead of TAC immunosuppression. For prediction over 7 days, both models demonstrated very similar performance on retrospective and prospective data, and a trend toward higher performance for patients with CSA immunosuppression. All CMV models remained well calibrated on prospective data, concluding the successful prospective validation (Figure S10).

In a pilot study, we compared the predictive performance of the ML models to the risk of CMV reactivation estimated by experienced HCT physicians. The results are shown in Table 3B. For 21-day prediction, the GBM model had the best performance, with an MCC of 0.329 ± 0.062 compared with 0.266 ± 0.023 and 0.234 ± 0.051 for LR model and physicians, respectively. On the other hand, the physicians had a small lead over both ML models for prediction over 7 days. In both cases, these differences in average performance were not decisive, given the limited dataset for this comparison.

4 | DISCUSSION

In response to persisting difficulties to predict relevant complications in HCT patients and to support clinical assessment, we developed and validated the first ML models for time-dependent prediction of mortality and CMV reactivation after HCT. These ML models accurately predict patient-specific event risks within a specified time window

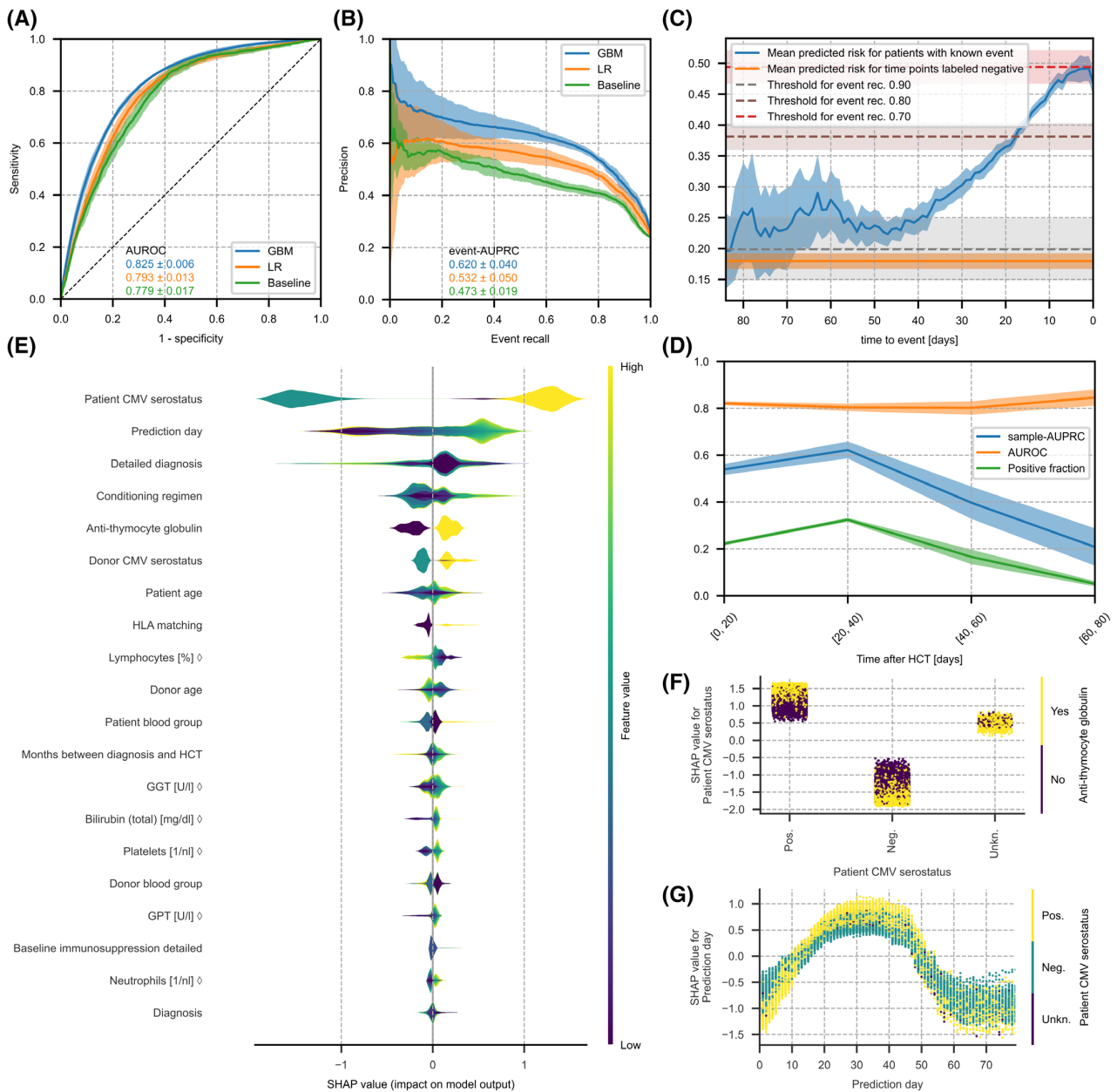


FIGURE 3 Performance and feature importance of the GBM model for 21-day prediction of CMV reactivation. (A) Receiver operating characteristic of GBM and LR model, which received a combination of static and time-dependent input features, and a baseline model which received only static features. (B) Precision-recall curve for the same models shown in (A) based on event recall, i.e. the fraction of events that were correctly predicted on any of the previous 21 days. (C) Mean predicted risk of the GBM model as a function of time to event. For reference, the orange horizontal line indicates the mean predicted risk over all time points labeled negative. Dashed horizontal lines as in Figure 2. (D) AUROC and sample-AUPRC of the GBM model and fraction of samples with positive label as functions of time after HCT. (A–D) Lines and shaded areas show the mean \pm SD on the test set over 10 random splits into training and test data. (E) Layered violin plot of SHAP values of the GBM model for the 20 features with highest mean absolute SHAP value. The thickness of the violins corresponds to the estimated density of each feature's SHAP values, colors show the magnitude of feature values (percentiles). For features marked with \diamond , the feature value is the time-normalized score that the model received as input, not the raw value in its original unit. For categorical features, the colors are based on an integer representation and should not be interpreted as ordered. All SHAP values were computed based on raw model output in log-odds space. (F–G) Scatter plots of SHAP values over feature values. Samples are colored by the value of a second feature to reveal interactions, which show as vertical color patterns. Displayed are plots for the feature patient CMV serostatus colored by anti-thymocyte globulin (F) and prediction day after HCT colored by patient CMV serostatus (G).

and at multiple time points after HCT and pave the way toward clinical decision support systems for transplantation medicine. While existing predictive models^{18–20} and scores^{8,9,32} for HCT-specific risk assessment predominantly focus on pre-HCT assessment to support treatment and donor selection, time-dependent risk assessment may enable physicians to refine and individually adjust treatments and preventive measures after HCT to obtain the best possible outcome for each patient.

Our ML models combine static patient information as used in previous HCT ML models¹⁸ with longitudinal laboratory data and update their predictions whenever new time-dependent data become available. Although this study builds on previous research on ICU data,¹⁷ our ML models prove the applicability of this new approach in the field of HCT and on a much larger time scale with varying data granularity, which underlines the relevance of this study beyond the field of transplantation.

Recent ML models in patients with leukemia combined static patient data at diagnosis with time series of laboratory measurements to predict patient outcome at a single point in time.³³ While these models included HCT as an input parameter, they neither predicted the outcome after HCT nor at multiple time points. Another ML study using longitudinal HCT data integrated patients' vital signs and predicted graft-versus-host disease by day +100 with a modest AUROC of 0.66,³⁴ allowing for a single prediction on day +10 after HCT. Personalized ML survival models for HCT patients refined prognosis at the time of HCT but exclusively relied on static pre-HCT data as input parameters without adapting to complications occurring after HCT.³⁵ Most recently, the integration of multiple time-dependent variables into an ML model improved the prediction of acute GVHD (AUROC 0.78) in HCT recipients.³⁶

