A FANCY UNIVERSITY

AN EVEN FANCIER LAB

Christian Stentoft Michelsen

`tufte-style-thesis`, a Tufte-styled LaTeX class for theses

Actually more of a mix
between Edward Tufte and Robert Bringhurst

Doctoral thesis

November 10, 2022

Supervisor	their name	their job
Cosupervisor	also their name	also their job
Jury members	jury 1	jobs
	jury 2	…

Christian Stentoft Michelsen, *tufte-style-thesis,*
a Tufte-styled LaTeX class for theses, Actually more of a mix
between Edward Tufte and Robert Bringhurst, November 10, 2022.

Til kvinderne i mit liv

Contents

Preface

This Ph.D. thesis summarizes my scientific research in collaboration with the Niels Bohr Institute (NBI) and the Globe Institute, University of Copenhagen, and was funded by the Lundbeck Foundation. The research was supervised by Associate Professor Troels C. Petersen (NBI) and Assistant Professor Thorfinn S. Korneliussen (Globe Institute).

Being a multi-disciplinary project, the research presented in this thesis is multi-faceted and covers a wide range of topics with the main scope being the development and integration of a novel statistical methods and machine learning models for the analysis of large-scale biological data. The thesis is organized as follows. First I present a brief introduction to the statistical methods and machine learning models used in the thesis. Then I present the research in the form of four papers, each of which reflects a different aspect of the research.

The first paper presents a novel method I developed for detecting and classifying ancient DNA damage in metagenomic samples taking the full taxanomic information into account. While the first paper focuses on the development of the statistical model in the field of ancient genomics, the second paper focuses on the use of modern machine learning models in medicine and how advanced boosted decision trees can not only improve the accuracy of identifying patients at risk of being readmitted after knee or hip surgery, but doing so in a way that is interpretable as well. In the beginning of 2020 we all experienced how COVID-19 suddenly changed our lives and impacted our societies in dramatic ways. During this time, I ended up working for Statens Serum Institut (SSI), the Danish CDC, on a project to develop an agent based model capable of simulating the spread of COVID-19 in Denmark. This model is presented in the third paper and was used to inform the Danish government on how to best handle the pandemic in the early stages and the effect of lockdowns and other measures. Finally, in the fourth paper we show how advanced Bayesian methods can be utilized to better estimate the diffusion coefficients of molecules in the cell nucleus in XXX experiments.

Acknowledgements

First of all I want to express my sincere gratitude to my long time supervisor, captain, and friend: Troels. You are truly an inspiration to work with. I want to thank you for opening so many doors for me, both academically, professionally, and nautically. I am looking forward to our future adventures together. I also want to thank my co-supervisor, Thorfinn. I want to thank you for introducing me to the field of bioinformatic and helping me to develop my skills in this area. I also want to thank you for your patience and guiding me through the endless amount of (near) identical biological concepts and helping me to understand the minute differences.

I have been fortunate to work with people from a wide range of backgrounds and disciplines during my Ph.D. The author lists on the papers in this thesis include a particle physicist, bioinformaticians, a clinical professor, epidemiologists, a medical doctor, a bio-physicist, a mathematician, a biologist, and the president of the Royal Danish Academy of sciences and letters. Before anything else, I want to thank all of my co-authors for their work and contributions to these papers and for allowing me to be a part of their projects. I have learned a lot from all of you, and I hope that I have been able to contribute something to your work as well.

I am thankful for all the people who have helped me with my work and listened to my complaints when I was stuck, when the code did not compile, or when the small bug was almost impossible to find (which was not a small amount of time). In particular I want to thank the people at Globe who I have spent the most time with; Rasa, Alba, and Rasmus. I also want to thank the Korneliussen Group and the people in my office for helpful advice, suggestions and discussions. This also includes Daniel Nielsen and Rasmus Ørsøe from NBI. Finally, I want to thank Mathias Heltberg for many years of fruitful collaboration and for including me in his projects.

This project would not have been possible had it not been for the Lundbeck Foundation which funded my Ph.D. In addition to the funding itself, I am grateful for the inter-disciplinary aspect of project which has allowed me to meet so many inspiring and talented people and for the freedom to pursue my own interests within the project.

I would also like to express my gratitude to Professor Guido Sanguinetti from the International School for Advanced Studies, SISSA, in Trieste, Italy, for hosting me in his group during the Winter of 2021/2022. My gratitude also goes out to Kosio, Sara, Max, Romina, Noor, Viplove, Anne-Marie, and all the other wonderful people that I met during in Trieste. Thanks for making my stay in Italy so enjoyable and for welcoming me in a way that only non-Danes can do.

I want to thank my friends for always being there for me. A special thanks to my friends from NBI and Borchen who I know that I can always count on, whether or not that includes a trip in the party bus of the Sea, taking Artemis out for a sail, or board games and beer. Thank you for always being there. I also want to thank my family, especially my parents for their unconditional support and encouragement. I am grateful for the opportunities that they have given me and for the sacrifices that they have made for me.

Lastly, I want to thank my future wife and mother of our child, Anna. I would not have been able to do this without you. Thank you for your patience and support. I am looking forward to our future together. I love you to the moon and back.

Abstract

Basically a thesis (book?) class for Tufte lovers like myself. I am aware that `tufte-latex` already exists but I just wanted to create my own thing.

Dansk Abstract

Her et dansk abstract.

Publications

The work presented in this thesis is based on the following publications:

Christian S. Michelsen, Mikkel W. Pedersen, Antonio Fernandez-Guerra, Lei Zhao, Troels C. Petersen, Thorfinn S. Korneliussen (2022). "metaDMG: An Ancient DNA Damage Toolkit".

Christian Michelsen, Christoffer C. Jorgensen, Mathias Heltberg, Mogens H. Jensen, Alessandra Lucchetti, Pelle B. Petersen, Troels C. Petersen, Henrik Kehlet (2022). "Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty - a machine learning based approach.".

Mathias S. Heltberg, Christian Michelsen, Emil S. Martiny, Lasse E. Christensen, Mogens H. Jensen, Tariq Halasa and Troels C. Petersen (2022). "Spatial Heterogeneity Affects Predictions from Early-Curve Fitting of Pandemic Outbreaks: A Case Study Using Population Data from Denmark". In: Royal Society Open Science 9.9. issn: 2054-5703. doi: 10.1098/rsos.220018.

Susmita Sridar, Mathias S. Heltberg, Christian S. 6 Michelsen Judith M. Hattab, Angela Taddei (2022). "Microscopic single molecule dynamics suggest underlying physical properties of the silencing foci".

1 *Notes on the design*

This class is my personal mix of different book design influences: mainly the works of Edward R. Tufte, (Heltberg et al., 2022; Korneliussen, Albrechtsen, and Nielsen, 2014) known for the big margin and the plentiness of sidenotes and sidecaptions. The margins are however not as prominent as in Tufte's works, the main text takes a bit more space, more like in Robert Bringhurt's typographer's bible (Heltberg et al., 2022).

So it is a bit of a mix of Tufte and Bringhurst, with some of my own choices for other design features, as we will see through this chapter.

1.1 *Document layout*

While `tufte-style-thesis` is a class for typesetting theses, the general layout is pretty much the same as in a regular book. A book is traditionally divided into three major sections: the front matter, the main matter and the back matter.

Bibliography

Heltberg, Mathias Spliid et al. (2022). "Spatial Heterogeneity Affects Predictions from Early-Curve Fitting of Pandemic Outbreaks: A Case Study Using Population Data from Denmark". In: *Royal Society Open Science* 9.9. ISSN: 2054-5703. DOI: 10.1098/rsos.220018.

Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen (2014). "ANGSD: Analysis of Next Generation Sequencing Data". In: *BMC Bioinformatics* 15.1, p. 356. ISSN: 1471-2105. DOI: 10.1186/s12859-014-0356-4. URL: https://doi.org/10.1186/s12859-014-0356-4 (visited on 2019).

2 *Paper I: metaDMG: An Ancient DNA Damage Toolkit*

The following pages contain the article:

Christian S. Michelsen, Mikkel W. Pedersen, Antonio Fernandez-Guerra, Lei Zhao, Troels C. Petersen, Thorfinn S. Korneliussen (2022). “metaDMG: An Ancient DNA Damage Toolkit”.

metaDMG:

An Ancient DNA Damage Toolkit

Christian Stentoft Michelsen [1 †] ✉, **Mikkel Winther Pedersen** [2 †] ✉, **Antonio Fernandez-Guerra** [2], **Lei Zhao**[2], **Troels C. Petersen** [1], **Thorfinn Sand Korneliussen** [2]

[1]Niels Bohr Institute, University of Copenhagen; [2]Globe Institute, University of Copenhagen

✉ **For correspondence:** christianmichelsen@gmail.com (CM); mwpedersen@sund.ku.dk (MW)

[†]Authors contributed equally.

Present address: Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark

Data availability: Data is available on Zenodo or the Github repository.

Funding: This work was supported by Carlsberg Foundation Young Researcher Fellowship awarded by the Carlsberg Foundation [CF19-0712], and the Lundbeck Foundation Centre for Disease Evolution: [R302-2018-2155 to L.Z]. The funders had no role in the decision to publish.

Competing interests: The author declare no competing interests.

Abstract

1. **Motivation** Under favourable conditions DNA molecules can survive for more than two million years (Kjaer et al in press). Such genetic remains can give unique insights to past assemblages, populations and evolution of species. However, DNA is degraded over time, and are therefore found in ultra low concentrations making it highly prone to contamination from modern DNA sources. Despite strict precautions implemented in the field (Llamas et al., 2017), DNA from modern sources does appear in the final output data. One authenticity criteria used in all ancient DNA studies are the high nucleotide mis-incorporation rates that can be observed as a result of chemical post-mortem DNA damage, in fact misincorporation patterns have become instrumental to authenticate ancient sequences. To date this has primarily been possible for single organisms (Jónsson et al., 2013)) and recently for assemblies (Borry et al., 2021), but these methods have not been designed, nor can they be computationally upscaled to calculate the thousands of taxonomic species that occur in just one metagenome.
2. **Methods** We present `metaDMG`, a novel framework that takes advantage of the information already contained in the alignment files to compute and statically evaluate post-mortem DNA damage, thus bypassing the need for classifying and splitting reads into individual organisms and realigning these to parse data to mapDamage2.0 (Jónsson et al., 2013). It

uses a Bayesian approach that combines a geometric damage profile with a beta-binomial model to fit the entire model to the misincorporations which drastically improve the damage estimates compared to previous methods.

3. **Results** Using a two-tier simulation setup, we find `metaDMG` to not only be a factor of 10 faster than previous methods but it is also more accurate and able to evaluate even complex metagenomes with tens of thousands of species. Even with very few number of reads, down to even below 1000 reads. BLABLA, more results here.
4. **Conclusion** `metaDMG` includes state-of-the-art statistical methods for computing nucleotide misincorporation and fragmentation patterns of even highly complex samples along with re-implementation of the current statistics used within the field such as PMDtools (Skoglund et al., 2014). This suite of programs is freely available and consists of computational parts implemented as multi threaded C++ programs as well as computationally optimized modern python libraries, and an interactive dashboard for displaying the results. `metaDMG` is furthermore flexible, compatible with custom databases, can output nucleotide misincorporation and fragmentation patterns at different taxonomic ranks as well as per reference ID.

keywords: ancient DNA, damage estimation, DNA damage, lowest common ancestor, metaDMG, metadamage, metagenomics, statistics.

1 | INTRODUCTION

Throughout the life of an organism, it contaminates its surrounding environment with cells or tissue and hence its DNA contained within. As the cell leaves its host, DNA repair mechanisms stops and the DNA is now subjected to chemical and mechanical degradation, resulting in fragmented molecules and chemical damages, characteristic for ancient DNA (Briggs et al., 2007; Dabney, Meyer, and Pääbo, 2013). Ancient DNA has been shown to be able to survive in the environment for thousands and even up to two million years (Kjaer et al in review), and have been widely used to study past organisms and organism composition (Cappellini et al., 2018). Particularly misincorporations of cytosines on thymines as a result of deamination has been found to independently authenticate ancient DNA origin (Dabney, Meyer, and Pääbo, 2013; Ginolhac et al.,

2011). Postmortem damage with regards to DNA is characterized by the four Briggs parameters (Briggs et al., 2007). A damaged dna fragment tend to be short, and is likely to be single stranded at the termini of the fragment. There is an high proportion of C→T substitutions at the single stranded part ϵ_{ss}, a somewhat higher C→ T at the double stranded part ϵ_{ds}. The length of the single stranded part (*overhang*) follows a geometric distribution λ, and finally there might be breaks at the backbone in the double stranded part ν. It is possible to estimate these four Briggs parameters Jónsson et al., 2013 but these four parameters are rarely used directly for asserting "ancientness", and researchers working with ancient DNA tend to simply use the empirical C→T on the first position of the fragment together with other supporting summary statistic of the experiment. This ancient DNA (aDNA) authenticity approach, were initially performed on single individual sources such as hair, bones, teeth and later on ancient environmental samples such as soil sediments CITE SOMETHING. While this is a relatively fast process for single individuals it becomes increasingly demanding, iterative and time consuming as the samples and the diversity within increases, as in the case for metagenomes from ancient soil, sediments, dental calculus, coprolites and other ancient environmental sources. It has therefore been practice to estimate damage for only the key taxa of interest in a metagenome, as a metagenomic sample easily includes thousands of different taxonomic entities, that would make a complete estimate an impossible task.

We have devised a novel test statistic in `metaDMG` which takes into account all relevant information in single scalar. For these reasons, we present here `metaDMG`, a tool that enables fast and accurate DNA damage estimation of whole metagenomes within hours. `metaDMG` is designed and upscales equally for the increasingly large datasets that are generated in the field of ancient environmental DNA, but can also with advantage be used to estimate DNA damage of single genomes and samples with low complexity, it can even compute an global damage estimate for a given sample. `metaDMG` is compatible with the NCBI taxonomy and can use ngsLCA to perform a naïve last common ancestor of the aligned reads to get precise damage estimates for the reads classified to different taxonomic nodes. In addition, it is also designed to be used with custom taxonomies and metagenomic assembled genomes.

After defining the method and notation used throughout this paper, we show through multiple sets of simulations that `metaDMG` not only improves on existing methods in the case of single-genome damage estimation but also work for metagenomic samples. Finally, we apply our method on a representative mix of nine different metagenomic samples to show the real life performance

of metaDMG.

2 | METHODS & MATERIALS

Perhaps the most basic bioinformatic analyses is the difference between two nucleotide sequences. This assumes that we have a haploid representation of our target organisms and larger differences can be interpreted as larger genetic differences. Obtaining a haploid representation is none trivial, firstly our target organism might not be haploid and we need to construct a consensus genome, secondly data from modern day sequencers are essentially a sampling with replacement process and we need to infer the relative location of each of the possible millions or even billions of short DNA fragments, this is the process which is called mapping or alignment. Thirdly, and the focus for this manuscript, is the quantification of the presence of postmortem damage (PMD) in DNA. PMD mainly manifests as an excess of cytosine to thymine substitutions at the termini of fragments that has been prepared for sequencing. A priori we can not directly observe these actual biochemical changes but we can align each fragment and consider the difference between reference and read as possible PMD, and it is even possible to use the excess of C to T at the single fragment level to separate modern from ancient (data with PMD) (Skoglund et al., 2014). Expanding from the single read all reads for a sequencing experiment and genome to tabulate the overall substitution or mismatch rates to obtain a stastistic of the damage (Borry et al., 2021) or even estimate the four Briggs parameters that is traditionally used to characterize the damage signal (Jónsson et al., 2013).

We have devised a general ancient DNA damage toolkit with a special emphasis in a metagenomic setting which implements and expands existing relevant methods but also expands with several state of the art novel methodologies. At the most basic level we have reimplemented the approach given in (Skoglund et al., 2014) which allows for the extracting and separation of highly damaged DNA reads. Secondly under the assumption of vast amounts of data we have defined a full multinomial regression model building on the method in (Cabanski et al., 2012), we show that this will give superior and stable results if it is possible to obtain high depth and coverage data. However in standard ancient DNA context it is generally not possible to obtain vast amounts of data and we propose two novel tests statistics that is especially suited for this scenario. To our knowledge there are no currently available methods that is geared towards damage analysis in a metagenomic setting and existing approaches are essentially based on remapping against the sin-

gle target organism and does not take into account any possible issues with regards to reads being well assigned or specified. Our solution called `metaDMG` (pronounced metadamage), estimates the damage patterns in metagenomic samples in a three step approach. First, the lowest common ancestor (LCA) for each read (mapped to a multi-species reference database) is computed and the the mismatch matrix for each leaf node (e.g. taxonomic ID or contig, depending on the database used) is computed based on the mapped reads. Second, `metaDMG` fits a damage model to each leaf node to compute the ancient damage estimates. Finally, the results are visualized in the `metaDMG` dashboard, which is a state of the art graphical user interface that allows for fast and user-friendly interaction with the results for further downstream analysis and visualization.

2.1 | Lowest Common Ancestor and Mismatch matrices

For environmental DNA (eDNA) studies we routinely apply a competitive alignment approach where we consider all possible alignments for a given read. Each read is mapped against a multi species reference databases (e.g. nucleotide or RefSeq from NCBI or custom downloaded). A single read might map to a highly conserved gene that is shared across higher taxonomic ranks such as class or even domains. This read will not provide relevant information due to the generality, whereas a read that maps solely to a single species or species from a genus would be indicative of the read being well classified. We seek to obtain the pattern or signal of damage which is done by the tabulation of the cycle specific mismatch rates between our reference and observed sequence for all well classified reads.

In details we compute the lowest common ancestor (lca) for all alignments for for each read, this is done using (Wang et al., 2022) and if a read is well classified or properly assigned based on a user defined threshold (specics, genus or family) we tabulate the mismatches for each cycle, if a read is not well assigned it is discarded. Pending on the run mode we allow for the construction of these mismatch tables on three different lavels. Either we obtain a basic single global mismatch matrix, which could be relevant in a standard single genome aDNA study and similar to the tabulation used in (Jónsson et al., 2013). Secondly we can obtain per reference counts or if a taxonomy database has been supplied we allow for the aggregation from leaf nodes to the internal taxonomic ranks towards the root.

To suit as many users as possible, `metaDMG` takes as input an alignment file (.bam, .sam, or .sam.gz), where Each read is hereafter allowed an equal chance to map against the multiple refer-

ences. One read can therefore attract multiple alignments, and we thus first seek to find the lowest common ancestor (LCA) among the alignments based on the tree structure from the databases and a user defined read-reference similarity interval (Wang et al., 2022). Note that `metaDMG` is not limited to the NCBI database and allow for custom databases as well.

Regardless of runmode or weighing scheme used in the possible aggregation we obtain the nucleotide substitution frequencies across reads which provides us with the position dependent mismatch matrices, $\underline{\underline{M}}(x)$, with x denoting the position in the read, starting from 1. At a specific position, $M_{\text{ref}\to\text{obs}}(x)$ describes the number of nucleotides that was mapped to a reference base $\mathcal{B}_{\text{ref}}$ but observed to be $\mathcal{B}_{\text{obs}}$, where $\mathcal{B} \in \{A, C, G, T\}$. The number of C to T transitions, e.g., is denoted as $M_{C\to T}(x)$.

When calculating the mismatch matrix, two different approaches can be taken. Either all alignments of the read will be counted, which we will refer to as weight-type 0, or the counts will be normalized by the number of alignments of each read; weight-type 1 (default).

2.2 | Damage Estimation

The damage pattern observed in aDNA has several features which are well characterized. By modelling these, one can construct observables sensitive to aDNA signal. We model the damage patterns seen in ancient DNA by looking exclusively at the $C \to T$ transitions in the forward direction (5') and the $G \to A$ transitions in the reverse direction (3'). For each LCA, we denote the number of transitions $k(x)$ as:

$$k(x) = \begin{cases} M_{C\to T}(x) & \text{for } x > 0 \quad \text{(forward)} \\ M_{G\to A}(x) & \text{for } x < 0 \quad \text{(reverse)} \end{cases} \tag{1}$$

and the number of the reference counts $N(x)$:

$$N(x) = \begin{cases} \sum_{i\in\mathcal{B}} M_{C\to i}(x) & \text{for } x > 0 \quad \text{(forward)} \\ \sum_{i\in\mathcal{B}} M_{G\to i}(x) & \text{for } x < 0 \quad \text{(reverse)}, \end{cases} \tag{2}$$

where the sum is over all four bases. The damage frequency is thus $f(x) = k(x)/N(x)$. A natural choice of likelihood model would be the binomial distribution. However, we found that a binomial likelihood lacks the flexibility needed to deal with the large amount of variance (overdispersion) we found in the data due to bad references and misalignments.

To accommodate overdispersion, we instead apply a beta-binomial distribution, $\mathcal{P}_{\text{BetaBinomial}}$, which treats the probability, p, as a random variable following a beta distribution[1] with mean μ and concentration ϕ: $p \sim \text{Beta}(\mu, \phi)$. The beta-binomial distribution has the the following probability density function:

$$\mathcal{P}_{\text{BetaBinomial}}(k \mid N, \mu, \phi) = \binom{N}{k} \frac{B\left(k + \mu\phi,\ N - k + \phi(1-\mu)\right)}{B\left(\mu\phi,\ \phi(1-\mu)\right)}, \tag{3}$$

where B is defined as the beta function:

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{4}$$

with $\Gamma(\cdot)$ being the gamma function (Cepeda-Cuervo and Cifuentes-Amado, 2017).

The close resemblance to a binomial model is most easily seen by comparing the mean and variance of a random variable k following a beta-binomial distribution, $k \sim \mathcal{P}_{\text{BetaBinomial}}(N, \mu, \phi)$:

$$\begin{aligned} \mathbb{E}[k] &= N\mu \\ \mathbb{V}[k] &= N\mu(1-\mu)\frac{\phi + N}{\phi + 1}. \end{aligned} \tag{5}$$

The expected value of k is similar to that of a binomial distribution and the variance of the beta-binomial distribution reduces to a binomial distribution as $\phi \to \infty$. The beta-binomial distribution can thus be seen as a generalization of the binomial distribution.

Note that both equation (3) and (5) relates to damage at a specific base position, i.e. for a single k and N. To estimate the overall damage in the entire read using the position dependent counts, $k(x)$ and $N(x)$, we model μ as position dependent, $\mu(x)$, and assume a position-independent concentration, ϕ. We model the damage frequency with a modified geometric sequence, i.e. exponential decreasing for discrete values of x:

$$\tilde{f}(x; A, q, c) = A(1-q)^{|x|-1} + c. \tag{6}$$

Here A is the amplitude of the damage and q is the relative decrease of damage pr. position. A background, c, was added to reflect the fact that the mismatch between the read and reference might be due to other factors than just ancient damage. As such, we allow for a non-zero amount of damage, even as $x \to \infty$. This is visualized in Fig. 1 along with a comparison between the classical binomial model and the beta-binomial model.

