Validating the Knowledge Bank Approach for Personalized Prediction of Survival in Acute Myeloid Leukemia: a Reproducibility Study

Yujun Xu (yujun.xu@hotmail.com)
University of Munich: Ludwig-Maximilians-Universitat Munchen https://orcid.org/0000-0002-8487-7584
Ulrich Mansmann
University of Munich: Ludwig-Maximilians-Universitat Munchen

Research Article

Keywords: Prognostic prediction, reproducibility, acute myeloid leukemia, genomics, precision oncology, multistage models

Posted Date: September 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-881649/v1

Version of Record: A version of this preprint was published at Human Genetics on April 16th, 2022. See the published version at https://doi.org/10.1007/s00439-022-02455-8.

Abstract

Reproducibility is not only essential for the integrity of scientific research, but is also a prerequisite of model validation and refinement for future application of (predictive) algorithms. However, reproducible research is becoming increasingly challenging, particularly in high-dimensional genomic data analyses with complex statistical or algorithmic techniques. Given that there are no mandatory requirements in most biomedical and statistical journals to provide the original data, analytical source code, or other relevant materials for publication, accessibility to these supplements naturally suggests a greater credibility of published work. In this study, we performed a reproducibility assessment of the notable paper by Gerstung et al. published in *Nature Genetics* (2017) by rerunning the analysis using their original code and data, which are publicly accessible. Despite a perfect open science setting, it was challenging to reproduce the entire research project; reasons included coding errors, suboptimal code legibility, incomplete documentation, intensive computations, and an R computing environment that could no longer be re-established. We learn that availability of code and data does not guarantee transparency and reproducibility of a study; in contrast, the source code is still liable to error and obsolescence, essentially due to methodological complexity, lack of editorial reproducibility checking at submission, and updates of software and operating environment. Building on the experience gained, we propose practical criteria for the conduct and reporting of reproducibility studies for future researchers.

Introduction

Acute myeloid leukemia (AML) is a myelopoiesis neoplasia characterized by pathological proliferation and accumulation of clonal myeloid cells. Primarily due to biological heterogeneity, patients with AML are expected to have varying post-therapeutic prognoses. Modern molecular techniques today have made the cytogenetic and genetic information of AML available, and it has become standard in clinical settings to initiate therapy by incorporating these individual profiles into a risk stratification. Compared to the current WHO or ELN classifications (Arber et al. 2016; Döhner et al. 2017), which define risk groups according to the presence of a long list of genetic aberrations, statistical algorithms are thought to be more capable of processing the high-dimensional AML data comprehensively, leading to a more accurate prediction of prognosis among risk subgroups given a specific treatment strategy.

One promising example is the personalized, therapeutic decision support tool for AML patients proposed by Gerstung et al. (2017), subsequently referred to as *the knowledge bank approach*, or the KBA. The KBA considered the choice between allogeneic hematopoietic cell transplant (allograft) in first complete remission (CR1), on the one hand, and standard chemotherapy in CR1, followed by salvage treatments (either allograft or more intensive chemotherapy) after relapse, on the other. They used a training database of 1540 patients from three clinical trials (AMLHD98A, AMLHD98B, and AMLSG0704) of the German–Austrian AML Study Group (Schlenk et al. 2004, 2010, 2016), to construct the KBA prediction algorithm. It integrated 231 clinical, cytogenetic, and genetic factors to predict three-year overall survival as the primary endpoint, and demonstrated an improved predictive accuracy as opposed to the ELN classification (Harrell's C-index: 0.72 vs 0.64). The KBA also allowed individual prediction of different survival rates (e.g., alive without CR1, alive in CR1, and alive after relapse) in possible therapeutic scenarios: receiving allograft in CR1, salvage allograft after relapse, or standard chemotherapy only.

Although of high clinical relevance, their results do not seem to have been translated into clinical practice. The improved predictive value of the KBA was validated later by Huet et al. (2018) using a retrospective cohort of 155 AML patients, based solely on outputs obtained from the interactive portal provided by Gerstung et al. (https://cancer.sanger.ac.uk/aml-multistage). Recently, Fenwarth et al. (2021) took a further step by looking at *NPM1* minimal residual disease and using the KBA to reclassify the current ELN rule. They emphasized the potential applicability of the KBA among a younger cohort (C-index for five-year overall survival using the modified KBA compared to the ELN: 68.9 vs 63.0). Yet neither study investigated the black box of the KBA.

Notwithstanding the fact that Gerstung et al.'s web portal allows readers to generate outcome predictions effortlessly, it hides the complexity of their algorithm and hence leaves limited information about the implicit statistical assumptions and the derivation details behind the multistage KBA. To find these, one must reference the 135-page Supplementary Note (https://static-content.springer.com/esm/art%3A10.1038%2Fng.3756/MediaObjects/41588_2017_BFng3756_MOESM10_ESM.pdf). Their main article published in *Nature Genetics* remains, however, the tip of a research iceberg. Therefore, reproducing their research is a precondition for a better understanding of their methods and for performing credible external validation (preferably with newer data that reflect substantial therapeutic advances in recent decades). Only then can the refinement and application of the KBA (e.g., combining time-dependent, treatment-related information throughout different stages of the prognosis of AML) be possible.

Fortunately, the data and analysis codes of Gerstung et al.'s paper were publicly accessible and allowed deeper inspection. It enabled us to perform a reproducibility evaluation through accessibility evaluation, code testing, error modification (if needed), and algorithm assessment. This work can be seen as the first step towards a larger project, with the goal of validating and modifying the KBA with a more current, external AML database in Germany. In the following sections, we describe the results and findings of our reproducibility study.