Although our final models update their predictions whenever new data become available, they use only the most recent laboratory result for each prediction. On large EHR databases, recurrent deep neural networks, for example, using long short-term memory (LSTM) units, have demonstrated high prediction performances utilizing entire time series as model input.^{16,37,38} A limitation of LSTMs is, however, their dependence on very large training data, which are not available in all medical domains. For instance, LSTMs did not outperform GBM models for the time-dependent prediction of circulatory failure based on a large single-center ICU dataset.¹⁷ Since additional features describing the history of laboratory values did not improve the performance of our GBM models (Figure S15), we did not pursue more complex approaches for time series data.

In this article, we considered multiple endpoints and time windows for prediction. Across these tasks, GBM models consistently outperformed LR and provided well-calibrated time-dependent risk predictions. Prediction performance was best for prediction of 21-day mortality, where we obtained very high AUROC and high event-AUPRC. High predictive performance, in addition to validity and independent replication, is a core requirement for the clinical use of predictive models in decision support systems³⁹ since it is the first indicator of health impact and effectiveness. Yet, identifying the optimal performance threshold for effectiveness and impact is also subject

to medical,⁴⁰ technical, and ethical⁴¹ considerations relating to the predicted outcome, potential consequences of false predictions, and implementation issues. Our pilot comparison to physicians' expectations indicates that the developed models will likely provide relevant practical use, for example, as a risk screening tool for post-HCT outpatients. Given the possibilities of intervening via anti-infective or immunosuppressive drugs and hospitalization, such warning systems might prevent fatal outcomes. The immediate availability of the features used by our models in most HCT centers, including both the static HCT parameters and the continuously measured standardized laboratory variables, is a major advantage for its clinical application for decision support. Finally, successful implementation in clinical practice can also be influenced by physicians' trust in ML models, which may be increased by providing understandable explanations for individual predictions,⁴⁰ for example, via SHAP values.

Since a direct comparison to existing scores designed for pre-HCT risk assessment is not possible, we compared our models to a baseline model, which was trained for the time-dependent prediction task but used only static input features. Interestingly, time-dependent input features proved highly valuable for mortality prediction but only offered modest improvement for CMV prediction, indicating that time-dependent outcome prediction may improve HCT-specific risk assessment beyond current standards, but possibly not for all endpoints in equal measure.

The final ML models were successfully validated on an independent, prospectively recruited cohort, as shown by the overall high predictive performance of the developed models on prospective data. For mortality prediction, model performance decreased slightly compared with the retrospective cohort, which was in part explained by changes in immunosuppression strategies. However, the slight performance drop also in patients with identical baseline immunosuppression indicates a dataset shift over time. This is well in line with a recent EBMT analysis of HCT data up to the year 2016, showing decreasing NRM over time.³ Given the small differences in prediction performance between the retrospective and prospective cohort, the applicability of the mortality prediction models remains unaffected. The importance of prospective validation has been previously shown⁴² and is also reflected in our study design. Indeed, predictive models developed for use in clinical practice require continuous monitoring and, if necessary, refinement. Possibly due to the large impact of static features, the performance of models predicting CMV reactivation was not affected by this dataset shift and remained stable.