To estimate the fit parameters, A, q, c, and ϕ, we apply Bayesian inference to utilize domain specific knowledge in the form of priors. We assume weakly informative beta-priors[2] for both A, q,

[1] Note that we do not parameterize the beta distribution in terms of the common (α, β) parameterization, but instead using the more intuitive (μ, ϕ) parameterization. One can reparameterize $(\alpha, \beta) \to (\mu, \phi)$ using the following two equations: $\mu = \frac{\alpha}{\alpha+\beta}$ and $\phi = \alpha + \beta$ (Cepeda-Cuervo and Cifuentes-Amado, 2017).

[2] Parameterized as (μ, ϕ)

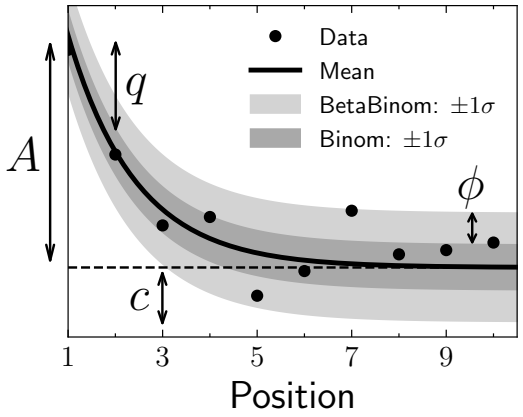


Figure 1. Illustration of the damage model. The figure shows data points as circles and the damage, $f(x)$, as a solid line. The amplitude of the damage is A, the offset is c, and the relative decrease in damage pr. position is given by q. The damage uncertainty for a binomial model is shown in dark grey and the uncertainty for a beta-binomial model in light grey. The additional uncertainty of the beta-binomial model, compared to the binomial model, is related to ϕ, see equation (5).

and c. In addition to this, we assume an exponential prior on ϕ with the requirement of $\phi > 2$ to avoid too much focus on 0-or-1 probabilities (McElreath, 2020). The final model is thus:

$$
\begin{aligned}
&[A \text{ prior}] && A \sim \text{Beta}(0.1, 10) \\
&[q \text{ prior}] && q \sim \text{Beta}(0.2, 5) \\
&[c \text{ prior}] && c \sim \text{Beta}(0.1, 10) \\
&[\phi \text{ prior}] && \phi \sim 2 + \text{Exponential}(1000) \\
&[\text{likelihood}] && k_i \sim \mathcal{P}_{\text{BetaBinomial}}\left(N_i, \tilde{f}\left(x_i; A, q, c\right), \phi\right),
\end{aligned}
\tag{7}
$$

where i is an index running over all positions.

We define the damage due to deamination, D, as the background-subtracted damage frequency at the first position: $D \equiv \tilde{f}(|x| = 1) - c$. As such, D is the damage related to ancientness. Using the properties of the beta-binomial distribution, eq. (5), we find the mean and variance of the damage, D:

$$
\begin{aligned}
\mathbb{E}\left[D\right] &\equiv \bar{D} = A \\
\mathbb{V}\left[D\right] &\equiv \sigma_D^2 = \frac{A(1-A)}{N}\frac{\phi + N}{(\phi + 1)}.
\end{aligned}
\tag{8}
$$

Since D estimates the overexpression of damage due to ancientness, not only the mean is relevant but also the certainty of $D > 0$. We quantify this through the significance $Z = \bar{D}/\sigma_D$

which is thus the number of standard deviations ("sigmas") away from zero. Assuming a Gaussian distribution of D, $Z > 2$ would indicate a probability of D being larger than zero, i.e. containing ancient damage, with more than 97.7% probability. These two values allows us to not only quantify the amount of ancient damage (ie. $\bar{D}$) but also the certainty of this damage (Z) without even having to run multiple models and comparing these.

We perform the Bayesian inference of the parameters models using Hamiltonian Monte Carlo (HMC) sampling which is a particular of Monte Carlo Markov Chain (MCMC) algorithm (Betancourt, 2018). Specifically, we use the NUTS implementation in NumPyro (Phan, Pradhan, and Jankowiak, 2019), a Python package which uses JAX (Bradbury et al., 2018) under the hood for automatic differentiation and JIT compilation. We treat each leaf node of the LCA as being independent and generate 1000 MCMC samples after an initial 500 samples as warm up.

Since running the full Bayesian model is computationally expensive, we also allow for a faster, approximate method by just fitting the maximum a posteriori probability (MAP) estimate. We use iMinuit (Dembinski et al., 2021) for the MAP optimization with Numba acceleration (Lam, Pitrou, and Seibert, 2015) for even faster run times. On a Macbook M1 Pro model from 2021, the timings for running the full Bayesian model is $1.41 \pm 0.04\,\mathrm{s/fit}$ and for the MAP it is $4.34 \pm 0.07\,\mathrm{ms/fit}$, showing more than a 2 order increase in performance (around 300x) for the approximate model. Both models allow for easy parallelisation to decrease the computation time.

2.3 | Visualisation

We provide an interactive dashboard to properly visualise the results from the modelling phase, see https://metadmg.onrender.com/ for an example. The dashboard allows for filtering, styling and variable selection, visualizing the mismatch matrix related to a specific leaf node, and exporting of both fit results and plots. By filtering, we include both filtering by sample, by specific cuts in the fit results (e.g. requiring D to be above a certain threshold), and even by taxonomic level (e.g. only looking tax IDs that are part of the Mammalia class). We greatly believe that a visual overview of the fit results increase understanding of the data at hand. The dashboard is implemented with Plotly plots and incorporated into a Dash dashboard (Plotly, 2015).

3 | SIMULATION STUDY

We conducted two sets of simulations, one to gauge the performance of the damage model itself and one to determine the performance of the full `metaDMG` pipeline, i.e. both LCA and damage model.

3.1 | Single-genome Simulations

The first set of simulations was performed by taking a single, representative genome and adding deamination and sequencing noise to it followed by a mapping step and finally damage estimation using `metaDMG`. The deamination was applied using NGSNGS (XXX, ref here) which is a recent implementation of the original Briggs model similar to Gargammel (Neukamm, Peltzer, and Nieselt, n.d.) but with better performance and more accurate deamination patterns. In this step we vary the simulated amount of damage added (in particular the single-stranded DNA deamination, δ_{SS} in the original Briggs model (Briggs et al., 2007)), the number of reads, and the fragment length distribution.

We chose five different, representative genomes, in each of these varying the three simulation parameters. These genomes where the homo sapiens, the betula, and three microbial organisms with respectively low, median, and high amount of GC-content. For each of these simulations, we performed 100 independent runs to measure the variability of the parameter estimations and quantify the robustness of the estimates. We simulated eight different sets of damage (approximately 0%, 1%, 2%, 5%, 10%, 15%, 20%, 30%), 13 sets of different number of reads (10, 25, 50, 100, 250, 500, 1.000, 2.500, 5.000, 10.000, 25.000, 50.000, 100.000), three sets of different fragment length distributions (samples from a lognormal distributions with mean 35, 60, and 90, each with a standard deviation of 10), and five different genomes, each simulation set repeated 100 times.

In addition to this, we also create 1000 repetitions of the non-damaged simulations for Homo Sapiens to be able to gauge the risk of finding false positives. Finally, to show that the damage estimates that `metaDMG` provides are independent of the contig size, we artificially create three different genomes by sampling 1.000, 10.000 or 100.000 different basepairs from a uniform categorical distribution of $\{A, C, G, T\}$.

To be able to compare our estimates to a known value, we generate 1.000.000 reads from NGSNGS without any added sequencing noise for each of other sets of simulation parameters.

The difference in damage frequency at position 1 and 15 is then the value to compare to:

$$D_{\text{known}} = \frac{f(x=1) - f(x=15)}{2} + \frac{f(x=-1) - f(x=-15)}{2}, \tag{9}$$

where we take the average of the C to T damage frequency difference and the G to A damage frequency difference.

The fastq files were simulated with NGSNGS using the above mentioned simulation parameters, all with the same quality scores profiles as used in ART (Huang et al., 2012), based on the Illumina HiSeq 2500 (150 bp). The mapping was performed using Bowtie-2 with the `--no-unal` flag (Langmead and Salzberg, 2012).

3.2 | Metagenomic Simulations

While the previously mentioned simulation study is perfectly aimed at quantifying the performance of the damage model in the case of single-reference genomics it does lack the complexity related to metagenomic samples. Therefore, we also conduct a more advanced simulation study to determine the accuracy of the full `metaDMG` pipeline.

The previously mentioned simulation study quantifies the damage model's performance for single-reference genomics, but it does not address the complexity of metagenomic samples. Therefore, we also conducted a more advanced simulation study to determine the accuracy of the full `metaDMG` pipeline. Based on an ancient metagenome, we created a synthetic dataset that reproduces the composition, fragment length distribution, and damage patterns for each genome. We selected X metagenomes (Supp table XXX) covering several environmental conditions and ages. First, we mapped the reads of each metagenome with bowtie2 against a database that contained the GTDB r202 (Parks et al., 2018) species cluster reference sequences, all organelles from NCBI RefSeq (NCBI Resource Coordinators, 2018), and the reference sequences from CheckV (Nayfach et al., 2021). We used bam-filter 1.0.11 with the flag `--read-length-freqs` to get the mapped read length distribution for each genome and their abundance. The genomes with an observed-to-expected coverage ratio greater than 0.75 were kept. The filtered BAM files were processed by `metaDMG` to obtain the misincorporation matrices. The abundance tables, fragment length distribution, and misincorporation matrices were used in aMGSIM-smk v0.0.1 (Fernandez-Guerra, 2022), a Snakemake workflow (Mölder et al., 2021) that facilitates the generation of many synthetic ancient metagenomes. The data used and generated by the workflow can be obtained from Figshare link

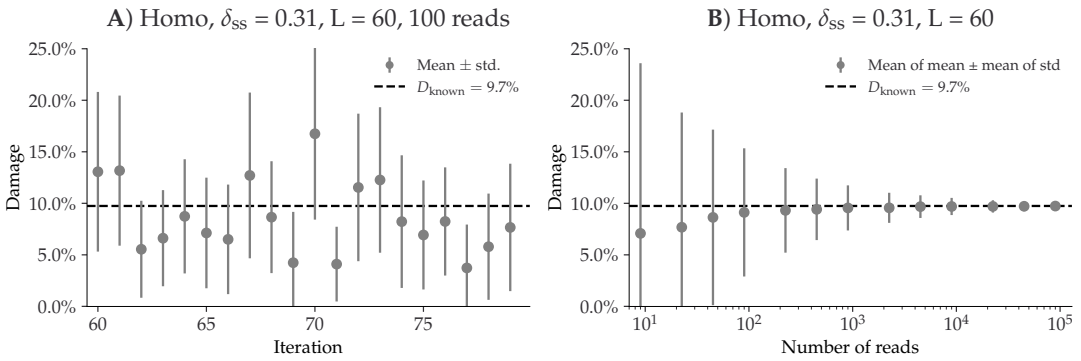


Figure 2. Overview of the single-genome simulations based on the homo sapiens genome with the Briggs parameter $\delta_{ss} = 0.065$ and a fragment length distribution with mean 60. **A**) This plot shows the estimated damage (D) of 10 simulations with 100 simulated reads. The grey points shows the mean damage (with its standard deviation as errorbars). The known damage (D_{known}) is shown as a dashed line, see eq. (9). **B**) This plot shows the average damage as a function of the number of reads. The grey points show the average of the individual means (with the average of the standard deviations as errors).

(XXX). We then performed taxonomic profiling using the same parameters used for the synthetic reads generated by aMGSIM-smk.

4 | RESULTS

The accuracy of all methods in `metaDMG` was tested in various simulation scenarios. In general we find that `metaDMG` yields accurate, precise damage estimates even in extreme low-coverage data.

4.1 | Single-genome Simulations

The results of the single-genome simulations can be seen in Figure 2. The left part of the figure shows `metaDMG` damage estimates based on the homo sapiens genome with the Briggs parameter $\delta_{ss} = 0.31$ and a fragment length distribution with mean 60, each of the 10 simulations generated with 100 simulated reads for 10 representative simulations. When the damage estimates are low, the distribution of D is highly skewed (restricted to positive values) leading to errorbars sometimes going into negative damage, which of course represents un-physical values. The right hand side of the figure visualizes the average amount of damage across a varying number of reads. This shows that the damage estimates converge to the known value with more data, and that one needs more than 100 reads to even get stricly positive damage estimates (when including uncertainties).

Across more than 5 different species, 3 different fragment length distributions, and 3 different

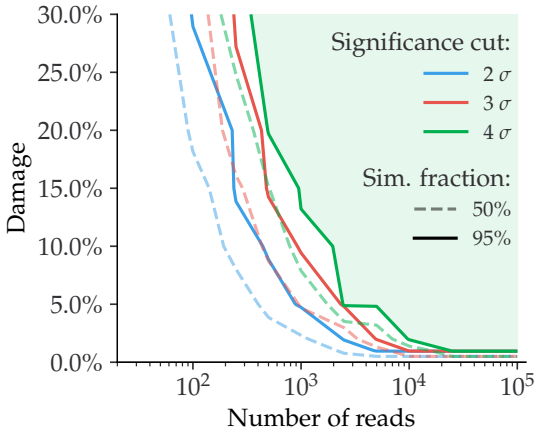


Figure 3. Given a specific significance cut, the solid contour line shows the relationship between the amount of damage and the number of reads required to be able to correctly infer damage in 95% of the species. The dashed line shows the similar value for a simulation fraction of 50%. The green part of the figure shows the "good" region of number of reads and estimated damage, given than one wants to be more than 95% certain of correctly identifying damage with more than 4σ confidence.

contig length distributions, each with 100 simulations for 104 different sets of simulation parameters, the only difference we note in the damage estimates is between species with low, median, and high GC-levels. In general, species with higher GC-levels exhibit lower variations in their damage estimates compared to species with lower GC-levels, leading to high-GC species requiring fewer reads to establish damage estimates.

Based on the single-genome simulations, we can compute the relationship between the amount of damage in a species and the number of reads required to correctly infer that the given species is damaged, see Figure 3. If we want to find damage with a significance of more than 2 (solid blue line) in a sample with around 5% expected damage, it requires about 1000 reads to be 95% certain that we will find results this good. Said in other words: given 100 different samples, each with 1000 reads and around 5% damage, one would expect to find damage (with a $Z > 2$) in 95 of the total 100 samples, on average. If we loose the requirement such that it is okay to only find it in every second sample, it would be enough with only around 250 reads in each sample (dashed blue line).

Finally, to quantify the risk of incorrectly assigning damage to a non-damaged species, we created 1000 independent simulations for a varying number of reads, where none of them had any artificial ancient damage applied, only sequencing noise. Figure 4 shows the damage (D) as a function of the significance (Z) for the case of 1000 simulated reads. Even though the estimated damage is larger than zero, the damage is non-significant since the significance is less than one. When looking

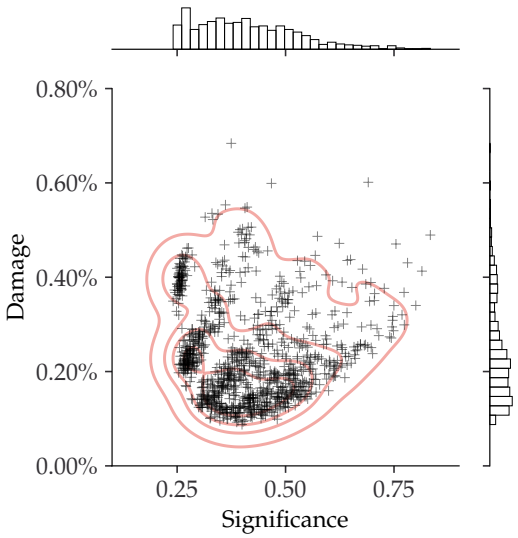


Figure 4. This figure shows the inferred damage estimates of 1000 independent simulations, each with 1000 reads and no artificial ancient damage applied, with the inferred damage shown on the y-axis and the significance on the x-axis. Each simulation is shown as a single cross and the red lines show the kernel density estimate (KDE) of the damage estimates. The marginal distributions are shown as histograms next to the scatter plot.

at all the figures across the different number of reads, see Figure `bayesian_zero_damage_plots.pdf`, we note that a loose cut requiring that $D > 1\%$ and $Z > 2$ would filter out all of non-damaged points.

4.2 | Metagenomic Simulations

With the full metagenomic simulation pipeline we can further probe the performance of `metaDMG`. By looking at the six different samples at different steps in the pipeline we are able to show that `metaDMG` provides relevant, accurate damage estimates. First of all, we run `metaDMG` on the six samples after fragmentation with FragSim. Since no deamination has yet been added at this step in the pipeline, this is also a test of the risk of getting false positives. The results can be seen in Figure 5 where we see the damage estimates for both the species that we simulate to be ancient and the species that we do not add deamination to. We see that the damage estimates are quite similar, as expected, and that our previously established loose cut of $D > 1\%$ and $Z > 2$ still filters out all of non-damaged points.

In comparison we can look at Figure 6 which shows the same plot, but after the deamination (deamSim) and sequencing errors (ART) has been added. Here we see a clear difference between

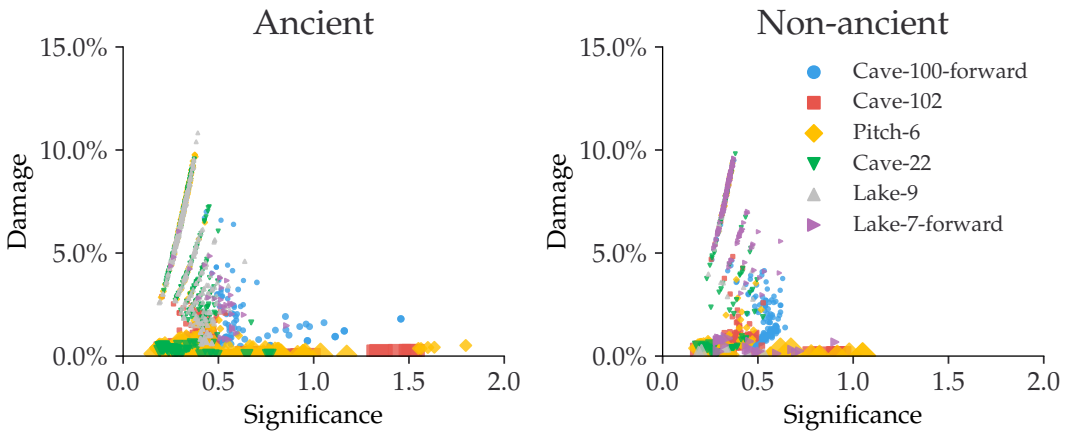


Figure 5. Estimated amount of damage as a function of significance using the fragSim data. The left figure shows the damage of the species that we simulated to be ancient (however with no deamination added yet) and the right figure shows the same for the species that are not going to have deamination added.

the ancient and the non-ancient ones, as expected. The non-ancient species would still not pass the loose cut, however, we note that a large number of the ancient samples would. By looking at Figure 6 we see that not all of the samples show similar amount of damage. These observations are summarised in Table 1 where we see that Cave-100-forward, Cave-102, Pitch-6 all have more than 60% of their ancient species labelled as damaged according to the loose cut, Cave-22 (18%) and Lake-7-forward (12%) a bit lower, while Lake-9 (0.5%) does not show any clear signs of damage. However, once we condition on the requirement of having more than 100 reads, the fraction of ancient species correctly identified as ancient increases to more than 90% for most the samples.

To better understand the damage estimates, we can look a them individually. Figure 7 shows the Stenotrophomonas maltophilia species from the Pitch-6 sample. We see that none of the fragmentation-only files were estimated to have damage and that most of the deamination and final files including sequencing errors have damage – at a simulation size of 1 million, the significance of both are $Z \approx 1.9$, so this one of the few fits with more than 100 reads that does not pass the loose cut. Furthermore, we notice that the error bars decrease with simulation size, as expected.

The rest of the metagenomic simulation results are shown in Figure XXX.

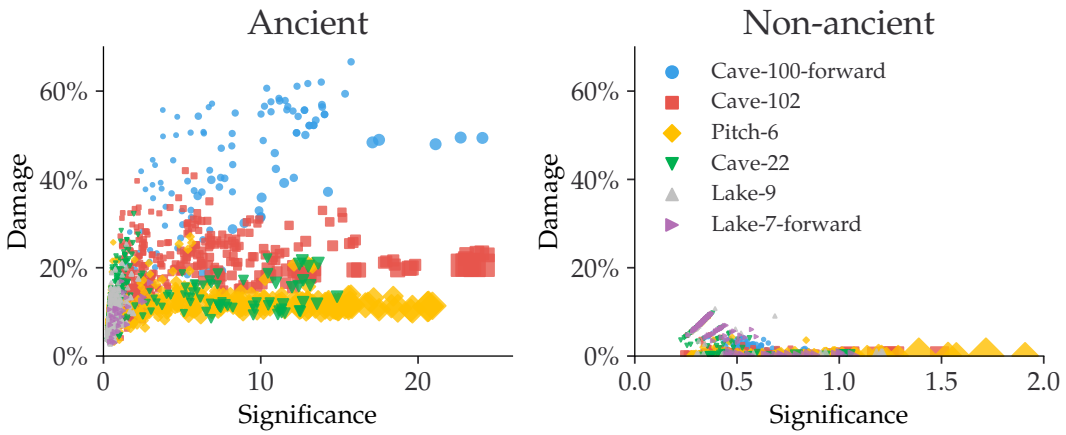


Figure 6. Estimated amount of damage as a function of significance using the ART data. The left figure shows the damage of the species that we simulated to be ancient and the right figure shows the same for the species that have not had deamination added.

Table 1. Number of ancient species for each of the six simulated samples. The first column is the total number of species, the second column is the total number of species that would pass the loose cut of $D > 1\%$ and $Z > 2$, the third column is the number of species with more than 100 reads, and the final column is the number of species with more than 100 reads that also do pass the cut.

Sample	Total	Pass		+100 Reads	+100 Reads and Pass	
Cave-100-forward	135	107	79.3%	88	87	98.9%
Cave-102	500	326	65.2%	309	285	92.2%
Pitch-6	415	260	62.7%	274	260	94.9%
Cave-22	393	71	18.1%	73	69	94.5%
Lake-9	410	2	0.5%	8	0	0%
Lake-7-forward	32	4	12.5%	6	4	66.7%

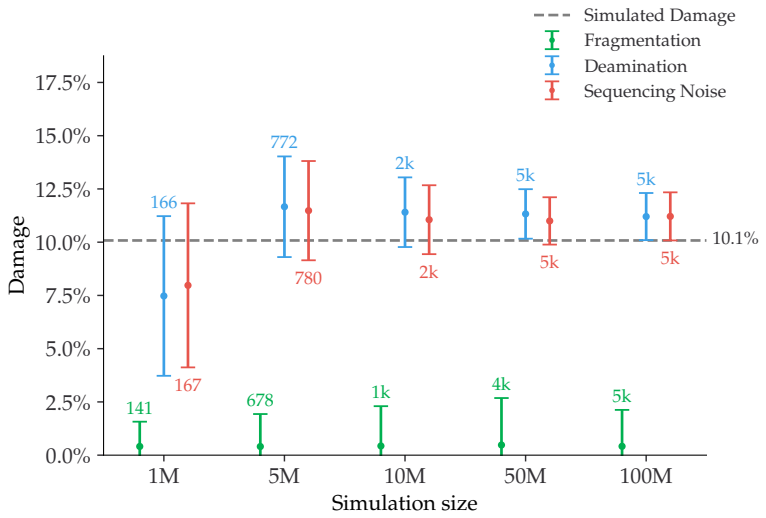


Figure 7. Damage estimates of the Stenotrophomonas maltophilia species from the Pitch-6 sample. Damage is shown as a function of simulation size, with the fragmentation files in green, the deamination files in blue and the final files including sequencing errors in red. All errors are 1σ error bars (standard deviation). The number of reads for each fit is shown as text and since this was a species simulated to have ancient damage, the simulated amount of damage is shown as a dashed grey line.