Materials And Methods

Reproducibility criteria

To date, there are no established guidelines available concerning how to perform and report a reproducibility study, and many scientists report their attempts to reproduce published results as case studies in a rather ad hoc fashion (Gentleman 2005; Hothorn and Leisch 2011; Kitzes et al. 2017). In the light of Seibold et al.'s work (2021) as well as Hofner and Scheipl's (2016) editorial guidance on reproducible research in the Biometrical Journal – one of the few (if not the only) statistical and biomedical journals requiring mandatory reproducibility checks – we propose the following criteria (summarized as a checklist in Table 1) and processes to understand the extent to which Gerstung et al.'s findings could be reproduced.

Evaluation processes

Two researchers (Yujun Xu is an epidemiologist and Ulrich Mansmann is a mathematician) undertook the rerun collaboratively. Firstly, the accessibility of materials essential for re-computation was evaluated, including data, analysis code, relevant documents, and computing environment (i.e., the specific version of R with accompanying R packages and the corresponding operating system (OS) used in the original paper). Secondly, during the run of the code, checks were performed to ascertain whether there were any warnings or errors occurring and, if so, whether they could be eliminated. Thirdly, code legibility was assessed to determine if the source code was self-explanatory and whether the comments in the code, as well as the documentation of custom R packages, were understandable. Fourthly, to check the reproducibility, we strictly followed the data processing strategies in Gerstung et al.'s paper, with respect to the preparation of genetic covariates, and statistical analyses, such as rules for covariate selection. Only necessary modifications were made so that computation could proceed smoothly.

The rerun of the code allowed for insights into the conceptual ideas behind the KBA so that the appropriateness of implemented analytical methods could be clarified, and thus to allow possible future refinement of the KBA.

This study was conducted primarily with the most recent version of R (v.4.1.1, R Core Team, 2021) on a standard personal computer (Mac). This decision was made due to software updates and compatibility conflicts that occurred at the time of our study (2021) when using R (v.3.1.2) and the accompanying R packages (the version Gerstung et al.'s original analysis

would have been running in 2016). In general, we believe the source code of scientific work should be robust enough against the upgrade of a computing environment to at least generate results leading to comparable conclusions, even if identical numbers cannot be obtained. In addition, we repeated the rerunning in an RStudio Server Pro (Linux with R v.4.0.4) to examine the robustness of our study. Finally, as Gerstung et al. also provided a Dockerfile in addition to the source code, we tried to use the Docker container to build an identical computing environment to increase the chance of reproducibility.

Data and materials

Gerstung et al.'s original paper states: "To maximize reproducibility, details of statistical methods and all of the analysis code used are provided in the Supplementary Note and as a git repository online". Specifically, the original data in an anonymized form, accompanied with source R code and other supplementary materials are publicly available in the online GitHub repository (https://github.com/gerstung-lab/aml-multistage). This repository is licensed under the GNU General Public License v3.0, which grants end users the freedom to distribute and modify the published content for commercial, patent, or private use (https://github.com/gerstung-lab/AML-multistage/blob/master/LICENSE).

Results

Reproducibility assessment

Overall, rerunning Gerstung et al.'s analysis was challenging and required considerable time commitment, sufficient expertise in biomedicine, statistics, and R programming, and even some background knowledge in systems engineering. In the end our efforts to reproduce their research project were only partly successful.

1. Accessibility As we noted, the GitHub repository provided a folder containing anonymized data, a Supplementary Note with detailed descriptions of statistical methods and codes used, together with other relevant materials (Dockerfile, README, etc.) sorted in a readily comprehensible way.

However, two different versions of the Supplementary Note were found: one that went with the main paper on *Nature Genetics'* website (version: Wed Sep 7 14:26:11 2016, see https://www.nature.com/articles/ng.3756) and another available in the repository, which was incomplete (version: Tue Dec 15 17:54:15 2015). The more recent version was referred to in the present study. After running through the provided R script, we noticed that the TCGA data from the cancer genome atlas, used in the original paper for external validation, were not provided. Thus, the corresponding results could not be reproduced. A data dictionary was also attached, yet it was found to be incomplete (see Online Resource 1), and one needed to refer to additional papers (Schlenk et al. 2004, 2010, 2016) before making sense of the Supplementary Note. For instance, the Note mentioned the abbreviations *CIR, MUD,* and *RD* without explaining their meanings, which are in fact short for *Cumulative Incidence of Relapse*, *Matched Unrelated Donors*, and *Refractory Acute Myeloid Leukemia*, respectively.

The published results were originally obtained using R (v.3.1.2); however, many packages used in the analysis were no longer supported by R (v.3.1.2). Of note, the latest available R (v.4.1.1) did not support a few packages either, such as *graph* or *hilbertVis*. As a result, other than the common way of executing the command *install.packages()* for the installation, additional searching was needed to avoid errors. Some intensive computations requiring parallel tasks, such as leave-one-out cross-validation, were performed by Gerstung et al. in a Load Sharing Facility (LSF) environment. Without access to such a platform, one must modify the code considerably to proceed, while such modifications are prone to unexpected errors.

The Dockerfile was once deemed an opportunity by which we could re-establish an identical computing environment with R (v.3.1.2) and the corresponding packages. However, our attempts were unsuccessful, since the given Dockerfile

assumed the existence of a Docker-based R (v.3.1.2), and only described how their author-customized R packages (i.e., *mg14* and *CoxHD*) could be built on top of that. Today, this is insufficient to build a reproducible environment due to compatibility issues. Specifically, the OS (i.e., *DebianWheezy*) specified in the official Dockerfile (from the Rocker project, see https://www.rocker-project.org), upon which the R (v.3.1.2) is to be built, has been too old to support essential R packages for this analysis (Fig. 1 depicts the layered structure of an R computing environment).