Our exploratory head-to-head comparison with experienced HCT physicians revealed that GBM models performed approximately on par for 21-day prediction of mortality and CMV reactivation. Despite the limitations of this pilot comparison, trends showed that the physicians performed slightly better in mortality prediction while the GBM model was better in predicting CMV reactivation. Since the physicians had direct contact with their patients, and therefore access to more information than the 60 input features of the ML models, these results underline the promising potential for future use of such GBM models in clinical practice. Integrating additional features, such as vital signs or current medication, could potentially increase model

performance further. However, the current feature set used by our final models is readily available in most HCT centers, which is a prerequisite for the implementation as a clinical decision support system.

Although this is a topic of active discussion in the scientific community,⁴³ better interpretability or explainability of ML models in healthcare may improve trust into model predictions⁴⁴ and even the quality of decision support systems.⁴⁵ Here, SHAP values provide insight into the impact of specific features on model predictions and offer a comprehensive approach to explore underlying biological mechanisms. In the GBM models for mortality prediction, mainly features related to organ function and inflammation (CRP, urea nitrogen, GOT, protein) affected the predicted risk. In contrast, the GBM models predicting CMV reactivation strongly relied on static patient data (CMV serostatus, diagnosis, conditioning regimen). For both endpoints, the prediction day after HCT had a large impact on the predicted risks indicating a typical time period for potential complications after HCT, which is in line with previous reports.¹ While SHAP values can provide valuable insight into the features contributing to individual model predictions, it is important to note that they do not represent causal relationships.

The time-dependent prediction problems we considered were imbalanced, meaning that our data contained few samples with a positive label. In this situation, AUPRC is a more informative performance measure than AUROC.¹⁷ However, the exact positive fraction in our data varied across prediction tasks, and we observed that event- and sample-AUPRC were strongly correlated with it. This made it difficult to compare models for different endpoints and time windows directly. Sampling methods could be used to adjust the positive fraction for such comparisons, but then performances would no longer be measured on the data distribution of a realistic application scenario, where the positive fraction is determined by the prevalence of events. By design, the positive fraction for 21-day prediction tasks was higher than for 7-day prediction. Quite unexpectedly, this made 21-day prediction the easier task for ML methods, leading to more robust results even though the distance from positive labeled prediction days to the event was longer. In addition, the 21-day prediction models have a greater potential clinical applicability because they may enable an earlier intervention to prevent or treat complications.

This study has limitations and strengths. It included only data from a single center, which may limit the general applicability of the developed models. However, the models were built on a homogeneous and large dataset of several million data points, and the patient characteristics and HCT practice standards reflected those of major international centers. The precise predictions of our models using standard laboratory features available in all HCT centers pave the way toward the implementation of decision support systems in HCT. Ultimately, its routine use as a medical device requires a prospective clinical trial for safety and efficacy, according to, for example, the EU medical device regulation (EU 2017/745). As in many previous studies,^{46,47} we defined CMV reactivation events only based on detectability, combining data of different quantitative and qualitative CMV tests. However, more recent studies have demonstrated that the severity of CMV disease may be revealed by viral load

kinetics.^{48,49} It would be interesting for future work to attempt time-dependent prediction of CMV reactivation with a narrower event definition based on a threshold for the viral load.

The developed ML models predict mortality and CMV reactivation for HCT patients reliably and in a time-dependent manner, and therefore may potentially improve patient outcomes once implemented as decision support systems in post-HCT care.

AUTHOR CONTRIBUTIONS

Lisa Eisenberg, Nico Pfeifer, and Amin T. Turki conceived the study and wrote the manuscript. Amin T. Turki and Dietrich W. Beelen prepared and provided retrospective data. Amin T. Turki coordinated the prospective XplOit study and Hellmut Ottinger and Dietrich W. Beelen recruited the participants. Christian Brossette and Norbert Graf pseudonymized the data and supported data preparation. Jürgen Rissland provided virology expertise. Lisa Eisenberg, Jürgen Rissland, and Amin T. Turki developed the validation concept. Lisa Eisenberg, Amin T. Turki, Jochen Rauch, Ulf Schwarz, and Andrea Grandjean contributed to data preprocessing. Lisa Eisenberg trained the predictive models and analyzed the data. Nico Pfeifer and Amin T. Turki supervised model development. Lisa Eisenberg, Nico Pfeifer, and Amin T. Turki interpreted the data. Stephan Kiefer coordinated the XplOit consortium.

ACKNOWLEDGMENTS

Research of the XplOit consortium and the conduct of the prospective XplOit study were funded by the German Federal Ministry of Education and Research (BMBF) grants No. 031L0027A-F (Nico Pfeifer, Stephan Kiefer, Norbert Graf, Dietrich W. Beelen). In part, the completion of this analysis was supported by the German Research Foundation (DFG)-UDE-UMEA grant No. FU 356/12-1 (Amin T. Turki). Nico Pfeifer is supported by the DFG Cluster of Excellence “Machine Learning—New Perspectives for Science”, EXC 2064/1, project No. 390727645. The authors thank the study nurses of the XplOit study team, in particular Aleksandra Pillibeit, for their support. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare the following competing interests: Amin T. Turki received consulting fees from CSL Behring, MSD, JAZZ Pharmaceuticals, and MaaT Pharma; travel subsidies from Neovii Biotech. Dietrich W. Beelen received travel subsidies from Medac. The other authors declare no competing financial interests.