4.3 | Real Data

The results from running the full `metaDMG` pipeline on real data can be seen in Figure 8. The figures shows Blablabla, real life data here. We find that the loose cut ($D > 1\%$, $Z > 2$) accepts only one of the fits from the control test Library-0, which would not have been accepted by more conservative cut ($D > 2\%$, $Z > 3$, more than 100 reads).

4.4 | Bayesian vs. MAP

Due to increased computational burden of running the full Bayesian model compared to faster, approximate MAP model, in samples with several thousand species, the MAP model is often the most realistic model to use due to time constrains. In this case, it is of course important to know that the damage estimates are indeed trustworthy. Figure 9 compares the estimated damage between the Bayesian model and the MAP model and the estimated significances for species with $D > 1\%$, $Z > 2$ and more than 100 reads. The figure shows that the vast majority of species map 1:1 between the Bayesian and the MAP model. One should note, though, that the few species with the highest mismatch, all are based on forward-only fits, i.e. with no information from the reverse strand, which thus leads to less data to base the fits on. For the comparison with no cuts, see

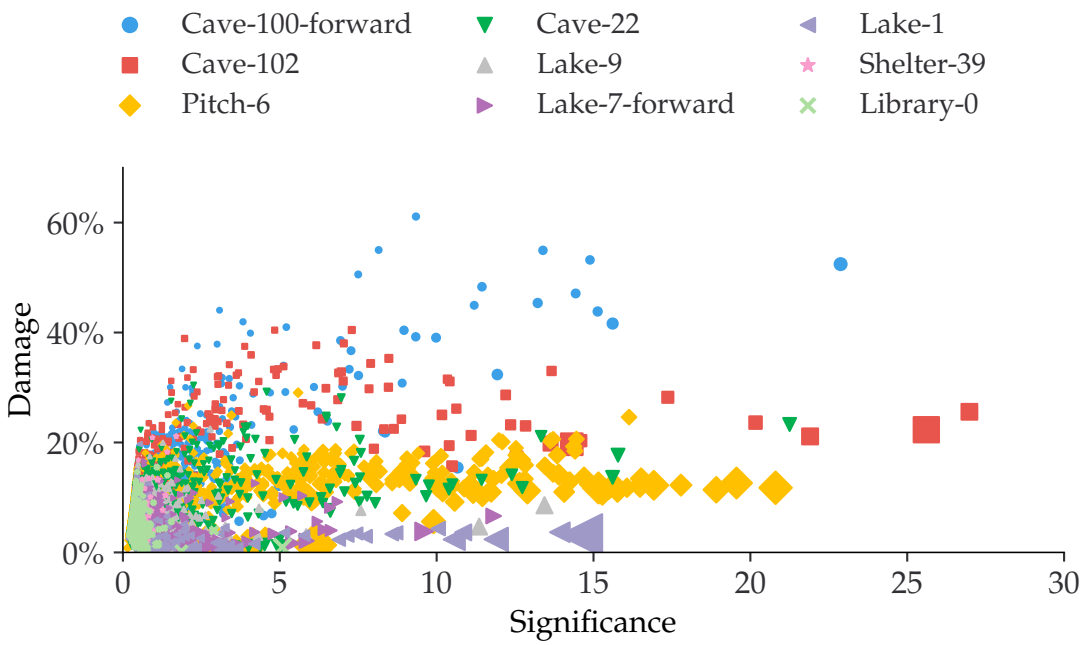


Figure 8. Estimated amount of damage as a function of significance using the real data.

Figure 1 in appendix.

4.5 | Existing Methods

We have also compared `metaDMG` to existing methods such as PyDamage (Borry et al., 2021). Since PyDamage does not include the LCA step, this comparison is based on the non-LCA mode (local-mode) of `metaDMG`. This mode iterates through the different assigned species for all mapped reads and estimates the damage for each. In general, we find that `metaDMG` is more conservative, accurate and precise in its damage estimates.

On example of this is can be found in Figure 10, which shows both the `metaDMG` and PyDamage results of the 100 Homo Sapiens single-genome simulations with 100 reads and 15% added artificial damage (and a fragment length distribution with mean 60).

To compare the computational performance, we use the Pitch-6 sample which has 11.433 unique taxa. When using only a single core, PyDamage took 1105 s to compute all fits, while `metaDMG` took 88 s, a factor of 12.6x faster. Out of the 88 s, `metaDMG` spent 53 s on the actual fits, the rest was for loading and reading the alignment file and computing the mismatch matrices. This makes `metaDMG` more than 20x faster than PyDamage for the fit computation. For the rest of the timings, see Table 2. PyDamage requires the alignment file to be sorted by chromosome position and be supplied with an index file, allowing it to iterate fast through the alignment file, at the expense of computa-

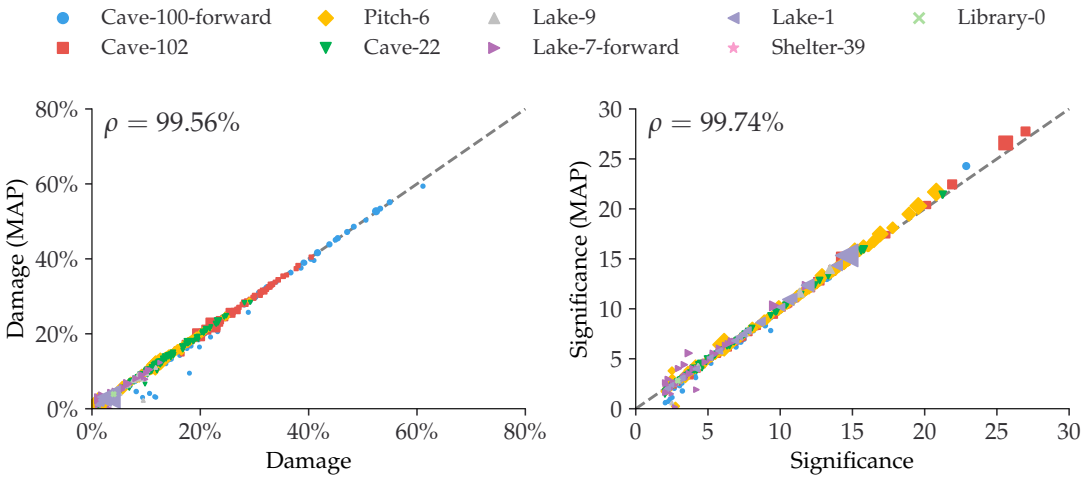


Figure 9. Comparison between the full Bayesian model and the fast, approximate, MAP model for the estimated damage and significance. The figure shows data after a loose cut of $D > 1\%$, $Z > 2$ and more than 100 reads. The dashed, grey line shows the 1:1 ratio and the correlation, ρ, is shown in the upper right corner.

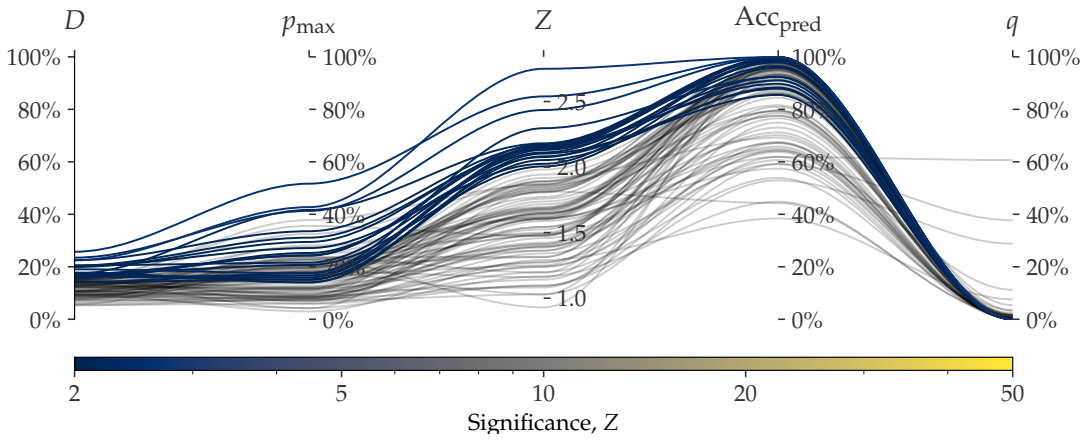


Figure 10. Parallel Coordinates plot comparing `metaDMG` and PyDamage for the Homo Sapiens single-genome simulation with 100 reads and 15% added artificial damage. The different axis shows the five different variables: metaDMG-damage (D, by `metaDMG`), PyDamage-damage ($p_{\max}$, by PyDamage), significance (Z, by `metaDMG`), predicted accuracy (Acc_{pred}, by PyDamage), and the p-value (q, by PyDamage). Each of the 100 simulations are plotted as single lines showing the values of the different dimensions. Simulations with $D > 1\%$ and $Z > 2$, i.e. damaged according to the loose `metaDMG` cut, are shown in color proportional to their significance. Non-damaged simulations are shown in semi-transparent black lines.

Table 2. Computational performance of PyDamage and `metaDMG`. The table contains the times it takes to run either PyDamage or `metaDMG` on the full Pitch-6 sample containing 11.433 species. The timings are shown for both single-processing case (1 core) and multi-processing (2 and 4 cores). The timings were performed on a Macbook M1 Pro model from 2021. "12.6x" means that `metaDMG` was 12.6 times faster than PyDamage for that particular test.

Pitch-6	Pydamage		metaDMG			
Cores	Total	Fits	Total		Fits	
1	1105 s	1102 s	88 s	12.6x	53 s	20.8x
2	592 s	590 s	66 s	9.0x	25 s	23.6x
4	398 s	397 s	54 s	7.4x	14s	28.4x

tional load before running the actual damage estimation. `metaDMG` on the other hand requires the reads to be sorted by name to minimize the time it takes to run the LCA, which however, is not tested in this comparison.

5 | DISCUSSION

Preliminary work indicates that the computational performance of the models can be even further optimized by using Julia (Bezanson et al., 2017), which shows around $7\times$ optimization for the Bayesian model ($\sim 0.2\,\mathrm{s/fit}$) and $4\times$ for the MAP model ($\sim 1.1\,\mathrm{ms/fit}$).

5.1 | Acknowledgment

Acknowledgements here

5.2 | Data availability

Source code is hosted at GitHub: https://github.com/metaDMG-dev. Sequencing data can be found at: https://somewhere.com XXX.

REFERENCES

Betancourt, Michael (2018). "A Conceptual Introduction to Hamiltonian Monte Carlo". In: *arXiv:1701.02434 [stat]*. arXiv: 1701.02434.

Bezanson, Jeff et al. (2017). "Julia: A fresh approach to numerical computing". In: *SIAM review* 59.1. Publisher: SIAM, pp. 65–98. URL: https://julialang.org/.

Borry, Maxime et al. (2021). "PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA de novo assembly". en. In: *PeerJ* 9. Publisher: PeerJ Inc., e11845. ISSN: 2167-8359. DOI: 10.7717/peerj.11845. URL: https://peerj.com/articles/11845 (visited on 2022).

Bradbury, James et al. (2018). *JAX: composable transformations of Python NumPy programs*. URL: http://github.com/google/jax.

Briggs, Adrian W. et al. (2007). "Patterns of damage in genomic DNA sequences from a Neandertal". en. In: *Proceedings of the National Academy of Sciences* 104.37. Publisher: National Academy of Sciences Section: Biological Sciences, pp. 14616–14621. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0704665104. URL: https://www.pnas.org/content/104/37/14616.

Cabanski, Christopher R. et al. (2012). "ReQON: a Bioconductor package for recalibrating quality scores from next-generation sequencing data". In: *BMC Bioinformatics* 13.1, p. 221. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-221. URL: https://doi.org/10.1186/1471-2105-13-221 (visited on 2022).

Cappellini, Enrico et al. (2018). "Ancient Biomolecules and Evolutionary Inference". In: *Annual Review of Biochemistry* 87.1. _eprint: https://doi.org/10.1146/annurev-biochem-062917-012002, pp. 1029–1060. DOI: 10.1146/annurev-biochem-062917-012002. URL: https://doi.org/10.1146/annurev-biochem-062917-012002 (visited on 2022).

Cepeda-Cuervo, Edilberto and MARíA VICTORIA Cifuentes-Amado (2017). "Double Generalized Beta-Binomial and Negative Binomial Regression Models". en. In: *Revista Colombiana de Estadística* 40.1. Publisher: Universidad Nacional de Colombia., pp. 141–163. ISSN: 0120-1751. DOI: 10.15446/rce.v40n1.61779.

Dabney, Jesse, Matthias Meyer, and Svante Pääbo (2013). "Ancient DNA Damage". In: *Cold Spring Harbor Perspectives in Biology* 5.7, a012567. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a012567. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3685887/ (visited on 2022).

Dembinski, Hans et al. (2021). *scikit-hep/iminuit: v2.8.2*. DOI: 10.5281/ZENODO.3949207. (Visited on 2021).

Fernandez-Guerra, Antonio (2022). *genomewalker/aMGSIM-smk: v0.0.1*. DOI: 10.5281/zenodo.7298422. URL: https://doi.org/10.5281/zenodo.7298422.

Ginolhac, Aurelien et al. (2011). "mapDamage: testing for damage patterns in ancient DNA sequences". In: *Bioinformatics* 27.15, pp. 2153–2155. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr347. URL: https://doi.org/10.1093/bioinformatics/btr347 (visited on 2022).

Huang, Weichun et al. (2012). "ART: a next-generation sequencing read simulator". eng. In: *Bioinformatics (Oxford, England)* 28.4, pp. 593–594. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr708.

Jónsson, Hákon et al. (2013). "mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters". en. In: *Bioinformatics* 29.13, pp. 1682–1684. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btt193.

Lam, Siu Kwan, Antoine Pitrou, and Stanley Seibert (2015). "Numba: a LLVM-based Python JIT compiler". In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*. LLVM '15. New York, NY, USA: Association for Computing Machinery, pp. 1–6. ISBN: 978-1-4503-4005-2. DOI: 10.1145/2833157.2833162. URL: https://github.com/numba/numba.

Langmead, Ben and Steven L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2". en. In: *Nature Methods* 9.4. Number: 4 Publisher: Nature Publishing Group, pp. 357–359. ISSN: 1548-7105. DOI: 10.1038/nmeth.1923. URL: https://www.nature.com/articles/nmeth.1923 (visited on 2022).

Llamas, Bastien et al. (2017). "From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era". en. In: *STAR: Science & Technology of Archaeological Research* 3.1, pp. 1–14. ISSN: 2054-8923. DOI: 10.1080/20548923.2016.1258824. URL: https://www.tandfonline.com/doi/full/10.1080/20548923.2016.1258824 (visited on 2022).

McElreath, Richard (2020). *Statistical rethinking: a Bayesian course with examples in R and Stan*. 2nd ed. CRC texts in statistical science. Boca Raton: Taylor and Francis, CRC Press. ISBN: 978-0-367-13991-9.

Mölder, Felix et al. (2021). *Sustainable data analysis with Snakemake*. en. Tech. rep. 10:33. Type: article. F1000Research. DOI: 10.12688/f1000research.29032.2. URL: https://f1000research.com/articles/10-33 (visited on 2022).

Nayfach, Stephen et al. (2021). "CheckV assesses the quality and completeness of metagenome-assembled viral genomes". en. In: *Nature Biotechnology* 39.5. Number: 5 Publisher: Nature Publishing Group, pp. 578–585. ISSN: 1546-1696. DOI: 10.1038/s41587-020-00774-7. URL: https://www.nature.com/articles/s41587-020-00774-7 (visited on 2022).

NCBI Resource Coordinators (2018). "Database resources of the National Center for Biotechnology Information". In: *Nucleic Acids Research* 46.D1, pp. D8–D13. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1095. URL: https://doi.org/10.1093/nar/gkx1095 (visited on 2022).

Neukamm, Judith, Alexander Peltzer, and Kay Nieselt (n.d.). "DamageProfiler: Fast damage pattern calculation for ancient DNA". en. In: (), p. 10.

Parks, Donovan H. et al. (2018). "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life". en. In: *Nature Biotechnology* 36.10. Number: 10 Publisher: Nature Publishing Group, pp. 996–1004. ISSN: 1546-1696. DOI: 10.1038/nbt.4229. URL: https://www.nature.com/articles/nbt.4229 (visited on 2022).

Phan, Du, Neeraj Pradhan, and Martin Jankowiak (2019). "Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro". In: *arXiv:1912.11554 [cs, stat]*. arXiv: 1912.11554.

Plotly, Technologies (2015). *Collaborative data science*. Place: Montreal, QC Publisher: Plotly Technologies Inc. URL: https://plot.ly.

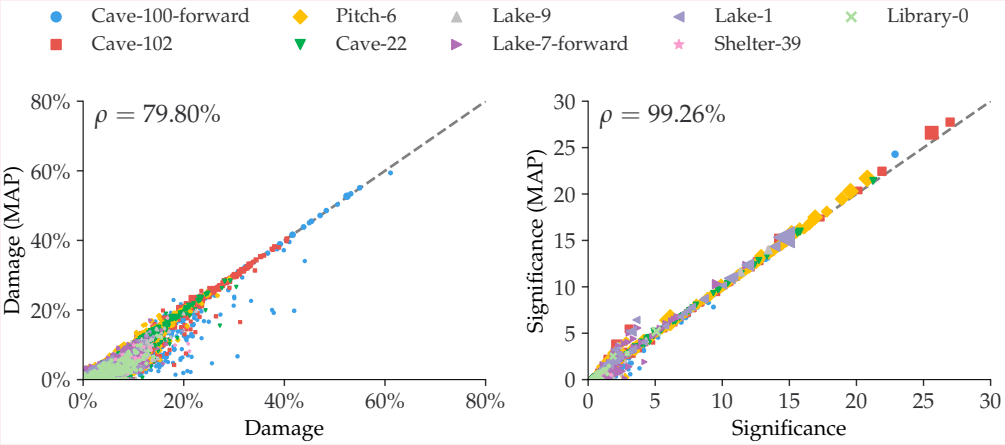
Skoglund, Pontus et al. (2014). "Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal". In: *Proceedings of the National Academy of Sciences* 111.6. Publisher: Proceedings of the National Academy of Sciences, pp. 2229–2234. DOI: 10.1073/pnas.1318934111. URL: https://www.pnas.org/doi/10.1073/pnas.1318934111 (visited on 2022).

Wang, Yucheng et al. (2022). "ngsLCA—A toolkit for fast and flexible lowest common ancestor inference and taxonomic profiling of metagenomic data". en. In: *Methods in Ecology and Evolution* n/a.n/a (). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14006. ISSN: 2041-210X. DOI: 10.1111/2041-210X.14006. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14006 (visited on 2022).

Appendix 1

A | EXAMPLE FIGURE

This is an example of including a figure in the appendix.



Appendix 1—figure 1. Comparison between the full Bayesian model and the fast, approximate, MAP model for the estimated damage and significance. The dashed, grey line shows the 1:1 ratio.

B | EXAMPLE TABLE

This is an example of including a table in the appendix.

Appendix 2—table 1. An example table.

Speed (mph)	Driver	Car	Engine	Date
407.447	Craig Breedlove	Spirit of America	GE J47	8/5/63
413.199	Tom Green	Wingfoot Express	WE J46	10/2/64
434.22	Art Arfons	Green Monster	GE J79	10/5/64
468.719	Craig Breedlove	Spirit of America	GE J79	10/13/64
526.277	Craig Breedlove	Spirit of America	GE J79	10/15/65
536.712	Art Arfons	Green Monster	GE J79	10/27/65
555.127	Craig Breedlove	Spirit of America, Sonic 1	GE J79	11/2/65
576.553	Art Arfons	Green Monster	GE J79	11/7/65
600.601	Craig Breedlove	Spirit of America, Sonic 1	GE J79	11/15/65
622.407	Gary Gabelich	Blue Flame	Rocket	10/23/70
633.468	Richard Noble	Thrust 2	RR RG 146	10/4/83
763.035	Andy Green	Thrust SSC	RR Spey	10/15/97

C | LIKELIHOOD CALCULATION

C.1 | Full multinomial logistic Regression models

Postmortem damages will have impacts on the NGS (next generation sequencing) reads. A common phenomenon is the calling error rates increases from nucleotide C to T due to the cytosine deamination process. Unawareness of this will lead to inaccurate inferences. Evidences show that the magnitude of such changes are related to the positions the site is within a read (the fraction of the ancient DNA). Here we present 3 slightly different ways to unveil the relationship between the calling error rates and the mismatching reference/read pairs as well as the site positions within a read. The methods are based on the multinomial logistic regressions.

Data Description

We perform the regression based on the summary statistic of the mismatch matrix which is a table which contains the counts of reads of different reference/read categories (in total 16) and positions on the forward/reversed strand (15 positions on each direction). Table 1 and Table 2 give an example of the data format we use for the inference.