2. Clarity Generally, the source code followed a consistent style and used relative paths to allow data being read into R on different devices without manual alterations. Comments in the code, although with a few insignificant inaccuracies and unnecessary alternative commands, were very helpful for readers to understand the code. Documentation of user-defined packages and functions could also be easily acquired.

Still, as the research project per se is very complex (needing more than 5000 code lines in total), the source code appeared not clear enough to us. Inconsistencies were identified throughout the source code, which inevitably impacted the code legibility, such as different names created for the same concept (e.g., *Time_Diag_TPL* and *TPL_date*, *Cir* and *Rel*, *kmPrs* and *kmPrd*); different concepts with the same name (e.g., *CIR* could mean either *cumulative incidence of relapse* or *Kaplan-Meier survival estimate for relapse*); incorrectly-called variables in the code lines (e.g., variable *TPL_efs* was not found in R data frames, but appeared in the code); wrongly-created variables which did not serve their purposes as suggested by their assignment commands; and unclear data frames created without further explanation (see details in Online Resource 2). In addition to Gerstung's Supplementary Note, a folder named *Code* was provided, containing 14 R scripts and an R data file, yet an accompanying README file was missing.

3. Code execution A mouse-clicking rerun through the code to reproduce the results was impossible. The first coding error appeared due to a command line (i.e., *dataList$Genetics = dataList$Genetics + 0*) in the 8th code chunk of the Supplementary Note (p. 13), which introduced undue factor variables to the list *dataList$Genetics* and stopped one from moving forward. More errors occurred throughout the Note and aborted their executions within each section. As a result, anyone inexperienced in R would find it challenging to debug and proceed. Furthermore, sections 3.6.5.1–3.6.5.9, 3.6.6.4, 3.6.6.6–3.6.6.7, 3.6.7.2–3.6.7.3, 4.4.1.3, 5.4.2.0.4, 5.4.2.0.5, 5.4.2.1.1, 5.4.2.2.1, and 5.4.2.3.1 contained parallel processing which were originally executed on an LSF platform that was different from our environments. Overall, we were able to tailor only parts of the R script, as the computations are very intensive with many custom functions unclear to us. Even so, our alterations were liable to unknown mistakes and might exacerbate the irreproducibility. We reported the errors and our modifications in Online Resource 2.

4. Implementation of the methods described The multistage KBA algorithm based on the provided code reflected the ideas described in the paper. Still, it is noteworthy that random-effect Cox models – the building blocks for the multistage KBA – assumed the parameters within each predictor group followed a normal distribution, which computationally led to a ridge regularization (Therneau et al., 2003). This simplified computations and was realized by specifying the ridge regression function argument in *coxph()* integrated in a user-defined function *CoxRFX()* (to learn this, we used the function *debug()* to step through the execution of *CoxRFX()*). It should be noted, on the one hand, normal distributed random effects are a strong assumption, on the other, Gerstung et al. claimed, in the Supplementary Note and in their function help file, that the parameter estimation was done via an Expectation–Maximization algorithm as suggested by a simulation study of Perperoglou (2014), which, in fact, favored the Restricted Maximum Likelihood-type method for the handling of random effects. Therefore, we could not fully understand how the *CoxRFX()* fits the random-effect Cox model.

5. Matching of outputs We could partly execute the R script after necessary modifications. Although with random seed fixed, deviations from the published results appeared at different locations, yet most of them are negligible, with numbers differing in the last few decimal places, or figures with minor alterations. Notably, our two rerunning attempts had the

same warnings and errors occurring at the same locations, and observed fewer discrepancies in numbers between the two set of results obtained.

We compared the outputs from the rerunning on the personal computer and the published results in Online Resource 2.

6. Further notes regarding the published results

After the rerun of the analysis, we established the full picture of prognostic trajectories among the 1540 patients in the KBA database (Fig. 2), of those, only 995 patients were considered eligible for allograft in CR1.

Gerstung et al. presented a global treatment efficacy of different therapeutic strategies by using the three-year mortality reduction after diagnosis, which represented the comparison between receiving allograft in CR1 and salvage allograft after relapse (y-axis), against the three-year mortality with standard chemotherapy only (x-axis), shown in our Fig. 3a (adapted from Fig. 5a in their original paper). We learned from this reproducibility study that asserted causal efficacy was calculated by applying the KBA to the 995 eligible patients. This had naturally led to the question of whether confounding effects were handled appropriately for two reasons. Firstly, the KBA was constructed using 1540 patients, among whom more than 1/3 of the patients were in fact ineligible for allograft. For illustrative purposes, we rebuilt the KBA after restricting it to the 995 patients and summarized the changes in our Fig. 3 and Table 2. Notably, among those who were predicted (by the original KBA) to have a lower 3-year mortality (<40% if receiving standard chemotherapy), the estimated benefits from timely allograft in CR1 disappeared, as indicated by the rebuilt KBA (Fig. 3). This change was also captured in the average survival rates of the Favorable group in Table 2.

Secondly, the KBA was developed for predictive purposes, in that the algorithm was trained and assessed empirically based on their predictive performance without a priori clinical information, hence was ill-suited to address the population-level treatment efficacy, which was, in essence, a causal question where the theoretical causal structure and confounding effects should be taken into account – this has a profound impact on each step of the modeling process. This topic is already beyond the scope of our study and has been elaborated in full detail by Shmueli (2010).