DATA AVAILABILITY STATEMENT

The data used in this article contains sensitive personal health information. Due to the high dimensionality and the inclusion of longitudinal data, it cannot be fully anonymized and published without the risk of re-identification. Requests for access to the data may be submitted to the University Hospital Essen and are subject to approval by data protection officer and ethics committee. Source code may be obtained from the corresponding authors upon request. To enable independent replication of our methods, we included detailed

descriptions of preprocessing and model development in the Methods section and in the supplementary material.

ORCID

Lisa Eisenberg  <https://orcid.org/0000-0002-1041-7948>

Norbert Graf  <https://orcid.org/0000-0002-2248-323X>

Dietrich W. Beelen  <https://orcid.org/0000-0001-5050-220X>

Nico Pfeifer  <https://orcid.org/0000-0002-4647-8566>

Amin T. Turki  <https://orcid.org/0000-0003-1347-3360>

REFERENCES

- Copelan EA. Hematopoietic stem-cell transplantation. *N Engl J Med*. 2006;354:1813-1826.
- Gooley TA, Chien JW, Pergam SA, et al. Reduced mortality after allogeneic hematopoietic-cell transplantation. *N Engl J Med*. 2010;363:2091-2101.
- Penack O, Peczynski C, Mohty M, et al. How much has allogeneic stem cell transplant-related mortality improved since the 1980s? A retrospective analysis from the EBMT. *Blood Adv*. 2020;4:6283-6290.
- Kröger N, Solano C, Wolschke C, et al. Antilymphocyte globulin for prevention of chronic graft-versus-host disease. *N Engl J Med*. 2016;374:43-53.
- Marty FM, Ljungman P, Chemaly RF, et al. Letermovir prophylaxis for cytomegalovirus in hematopoietic-cell transplantation. *N Engl J Med*. 2017;377:2433-2444.
- Gratwohl A, Mohty M, Apperley J. The EBMT: history, present, and future. In: Carreras E, Dufour C, Mohty M, Kröger N, eds. *The EBMT Handbook*. Springer; 2019:11-17.
- Phelan R, Arora M, Chen, M. *Current Use and Outcome of Hematopoietic Stem Cell Transplantation: CIBMTR US Summary Slides*. Center for International Blood and Marrow Transplant Research, Milwaukee, WI, USA; (2020).
- Sorror ML, Sandmaier BM, Storer BE, et al. Comorbidity and disease status based risk stratification of outcomes among patients with acute myeloid leukemia or myelodysplasia receiving allogeneic hematopoietic cell transplantation. *J Clin Oncol*. 2007;25:4246-4254.
- Gratwohl A, Stern M, Brand R, et al. Risk score for outcome after allogeneic hematopoietic stem cell transplantation: a retrospective analysis. *Cancer*. 2009;115:4715-4726.
- Armand P, Kim HT, Logan BR, et al. Validation and refinement of the disease risk index for allogeneic stem cell transplantation. *Blood*. 2014;123:3664-3671.
- Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118.
- Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, et al. Advanced machine learning in action: Identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *npj Digit Med*. 2018;1:9.
- Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;25:65-69.
- McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89-94.
- Gao Y, Cai GY, Fang W, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat Commun*. 2020;11:5033.
- Rank N, Pfahringer B, Kempfert J, et al. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. *npj Digit Med*. 2020;3:139.
- Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med*. 2020;26:364-373.
- Shouval R, Labopin M, Bondi O, et al. Prediction of allogeneic hematopoietic stem-cell transplantation mortality 100 days after transplantation using a machine learning algorithm: a European Group for Blood and Marrow Transplantation Acute Leukemia Working Party retrospective data mining study. *J Clin Oncol*. 2015;33:3144-3152.
- Arai Y, Kondo T, Fuse K, et al. Using a machine learning algorithm to predict acute graft-versus-host disease following allogeneic transplantation. *Blood Adv*. 2019;3:3626-3634.
- Logan BR, Maiers MJ, Sparapani RA, et al. Optimal donor selection for hematopoietic cell transplantation using Bayesian machine learning. *JCO Clin Cancer Inform*. 2021;5:494-507.
- Luft T, Benner A, Terzer T, et al. EASIX and mortality after allogeneic stem cell transplantation. *Bone Marrow Transplant*. 2020;55:553-561.
- Luft T, Benner A, Jodele S, et al. EASIX in patients with acute graft-versus-host disease: a retrospective cohort analysis. *Lancet Haematol*. 2017;4:e414-e423.
- Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.
- Semler SC, Wissing F, Heyder R. German medical informatics initiative. *Methods Inf Med*. 2018;57:e50-e56.
- Kolitsi Z, Dipak K, Petra W, et al. *DigitalHealthEurope Recommendations on the European Health Data Space: Supporting Responsible Health Data Sharing and Use through Governance, Policy and Practice*. Digital-HealthEurope, 2021.
- Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30:3146-3154.
- Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
- Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with numpy. *Nature*. 2020;585:357-362.
- McKinney, W. Data structures for statistical computing in python. In *Proc. 9th Python Sci. Conf.*, 56-61 (2010).
- Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2:56-67.
- Weiler G, Schwarz U, Rauch J, et al. XplOit: an ontology-based data integration platform supporting the development of predictive models for personalized medicine. *Stud Health Technol Inform*. 2018;247:21-25.
- Beauvais D, Drumez E, Blaise D, et al. Scoring system for clinically significant CMV infection in seropositive recipients following allogeneic hematopoietic cell transplant: an SFGM-TC study. *Bone Marrow Transplant*. 2021;56:1305-1315.
- Lu C-C, Li J-L, Wang Y-F, Ko B-S, Tang J-L, Lee C-C. A BLSTM with attention network for predicting acute myeloid leukemia patient's prognosis using comprehensive clinical parameters. In *Proc. 2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2455-2458 (2019).
- Tang S, Chappell GT, Mazzoli A, Tewari M, Choi SW, Wiens J. Predicting acute graft-versus-host disease using machine learning and longitudinal vital sign data from electronic health records. *JCO Clin Cancer Inform*. 2020;4:128-135.
- Okamura H, Nakamae M, Koh S, et al. Interactive web application for plotting personalized prognosis prediction curves in allogeneic hematopoietic cell transplantation using machine learning. *Transplantation*. 2021;105:1090-1096.
- Liu X, Cao Y, Guo Y, et al. Dynamic forecasting of severe acute graft-versus-host disease after transplantation. *Nat Comput Sci*. 2022;2:153-159.
- Rajkumar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj Digit Med*. 2018;1:18.
- Ayala Solares JR, Diletta Raimondi FE, Zhu Y, et al. Deep learning for electronic health records: a comparative review of multiple deep neural architectures. *J Biomed Inform*. 2020;101:103337.