Read Counts								
Ref.	A				C			
Read	A	C	G	T	A	C	G	T
1	12794053	8325	28769	16073	10404	8045811	8020	2092619
2	13480290	6812	21107	12102	9151	8260185	6531	1145605
3	12760253	6131	18859	10327	7772	8385423	5899	914709
4	12995572	5240	17671	8940	7880	8345892	5252	767237
5	12930102	4601	17021	8188	8374	8474964	5161	703283
6	12879355	4684	16435	7536	8726	8571141	4811	643607
7	12684349	4557	15298	7394	8835	8727254	4762	586674
8	12585563	4454	15497	7236	8898	8888173	5058	527691
9	12468622	4309	14704	6942	8948	9076851	4673	481170
10	12491183	4437	14567	6912	9103	9237982	4702	443329
11	12430899	4296	14083	6515	9313	9364121	4609	404431
12	12419506	4226	13985	6503	9342	9357468	4367	371475
13	12469412	4147	13851	6375	9586	9386737	4588	345390
14	12549936	4045	13650	6246	9673	9324488	4628	322294
15	12566555	4174	13499	6213	9735	9305820	4518	301360
-1	11599167	8800	16164	14851	90888	9613102	10843	19810
-2	11985637	8769	14044	12040	28799	9561124	7184	18424
-3	12941743	7805	13861	12001	24988	9400151	6368	15466
-4	12808985	7141	12885	9889	23067	9509723	5421	14901
-5	12869585	6954	12100	9428	22349	9464831	5789	13987
-6	12784911	6440	12080	8735	20556	9566794	6544	14021
-7	12878349	5946	12311	8225	19480	9566359	6478	16419
-8	12719722	9521	12156	8131	19226	9725468	6709	23434
-9	12652860	5634	11940	7671	18035	9762224	6321	31667
-10	12566817	5448	11850	7178	17353	9701382	6306	37831
-11	12702498	5309	12092	7568	16121	9526031	6035	43215
-12	12731940	5207	11933	6856	15637	9533858	5557	47650
-13	12697647	4989	12199	7153	15072	9508117	5434	51614
-14	12689924	4944	11891	6816	15050	9525285	5237	55598
-15	12660634	4746	11753	6732	14815	9561359	5184	59633

Appendix 3—table 1. The read counts per position given the reference nucleotides are A or C of an ancient human data. The negative position indices are the position on the reversed strand. In the manuscript, the elements (the values of a specific nucleotide read counts per position given the reference nucleotide is A or C) in this table are denoted as $o_{\mathrm{A}\to i,p}$ or $o_{\mathrm{C}\to i,p}$.

Read Counts								
Ref.	G				T			
Read	A	C	G	T	A	C	G	T
1	16389	8976	9639767	86584	11733	15878	8351	11718463
2	17614	6483	9510149	26655	10761	13958	7011	11974947
3	15164	5949	9488917	23374	9509	13767	6046	12839015
4	14844	5186	9566468	21960	8170	12509	5585	12721790
5	14005	5612	9497118	20468	7186	11991	5233	12795244
6	13671	6195	9622572	19096	6948	11683	4790	12686645
7	16648	6394	9609855	18594	6203	12122	4780	12794172
8	23659	6405	9768666	17341	6131	11847	4758	12626614
9	31680	6139	9785449	17034	5998	12040	4469	12579260
10	38484	5982	9700857	16235	5487	11546	4175	12513653
11	44665	5722	9536341	15284	5651	12044	4176	12646627
12	48949	5371	9547134	14569	5449	11663	4060	12684645
13	53076	5234	9543953	14090	5262	11785	4046	12631297
14	57343	5186	9551477	13855	5257	11768	4006	12624840
15	61236	5137	9583481	13667	5122	11733	3947	12612416
-1	2078554	7947	8096447	11847	15732	28461	8551	12890628
-2	1138478	6656	8232666	10760	12299	20759	6999	13446882
-3	921712	5970	8399013	8643	10514	18226	6564	12718084
-4	775038	5720	8319235	8416	9415	17800	5388	12977322
-5	710955	5499	8462058	8926	8526	17088	4911	12886576
-6	647761	5052	8545455	9193	7640	16351	4879	12852322
-7	593854	4872	8693834	9318	7600	15523	5048	12664576
-8	535542	7828	8889921	9399	7163	18704	4718	12510123
-9	486549	4696	9075263	9522	7109	14547	4611	12409220
-10	448895	4622	9226758	9432	6816	14567	4668	12438344
-11	409027	4654	9352528	9544	6575	14019	4611	12388650
-12	376069	4637	9344701	9419	6511	13874	4486	12390148
-13	350609	4655	9384853	9885	6197	13877	4327	12432024
-14	326760	4595	9337266	9889	5986	13928	4403	12490990
-15	305014	4541	9310617	10065	5919	13442	4232	12529684

Appendix 3—table 2. The read counts per position given the reference nucleotides are G or T of the same human data as in Table 1. The negative position indices are the position on the reversed strand. In the manuscript, the elements (the values of a specific nucleotide read counts per position given the reference nucleotide is G or T) in this table are denoted as $o_{\mathrm{G}\to i,p}$ or $o_{\mathrm{T}\to i,p}$.

The terminology used here might not be standard. The term full regression here is to distinguish itself from the folded regression discussed later, which simply means inferring the coefficients of forward strand and reversed strand separately. Full regression includes both unconditional regression and conditional regression. The unconditional regression's objective is to infer the probability of observing a read of nucleotide j and its reference is i at position p, i.e., $\mathrm{P}(\mathrm{Obs}:i\to j|\mathrm{Pos}:p)$; while the unconditional regression's target is to estimate the probability of observing a read of nucleotide j given its reference is i at position p, i.e., $\mathrm{P}(\mathrm{Obs}:j|\mathrm{Ref}:i,\mathrm{Pos}:p)$. Their relationship is as follows:

$$\mathrm{P}(\mathrm{Obs}:j|\mathrm{Ref}:i,\mathrm{Pos}:p)=\frac{\mathrm{P}(\mathrm{Obs}:i\to j|\mathrm{Pos}:p)}{\sum_j \mathrm{P}(\mathrm{Obs}:i\to j|\mathrm{Pos}:p)}.$$

So in fact, unconditional regression can give us more detailed inferred results (extra information the nucleotide composition per position of the reference, which may be related to the prepared libraries).

Unconditional Regression likelihood

$$\begin{aligned} l_1 &= \sum_p \sum_{i,j\in\{\mathrm{A,C,G,T}\}} o_{i\to j,p}\log \mathrm{P}_{i,j|p} \\ &= \sum_p \left[o_p \log \mathrm{P}_{\mathrm{TT}|p} + \sum_{(i,j)\neq \mathrm{TT}} o_{i\to j,p}\log\frac{\mathrm{P}_{ij|p}}{\mathrm{P}_{\mathrm{TT}|p}} \right], \end{aligned} \tag{10}$$

where $\mathrm{P}_{ij|p}=\mathrm{P}(\mathrm{Obs}:i\to j|\mathrm{Pos}:p)$, and $o_p=\sum_{i,j\in\{\mathrm{A,C,G,T}\}} o_{i\to j,p}$.

$$\log\frac{\mathrm{P}_{ij|p}}{\mathrm{P}_{\mathrm{TT}|p}}=\sum_{n=0}^{\mathrm{order}}\alpha_{i,j,p,n}p^n \tag{11}$$

$$\begin{aligned} l_1 &= \sum_p \left\{ -o_p \log\left[1+\sum_{(i,j)\neq\mathrm{TT}}\exp\left(\sum_{n=0}^{\mathrm{order}}\alpha_{i,j,p,n}p^n\right)\right] + \sum_{(i,j)\neq\mathrm{TT}} o_{i\to j,p}\sum_{n=0}^{\mathrm{order}}\alpha_{i,j,p,n}p^n \right\} \\ &= l_{1,5'}+l_{1,3'}. \end{aligned} \tag{12}$$

The number of inferred parameters for the full conditional regression is $30\,(\text{order}+1)$.

$$\frac{\partial l_1}{\partial \alpha_{i,j,p,n}} = -o_p \frac{p^n \exp\left(\sum_{n=0}^{\text{order}} \alpha_{i,j,p,n} p^n\right)}{1 + \sum_{(i,j)\neq \text{TT}} \exp\left(\sum_{n=0}^{\text{order}} \alpha_{i,j,p,n} p^n\right)} + o_{i\to j,p} p^n. \tag{13}$$

Conditional Regression likelihood

$$\begin{aligned} l_2 &= \sum_{i\in\{\text{A,C,G,T}\}} \sum_p \sum_{j\in\{\text{A,C,G,T}\}} o_{i\to j,p} \log \text{P}_{j|p,i} \\ &= \sum_{i\in\{\text{A,C,G,T}\}} \sum_p \left[o_{i,p} \log \text{P}_{\text{T}|p,i} + \sum_{j\neq \text{T}} o_{i\to j,p} \log \frac{\text{P}_{j|p,i}}{\text{P}_{\text{T}|p,i}} \right], \end{aligned} \tag{14}$$

where $\text{P}_{j|p,i} = \text{P}(\text{Obs}: j|\text{Ref}: i, \text{Pos}: p)$, and $o_{i,p} = \sum_{j\in\{\text{A,C,G,T}\}} o_{i\to j,p}$.

$$\log \frac{\text{P}_{j|p,i}}{\text{P}_{\text{T}|p,i}} = \sum_{n=0}^{\text{order}} \beta_{i,j,p,n} p^n \tag{15}$$

$$\begin{aligned} l_2 &= \sum_{i\in\{\text{A,C,G,T}\}} \sum_p \left\{ -o_{i,p} \log\left[1 + \sum_{j\neq \text{T}} \exp\left(\sum_{n=0}^{\text{order}} \beta_{i,j,p,n} p^n\right)\right] + \sum_{j\neq \text{T}} o_{i\to j,p} \sum_{n=0}^{\text{order}} \beta_{i,j,p,n} p^n \right\} \\ &= l_{2,\text{A},5'} + l_{2,\text{C},5'} + l_{2,\text{G},5'} + l_{2,\text{T},5'} + l_{2,\text{A},3'} + l_{2,\text{C},3'} + l_{2,\text{G},3'} + l_{2,\text{T},3'}. \end{aligned} \tag{16}$$

The number of inferred parameters for the full unconditional regression is $24\,(\text{order}+1)$.

$$\frac{\partial l_2}{\partial \beta_{i,j,p,n}} = -o_{i,p} \frac{p^n \exp\left(\sum_{n=0}^{\text{order}} \beta_{i,j,p,n} p^n\right)}{1 + \sum_{j\neq \text{T}} \exp\left(\sum_{n=0}^{\text{order}} \beta_{i,j,p,n} p^n\right)} + o_{i\to j,p} p^n. \tag{17}$$

Folded Regression

The folded regressions use the same log-likelihood function as the full regression (i.e., Equation) but are conducted based on the assumptions that,

$$\alpha_{i,j,p,n} = \alpha_{c(i),c(j),-p,n}, \tag{18}$$

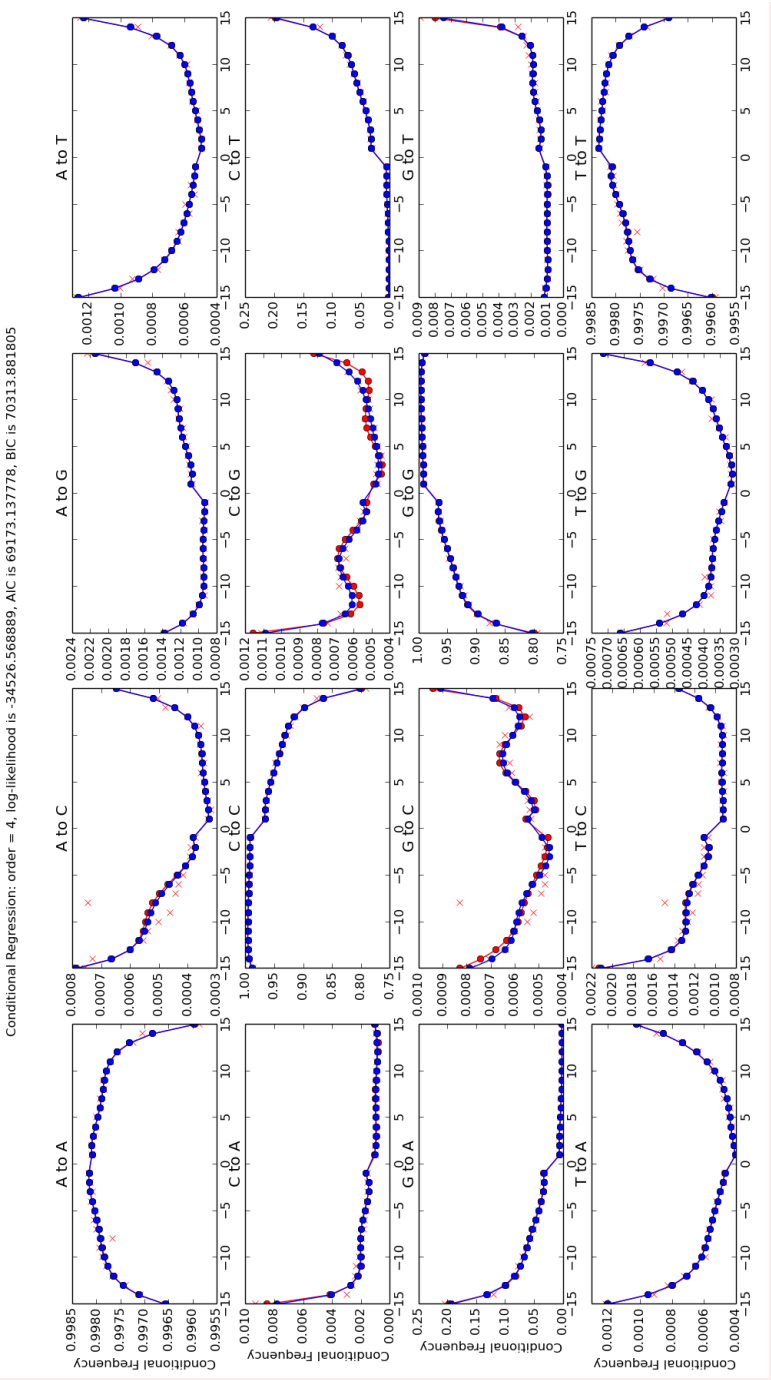
$$\beta_{i,j,p,n} = \beta_{c(i),c(j),-p,n}, \tag{19}$$

where $c(i)$ means the complimentary nucleotide of the nucleotide i, e.g., $c(\text{A}) = \text{T}$ and $c(\text{G}) = \text{C}$. Our data (both taxon and human data) and some models studies seem to support this assumption.

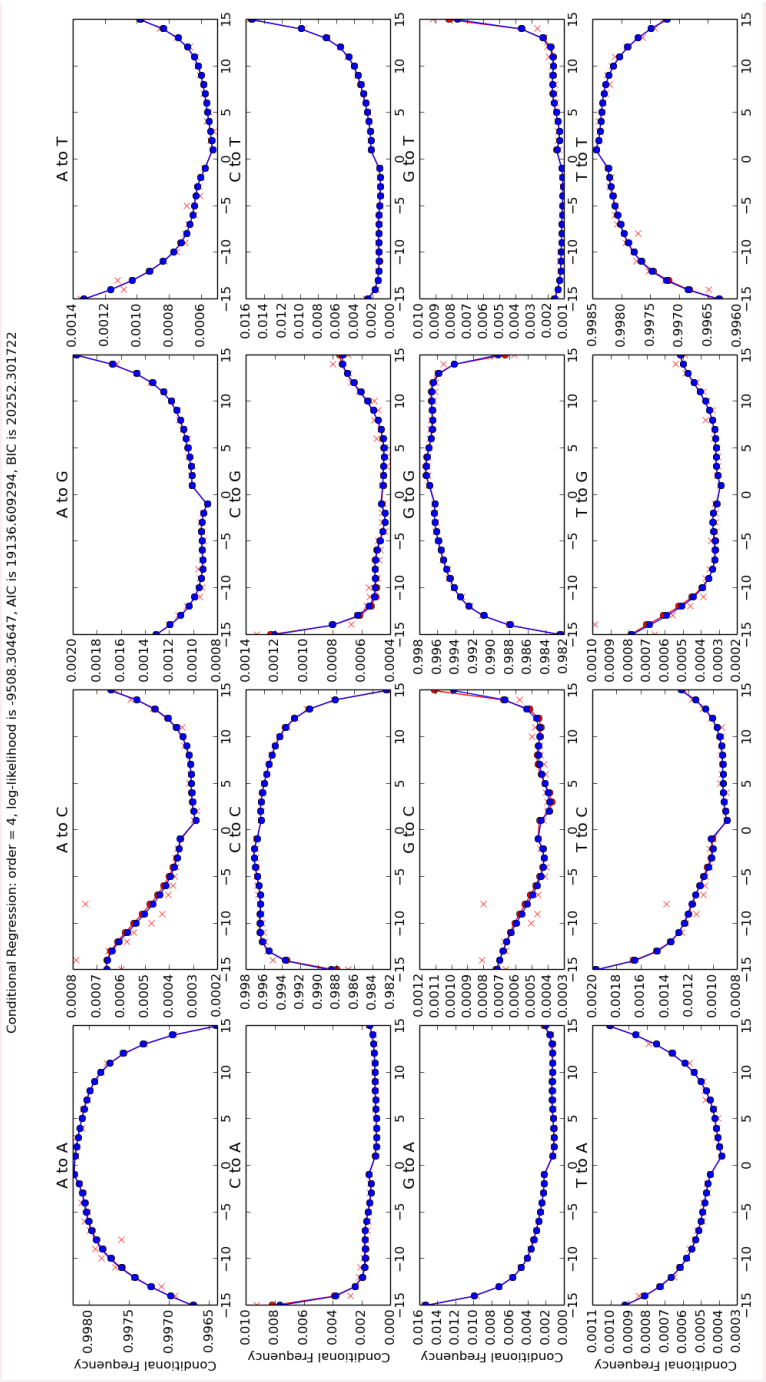
By doing the folded regression, we halve the number of inferred parameters. Hence The number of inferred parameters for the folded unconditional regression is $15\,(\text{order}+1)$, and that of folded conditional regression is $12\,(\text{order}+1)$.

Results for multinomial logistic regression

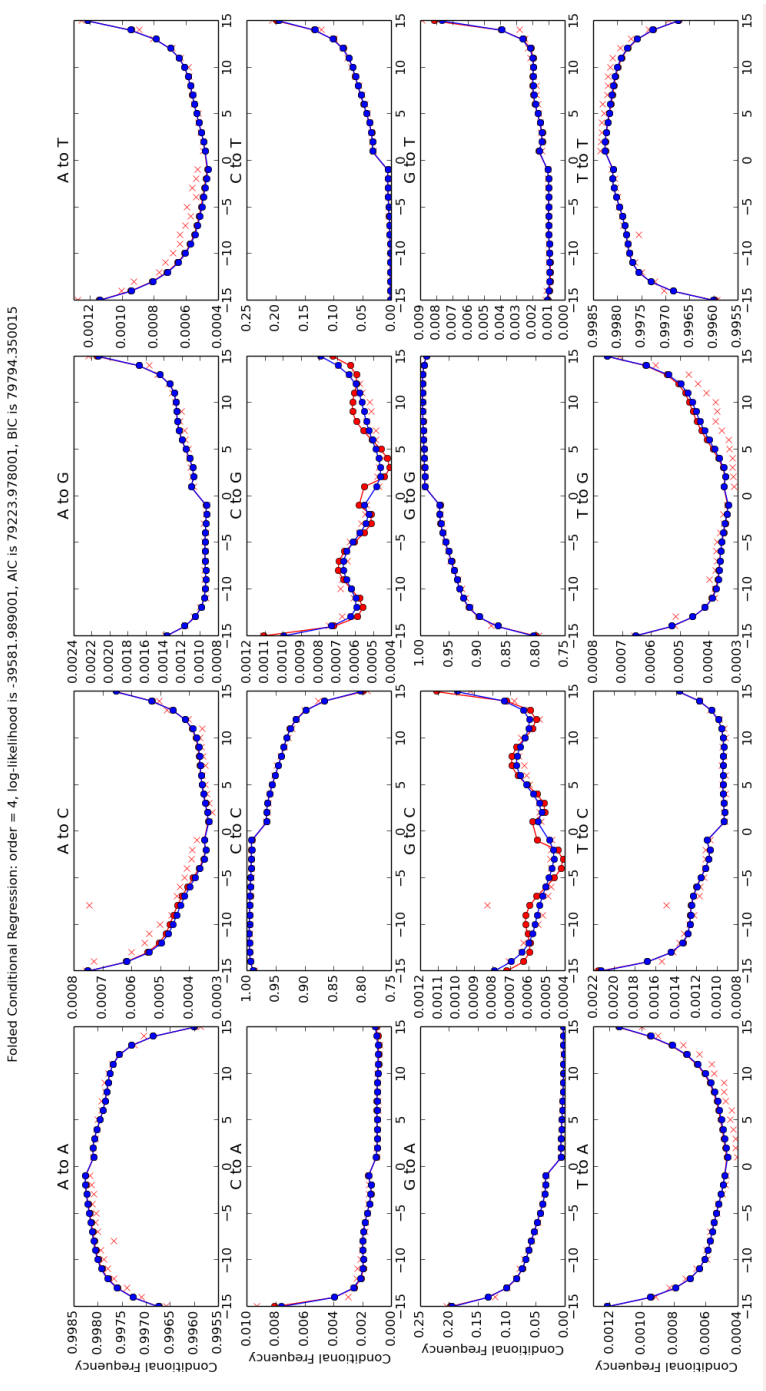
Currently, the optimization of the likelihood functions are based on the C++ library of gsl and use the function *gsl_multimin_fminimizer_nmsimplex2*. with the initial searching point is set to be the results of logistic regression. We here present here 4 figures pertaining to showcase the performance of our model. The regression methods are based on the summary statistic of the counts of mismatches and the optimization is therefore in the scale of miliseconds. Fig. 1 and Fig. 2 are the conditional regression results of the ancient and control human data correspondingly. And Fig. 3 and Fig. 4 are the folded conditional regression results of the same data as above. Our codes can also do the unconditional regression, but I have not generated the results for now.



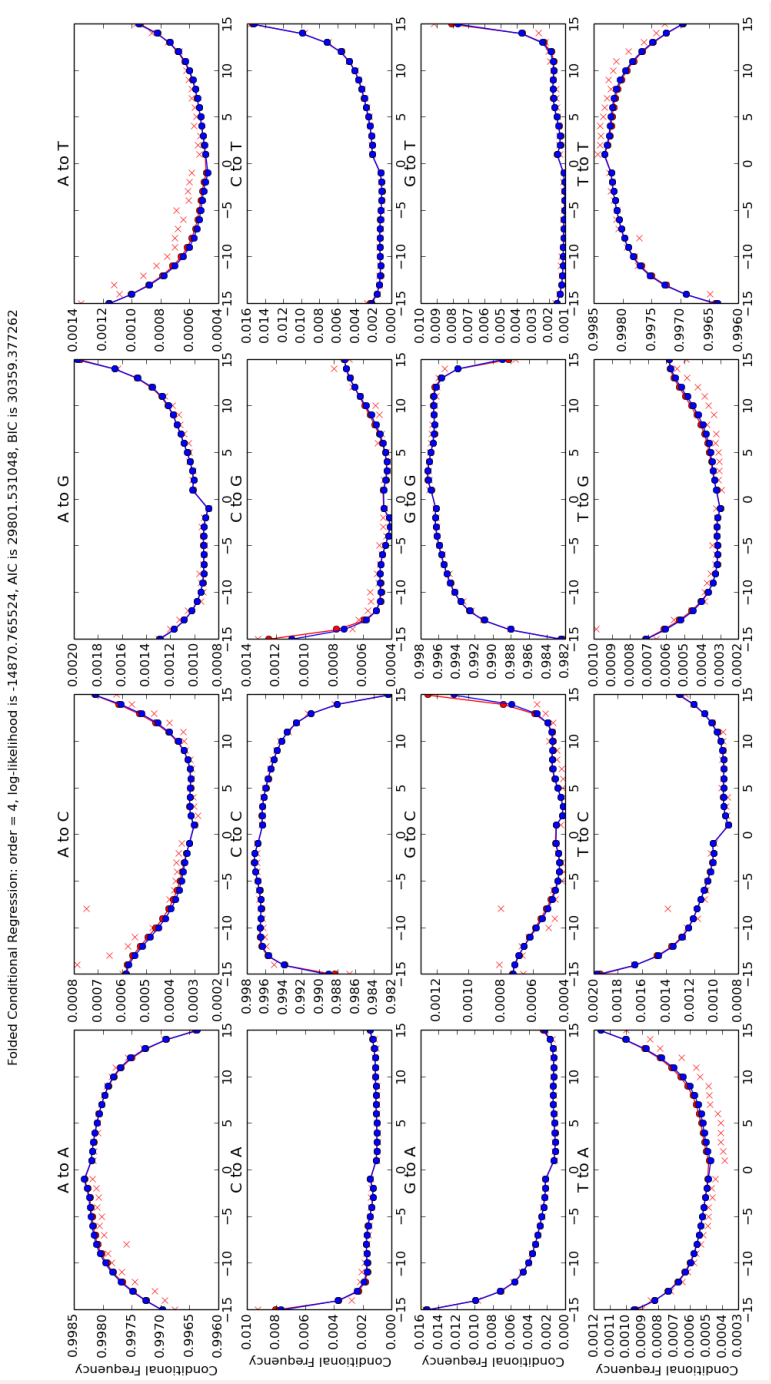
Appendix 3—figure 1. Conditional regression results with the order 4 of the ancient human data. Each panel of figure represents a specific reference/read pair and plots its frequency across different positions. The positions from left to right are −1 to −15 and 15 to 1.