Discussion

Genomic-based precision oncology has attracted research funding enthusiasm over the decades, leading to very high expectations of its clinical application and healthcare impacts. Here we have evaluated the reproducibility of Gerstung et al.'s research published in *Nature Genetics* – the KBA that generated personally tailored predictions using individual clinical and genomic profiles, for which reproducibility and external validation are indispensable for its potential utility in future. Our evaluation went a long way to explaining the reasons why a study published in a top medical journal with promising results did not seem to have been put into practice. Moreover, we were able to take a closer look inside the black box of the model's construction, as well as at the causal inference reported in their paper.

The scientific community arguably values the novelty of research more than reproducibility, which systematically discourages the practice of reproducibility studies (Atmanspacher and Maasen 2016). We noticed that most of the current publication standards, including individual journals' submission guidelines for authors and general reporting guidelines such as CONSORT or STROBE, do not require the scrutiny of reproducibility before a submission can be accepted. This study reaffirmed the scientific value of sound reproducibility practices; even a single reproducibility study can add to the credibility of published results (Ioannidis 2014; Seibold et al. 2021).

Open policies regarding data and relevant supplements (e.g., code and documents) have long been considered a remedy for irreproducibility. With full access to the study materials, however, our study recognized that openness is a starting point, rather than a catch-all solution for reproducible research. Based on the experience of the first author (Yujun Xu) serving as a reproducible research editor at Biometrical Journal and informal correspondence with a senior colleague, at

present, only about two out of every ten submissions (manuscripts with supplementary data and code) are compliant with the journal guidelines and found reproducible from the start, while for the rest, three to four rounds of communication and correction between authors and editors, just to improve reproducibility, are standard.

There are numerous causes of irreproducibility given open data and code policies; in this study, two points have particularly caught our attention. One is code legibility – partially as a result of the research complexity with its hallmarks of high-dimensional data, advanced statistical methods, and intensive computation. Enhancing code readability (so that a code is easy to follow in a logical way) has been discussed widely, with code review being a standard practice in the field of computer science (Jones 2009; Oliveira et al. 2020). Yet little emphasis on this issue is seen in today's programming courses for scientists and in daily research activities in statistics, epidemiology, and biomedicine. Rather, these researchers are likely to play a multifaceted role, and acquire their coding skills only through practice. Still, programming style guides, such as Google's R style guide (Wickham, see https://style.tidyverse.org), remain material references to facilitate communication with other researchers. As Gerstung et al. also put it, "knowledge banks facilitate personally tailored therapeutic decisions but require sustainable updating, inclusive cohorts and large sample sizes" – a task that unavoidably requires collaboration among scientific entities.

The second point relating to causes of irreproducibility is about compatibility issues. The outputs of a given analysis code, especially those that are computationally intensive, can be influenced not only by the code itself, but also the version of the analytics software (and its extensions), OS (and its dependencies), as can be seen in **Fig. 1**, and even aspects of lower-level computer architecture (e.g., central processing unit). Modern software like R or Python is continuously updated. In most cases, executing a code in an upgraded environment will only result in small deviations (if any) in the results due to different precision levels of mathematical libraries in the computing environment. However, in rarer cases, unpredicted compatibility issues may occur, aborting the execution process or preventing the code from being executed at all. Solving such problems may be infeasible or impractical.

Docker is an open-source containerization tool that provides an opportunity to address the compatibility problems discussed above and re-installs an identical and server-independent computing environment for specific tasks, increasing the chance of reproducibility (Boettiger and Eddelbuettel 2017). A Dockerfile includes the steps needed to build a custom, layer-structured Docker image (which contains, for instance, a specific version of R with necessary R packages for a research project) building up from the OS on the base level or from another Docker image that already contains OS and other deep level structures. In the latter case, we obtain an updated Docker image without rebuilding the deeper layers, but compatibility issues may occur, as the existing bottom layers specified by one Dockerfile may not agree with the upper layers added by another. To avoid such issues, authors should provide a Dockerfile describing the full picture from the bottom OS to build the Docker image – a prohibitively massive undertaking requiring great engineering expertise. Alternatively, an easier and more practical option would be for future researchers to directly provide a complete Docker image that has been tested offline successfully, so that end-users can run a process directly in the environment contained in the provided image.

Gerstung et al.'s code has taught us an innovative application of methodological concepts which are elsewhere undocumented. It was a very useful exercise to study the code, even without rerunning it successfully, to deepen our understanding of Cox proportional hazards regression models with grouped random effects and handling the multistage processes with these Cox models as the building blocks. Publishing and studying code offer access to a rich source of statistical wisdom.

Meanwhile, the lack of relevant guidelines for the conduct and reporting of reproducibility studies means that good practice in this matter remains an open issue. Therefore, the present paper proposes criteria (presented in Table 1 and are open for discussion) for reproducibility, in the hope that it could not only encourage the publication of more reproducibility studies, but also help future researchers to perform reproducibility checks for themselves.

List Of Abbreviations

AML, acute myeloid leukemia; **allograft**, allogeneic hematopoietic cell transplant; **CR1**, first complete remission; **ELN**, European LeukemiaNet; **KBA**, the knowledge bank approach; **LSF**, Load Sharing Facility; **OS**, Operating System

Declarations

Funding

Not applicable

Conflicts of interest/Competing interests

The authors declare that they have no competing interests

Availability of data and material

The datasets analyzed during the current study are available in an online Github repository provided by Gerstung et al., https://github.com/gerstung-lab/aml-multistage

Code availability

The source code is publicly accessible in the Github repository provided by Gerstung et al. and our analysis was conducted in R (R Core Team, 2021)

Authors' contributions

Yujun Xu and Ulrich Mansmann both contributed to the study conception and methodology. Yujun Xu undertook the literature research, statistical analysis, and wrote the first draft of the manuscript under the supervision of Ulrich Mansmann. Yujun Xu and Ulrich Mansmann both reviewed, edited, and approved the final manuscript.