39. Tcheng JE, ed. *Optimizing Strategies for Clinical Decision Support: Summary of a Meeting Series*. The learning health system series. National Academy of Medicine; 2017.
40. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digit Med*. 2020;3:17.
41. McLennan S, Fiske A, Celi LA, et al. An embedded ethics approach for AI development. *Nat Mach Intell*. 2020;2:488-490.
42. Buturovic L, Shelton J, Spellman SR, et al. Evaluation of a machine learning-based prognostic model for unrelated hematopoietic cell transplantation donor selection. *Biol Blood Marrow Transplant*. 2018;24:1299-1306.
43. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. *Science*. 2021;373:284-286.
44. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655.
45. Bruckert S, Finzel B, Schmid U. The next generation of medical decision support: a roadmap toward transparent expert companions. *Front Artif Intell*. 2020;3:507973.
46. Teira P, Battiwalla M, Ramanathan M, et al. Early cytomegalovirus reactivation remains associated with increased transplant-related mortality in the current era: a CIBMTR analysis. *Blood*. 2016;127:2427-2438.
47. Elmaagacli AH, Steckel NK, Koldehoff M, et al. Early human cytomegalovirus replication after transplantation is associated with a decreased relapse risk: evidence for a putative virus-versus-leukemia effect in acute myeloid leukemia patients. *Blood*. 2011;118:1402-1412.
48. Leser S, Bayraktar E, Trilling M, et al. Cytomegalovirus kinetics after hematopoietic cell transplantation reveal peak titers with differential impact on mortality, relapse and immune reconstitution. *Am J Hematol*. 2021;96:436-445.
49. Duke ER, Williamson BD, Borate B, et al. CMV viral load kinetics as surrogate endpoints after allogeneic transplantation. *J Clin Investig*. 2021;131:e133960.
50. Lablans M, Borg A, Ückert F. A RESTful interface to pseudonymization services in modern web applications. *BMC Med Inform Decis Mak*. 2015;15:2.
51. Seuss H, Dankerl P, Ihle M, et al. Semi-automated de-identification of German content sensitive reports for big data analytics. *Rofo*. 2017;189:661-671.
52. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53:457-481.
53. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94:496-509.
54. Therneau, T. survival: A Package for Survival Analysis in R (2021). R package version 3.2-11.
55. Kassambara, A., Kosinski, M. & Biecek, P. survminer: Drawing Survival Curves using 'ggplot2' (2021). R package version 0.4.9.
56. Gray, B. cmprsk: Subdistribution Analysis of Competing Risks (2020). R package version 2.2-10.
57. Malone B, Garcia-Duran A, Niepert M. Learning representations of missing data for predicting patient outcomes (2018). Preprint at <http://arxiv.org/pdf/1811.04752v1>.
58. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks (2015). Preprint at <http://arxiv.org/pdf/1511.03677v7>.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Eisenberg L, the XploIt consortium, Brossette C, et al. Time-dependent prediction of mortality and cytomegalovirus reactivation after allogeneic hematopoietic cell transplantation using machine learning. *Am J Hematol*. 2022;1-15. doi:10.1002/ajh.26671