Appendix 3—figure 2. Conditional regression results with the order 4 of the control human data. Each panel of figure represents a specific reference/read pair and plots its frequency across different positions. The positions from left to right are −1 to −15 and 15 to 1.



Appendix 3—figure 3. Folded conditional regression results with the order 4 of the ancient human data. Each panel of figure represents a specific reference/read pair and plots its frequency across different positions. The positions from left to right are −1 to −15 and 15 to 1.



Appendix 3—figure 4. Folded conditional regression results with the order 4 of the control human data. Each panel of figure represents a specific reference/read pair and plots its frequency across different positions. The positions from left to right are −1 to −15 and 15 to 1.

3 *Paper II*

The following pages contain the article:

Christian Michelsen, Christoffer C. Jorgensen, Mathias Heltberg, Mogens H. Jensen, Alessandra Lucchetti, Pelle B. Petersen, Troels C. Petersen, Henrik Kehlet (2022). “Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty - a machine learning based approach.”.

Anesthesiology

Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty - a machine learning based approach.

--Manuscript Draft--

Manuscript Number:	
Full Title:	Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty - a machine learning based approach.
Short Title:	Machine-learning models in joint replacement
Article Type:	Original Investigation: Perioperative Medicine
Section/Category:	
Corresponding Author:	Christoffer Calov Jorgensen, M.D. Rigshospitalet Copenhagen, DENMARK
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Rigshospitalet
Corresponding Author's Secondary Institution:	
First Author:	Christian Michelsen, M.Sci
First Author Secondary Information:	
Order of Authors:	Christian Michelsen, M.Sci
	Christoffer Calov Jorgensen, M.D.
	Mathias Heltberg, M.Sci
	Mogens H. Jensen, D.Sci
	Alessandra Lucchetti, M.Sci
	Pelle Baggesgaard Petersen, M.D., Ph.D
	Troels Christian Petersen, M.Sci
	Henrik Kehlet, M.D., Ph.D
Order of Authors Secondary Information:	
Suggested Reviewers:	Stavros Memtsoudis, M.D. Director of Critical Care Services, HSS: Hospital for Special Surgery memtsoudiss@hss.edu Dr. Memtsoudis is an international expert on perioperative care in hip and knee arthroplasty.
	Asokumar Buvanendran, M.D. Rush Medical College of Rush University: Rush University Rush Medical College asokumar@aol.com
	Lee Fleisher, M.D. University of Pennsylvania Perelman School of Medicine Lee.Fleisher@pennmedicine.upenn.edu
Opposed Reviewers:	

Dear Dr. Kharasch
Enclosed is our manuscript entitled:" Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty - a machine learning based approach." This study investigates potential advantages of a state of the art machine-learning model comprising 33 preoperative variables and including novel use of preoperative dispensed prescriptions within 3 months preoperatively, for prediction of postoperative medical complications resulting in prolonged hospitalizations or readmissions in fully implemented fast-track total hip and knee replacement. We believe that our results are an important contribution within the field of perioperative medicine and risk-prediction, especially the novel analyzes on the specific contributions of individual risk-factors in the machine-learning model. Consequently, we hope you will consider our study for publication in Anesthesiology.
We are aware of the large number of figures, most of which are Supplemental Digital Content. However, due to difficulties in combining the 4 panels of Figure 3a-d into a single PDF-file these have been submitted as separate files. We hope that you are able to assist in combining these panels into a single figure during the editorial process in case of acceptance.

Kind regards, on behalf of the authors
Christoffer Jørgensen and Henrik Kehlet

Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty - a machine learning based approach.

2. Author information:

Christian Michelsen, M.Sci., Research Fellow, The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17 2100 Copenhagen, Denmark

Christoffer C Jørgensen, M.D., Senior Researcher, Section of Surgical Pathophysiology and Centre for Fast-track Hip and Knee Replacement, 7621, Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen, Denmark

Mathias Heltberg, M.Sci, Research Fellow, The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17 2100 Copenhagen, Denmark

Mogens H. Jensen, D.Sci., Prof., The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17 2100 Copenhagen, Denmark

Alessandra Lucchetti, M.Sci., Research Fellow., The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17 2100 Copenhagen, Denmark

Pelle B Petersen, M.D., Ph.D, Section of Surgical Pathophysiology and Centre for Fast-track Hip and Knee Replacement, 7621, Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen, Denmark

Troels Petersen, M.Sci, Ass.Prof., The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17 2100 Copenhagen, Denmark

Henrik Kehlet, M.D., Ph.D., Prof. Section of Surgical Pathophysiology and Centre for Fast-track Hip and Knee Replacement, 7621, Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen, Denmark

*This is a joint first-authorship between CM and CJ

3. Corresponding author:

Dr. Christoffer Calov Jørgensen

Section for Surgical Pathophysiology 7621

Rigshospitalet, Blegdamsvej 9,

DK-2100 Copenhagen, Denmark

Phone +45 3545 4616 Fax: +45 3545 6543

E-mail: christoffer.calov.joergensen@regionh.dk

4. Clinical Trial Number: The Centre for Fast-track Hip and Knee Replacement Database was registered as a study registry on ClinicalTrials,gov:NCT01515670

5. Prior presentations: Not applicable

6. Acknowledgements: The members of the Centre for Fast-track Hip and Knee Replacement Database collaborative group all contributed by implementing the fast-track protocol at their respective departments and reviewing the final manuscript.

Frank Madsen M.D. Consultant, Department of Orthopedics, Aarhus University Hospital, Aarhus, Denmark

Torben B. Hansen M.D., Ph.D., Prof. Department of Orthopedics, Holstebro Hospital, Holstebro, Denmark

Kirill Gromov, M.D., Ph.D., Ass.Prof. Department of Orthopedics Hvidovre Hospital, Hvidovre Denmark

Thomas Jakobsen, M.D., Ph.D., DM.Sci., Ass. Prof. Department of Orthopedics, Aalborg University Hospital, Farsø, Denmark

Claus Varnum, M.D., Ph.D., Ass. Prof. Department of Orthopedic Surgery, Lillebaelt Hospital - Vejle, University Hospital of Southern Denmark, Denmark

Soren Overgaard, M.D., DM.Sci., Prof, Department of Orthopedics, Bispebjerg Hospital, Copenhagen, Denmark

Mikkel Rathsach, M.D., Ph.D., Ass. Prof. Department of Orthopedics, Gentofte Hospital, Gentofte, Denmark

Lars Hansen, M.D., Consultant, Department of Orthopedics, Sydvestjysk Hospital, Grindsted, Denmark

7. Word and Element Counts:

Abstract: 300/300 Introduction: 466/500 Discussion:1278/1500 Figures:3 Tables:2 Appendices:2 Supplementary Digital Files:4

8. Abbreviated title: Machine learning models in joint replacement

9. Summary Statement: Not applicable.

10. Funding: The study received funding from the Lundbeck Foundation, Denmark, as well as from institutional and departmental sources.

11. Conflict of interest: Prof. Kehlet is a board member of “Rapid Recovery”, by Zimmer Biomet. Mr. Heltberg is sponsored by a grant from the Lundbeck Foundation, independently of the present study. Dr. Petersen is an advisory member of Sanofi outside of the present study.

Abstract

Background: Introduction of machine-learning models has potentially improved prediction of postoperative hospitalization and morbidity after hip and knee replacement. However, few studies include enhanced recovery programs, and most rely on administrative coding with limited follow-up and information on perioperative care. Thus, benefits of machine-learning models for prediction of postoperative morbidity in enhanced recovery hip and knee replacement remain uncertain.

Methods: Multicenter cohort study from 2014-2017 in enhanced recovery total hip and knee replacement. Prospective recording of comorbidity and prescriptions. Information on length of stay and readmissions through the Danish National Patient Registry and medical records. Data was split into training (n:18013) and test sets (n:3913). A machine-learning model with 33 variables was used for predicting “medical” morbidity with a length of stay of >4 days or 90-days readmission and compared to a full logistic regression model. In addition, a machine-learning model excluding age, an age-only model and parsimonious machine-learning and logistic regression models using the ten most important variables were evaluated. Model performances were evaluated using several metrics, including precision, operating receiver (AUC) and precision recall curves (AUPRC). Variable importance was analyzed using Shapley Additive Explanations values.

Results: With 782 (20%) “risk-patients”, precision, AUC and AUPRC were 13.6%, 76.3% and 15.5% for the full and 12,8%, 75.9% and 17.1% for the parsimonious machine-learning models vs. 12.5%, 74.5% and 15.7% for the full logistic regression model. The machine-learning model excluding age and the Age-only model performed worse. Of the top ten variables, eight were shared between the full machine-learning and logistic regression models, and the importance of specific prescribed drugs varied considerably with age.

Conclusion: A machine-learning algorithm using preoperative characteristics and prescriptions likely improves identification of patients in high-risk of medical complications after fast-track hip and knee replacement. Such algorithms could help identify patients who benefit from intensified perioperative care.

INTRODUCTION

Prediction of postoperative morbidity and requirement for hospitalization is important for planning of health care resources. With regard to the common surgical procedures of primary total hip and knee arthroplasty, the introduction of enhanced recovery or fast-track programs has led to a significant reduction of postoperative length of stay (length of stay) as well as morbidity and mortality.[1-3] However, despite such progress, a fraction of patients still have postoperative complications leading to prolonged length of stay or readmissions.[1,3,4]

Consequently, in order to prioritize perioperative care, many efforts have been published to preoperatively predict length of stay and morbidity using traditional risk factors such as age, preoperative cardio-pulmonary disease, anemia, diabetes, frailty, etc.[4-8] These efforts have been based on traditional statistical methods, most often multiple regression analyses, and essentially concluding that it is “better to be young and healthy than old and sick”.

Consequently, despite being statistically significant, conventional risk-stratification based on such studies has had a relatively limited clinically relevant ability to predict and reduce potentially preventable morbidity and length of stay.[4-8]

More recently, machine-learning methods have been introduced with success in several areas of healthcare and where preliminary data suggest them to improve surgical risk prediction compared to traditional risk calculation in certain anesthetic and surgical conditions.[9,10] This is also the case in total hip replacement, total knee replacement and uni-compartmental knee replacement, where several publications on machine-learning algorithms for prediction of length of stay,[11,12] complications,[13] disability,[14] potential outpatient setup,[15] readmissions[16] or payment models,[17,18] have shown promising predictive value compared to conventional statistical methods.[19]

However, few papers have included enhanced recovery programs, and most are based on large database cohorts with the presence of risk factors and complications often relying on administrative coding with limited information on perioperative care, follow-up and discharge destination. In our previous study of 9512 total hip and knee replacements within an enhanced recovery protocol and including the above information, we did not find advantages of machine-learning methods compared to logistic regression in predicting a length of stay > 2 days.[20] However, this may have been due to data imbalance, lack of details on medication and the chosen outcome of length of stay of >2 days.[20] Thus, machine-learning models remain promising and could provide an improved basis for identifying a potential “high-risk” surgical

population who may benefit from more extensive preoperative evaluation and postoperative medical care.

Consequently, we analyzed whether an improved machine-learning model was better for preoperative prediction of medical complications resulting in prolonged length of stay and readmissions compared to a traditional logistic regression model, in a large consecutive cohort of patients undergoing fast-track total hip and knee replacement within a national public health-care system.[1] In addition to well-defined patient-reported preoperative risk-factors, we also included information on dispensed reimbursed prescriptions 6 months prior to surgery using a nationwide registry.[21]

Method

Reporting of the study is done in accordance with the Transparent reporting of multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement[22] and the Clinical AI Research (CAIR) checklist proposal.[23]

The study is based on the Centre for Fast-track Hip and Knee Replacement database which is a prospective database on preoperative patient characteristics and enrolling consecutive patients from 7 departments between 2010 and 2017. The database is registered on ClinicalTrials.gov as a study registry (NCT01515670). Permission to review and store information from medical records without informed consent was acquired from Center for Regional Development (R-20073405) and the Danish Data Protection Agency (RH-2007-30-0623). Patients completed a preoperative questionnaire with nurse assistance if needed. Additional information on reimbursed prescriptions 6 months prior to surgery was acquired using the Danish National Database of Reimbursed Prescriptions (DNDRP) which records all dispensed prescriptions with reimbursement in Denmark.[21] Finally, data were combined with the Danish National Patient Registry (DNPR) for information on length of stay (counted as postoperative nights spent in hospital), 90-days readmissions with overnight stay and mortality. In case of length of stay >4 days or readmission, patient discharge summaries were reviewed for information on postoperative morbidity and in case of insufficient information, the entire medical records were reviewed. Readmissions were only included if considered related to the surgical procedure, thus excluding planned procedures like cancer workouts, cataract surgery, etc. Readmissions due to urinary tract infection or dizziness after day 30 were also considered unrelated to the surgical procedure. In case of postoperative mortality the entire medical record, including potential

readmissions, was reviewed to identify cause of death. Evaluation of discharge and medical records was performed by PP supervised by CJ. In case of disagreement, records were conferred with HK. Subsequently, causes of length of stay >4, readmissions or mortality were classified as “medical” when related to perioperative care (renal failure, falls, pain, thrombosis, anemia, venous thromboembolism or infection etc.) and “surgical” if related to surgical technique (prosthetic infection, revision surgery, periprosthetic fracture, hip dislocation, etc.).[1] In case of a length of stay 4-6 days with a standard discharge summary describing a successful postoperative course, it was assumed that no clinically relevant postoperative complications had occurred. If length of stay was >6 days but with standard discharge summary, the entire medical record was evaluated to confirm that no relevant complications had occurred.

For the present study, only cases between 2014 and 2017 were used to provide the most up-to date data. All patients had elective unilateral total hip and knee replacement in dedicated arthroplasty departments with similar fast-track protocols, including multimodal opioid sparing analgesia with high-dose (125mg) methylprednisolone, preference for spinal anesthesia, only in-hospital thromboprophylaxis when length of stay ≤5 days, early mobilization, functional discharge criteria and discharge to own home.[1] There was no selection criteria for the fast-track protocol as it is considered standard of care, but we excluded patients with previous major hip or knee surgery within 90-days of their total hip or total knee replacement and total hip replacement due to severe congenital joint disorder or cancer.

Outcomes

The primary outcome was to compare prediction quality when using a machine-learning model to predict the occurrence of “medical” complications resulting in a length of stay >4 days or readmission compared to a traditional logistic regression model (outcome A). Secondarily, we investigated how inclusion of cases with a length of stay >4 days but no reported “medical” complication as a positive outcome influenced the model (outcome B). For both outcomes, we also investigated whether a parsimonious model including only the top ten variables would perform equally well as the full model, and whether the effect of age per se would compare to the full machine-learning model. All figures and tables in the main text and Appendix are based on outcome A; the corresponding figures for outcome B are reported in the Supplemental Digital Content.

Statistical Analysis

Data was initially trimmed by removing 156 patients (1.7%) who were outliers with regards to weight (<30 kg or >250 kg) and height (<100 cm or >210 cm) or where these data were missing. To reduce the risk of overfitting, the dataset was subsequently split into a training set consisting of 18.013 (82.2%) procedures from 2014-2016 and a test set of 3913 (17.8%) procedures from 2017.

As a reference model, we used classical logistic regression using all 33 input variables (table 1). Cases of missing values in the logistic regression model were handled by imputing missing values with the median of present values. All variables were then normalized.

In addition, we used Boosted Decision Trees (LightGBM)[24] for the machine-learning models, as such methods work well with categorical data and missing values. We tried using both normal cross entropy and FocalLoss[25] as the objective function for the machine-learning model. The reason for testing FocalLoss was to allow the machine-learning model to focus more on the (few) positives.

The full machine-learning model was trained and hyperparameter optimized using state of the art machine-learning methods. The models were trained on the training data and then used for making predictions on the unseen test data (see supplementary for details). The classification threshold was calibrated such that no more than 20% of the total number of patients were predicted as positive by the model (a positive predictive fraction (PPF) of 20%). We also included results for PPF values of 25% and 30%. Furthermore, we trained two parsimonious models using machine-learning and logistic regression with only the 10 most important features. Finally, we specifically explored the influence of increasing age, by constructing a model based only on age (Age), and a machine-learning model based on all variables except for age.

To investigate the importance of the included variables, we computed the SHapley Additive exPlanations (SHAP) values, which provide estimates on which variables contribute most to the risk score predictions.[26,27] Finally, we investigated a potential relation between reimbursed prescribed cardiac drugs, anticoagulants, psychotropics and pulmonary drugs and age, as the relation between polypharmacy and postoperative outcomes have mainly been found in older patients.[28]

For evaluating model performance, we computed the number of true positives, false positives, false negatives, true negatives, sensitivity (true positive rate), precision (positive predictive value). Since the data was quite imbalanced (about a 1:20 positive:negative ratio) we also computed the Matthews Correlation Coefficient (MCC) which is independent of class

imbalance.[29,30] The MCC ranges between -1 (the 100% wrong classifier), 0 (the random classifier), and +1 (the perfect classifier). Finally, we computed the area under the receiver operating characteristic curve (AUC) and the area under the precision recall curve (AUPRC). To evaluate the statistical difference between the classifiers, we applied a Bayesian metric comparison P(sensitivity),[31] which is the probability that a model will perform better than the machine-learning model relative to the sensitivity. Thus, for two equally performing models P(sensitivity) is ≈ 50%.

Results

Median age in the 3913 patients was 70 years (IQR 62-76), 59% were female and 58% had total hip replacement (table 1). Details on prescribed drug types are shown in Appendix 1. Median length of stay was 2 (IQR: 1-2) days with 7.6% 90-days readmissions and outcome A occurring in 182 (4.7%) patients. When applying any model with a positive prediction fraction of 20% to the 3913 patients, 782 qualified as “risk-patients”. The results are summarized in figure 1 and table 2. When considering risk scores from the full machine-learning (figure 1a) and full logistic regression model leading to this risk-patient selection, 106 and 98 had outcome A, respectively. Correspondingly, the sensitivity and precision were 58.2% and 13.6% for the full machine-learning and 53.8% and 12.5% for the full logistic regression model, respectively. The full machine-learning model was superior (figure 1b) on essentially all parameters compared to any of the other models, although the differences were minor (table 2).The results were similar when using positive prediction fractions of 25% and 30%, but with the sensitivity for the full machine-learning model increasing to 64.4% and 69.2% and precision decreasing to 12.0% and 10.7%, respectively (Appendix table 2).

Both the machine-learning model excluding age and age-only model had significantly lower sensitivity than the full machine-learning model (figure 1b). Despite age being the single most important variable (figure 2), the machine-learning model excluding age performed as well as the age-only model.

When evaluating feature importance, we found a strong correlation between the full machine-learning and full logistic regression model, with age and use of walking aids being the most important variables in both (figure 2a). From the combined importance of variables outside the top ten, the machine-learning approach extracted more information with fewer variables than logistic regression (figure 1b).

For the full machine-learning model, there was a clear signal that increasing age, number of reimbursed prescriptions, and presence of comorbidity, all contributed to an increased risk score. In contrast, a recent date of surgery and an increased hemoglobin level seemed to reduce the calculated risk (figure 2b). Individual analysis of the SHAP interaction values for types of anticoagulant prescriptions revealed that prescriptions on vitamin-K antagonists (VKA) or adenosine diphosphate (ADP) antagonists increased, while acetylic salicylic acid and direct oral anticoagulants (DOAC) reduced the risk score of the full machine-learning model, regardless of age (figure 3a). The SHAP analysis of prescribed cardiac drugs revealed that prescriptions on Ca^{2+}-antagonists and betablockers in combination with one or two other antihypertensives increased the risk-score, as did prescriptions on nitrates, other antihypertensives and antiarrhythmics. For the remaining cardiac drugs, prescriptions either reduced or had minor influence, and with limited relation with age (figure 3b). Preoperative psychotropic prescriptions increased the risk-score except for antipsychotics (0.6%). For users of selective serotonin inhibitors there was a clear age-related distinction with the risk score being increased in elderly patients but decreased in those < 60 years (figure 3c). Finally, the risk score increased with prescriptions on inhalation steroid and β-blockers, and more accentuated in the younger patients (figure 3d).

The results including patients with a length of stay >4 days, but no reported postoperative complications (outcome B) were similar as for outcome A. In general, we found that the full machine-learning model was superior to the others, although the difference were smaller than for outcome A. (Supplemental Digital Content table S1 listing outcome parameters and Supplemental Digital Content 2 figure S1a-b showing distributions and ROC curves for outcome B). While the ten most important variables for the full machine-learning model remained unchanged, familiar disposition for venous thromboembolism replaced gender as one of the top ten important variables in the full logistic regression model (Supplemental Digital Content figure S2a-b showing SHAP values for outcome B). Furthermore, the SHAP analysis on specific prescribed drugs demonstrated that the machine-learning model found no benefits from information on prescriptions on respiratory drugs, why all SHAP values were zero. In addition, the reduced risk with acetylsalicylic acid and DOAC prescriptions, as well as the influence of practically all cardiac drugs except for nitrates, other antihypertensives and antiarrhythmics, was attenuated (Supplemental Digital Content 4 figure S3a-d showing SHAP-values of prescriptions of specific drugs for outcome B).