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent for publication

Not applicable

Acknowledgments

We thank Prof. Dr. Anne-Laure Boulesteix for her invaluable advice about reproducibility, Prof. Dr. med. Joerg Hasford for his constructive comments on this study, Mr Nikolaus von Bomhard for sharing his knowledge in systems engineering, Dr. Michael Medley, Dr. Graham Allen, and Ms Anna Jacob for their substantial help improving the manuscript.

References

1. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, Bloomfield CD, Cazzola M, Vardiman JW (2016) The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood 127(20):2391–2405. https://doi.org/10.1182/blood-2016-03-643544
2. Atmanspacher H, Maasen S (2016) Reproducibility: principles, problems, practices, and prospects. John Wiley & Sons
3. Boettiger C, Eddelbuettel D (2017) An Introduction to Rocker: Docker Containers for R. The R Journal 9(2):527–536. https://journal.r-project.org/archive/2017/RJ-2017-065/RJ-2017-065.pdf
4. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, Dombret H, Ebert BL, Fenaux P, Larson RA, Levine RL, Lo-Coco F, Naoe T, Niederwieser D, Ossenkoppele GJ, Sanz M, Sierra J, Tallman MS, Tien HF, Wei AH, Löwenberg B, Bloomfield CD (2017) Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood 129(4):424–447. https://doi.org/10.1182/blood-2016-08-733196
5. Fenwarth L, Thomas X, de Botton S, Duployez N, Bourhis JH, Lesieur A, Fortin G, Meslin PA, Yakoub-Agha I, Sujobert P, Dumas PY, Récher C, Lebon D, Berthon C, Michallet M, Pigneux A, Nguyen S, Chantepie S, Vey N, Raffoux E, Celli-Lebras K, Gardin C, Lambert J, Malfuson JV, Caillot D, Maury S, Ducourneau B, Turlure P, Lemasle E, Pautas C, Chevret S, Terré C, Boissel N, Socié G, Dombret H, Preudhomme C, Itzykson R (2021) A personalized approach to guide allogeneic stem cell transplantation in younger adults with acute myeloid leukemia. Blood 137(4):524–532. https://doi.org/10.1182/blood.2020005524
6. Hofner B, Scheipl F (2016) Guidelines for Code and Data Submission: Specific Guidance on Reproducible Research (RR), Document Version: 1.7. https://github.com/hofnerb/RR_Guideline/releases/download/v1.7-2/RR_Guideline.pdf. Accessed 11 Aug 2021
7. Hothorn T, Leisch F (2011) Case studies in reproducibility. Brief Bioinform 12(3):288–300. https://doi.org/10.1093/bib/bbq084
8. Huet S, Paubelle E, Lours C, Grange B, Courtois L, Chabane K, Charlot C, Mosnier I, Simonet T, Hayette S, Tigaud I, Thomas X, Salles G, Subtil F, Sujobert P (2018) Validation of the prognostic value of the knowledge bank approach to determine AML prognosis in real life. Blood 132(8):865–867. https://doi.org/10.1182/blood-2018-03-840348
9. Gentleman R (2005) Reproducible research: a bioinformatics case study. Statistical applications in genetics and molecular biology, 4, Article2. https://doi.org/10.2202/1544-6115.1034
10. Gerstung M (2016) AML multistage predictions (research only). https://cancer.sanger.ac.uk/aml-multistage. Accessed 11 Aug 2021
11. Gerstung M, Papaemmanuil E, Martincorena I, Bullinger L, Gaidzik VI, Paschka P, Heuser M, Thol F, Bolli N, Ganly P, Ganser A, McDermott U, Döhner K, Schlenk RF, Döhner H, Campbell PJ (2017) Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nat Genet 49(3):332–340. https://doi.org/10.1038/ng.3756
12. Gerstung M (2017) Supplementary Note. https://static-content.springer.com/esm/art%3A10.1038%2Fng.3756/MediaObjects/41588_2017_BFng3756_MOESM10_ESM.pdf. Accessed 11 Aug 2021
13. Gerstung M (2017) Code accompanying Precision oncology for acute myeloid leukemia using a knowledge bank approach. https://github.com/gerstung-lab/aml-multistage. Accessed 11 Aug 2021
14. Ioannidis JP (2014) How to make more published research true. PLoS Med 11(10):e1001747. https://doi.org/10.1371/journal.pmed.1001747
15. Jones DM (2009) The New C Standard: An Economic and Cultural Commentary. Knowledge Software Ltd, eBook (version 1.2). http://www.coding-guidelines.com/cbook/cbook1_2.pdf
16. Kitzes J, Turek D, Deniz F (2018) The practice of reproducible research: case studies and lessons from the data-intensive sciences. University of California Press, Oakland