APPENDIX A: XploIt Consortium

The members of the XploIt consortium were Lisa Eisenberg, Prof. Nico Pfeifer, Jochen Rauch, Kerstin Rohm, Dr Gabriele Weiler, Dr Stephan Kiefer, Dr Jürgen Rissland, Prof. Sigrun Smola, Dr Thorsten Pfuhl, Dr Pascal Feld, Dr Lise Lauterbach-Rivière, Dr Anna Marthaler, Dr Jörg Bittenbring, Dr Dominic Kaddu-Mulindwa, Katharina Götz, Katharina Och, Prof. Thorsten Lehr, Christian Brossette, Stefan Theobald, Yvonne Braun, Prof. Norbert Graf, Abdul Kadir, Dr Ulf Schwarz, Andrea Grandjean, Dr Matthias Ihle, Claudia Riede, Sonja Fix, Dr Amin T. Turki, MD PhD, Prof. Dietrich W. Beelen, MD, PD Dr Hellmut Ottinger, MD and The XploIt Study Team: Dr Nikolaos Tsachakis-Mück, MD, Dr Rashit Bogdanov, Prof. Michael Koldehoff, MD, Dr Nina Steckel, MD, Dr Ji-He Yi, MD, Aiste Fokaite, MD, Dr Vesna Klisanin, MD, PD Dr Lambros Kordelas, MD, Dr Diana Garay, Ximena Gavilanes, MD, Robert F. Lams, MD, Aleksandra Pillibeit, Saskia Leser and Theresa Graf. Stefan Hilbig and Joachim Weiß kindly provided EHR-related IT support.

A.6 Manuscript 6

Title:

XplOit: An ontology-based data integration platform supporting the development of predictive models for personalized medicine

Authors:

Gabriele Weiler, Ulf Schwarz, Jochen Rauch, Kerstin Rohm, Thorsten Lehr, Stefan Theobald, Stephan Kiefer, Katharina Götz, Katharina Och, Nico Pfeifer, Lisa Handl, Sigrun Smola, Matthias Ihle, Amin T. Turki, Dietrich W. Beelen, Jürgen Rissland, Jörg Bittenbring & Norbert Graf

Published in:

Studies in Health Technology and Informatics, Volume 247, Pages 21–25

<https://www.doi.org/10.3233/978-1-61499-852-5-21>

Time of Publication:

March 2018

License information:

This article was published under a Creative Commons CC-BY-NC license (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial reuse and reproduction in any medium, provided the original work is properly cited.

XplOit: An Ontology-Based Data Integration Platform Supporting the Development of Predictive Models for Personalized Medicine

Gabriele WEILER ^{a,1}, Ulf SCHWARZ ^b, Jochen RAUCH ^a, Kerstin ROHM ^a, Thorsten LEHR ^c, Stefan THEOBALD ^d, Stephan KIEFER ^a, Katharina GÖTZ ^c, Katharina OCH ^c, Nico PFEIFER ⁱ, Lisa HANDL ⁱ, Sigrun SMOLA ^e, Matthias IHLE ^h, Amin T. TURKI ^g, Dietrich W. BEELEN ^g, Jürgen RISSLAND ^c, Jörg BITTENBRING ^f and Norbert GRAF ^d

^a *Fraunhofer Institute for Biomedical Engineering, St. Ingbert, Germany*

^b *Institute for formal ontologies and medical information science*, ^c *Clinical Pharmacy*,

^d *Department of Pediatric Oncology and Hematology*, ^e *Institute of Virology*, ^f *Department of Internal Medicine 1, Saarland University, Germany*

^g *Department of Bone Marrow Transplantation, West-German Cancer Center, University Hospital Essen, Germany*

^h *Averbis GmbH, Freiburg, Germany*

ⁱ *University of Tübingen, Germany*

Abstract. Predictive models can support physicians to tailor interventions and treatments to their individual patients based on their predicted response and risk of disease and help in this way to put personalized medicine into practice. In allogeneic stem cell transplantation risk assessment is to be enhanced in order to respond to emerging viral infections and transplantation reactions. However, to develop predictive models it is necessary to harmonize and integrate high amounts of heterogeneous medical data that is stored in different health information systems. Driven by the demand for predictive instruments in allogeneic stem cell transplantation we present in this paper an ontology-based platform that supports data owners and model developers to share and harmonize their data for model development respecting data privacy.

Keywords. Predictive models, semantic data annotation, semantic integration

1. Introduction

Patterns in individual health data and personalized multiscale models of diseases can predict future events and outcome. Such predictive models are able to support decisions by physicians in all aspects of personalized diagnosis and treatment. Especially in the area of stem cell transplantation (SCT) predictive models are needed, since complica-

¹ Corresponding author, Fraunhofer IBMT, Ensheimer Str. 48, St. Ingbert, Germany; E-mail: Gabriele.weiler@ibmt.fraunhofer.de.

tions, e.g. viral infections, graft-versus-host disease (GvHD) or relapse, can be life-threatening. It is currently not possible to predict the course of SCT and therefore in a number of patients lifesaving interventions cannot be applied on time. Despite the popularity of predictive research, the development of required models lacks behind expectations [1].

Model developers need a reliable methodology to easily collect and correlate data from different hospitals and diverse sources (health information or laboratory systems, medical reports, etc.) in order to reach a sufficient study cohort. It is a tedious task to do so manually and it is estimated that currently 50%-80% of a data scientist's time is spent on data integration [2]. An expressive description of data using emerging standards is necessary to maximize the quality and applicability of the developed models. Hence, in the XplOit project we are developing a platform that enhances and accelerates development of predictive models with an innovative approach for semantic data integration. The XplOit Platform enables data owners to easily share and harmonize their data respecting data protection. Model developers can inspect and analyze cross-institutional harmonized data. In the following, we describe our approach to semantic data integration and the architecture of the XplOit Platform.