Discussion

We found that using a machine-learning algorithm including all 33 available variables and a parsimonious machine-learning-algorithm encompassing only the 10 most important predictors improved prediction of patients at increased risk of having a length of stay >4 days or readmissions due to medical complications compared to traditional logistic regression models. In contrast, when also including patients having a length of stay >4 days but without a well-defined complication as an outcome, the parsimonious machine-learning model was slightly worse than a traditional logistic regression model including all variables. We also found that although age was the single most important predictor of both outcome A and B, it was less suited for prediction of postoperative medical complications after fast-track total hip and knee replacement on its own. Finally, we demonstrated how the chosen classification threshold of the machine-learning algorithm influenced model performance through an increase in sensitivity at the cost of decreased precision.

A previous systematic review also found that machine-learning algorithms may provide better prediction of postoperative outcomes in THA and TKA.[32] However, the authors concluded that such models performed best at predicting postoperative complications, pain and patient reported outcomes and were less accurate at predicting readmissions and reoperations.[32] That machine-learning algorithms may improve prediction of complications after THA and TKA compared to traditional logistic regression was also found by Shah *et al.* who used an automated machine-learning framework to predict selected major complications after THA.[13] However, theirs was a retrospective study based on diagnostic and administrative coding and the selected complications occurred only in 0.61% of patients, potentially limiting clinical relevance. In contrast, we aimed at identifying a cohort which would comprise 20% of patients in which we found about 60% of all medical complications. This we believe, is within the means of the Danish socialized healthcare system to allocate additional resources for intensified perioperative care and with both patient-related and economic benefits due to potentially avoided complications and costs.

In contrast to many other machine-learning studies,[33] our dataset included not only preoperative data but also only one paraclinical variable, which was preoperative hemoglobin. Although the inclusion of other laboratory tests such as preoperative albumin, sodium and alkaline phosphatase has been found to be of importance in machine-learning algorithms for home discharge in UKA[12] and spine surgery,[9] they are not standard in our fast-track protocols and not easy to interpret from a pathophysiological point of view. As there is a need to prioritize the

limited health-care resources, most decisions on which patients may benefit from more extensive postoperative care will likely need to be conducted preoperatively. Thus, although postoperative information such as duration of surgery, perioperative blood length of stays or postoperative hemoglobin have been included in other studies[33], we decided against the use of peri- and postoperative data. The same approach has been used by Ramkumar *et al.* who used U.S. National Inpatient Sample data including 15 preoperative variables, to predict length of stay, patient charges and disposition after both TKA[34] and THA.[18] However, these studies were not conducted in a socialized health care system, and the main focus was on the need for differentiated payment bundles and without specific information on the reason for increased length of stay or non-home discharge.[34] Wei *et al.* used an artificial neural network model to predict same-day discharge after TKA, based on the NSQUIP database from 2018 and found that six of the ten most important variables were the same compared with logistic regression, similar to our findings.[35] However, patients with one-day length of stay were intentionally excluded due to variations in in-patient vs. out-patient registration.[35]

Age has traditionally been a major factor when predicting surgical outcomes which is why we choose to specifically evaluate its effect on our risk-prediction. That age is important for risk-prediction was further illustrated by the machine-learning model without age being comparable to the age-only model. Note that, although elderly patients had increased risk of postoperative complications, likely related to decline of physical reserves,[36] the use of chronological age alone as a selection criteria for being a “risk-patient” was inferior compared to both machine-learning and logistic regression models incorporating comorbidity and functional status.

We used the SHAP values for estimation of feature importance, thus providing a better understanding of the otherwise “black-box” machine-learning model. The SHAP values showed which variables contribute most to the risk-score predictions.

Our inclusion of specific data on reimbursed prescriptions 6 months prior to surgery based upon the unique Danish registries, unsurprisingly found increased risk-scores with increased number of prescriptions and with the majority being in elderly patients. Similarly, a Canadian study in elective non-cardiac surgery found decreased survival and increased length of stay and readmissions and costs in patients >65 years with polypharmacy.[28] However, this is a complex relationship where some patients benefit from their treatments, while other may suffer from undesirable side-effects. Consequently, the authors cautioned against altering perioperative practices based on current evidence.[28] However, the information from the included prescriptions with SHAP analysis may provide inspiration for new hypothesis-generating studies such as investigation of the potential differences in risk-profile between having preoperative prescribed

VKA and DOAKs. Also, the age-related differences in risk from SSRI's seen in our study could guide further studies on "deprescription".

Another requirement for machine-learning-algorithms to be clinically useful is user friendliness and not depending on excessive additional data collection by the attending clinicians. In this context, it was a bit disappointing that the parsimonious machine-learning algorithm with only the ten most important variables was slightly worse at predicting outcome B than the full logistic regression model. A reason for this could be that when including a length of stay >4 days but without described medical complications, the combination of all variables provide information not available by merely including the ten most important ones. This highlights the need for as much detailed, and preferably non-binary, data as possible to fulfill the true potential of machine-learning algorithms.

Our study has some limitations. First, one of the strengths of machine learning compared to logistic regression is the analysis of multilevel continuous data, whereas we included only a limited number of, often binary, preoperative variables. This could have limited the full realization of our machine learning model. As previously mentioned, we excluded intraoperative information, including type of anesthesia, surgical approach etc. all of which may influence postoperative outcomes. The observational design of this study means that we cannot exclude unmeasured confounding or confounding by indication. Also, despite that the DNDRP has a near complete registration of dispensed medicine in Denmark, some types or drugs, especially benzodiazepines, are exempt from general reimbursement and thus not sufficiently captured.[21] Furthermore, it is doubtful whether the patients used all types of drugs at the time of surgery (e.g. heparin which is rarely for long-term use). Finally, classification of a complication being "medical" depended on review of the discharge records which can also introduce bias. However, we believe our approach to be superior to depending only on diagnostic codes which often are inaccurate[37] and provide limited details on whether the complication may be attributed to a medical or surgical adverse event. The strengths of our study include the use of national registries with high degree of completion (>99% of all somatic admissions in case of the DNDRP),[38] prospective recording of comorbidity, extensive information on prescription patterns 6 months prior to surgery and similar established enhanced recovery protocols in all departments.

In summary, our results suggest that machine-learning-algorithms likely provide clinically relevant improved predictions for defining patients in high-risk of medical complications after fast-track THA and TKA compared to a logistic regression model. Future studies could benefit

from using such algorithms to find a manageable population of patients who benefit from intensified perioperative care.

References

1. Petersen PB, Kehlet H, Jorgensen CC, Lundbeck Foundation Centre for Fast-track H, Knee Replacement Collaborative G: Improvement in fast-track hip and knee arthroplasty: a prospective multicentre study of 36,935 procedures from 2010 to 2017. Sci Rep 2020; 10: 21233
2. Khan SK, Malviya A, Muller SD, Carluke I, Partington PF, Emmerson KP, Reed MR: Reduced short-term complications and mortality following Enhanced Recovery primary hip and knee arthroplasty: results from 6,000 consecutive procedures. Acta Orthop. 2014; 85: 26-31
3. Partridge T, Jameson S, Baker P, Deehan D, Mason J, Reed MR: Ten-Year Trends in Medical Complications Following 540,623 Primary Total Hip Replacements from a National Database. J Bone Joint Surg Am 2018; 100: 360-367
4. Jorgensen CC, Gromov K, Petersen PB, Kehlet H: Influence of day of surgery and prediction of LOS > 2 days after fast-track hip and knee replacement. Acta Orthop 2021; 92: 170-175
5. Jorgensen CC, Petersen MA, Kehlet H: Preoperative prediction of potentially preventable morbidity after fast-track hip and knee arthroplasty: a detailed descriptive cohort study. BMJ Open. 2016; 6: e009813
6. Johns WL, Layon D, Golladay GJ, Kates SL, Scott M, Patel NK: Preoperative Risk Factor Screening Protocols in Total Joint Arthroplasty: A Systematic Review. J Arthroplasty 2020; 35: 3353-3363
7. Adhia AH, Feinglass JM, Suleiman LI: What Are the Risk Factors for 48 or More-Hour Stay and Nonhome Discharge After Total Knee Arthroplasty? Results From 151 Illinois Hospitals, 2016-2018. J Arthroplasty 2020; 35: 1466-1473 e1
8. Shah A, Memon M, Kay J, Wood TJ, Tushinski DM, Khanna V, McMaster Arthroplasty Collective g: Preoperative Patient Factors Affecting Length of Stay following Total Knee Arthroplasty: A Systematic Review and Meta-Analysis. J Arthroplasty 2019; 34: 2124-2165 e1
9. Li Q, Zhong H, Girardi FP, Poeran J, Wilson LA, Memtsoudis SG, Liu J: Machine Learning Approaches to Define Candidates for Ambulatory Single Level Laminectomy Surgery. Global Spine J 2021: 2192568220979835
10. Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR: Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission. Ann Surg 2020; 272: 1133-1139
11. Li H, Jiao J, Zhang S, Tang H, Qu X, Yue B: Construction and Comparison of Predictive Models for Length of Stay after Total Knee Arthroplasty: Regression Model and Machine Learning Analysis Based on 1,826 Cases in a Single Singapore Center. J Knee Surg 2022; 35: 7-14
12. Lu Y, Khazi ZM, Agarwalla A, Forsythe B, Taunton MJ: Development of a Machine Learning Algorithm to Predict Nonroutine Discharge Following Unicompartmental Knee Arthroplasty. J Arthroplasty 2021; 36: 1568-1576
13. Shah AA, Devana SK, Lee C, Kianian R, van der Schaar M, SooHoo NF: Development of a Novel, Potentially Universal Machine Learning Algorithm for Prediction of Complications After Total Hip Arthroplasty. J Arthroplasty 2021; 36: 1655-1662 e1
14. Sniderman J, Stark RB, Schwartz CE, Imam H, Finkelstein JA, Nousiainen MT: Patient Factors That Matter in Predicting Hip Arthroplasty Outcomes: A Machine-Learning Approach. J Arthroplasty 2021; 36: 2024-2032
15. Kugelman DN, Teo G, Huang S, Doran MG, Singh V, Long WJ: A Novel Machine Learning Predictive Tool Assessing Outpatient or Inpatient Designation for Medicare Patients Undergoing Total Hip Arthroplasty. Arthroplast Today 2021; 8: 194-199

16. Mohammadi R, Jain S, Namin AT, Scholem Heller M, Palacholla R, Kamarthi S, Wallace B: Predicting Unplanned Readmissions Following a Hip or Knee Arthroplasty: Retrospective Observational Study. JMIR Med Inform 2020; 8: e19761
17. Ramkumar PN, Navarro SM, Haeberle HS, Karnuta JM, Mont MA, Iannotti JP, Patterson BM, Krebs VE: Development and Validation of a Machine Learning Algorithm After Primary Total Hip Arthroplasty: Applications to Length of Stay and Payment Models. J Arthroplasty 2019; 34: 632-637
18. Ramkumar PN, Karnuta JM, Navarro SM, Haeberle HS, Iorio R, Mont MA, Patterson BM, Krebs VE: Preoperative Prediction of Value Metrics and a Patient-Specific Payment Model for Primary Total Hip Arthroplasty: Development and Validation of a Deep Learning Model. J Arthroplasty 2019; 34: 2228-2234 e1
19. Haeberle HS, Helm JM, Navarro SM, Karnuta JM, Schaffer JL, Callaghan JJ, Mont MA, Kamath AF, Krebs VE, Ramkumar PN: Artificial Intelligence and Machine Learning in Lower Extremity Arthroplasty: A Review. J Arthroplasty 2019; 34: 2201-2203
20. Johannesdottir KB, Kehlet H, Petersen PB, Aasvang EK, Sørensen HBD, Jørgensen CC: Machine learning classifiers do not improve prediction of hospitalization > 2 days after fast-track hip and knee arthroplasty compared with a classical statistical risk model. Acta Orthop 2022; 93: 117-123
21. Johannesdottir SA, Horvath-Puho E, Ehrenstein V, Schmidt M, Pedersen L, Sorensen HT: Existing data sources for clinical epidemiology: The Danish National Database of Reimbursed Prescriptions. Clin.Epidemiol. 2012; 4: 303-313
22. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015; 162: W1-73
23. Olczak J, Pavlopoulos J, Prijs J, Ijpma FFA, Doornberg JN, Lundstrom C, Hedlund J, Gordon M: Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. Acta Orthop 2021; 92: 513-525
24. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T: LightGBM: a highly efficient gradient boosting decision tree, Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA, Curran Associates Inc, 2017, pp 3149-57
25. Lin T-Y, Goyal P, Girshick R, He K, Dollár P: Focal Loss for Dense Object Detection. http://arxiv.org/abs/1708.02002, ArXiv170802002 Cs 2018
26. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI: From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell 2020; 2: 56-67
27. Lundberg SMLSI: A Unified Approach to Interpreting Model Predictions. Edited by Guyon I. Adv Neural Inf Process Syst [Internet], Curran Associates, Inc., 2017
28. McIsaac DI, Wong CA, Bryson GL, van Walraven C: Association of Polypharmacy with Survival, Complications, and Healthcare Resource Use after Elective Noncardiac Surgery: A Population-based Cohort Study. Anesthesiology 2018; 128: 1140-1150
29. Chicco D: Ten quick tips for machine learning in computational biology. BioData Mining 2017; 10: 35 (2017)
30. Chicco D, Totsch N, Jurman G: The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining 2021; 14: 13 (2021)
31. Totsch N, Hoffmann D: Classifier uncertainty: evidence, potential impact, and probabilistic treatment. PeerJ Comput Sci 2021; 7: e398

32. Lopez CD, Gazgalis A, Boddapati V, Shah RP, Cooper HJ, Geller JA: Artificial Learning and Machine Learning Decision Guidance Applications in Total Hip and Knee Arthroplasty: A Systematic Review. Arthroplast Today 2021; 11: 103-112
33. Han C, Liu J, Wu Y, Chong Y, Chai X, Weng X: To Predict the Length of Hospital Stay After Total Knee Arthroplasty in an Orthopedic Center in China: The Use of Machine Learning Algorithms. Front Surg 2021; 8: 606038
34. Ramkumar PN, Karnuta JM, Navarro SM, Haeberle HS, Scuderi GR, Mont MA, Krebs VE, Patterson BM: Deep Learning Preoperatively Predicts Value Metrics for Primary Total Knee Arthroplasty: Development and Validation of an Artificial Neural Network Model. J Arthroplasty 2019; 34: 2220-2227 e1
35. Wei C, Quan T, Wang KY, Gu A, Fassihi SC, Kahlenberg CA, Malahias MA, Liu J, Thakkar S, Gonzalez Della Valle A, Sculco PK: Artificial neural network prediction of same-day discharge following primary total knee arthroplasty based on preoperative and intraoperative variables. Bone Joint J 2021; 103-B: 1358-1366
36. Griffiths R, Beech F, Brown A, Dhesi J, Foo I, Goodall J, Harrop-Griffiths W, Jameson J, Love N, Pappenheim K, White S: Peri-operative care of the elderly. Anaesthesia 2014; 69 Suppl 1: 81-98
37. Bedard NA, Pugely AJ, McHugh MA, Lux NR, Bozic KJ, Callaghan JJ: Big Data and Total Hip Arthroplasty: How Do Large Databases Compare? J Arthroplasty 2018; 33: 41-45.e3
38. Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sorensen HT: The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol 2015; 7: 449-90

Figure legends

Figure 1a-b

1a) Distribution of full machine learning model risk scores for patients +/- outcome A. The dashed line marks the classification threshold of 20% positive prediction fraction.

1b) Receiver operating curves (ROC) for the full machine learning model (F-MLM), full logistic regression model (F-LRM), parsimonious machine learning model (P-MLM), parsimonious logistic regression model (P-LRM), machine learning excluding age (MLM -age) and the age-only model (AM).

Figure 2a-b

2a) The overall importance of the 10 most important variables measured by the SHAP-values for the full machine-learning and full logistic regression models on outcome A (LOS >4 days or readmission due to “medical” morbidity). Only the importance of prescribed anticholesterols and gender differ between the models. The contributions of the remaining variables are summed in the bottom bar.

2b) The SHAP-values for the full machine-learning model on outcome A, where positive increase and negative values decrease the risk score. The color is related to the value of the variable with blue being lowest and red highest and each dot represents a patient.

Figure 3a-d

SHAP scatter-plot on the contributions to the full machine-learning model on outcome A (LOS >4 days or readmission due to “medical” morbidity), for individual types of prescribed anticoagulants, cardiac drugs, psychotropics and respiratory drugs stratified by age.

3a) Prescribed anticoagulants

VKA: vitamin K antagonists ASA: acetylsalicylic acid DOAC: direct oral anticoagulant ADP: Adenosine diphosphate ACE: angiotensin converting enzyme

3b) Prescribed cardiac drugs

ACE: angiotensin converting enzyme AHT: antihypertensive. Other AHT were defined as AHT different from diuretics ANG-II/ACE inhibitors or Ca^{2+}antagonists. IHD: Ischemic heart disease

3c) Prescribed psychotropics

SSRI: Selective serotonin inhibitor SNRI: Serotonin and norepinephrine reuptake inhibitor NaRI: Norepinephrine reuptake inhibitor NaSSA: Norepinephrine and specific serotonergic antidepressants. AD: antidepressants BZ: Benzodiazepines (likely underreported due to limited general reimbursement in Denmark). ADHD: Attention-deficit/hyperactivity disorder

3d) Prescribed respiratory drugs

SABA: Short-acting beta agonist LABA: long-acting beta agonist LAMA: Long-acting muscarinic antagonist.

Table 1. patient demographics with and without outcome A (length of stay >4 days or readmissions due to "medical" morbidity) in the combined test and training dataset.

Preoperative characteristics n (%) unless otherwise specified	+outcome A (n:1180)	-outcome A (n:20837)
mean age (SD)	75.0 (68.0-81.0)	69.0 (62.0-75.0)
mean number of reimbursed prescriptions[1] (SD)	3.0 (1.0-4.0)	2.0 (0.0-3.0)
female gender	755 (64.0)	12133 (58.2)
Total hip replacement	636 (53.9)	11542 (55.4)
mean weight in kg (SD)	78.0 (67.0-91.0)	81 (70.0-93.0)
mean height in cm (SD)	168 (162.0-175.0)	170.0 (164.0-178.0)
mean body mass index (SD)	27.3 (23.9-31.2)	27.5 (24.6-31.1)
regular use of walking aid	552 (46.8)	4398 (21.5)
missing	29 (2.5)	359 (1.7)
living alone	578 (49.0)	6717 (32.2)
with others	571 (48.4)	13869 (66.6)
institution	24 (2.0)	113 (0.5)
missing	7 (0.6)	138 (0.7)
hemoglobin	8.2 (7.7-8.8)	8.6 (8.1-9.2)
missing	11 (0.9)	314 (1.5)
>2 units of alcohol/day	79 (6.7)	1589 (7.6)
missing	10 (0.8)	174 (0.8)
active smoker	130 (11.0)	2751 (13.2)
missing	11 (0.9)	141 (0.7)
cardiac disease	306 (25.9)	2750 (13.2)
missing	8 (0.8)	153 (0.7)
hypercholesterolemia	467 (39.6)	6062 (29.1)
missing	8 (0.7)	120 (0.6)
hypertension	738 (62.5)	10141 (48.7)
missing	64 (5.4)	663 (3.2)
pulmonary disease	182 (15.4)	1841 (8.8)
missing	5 (0.4)	96 (0.5)
previous cerebral attack	165 (14.0)	1086 (5.2)
missing	25 (2.1)	282 (1.4)
previous VTE	133 (11.3)	1481 (7.1)
missing	26 (2.2)	325 (1.6)
malignancy (undefined)	557 (47.2)	8843 (42.4)
previous radically treated malignancy	127 (10.8)	2065 (9.9)
missing	14 (1.2)	162 (0.8)
chronic kidney disease	50 (4.2)	273 (1.3)
missing	35 (3.0)	292 (1.4)
family member with VTE	155 (13.1)	2510 (12.0)
missing	1190 (16.1)	2569 (12.3)
regular snoring	266 (22.5)	5522 (26.5)
uncertain about snoring	208 (17.6)	3781 (18.1)
missing	259 (21.9)	3309 (15.9)
not feeling rested	468 (39.7)	9340 (44.8)

uncertain about being rested	48 (4.1)	809 (3.9)
missing	105 (8.9)	1230 (5.9)
psychiatric disorder	156 (13.2)	1590 (7.6)
missing	62 (5.3)	703 (3.4)
Characteristic based on combination of questionnaire and DNDRP		
diabetes		
diet treated diabetes[2]	29 (2.5)	274 (1.3)
oral antidiabetics	137 (11.6)	1448 (6.9)
insulin treated diabetes[3]	60 (5.1)	413 (2.0)
missing	7 (0.6)	98 (0.5)

SD: standard deviation VTE: venous thromboembolic event DNDRP: Danish National Database of Reimbursed Prescriptions.

[1]Antirheumatica, steroids, anticoagulants, cardiac, cholesterol lowering, respiratory and psychotropic drugs.

[2]Reported diabetes but no registered prescriptions [3] +/- oral antidiabetics

Table 2: Performance of the six different models with a predefined positive prediction fraction of 20% for outcome A

Positive prediction fraction 20%	TP	FP	FN	TN	sensitivity	precision	MCC	AUROC	AUPRC	P (sensitivity)
Full machine-learning model	106	676	76	3055	58.2%	13.6%	21.1%	76.3%	15.5%	-
Full logistic regression model	98	684	84	3047	53.8%	12.5%	18.7%	74.5%	15.7%	19.7%
Parsimonious machine-learning model	100	682	82	3049	54.9%	12.8%	19.3%	75.9%	17.3%	26.1%
Parsimonious logistic regression model	95	687	87	3045	52.2%	12.1%	17.8%	73.7%	13.6%	12.4%
machine-learning model excluding age	88	694	94	3037	48.4%	11.3%	15.7%	72.3%	13.6%	3.1%
Age-only model	87	676	95	3055	47.8%	11.4%	15.8%	69.7%	12.1%	2.3%

TP: true positives FP: false positives FN: false negatives TN: true negatives MCC: Matthews correlation coefficient AUC: area under the ROC curve AUPRC: area under the precision recall curve P(sensitivity): probability that the model performs better than the machine-learning model relative to sensitivity.