17. Oliveira D, Bruno R, Madeiral F, Castor F (2020) Evaluating Code Readability and Legibility: An Examination of Human-centric Studies. IEEE International Conference on Software Maintenance and Evolution (ICSME), 2020, pp. 348–359, https://doi.org/10.1109/ICSME46990.2020.00041
18. Perperoglou A (2014) Cox models with dynamic ridge penalties on time-varying effects of the covariates. Statistics in medicine 33(1):170–180. https://doi.org/10.1002/sim.5921
19. R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org
20. Seibold H, Czerny S, Decke S, Dieterle R, Eder T, Fohr S, Hahn N, Hartmann R, Heindl C, Kopper P, Lepke D, Loidl V, Mandl M, Musiol S, Peter J, Piehler A, Rojas E, Schmid S, Schmidt H, Schmoll M, Schneider L, To XY, Tran V, Völker A, Wagner M, Wagner J, Waize M, Wecker H, Yang R, Zellner S, Nalenz M (2021) A computational reproducibility study of PLOS ONE articles featuring longitudinal data analyses. PloS one 16(6):e0251194. https://doi.org/10.1371/journal.pone.0251194
21. Schlenk RF, Fröhling S, Hartmann F, Fischer JT, Glasmacher A, del Valle F, Grimminger W, Götze K, Waterhouse C, Schoch R, Pralle H, Mergenthaler HG, Hensel M, Koller E, Kirchen H, Preiss J, Salwender H, Biedermann HG, Kremers S, Griesinger F, Benner A, Addamo B, Döhner K, Haas R, Döhner H, AML Study Group Ulm (2004) Phase III study of all-trans retinoic acid in previously untreated patients 61 years or older with acute myeloid leukemia. Leukemia 18(11):1798–1803. https://doi.org/10.1038/sj.leu.2403528
22. Schlenk RF, Döhner K, Mack S, Stoppel M, Király F, Götze K, Hartmann F, Horst HA, Koller E, Petzer A, Grimminger W, Kobbe G, Glasmacher A, Salwender H, Kirchen H, Haase D, Kremers S, Matzdorff A, Benner A, Döhner H (2010) Prospective evaluation of allogeneic hematopoietic stem-cell transplantation from matched related and matched unrelated donors in younger adults with high-risk acute myeloid leukemia: German-Austrian trial AMLHD98A. Journal of clinical oncology: official journal of the American Society of Clinical Oncology 28(30):4642–4648. https://doi.org/10.1200/JCO.2010.28.6856
23. Schlenk RF, Lübbert M, Benner A, Lamparter A, Krauter J, Herr W, Martin H, Salih HR, Kündgen A, Horst HA, Brossart P, Götze K, Nachbaur D, Wattad M, Köhne CH, Fiedler W, Bentz M, Wulf G, Held G, Hertenstein B, Salwender H, Gaidzik VI, Schlegelberger B, Weber D, Döhner K, Ganser A, Döhner H, German-Austrian Acute Myeloid Leukemia Study Group (2016) All-trans retinoic acid as adjunct to intensive treatment in younger adult patients with acute myeloid leukemia: results of the randomized AMLSG 07 – 04 study. Annals of hematology 95(12):1931–1942. https://doi.org/10.1007/s00277-016-2810-z
24. Shmueli G (2010) To explain or to predict? Statist Sci 25(3):289–310. https://doi.org/10.1214/10-STS330
25. Therneau TM, Grambsch PM, Pankratz VS (2003) Penalized survival models and frailty. Journal of computational graphical statistics 12(1):156–175. https://doi.org/10.1198/1061860031365

Tables

Table 1

Reproducibility checklist

Aspect	Item No	Item	Note
Accessibility (yes/partially/no)	1a	Data	
	1b	Is the data (if available) original, anonymized, or simulated?	Simulated data is usually provided when the original data is confidential
	1c	Data dictionary	A collection of names, definitions, descriptions, etc., of variables in the dataset(s) of the research project
	2	Source code	
	3	Documentation	Is there a README file or technical note?
	4	Statistical software and the version used	e.g., R (v.4.1.1)
	5	Software extensions and the versions used	e.g., Tidyverse (v.1.3.0)
	6	Operating system	e.g. x86_64, Debian GNU/Linux 11
	7	Format of the results	What types of tables/figures are shown?
Clarity (yes/partially/no)	8	Description of methods	e.g., Theoretical concepts, analytical strategies
	9	Code legibility	Is the code self-explanatory, regardless of comments (e.g., use consistent variable names, separate code chunks with different purposes; well-structured without short-cuts or cryptic codes)? Are compiled languages like C or C + + used?
	10	Comments in the code	e.g., Are there unnecessary comments such as alternative code lines?
	11a	Documentation of custom packages and functions, if applicable	e.g., R package vignette
	11b	When applicable, is there a validation of custom functions/packages?	
Code execution	12a	On mouse-clicks	
		Minor modifications required	
		Major modifications with expertise (e.g., reverse engineering of results) required	
		Impossible to rerun	
	12b	When applicable, reasons for irreproducibility	e.g., Critical information is missing for modifications

Aspect	Item No	Item	Note
Implementation of the theoretical methods described	13	Consistent/largely consistent/largely inconsistent/unable to identify	Does the code reflect the methods in the paper?
Matching of outputs	14	Identical with exactly the same results	
		Same interpretation with deviations in numbers	
		Inconsistent conclusions	
		Unable to reproduce the results	
Overall reproducibility	15	Reproducible/partially reproducible/irreproducible	
	16	Background of researcher(s) performing the assessment	e.g., Clinician, epidemiologist, biostatistician/statistician, engineer

Table 2

Comparison of estimated 3-year survival before and after the modification of the KBA

	Original KBA[a] (with 1540 patients)		Modified KBA[a] (with 995 patients)	
Survival predictions	Allograft in CR1	Standard Chemo in CR1[b]	Allograft in CR1	Standard Chemo in CR1[b]
Overall	58,1%	49,8%	63,2%	62,0%
Favorable[c]	75,7%	70,7%	73,8%	76,9%
Inter-2[c]	53,9%	45,8%	59,0%	57,5%
Inter-1[c]	55,0%	45,0%	58,1%	54,2%
Adverse[c]	35,6%	25,3%	49,4%	39,1%

a: the knowledge bank approach based on either 1540 patients (left), or 995 patients with ineligible patients excluded (right)
b: either chemotherapy only, or chemotherapy in CR1 with salvage allograft after relapse