2. Methods – Ontology-Based Data Integration

Establishment of reliable predictive models requires a profound understanding of the meaning and the correlation of cross-institutional data. Therefore, expressive data annotations and deep semantic data integration is required. Hence, we have chosen an ontology-based data integration approach [2] with the following features: (1) As global scheme, we use an ontology that is an expressive standardized description of the domain formalized in the Web Ontology Language OWL [3]. (2) Unlike most current medical data integration approaches, which use merely concepts from the ontology as annotations, we allow complex descriptions to realize deep expressiveness. (3) We aim to enable data owners themselves to perform the data integration tasks, i.e. extend ontology and create data annotations, since they know the meaning of the data and can decide which data is needed.

Our work exceeds approaches in most other medical data integration platforms as e.g. tranSMART [4] or i2B2 [5] in the expressivity of metadata. There are only few approaches, e.g. the p-medicine platform [6], with comparable expressivity in their metadata. However, in these approaches annotations have to be created mostly manually, browsing complex ontologies, a tedious and time-consuming task. We, in contrast, provide easy-to-use semi-automatic tools as described in the following.

2.1. VDOT-Ontology and Ontology Aggregator Tool

The global scheme of our data integration approach is the Viral Disease Ontology Trunk (VDOT)², a modular, domain ontology. It provides formal, human-and computer-understandable axiomatic semantic descriptions of concepts and expressions for the description of biomedical data and predictive models. VDOT is standardized by reusing parts of established ontologies relating them in an axiomatic new framework. It

² VDOT is stored in the library of biomedical ontologies: www.ifomis.org/vdot.

does not contain all concepts needed for data annotation purposes, but rather provides a framework, which can easily be extended by users to cover their individual annotation needs. VDOT extensions can be semi-automatically generated by end users with the Ontology Aggregator Tool. This tool searches a semantic repository, with standardized ontologies and can automatically relate needed concepts with others in VDOT.

2.2. Semantic Data Annotation

We realize data annotations as paths relating ontology concepts by axiomatically defined relations, allowing to describe the meaning of a data element with different information. The annotation service supports data owners to easily annotate a data object type (DOT) that describes similar data files with paths from the ontology: For each data element an ontology path is semi-automatically created in three steps as follows:

1. *Matching concepts.* A string matching algorithm searches matching concepts for the data element. The label of the data element is compared to labels and synonyms in the ontology. If no concept can be found, the data owner can specify aliases. If still nothing can be found the Ontology Aggregator Tool can be used to semi-automatically extend the ontology.
2. *Ontology paths.* For the matched concepts potential ontology paths are created. The starting point of the path is the patient, the ending point the matched concept. Automatically all possible paths can be created by iteratively searching VDOT utilizing axiomatic constraints of its concepts (ontological relations).
3. *Selection of path.* If more than one path is found, the paths are ranked according to their likelihood that grows when: 1) Same path is already used in other DOTs. 2) Path is similar to paths related to other data elements in the DOT. 3) Path contains concept with a high string matching similarity. The ranked paths are shown to the data owner with additional descriptions from the ontology, enabling him to select the right path.

From the annotations a formal annotation template is generated, storing the information to translate uploaded data matching the DOT into an RDF triple graph and integrate it with other data. The information stored in the graph is sufficient to provide model developers with extensive search and analysis functionality as described in the next chapter.

3. Results – The XplOit Platform

The XplOit Platform is a web-based platform for model developers and clinicians working in a modelling project together. For modelling projects, a community can be founded, which allows data sharing with their members respecting data protection. A community manager, who is in general a data owner, is responsible that only authorized persons can become a member of the community in order to guarantee data privacy. The data owners from different hospitals upload their pseudonymized data. After annotating as described above the data is harmonized by representing them as triples in a data store, and model developers can search and analyze it. In the following we describe the main components and services of the platform as depicted in Figure 1.

The **De-Identification and Pseudonymisation Services** are locally installed in the hospitals to semi-automatically pseudonymize and deidentify structured (Mainzelliste [7]) and unstructured data (deID-Tool [8]), as e.g. medical reports. The **Information Extraction Framework (IEF)** provides user friendly tools for data owners to share and harmonize their data. It processes structured and unstructured heterogeneous data to store the raw data as well as the processed triples data in the data warehouse. The IEF enables to integrate different pipelines for a flexible ETL (Extract, Transform and Load) processing of the provided data according to configured DOTs. These pipelines can be designed as Pentaho Data Integration [9] transformations and uploaded via the XplOit portal. The **Data Warehouse** is the central database of the XplOit Platform. It consists of an OpenLink Virtuoso quad store [10], for storing the RDF graph, and a MongoDB database [11], for storing application specific data as e.g. user information. The **Semantic Integration Framework** implements the described data integration approach. The **Modeling Workbench** enables model developers to search and analyze the integrated data to check data quality, possible correlations and generate first hypotheses. It is possible to search for parameters using standardized search terms from the ontology, while restricting the data ranges of values. For inspecting the data, it can be chosen between a tabular and various graphical views as e.g. histogram, box plots, scatterplots and parallel coordinates. The **Security Services** ensure data protection and data integrity. They manage secure data access and provide an audit trail.