Appendix table 1

Details on specific drugs with reimbursed prescriptions 6 months preoperatively.
Numbers are n (%)

	+Outcome A	-Outcome A
Anticoagulants		
none	679 (57.5)	15844 (76.0)
VKA	106 (9.0)	750 (3.6)
Heparin & Acetylsalicylic acid	0 (0.0)	7 (0.0)
DOAC	48 (4.1)	659 (3.2)
Acetylsalicylic acid	205 (17.4)	2492 (12.0)
Dipyradimol	5 (0.4)	29 (0.1)
ADP-antagonist	75 (6.4)	569 (2.7)
Acetylsalicylic acid & Dipyradimol	17 (1.4)	168 (0.8)
VKA & Acetylsalicylic acid	10 (0.8)	78 (0.4)
DOAC & Acetylsalicylic acid	6 (0.5)	41 (0.2)
VKA & ADP-antagonist	4 (0.3)	11 (0.1)
DOAC & ADP-antagonist	3 (0.3)	14 (0.1)
VKA & Heparin	1 (0.1)	21 (0.1)
DOAC & Acetylsalicylic acid & ADP-antagonist	1 (0.1)	3 (0.0)
Acetylsalicylic acid & ADP-antagonist	18 (1.5)	132 (0.6)
Acetylsalicylic acid & ADP-antagonist & Heparin	1 (0.1)	12 (0.1)
Acetylsalicylic acid & ADP-antagonist & Dipyradimol	1 (0.1)	7 (0.0)
Cardiac prescriptions		
none	321 (27.2)	9200 (44.2)
diuretics	77 (6.5)	1184 (5.7)
angiotensin-II/ACE-inhibitors	132 (11.2)	2683 (12.9)
Ca^{2+} antagonists	55 (4.7)	773 (3.7)
β-blocker	29 (2.5)	559 (2.7)
nitrates	1 (0.1)	18 (0.1)
other antihypertensives	0 (0.0)	12 (0.1)
other types of medication for IHD	2 (0.2)	21 (0.1)
2 antihypertensives	177 (15.0)	2696 (12.9)
β-blocker & 1 antihypertensive[1]	92 (8.1)	1069 (5.1)
3 antihypertensives	50 (4.2)	548 (2.6)
β-blocker & 2 antihypertensives[1]	95 (8.1)	975 (4.7)
β-blocker & 3 antihypertensives[1]	25 (2.1)	265 (1.3)
4 antihypertensives	2 (0.2)	18 (0.1)
β-blocker & 4 antihypertensives	2 (0.2)	19 (0.1)
other antihypertensive & antihypertensives[1]	9 (0.8)	87 (0.4)
nitrates & any hypertensive	49 (4.2)	331 (1.6)
other drugs for IHD & any antihypertensive and/or nitrate	5 (0.4)	15 (0.1)
other antiarrhythmics & any antihypertensives	57 (4.8)	364 (1.7)
Anticholesterols		
none	708 (60.0)	14719 (70.6)
statins	457 (38.7)	5866 (28.2)
other anti-lipids	7 (0.6)	135 (0.6)
Statins +other anti-lipids	8 (0.7)	117 (0.6)

Systemic steroids	123 (10.4)	1149 (5.5)
Antirheumatics		
none	1143 (96.9)	20388 (97.8)
disease-modifying antirheumatic drugs	37 (3.1)	446 (2.1)
other antirheumatics	0 (0.0)	3 (0.0)
Respiratory prescriptions		
none	1000 (84.7)	18754 (90.0)
SABA	13 (1.1)	276 (1.3)
LABA or LAMA	19 (1.6)	217 (1.0)
inhalation steroid only	8 (0.7)	211 (1.0)
SABA & Ipratropium (+/- others)	6 (0.5)	18 (0.1)
LABA & steroid	45 (3.8)	474 (2.3)
LABA & LAMA & steroid	19 (1.6)	122 (0.6)
LAMA & steroid	0 (0.0)	11 (0.1)
LABA & LAMA	7 (0.6)	80 (0.4)
other pulmonary drugs	3 (0.3)	32 (0.2)
other pulmonary drugs & steroid	9 (0.8)	98 (0.5)
SABA & LABA or LAMA	6 (0.5)	96 (0.5)
SABA & LABA or LAMA & steroid	45 (3.8)	448 (2.2)
Psychotropic prescriptions		
none	952 (80.7)	18657 (89.5)
SSRI/SNRI/NaRI	100 (8.5)	1164 (5.6)
other antidepressants	1 (0.1)	17 (0.1)
antipsychotics	8 (0.7)	116 (0.6)
benzodiazepines[2]	0 (0.0)	7 (0.0)
anti-cholinergics or memantine	6 (0.5)	27 (0.1)
anti-ADHD drugs	1 (0.1)	10 (0.0)
NaSSA	25 (2.1)	184 (0.9)
other psychotropics	28 (2.4)	182 (0.9)
SSRI + other antidepressants	4 (0.3)	6 (0.0)
SSRI + NaSSA	8 (0.7)	94 (0.5)
SRRI + antipsychotics	11 (0.9)	87 (0.4)
SRRI + other psychotropics	7 (0.6)	84 (0.4)
benzodiazepines + any psychotropic	3 (0.3)	12 (0.1)
antipsychotics + any psychotropic	20 (1.7)	149 (0.7)
anti-ADHD + any psychotropic	0 (0.0)	14 (0.1)
NaSSA + any psychotropic	4 (0.3)	18 (0.1)
other psychotropics + any specified psychotropic	2 (0.2)	9 (0.0)

VKA: vitamin K antagonists DOAC: direct oral anticoagulant ADP: Adenosine diphosphate ACE: angiotensin converting enzyme IHD: Ischemic heart disease SABA: Short-acting beta agonist LABA: long-acting beta agonist LAMA: Long-acting muscarinic antagonist SSRI: Selective serotonin inhibitor SNRI: Serotonin and norepinephrine reuptake inhibitor NaRI: Norepinephrine reuptake inhibitor NaSSA: Norepinephrine and specific serotonergic antidepressants

[1]either diuretics, ACE/ANG-II inhibitors or Ca^{2+}antagonists [2] likely underreported due to limited general reimbursement for benzodiazepines in Denmark

Appendix table 2

Performance of the six different models with a predefined positive prediction fraction of 25% and 30% for outcome A (LOS >4 days or readmission due to “medical” morbidity.

Positive prediction fraction 25%	TP	FP	FN	TN	sensitivity	precision	MCC	AUC	AUPRC	P (sensitivity)
Full machine-learning model	117	861	65	2870	64.3%	12.0%	20.0%	76.3%	15.5%	-
Full logistic regression model	110	868	72	2863	60.4%	11.2%	18.1%	74.5%	15.7%	23.1%
Parsimonious machine-learning model	115	863	67	2868	63.2%	11.8%	19.5%	75.9%	17.3%	41.2%
Parsimonious logistic regression model	106	872	76	2859	58.2%	10.8%	17.0%	73.4%	15.5%	11.8%
machine-learning model excluding age	106	872	76	2859	58.2%	10.8%	17.0%	72.3%	13.6%	11.8%
Age-model	94	824	88	2907	51.6%	10.2%	14.7%	69.7%	12.2%	0.7%
Positive prediction fraction 30%	**TP**	**FP**	**FN**	**TN**	**sensitivity**	**precision**	**MCC**	**AUC**	**AUPRC**	**P (sensitivity)**
Full machine-learning model	126	1047	56	2684	69.2%	10.7%	18.9%	76.3%	15.5%	-
Full logistic regression model	120	1053	62	2678	65.9%	10.2%	17.3%	74.5%	15.7%	25.2%
Parsimonious machine-learning model	124	1049	58	2682	68.1%	10.6%	18.4%	75.9%	17.3%	40.8%
Parsimonious logistic regression model	115	1058	67	2673	63.2%	9.8%	16.0%	73.7%	15.5%	11.1%
machine-learning model excluding age	116	1057	66	2674	63.7%	9.9%	16.3%	72.3%	13.6%	13.8%
Age-model	100	955	82	2776	54.9%	9.5%	13.9%	69.7%	12.2%	0.2%

TP: true positives FP: false positives FN: false negatives TN: true negatives MCC: Matthews correlation coefficient AUC: area under the ROC curve AUPRC: area under the precision recall curve P(sensitivity): probability that the model performs better than the machine-learning model relative to sensitivity.

Fig1

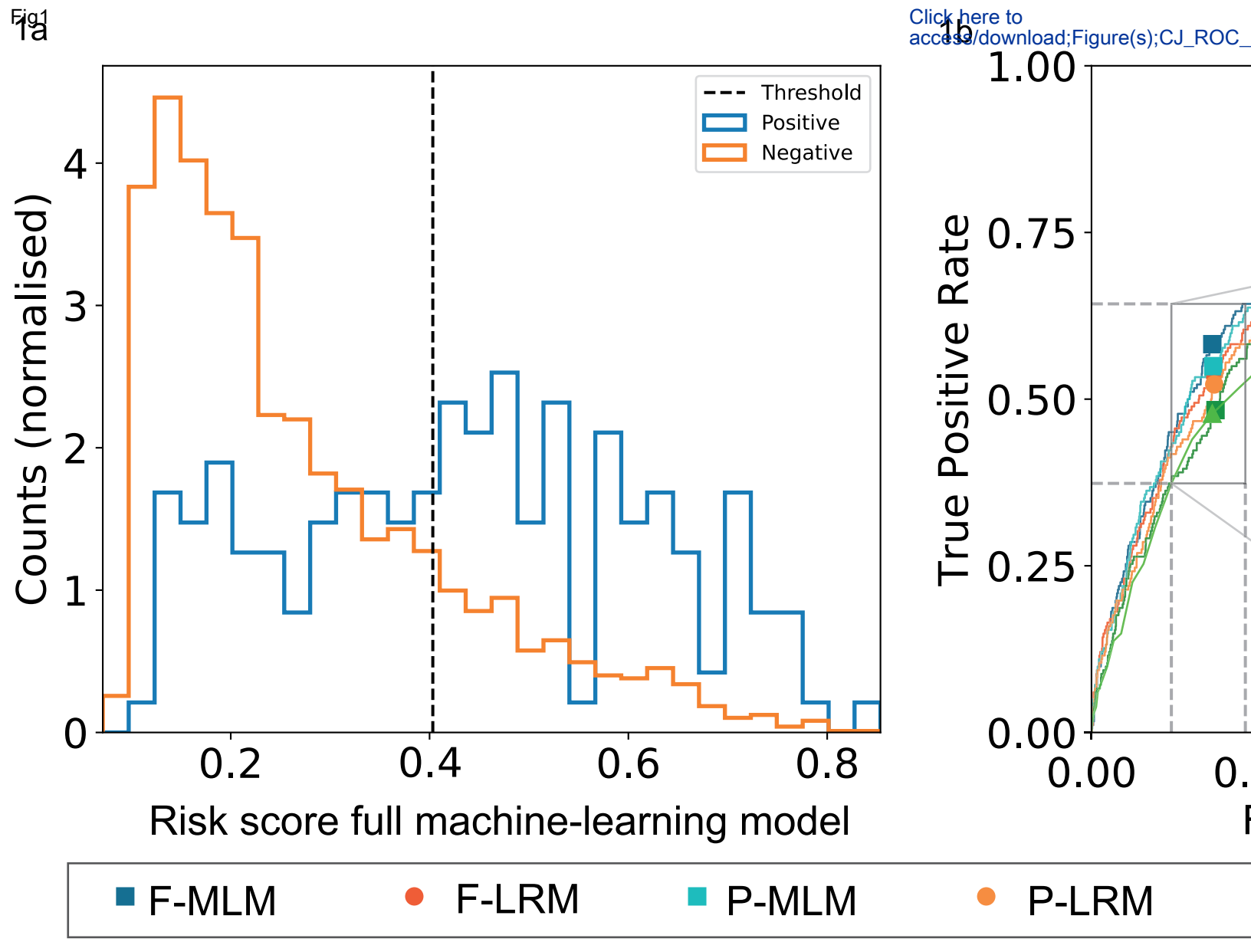

1a
Threshold
Positive
Negative
Counts (normalised)
0
1
2
3
4
0.2
0.4
0.6
0.8
Risk score full machine-learning model
1b
Click here to access/download;Figure(s);CJ_ROC_
1.00
0.75
0.50
0.25
0.00
True Positive Rate
0.00
F-MLM
F-LRM
P-MLM
P-LRM

Fig2

2a

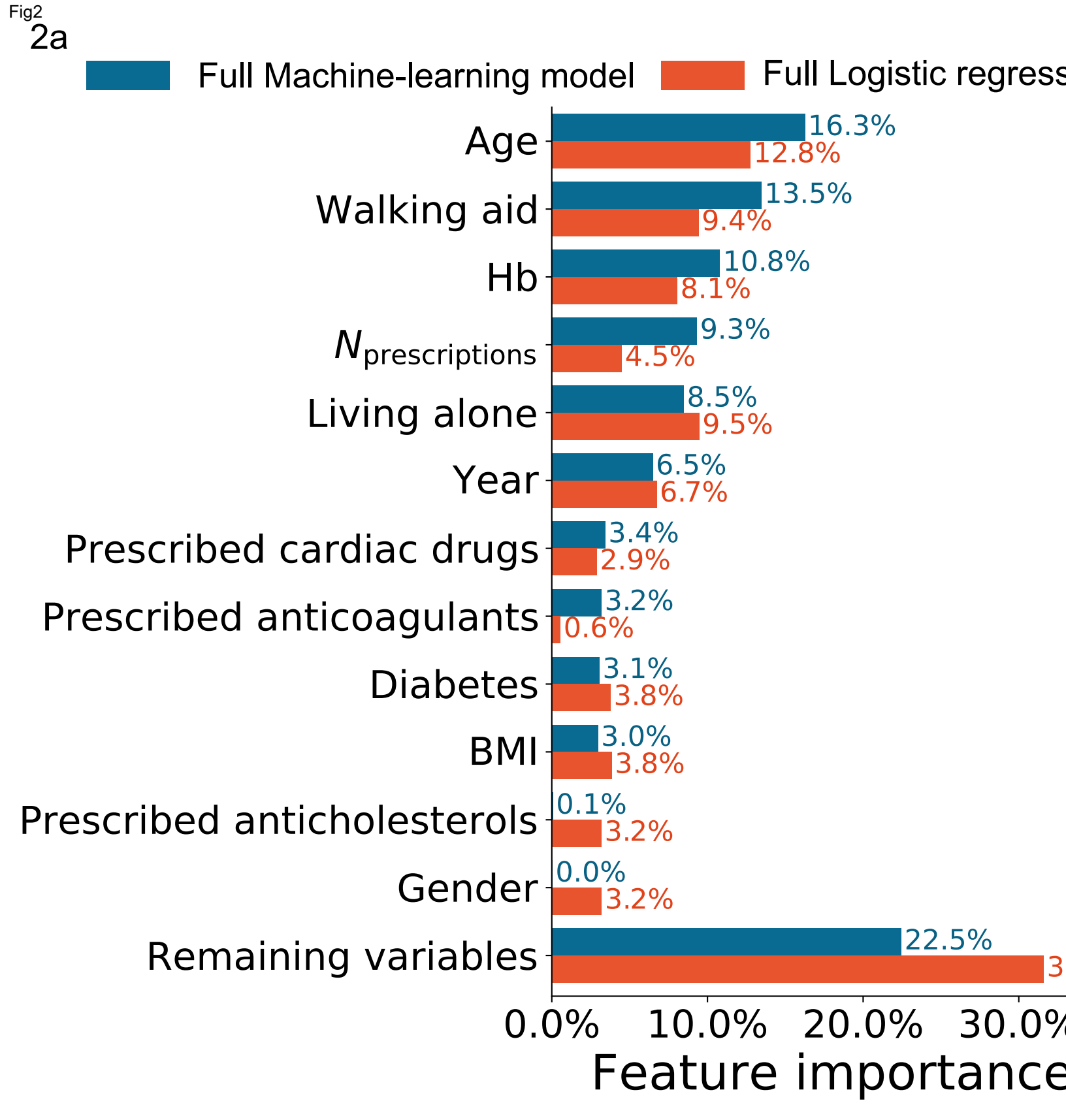

Full Machine-learning model
Full Logistic regress
Age
16.3%
12.8%
Walking aid
13.5%
9.4%
Hb
10.8%
8.1%
$N_{prescriptions}$
9.3%
4.5%
Living alone
8.5%
9.5%
Year
6.5%
6.7%
Prescribed cardiac drugs
3.4%
2.9%
Prescribed anticoagulants
3.2%
0.6%
Diabetes
3.1%
3.8%
BMI
3.0%
3.8%
Prescribed anticholesterols
0.1%
3.2%
Gender
0.0%
3.2%
Remaining variables
22.5%
3
0.0%
10.0%
20.0%
30.0%
Feature importance