c: risk classification according to the European LeukemiaNet recommendation

Figures

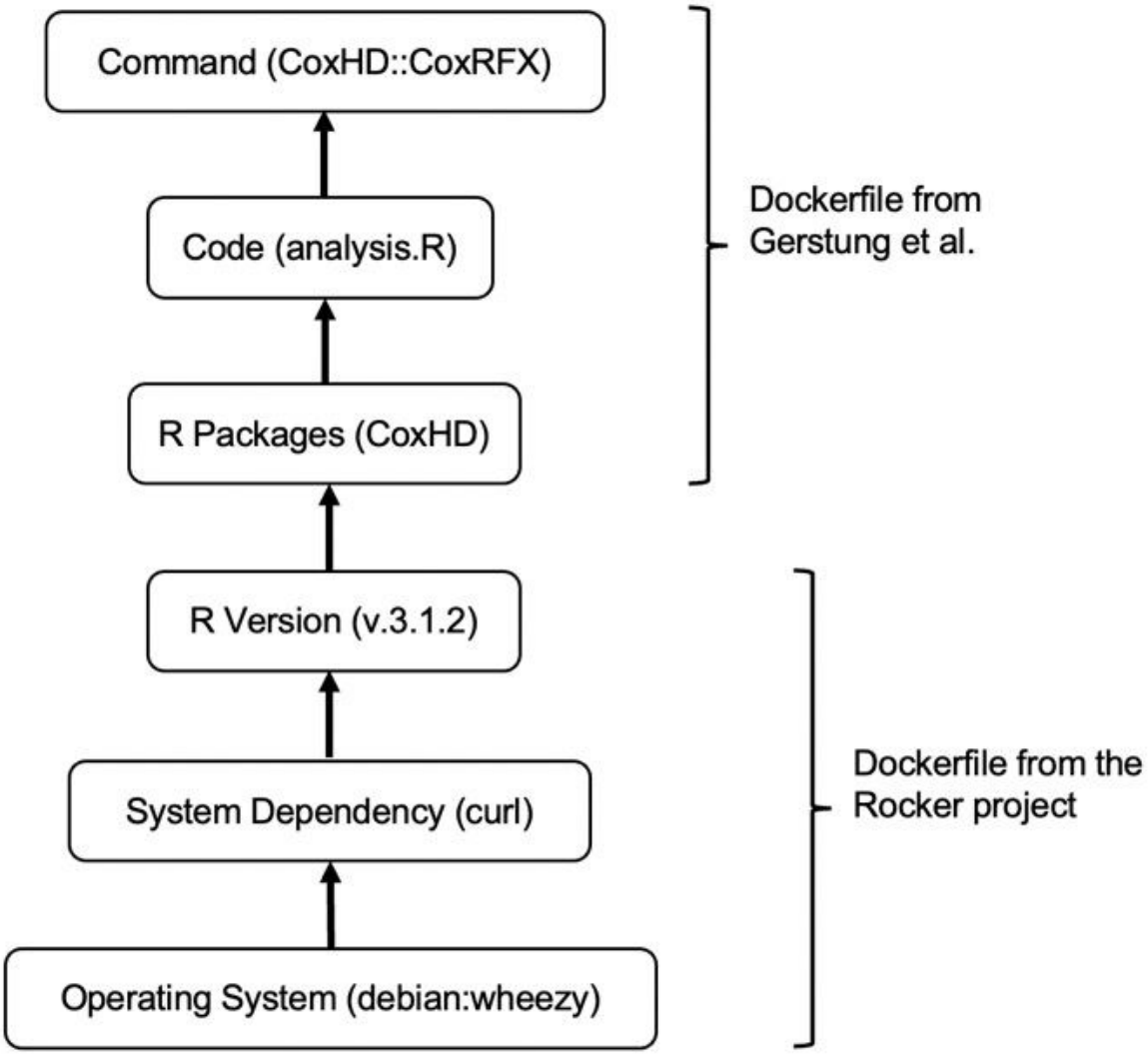


Figure 1

Six-layered structure of an R computing environment. The Dockerfile from Gerstung et al. contructs layers upon a pre-built Docker-based R (v.3.1.2); however, compatibility issues occur across these layers