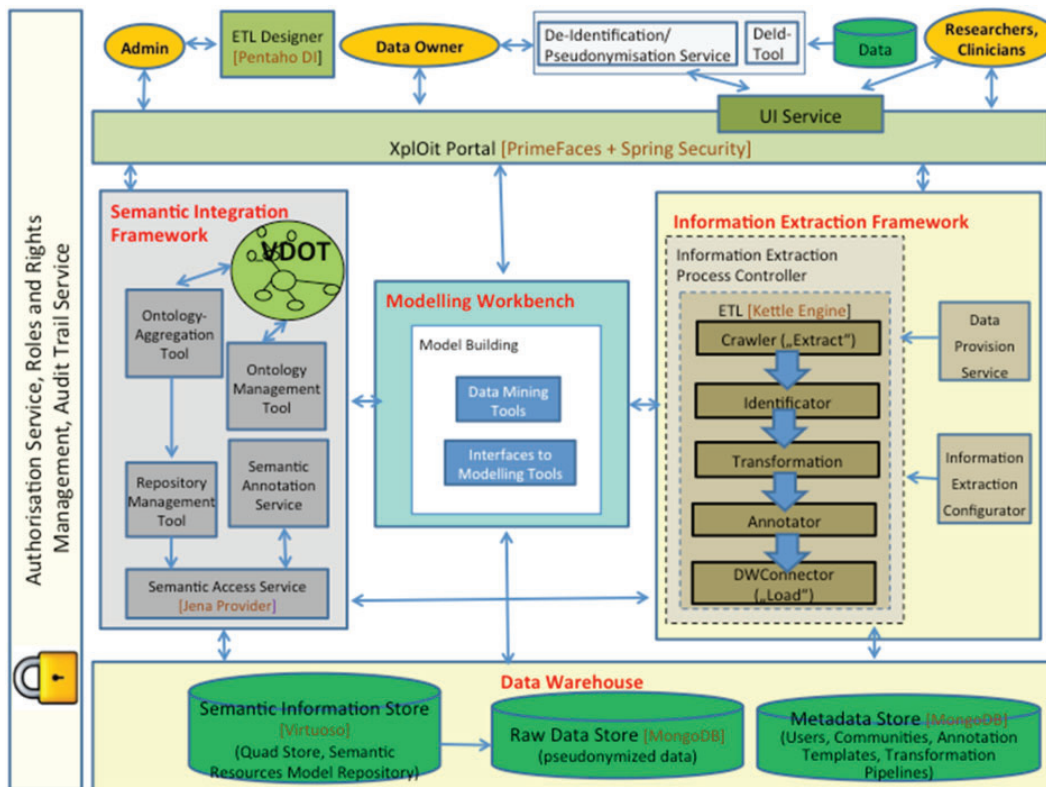


Figure 1. Architecture of the XplOit Platform.

4. Conclusion

We have presented a semantic data integration platform for enhancing and accelerating the development of predictive models. It allows clinicians and model developers to work together efficiently. Data owners can easily harmonize and share heterogeneous health data while respecting data privacy. This is achieved through an innovative semantic data integration approach that enables model developers to quickly gain a deep understanding of meaning and correlation of data providing them with data-inspecting, -analyzing and -export tools enhancing and accelerating their work.

First tests with clinicians and model developers are promising. They have shown that both user groups are able to efficiently work with the platform and confirmed that important preliminary work for model development can be conducted. Currently our platform is applied for developing predictive models for stem cell transplantation utilizing data from two university hospitals.

In the future, we will also support transfer of predictive models from bench to bedside. Model developers will be able to upload their models into a model repository. The models can be validated with prospective clinical trials using the ontology-based trial management system ObTiMA [12]. Following validation, clinicians can apply the models in patient treatment. Furthermore, data pipelines for miRNA and imaging data will be integrated. The presented ontology-based data integration approach can be also applied to other kinds of biomedical data integration scenarios.

Acknowledgement

The XplOit project is funded by the German Federal Ministry of Education and Research (BMBF, Grant id: 031L0027A).

References

- [1] A. K. Walijee, P. D. R. Higgins and A. G. Singal: A Primer on Predictive Models. *Clinical and Translational Gastroenterology*, 2014.
- [2] G. De Giacomo, D. Lembo, M. Lenzerini, A Poggi and R. Rosati: Using Ontologies for Semantic Data Integration. In: Flesca S., Greco S., Masciari E., Saccà D. (eds) *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years. Studies in Big Data*, vol 31. Springer, 2017.
- [3] OWL 2 Web Ontology Language, <https://www.w3.org/TR/owl2-overview/>, last accessed: 07.02.2018
- [4] E. Scheufele, D. Aronzon, R. Coopersmith, et al: tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform. *AMIA Summits on Translational Science Proceedings*. 2014.
- [5] I2B2, Informatics for Integrating Biology and the Bedside, <https://www.i2b2.org/index.html>, last accessed: 07.02.2018
- [6] J. Marés, L. Shamardin, G.Weiler, et al: p-medicine: A Medical Informatics Platform for Integrated Large Scale Heterogeneous Patient Data. *AMIA Annual Symposium Proceedings*. 2014.
- [7] M. Lablans, A. Borg, F. Ückert F: A RESTful interface to pseudonymization services in modern web applications. *BMC Medical Informatics Decision Making*. 2015.
- [8] H. Seuss, P. Dankerl, M. Ihle et al: Semi-automated De-identification of German Content Sensitive Reports for Big Data Analytics. *Fortschritte Röntgenstrahlen*. 2017.
- [9] OpenLink Virtuoso, <https://virtuoso.openlinksw.com>, last accessed: 07.02.2018
- [10] Mongo DB, Database as a service, <https://www.mongodb.com>, last accessed: 07.02.2018
- [11] Pentaho Data Integration, <http://www.pentaho.com/product/data-integration>, last accessed: 07.02.2018
- [12] ObTiMA, <https://obtima.org>, last accessed: 07.02.2018