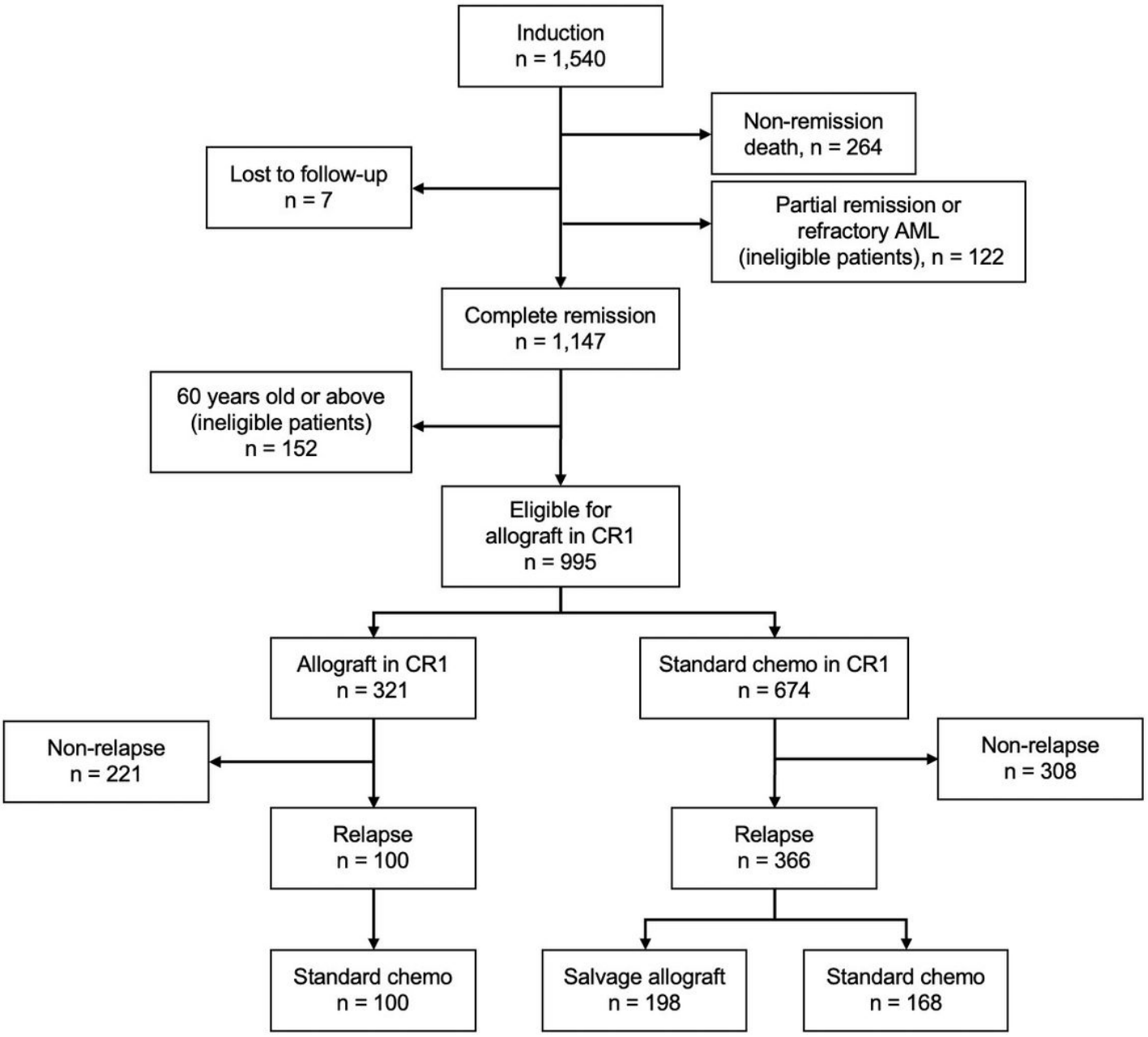
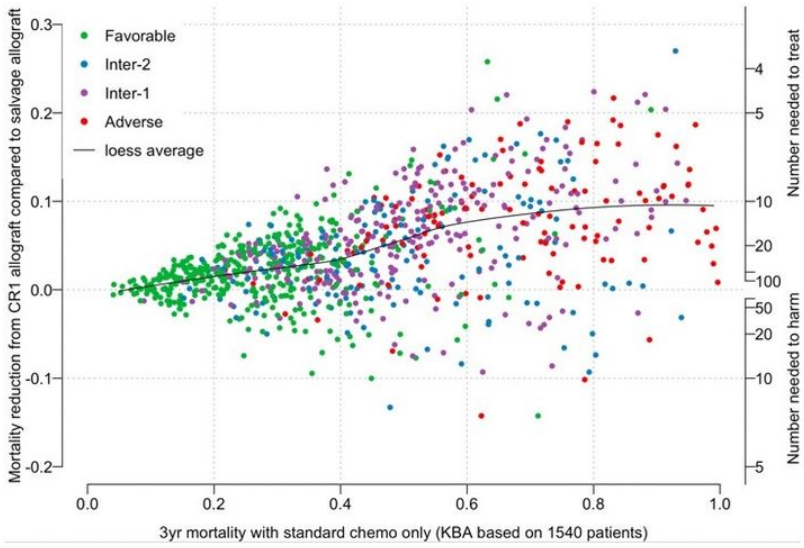
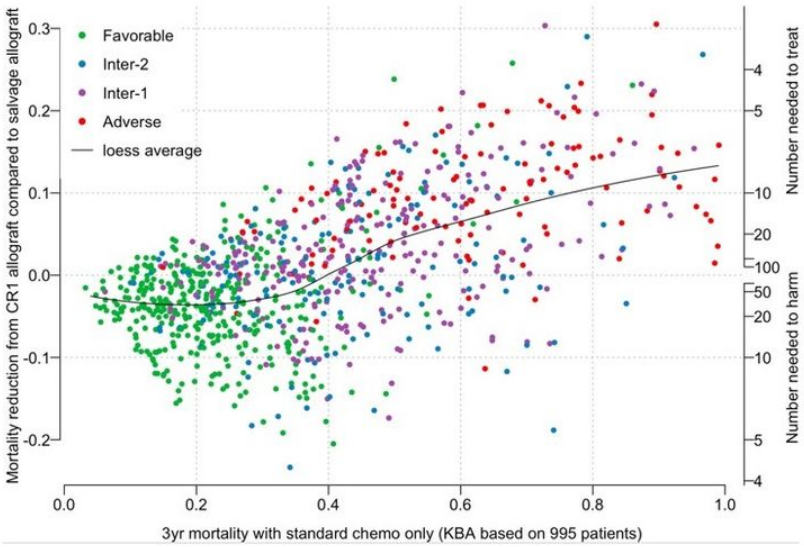


Figure 2

Flow chart showing treatments and prognoses of AML patients in the knowledge bank database



(a)



(b)

Figure 3

Predicted 3-year mortality reduction from allograft in CR1, as opposed to salvage allograft after relapse (y-axis), predicted 3-year mortality of standard chemotherapy only (x-axis). Diagnosis as the starting point. (a) The KBA was based on the entire 1540 patients, while predictions were calculated for 995 patients eligible for allograft, adapted from Gerstung et al. (2017); (b) Modified KBA based on the 995 eligible patients

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfile4.xlsx
- Supplementaryfiles13.pdf