fig3a

Click here to access/download;Figure(s);Fig3a.pdf

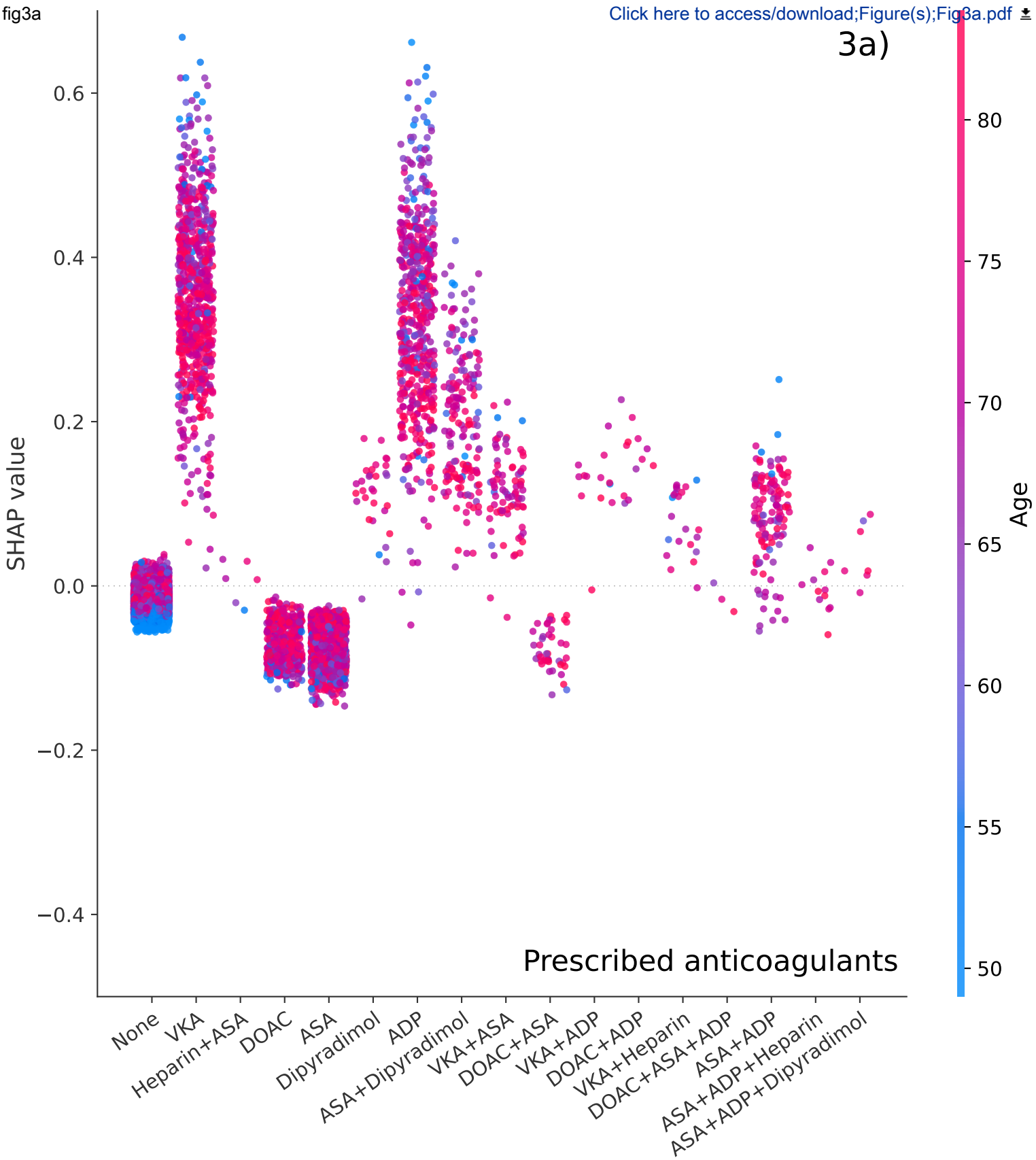


fig3b

Click here to access/download;Figure(s);Fig3b.pdf

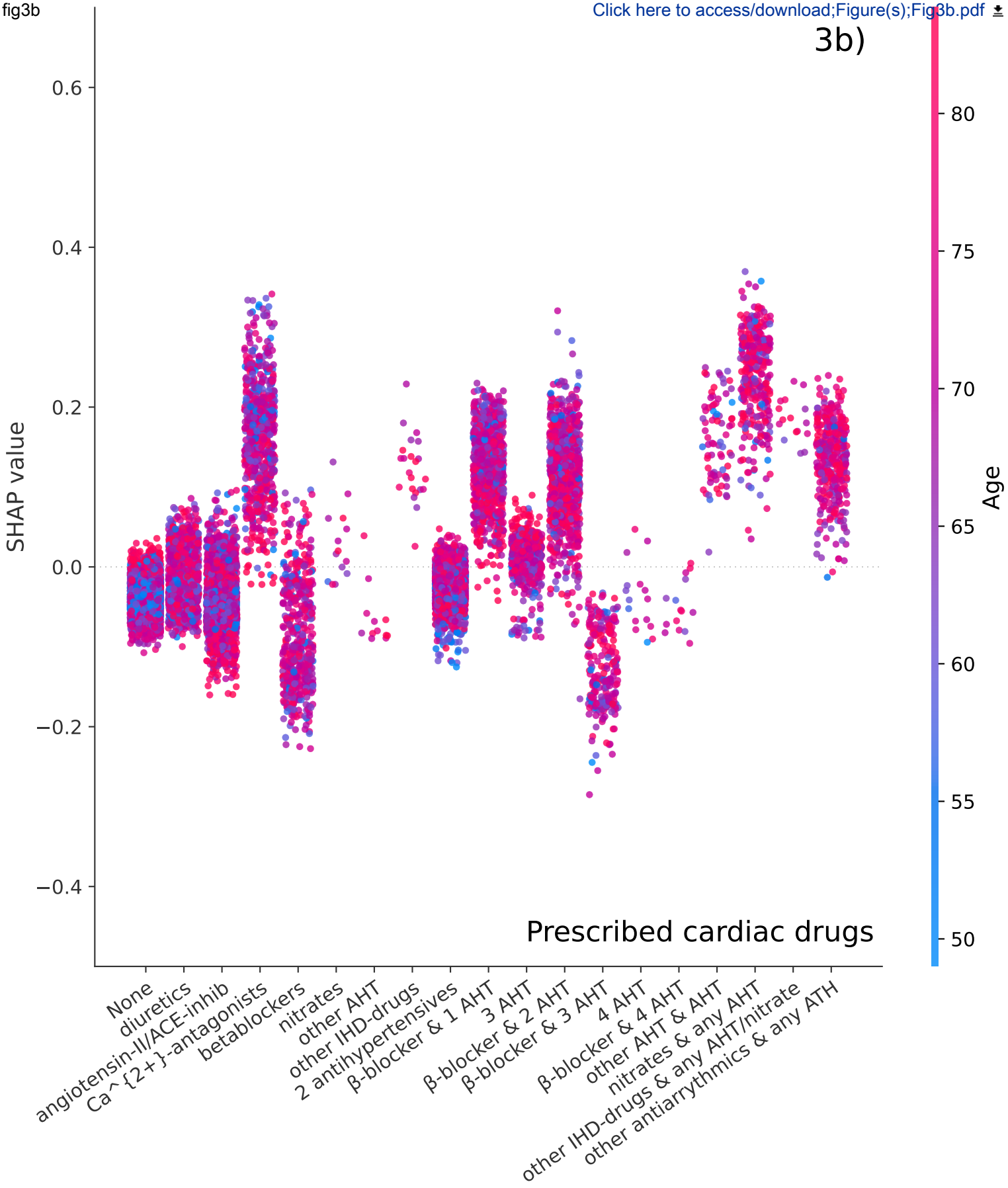


fig3c

Click here to access/download;Figure(s);Fig3c.pdf

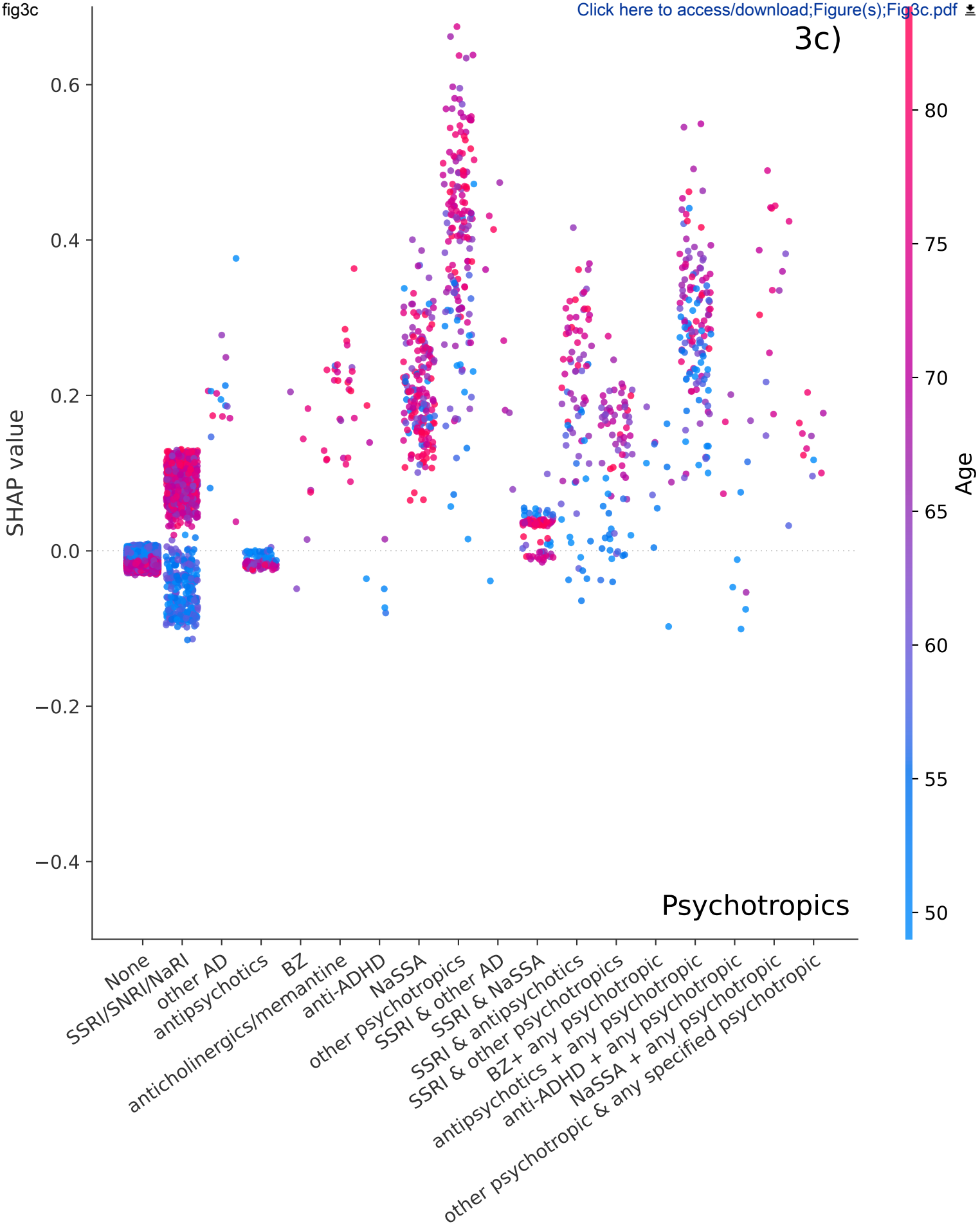
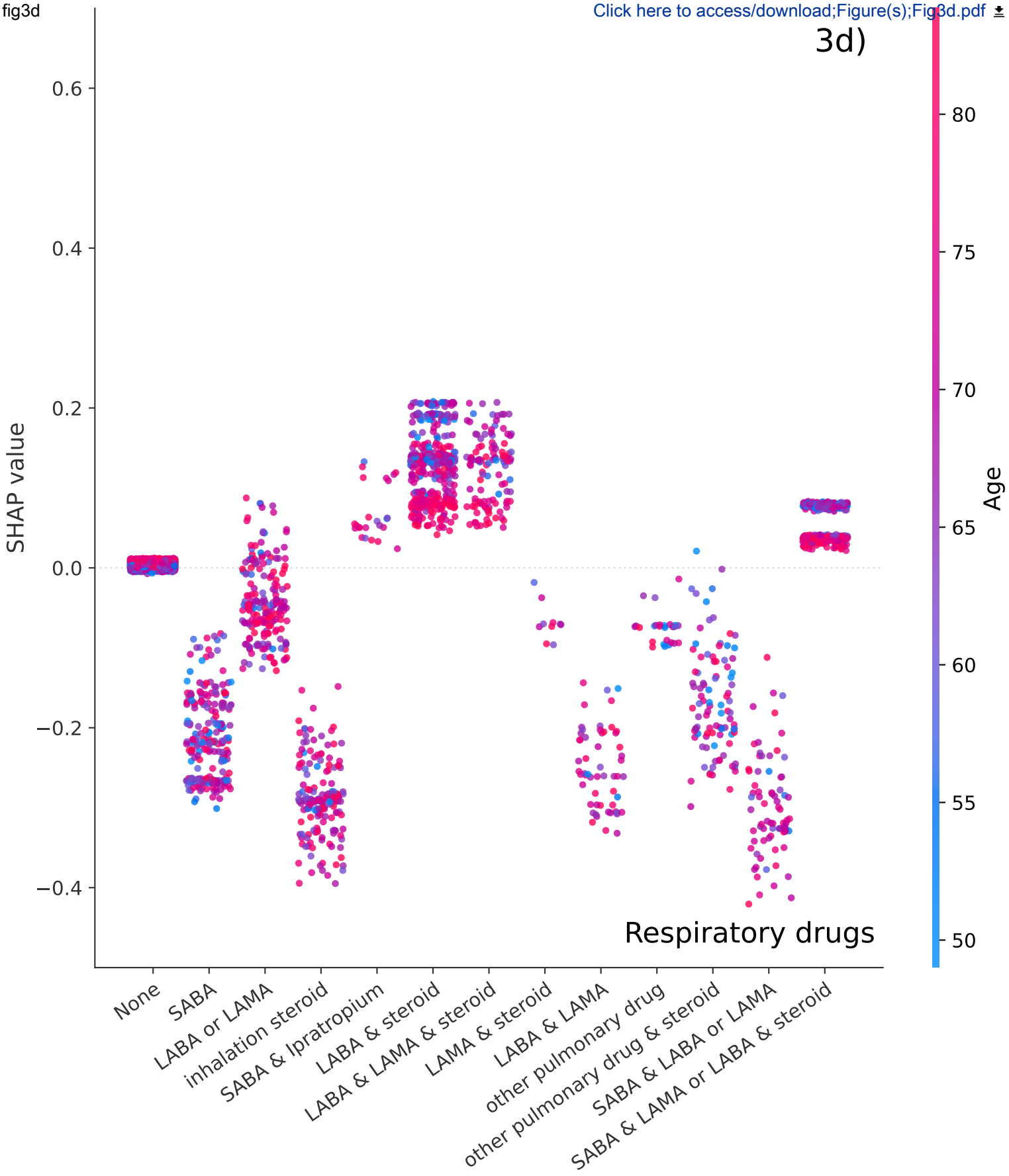


fig3d

Click here to access/download;Figure(s);Fig3d.pdf



Supplemental Digital Content 1

Table S1: Performance of different models for Outcome B (Los >4 days or readmissions due to “medical” morbidity or LOS >4 days but without recorded morbidity)

Positive prediction fraction 20%	TP	FP	FN	TN	sensitivity	precision	MCC	AUC	AUPRC	P (sensitivity)
Full machine-learning model	121	661	108	3023	52.8%	15.5%	20.5%	75.3%	17.1%	-
Full logistic regression model	115	667	114	3017	50.2%	14.7%	18.9%	74.1%	16.7%	28.3%
Parsimonious machine-learning model	111	671	118	3013	48.4%	14.2%	17.8%	74.4%	16.8%	17.2%
Parsimonious logistic regression model	109	673	120	3011	47.6%	13.9%	17.2%	73.1%	16.8%	12.9%
machine-learning model excluding age	110	672	119	3012	48.0%	14.1%	17.5%	72.8%	16.9%	15.1%
Age-model	102	661	127	3023	44.5%	13.4%	15.8%	68.7%	13.4%	3.8%
Positive prediction fraction 25%	TP	FP	FN	TN	sensitivity	precision	MCC	AUC	AUPRC	P (sensitivity)
Full machine-learning model	140	838	89	2846	61.1%	14.3%	20.8%	75.3%	17.1%	-
Full logistic regression model	136	842	93	2842	59.4%	13.9%	19.8%	74.1%	16.7%	35.3
Parsimonious machine-learning model	134	844	95	2840	58.5%	13.7%	19.3%	74.4%	16.8%	28.3
Parsimonious logistic regression model	125	853	104	2831	54.6%	12.8%	17.0%	73.1%	16.8%	7.8
machine-learning model excluding age	121	857	108	2827	52.8%	12.4%	16.0%	72.8%	16.9%	3.6
Age-model	113	805	116	2879	49.3%	12.3%	15.2%	68.7%	13.4%	0.5
Positive prediction fraction 30%	TP	FP	FN	TN	sensitivity	precision	MCC	AUC	AUPRC	P (sensitivity)
Full machine-learning model	153	1020	76	2664	66.8%	13.0%	20.0%	75.3%	17.1%	-
Full logistic regression model	147	1026	82	2658	64.2%	12.5%	18.6%	74.1%	16.7%	27.9

Parsimonious machine-learning model	147	1026	82	2658	64.2%	12.5%	18.6%	74.4%	16.8%	27.7
Parsimonious logistic regression model	145	1028	84	2656	63.3%	12.4%	18.1%	73.1%	16.8%	21.6
machine-learning model excluding age	140	1033	89	2651	61.1%	11.9%	17.0%	72.8%	16.9%	10.2
Age-model	122	933	107	2751	53.3%	11.6%	14.8%	69.8%	13.4%	0.1

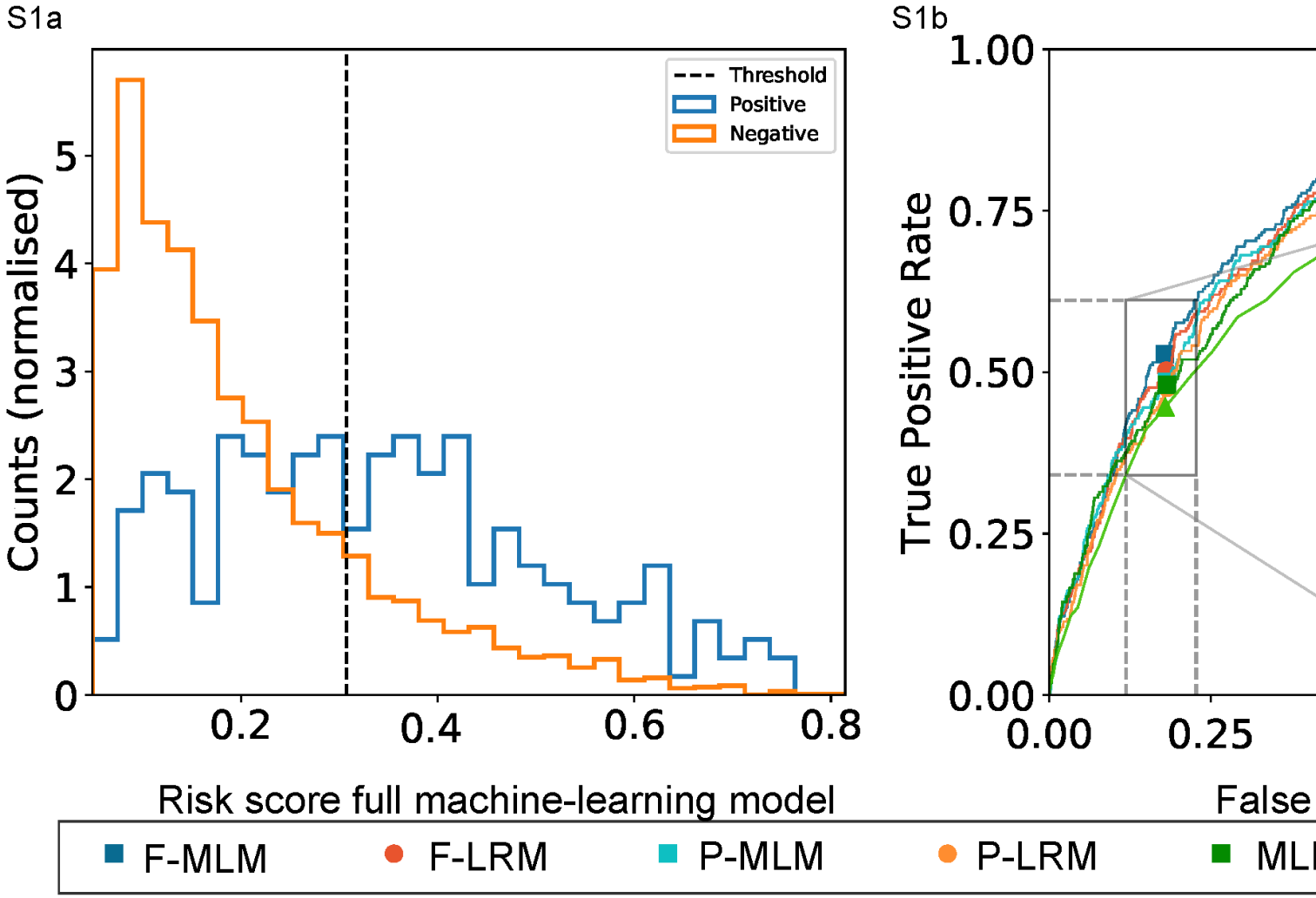
TP: true positives FP: false positives FN: false negatives TN: true negatives MCC: Matthews correlation coefficient AURC: area under the ROC curve AUPRC: area under the precision recall curve P(sensitivity): probability that the model performs better than the machine-learning model relative to sensitivity.

SDC2

Click here to access/do

Supplemental Digital Content 2

Figure S1a-b



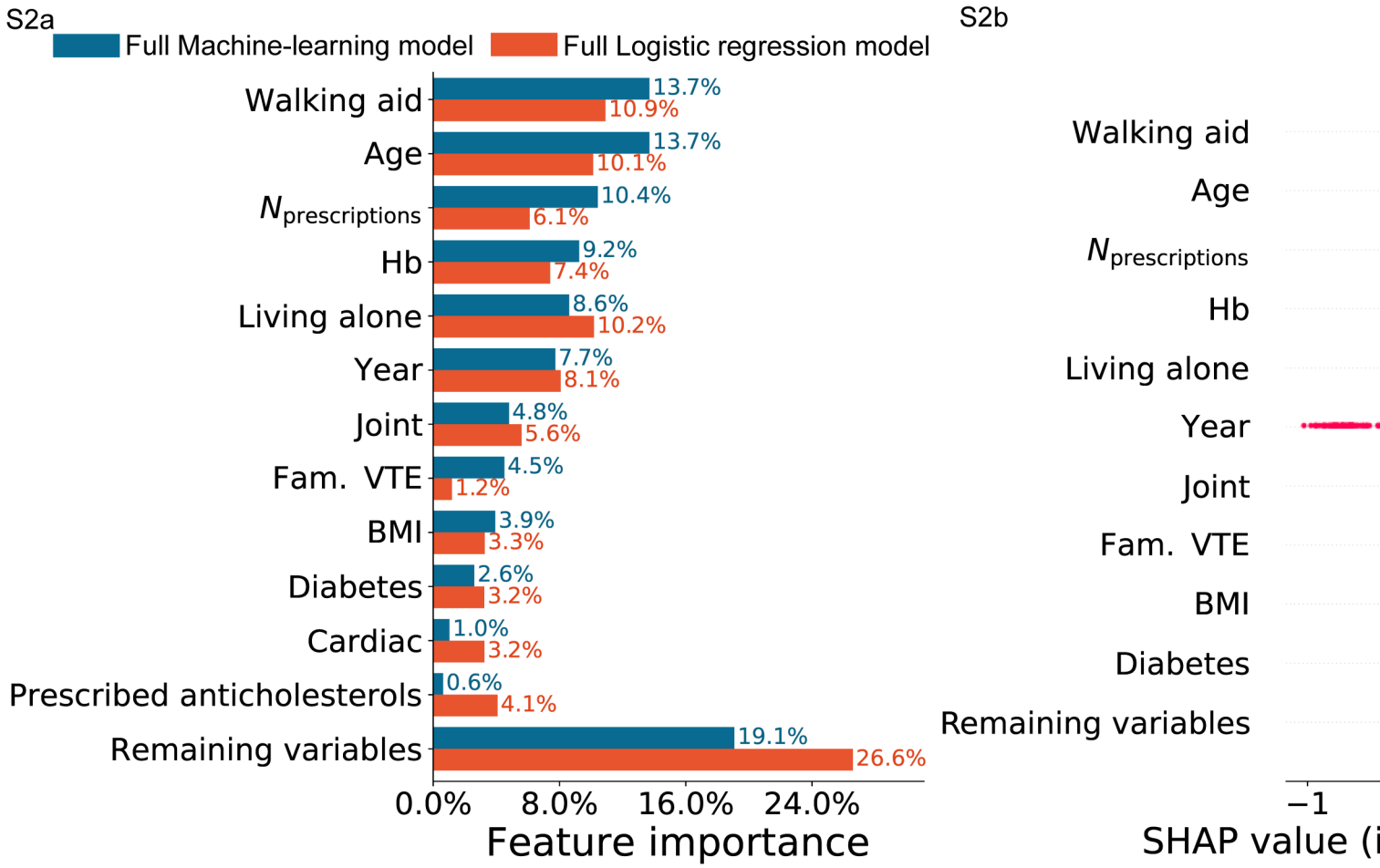
S1a) Distribution of full machine learning model risk scores for patients +/- outcome B(LOS >4 days or readmissions due to "me recorded morbidity). The dashed line marks the classification threshold of 20% positive prediction fraction.
S1b) Receiver operating curves (ROC) for the full machine learning model (F-MLM), full logistic regression model (F-LRM), pars MLM), parsimonious logistic regression model (P-LRM), machine learning excluding age (MLM -age) and the age-only model (A

SDC3

Click here to access/do

Supplemental Digital Content 3

Figure S2a-b

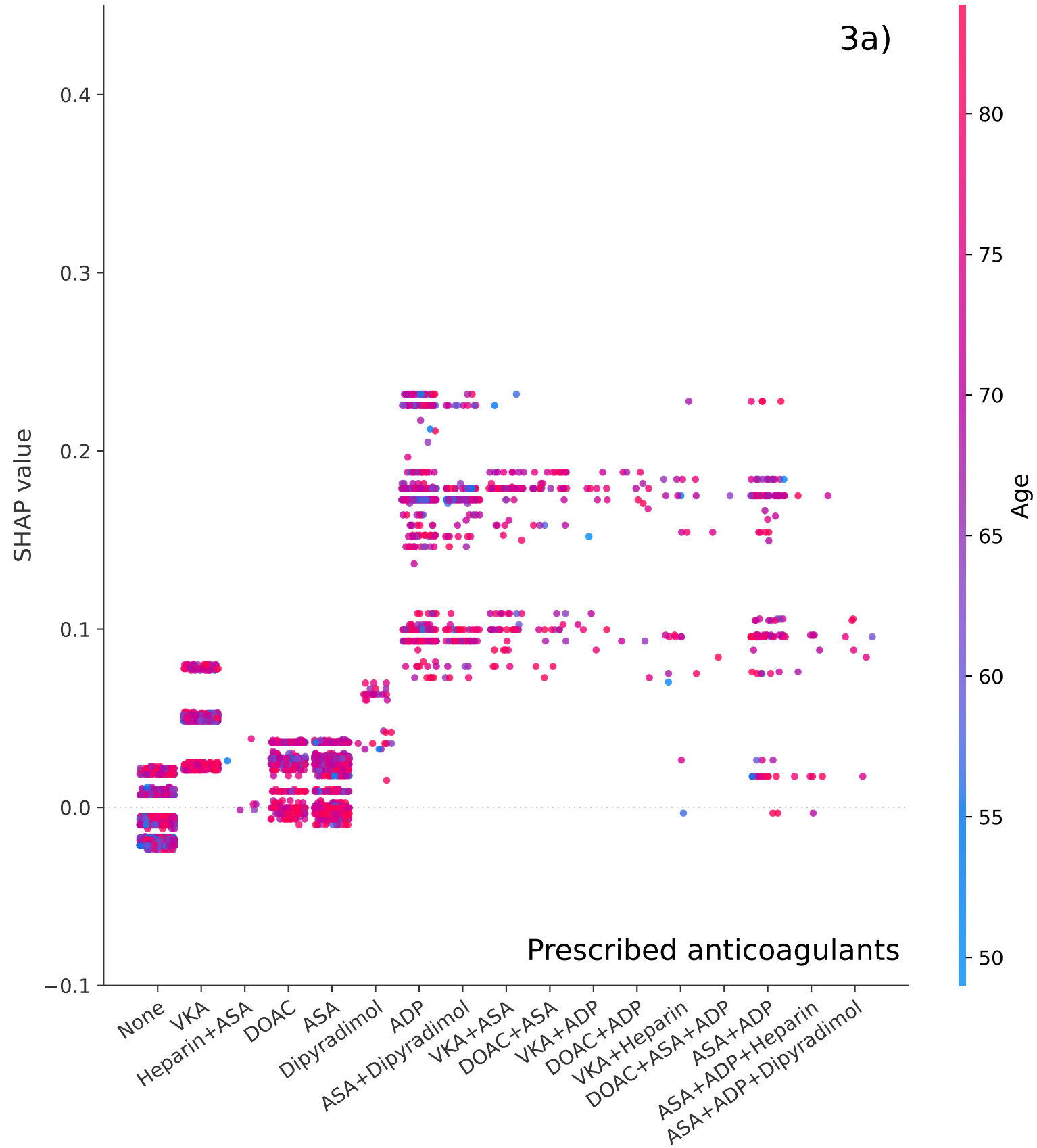


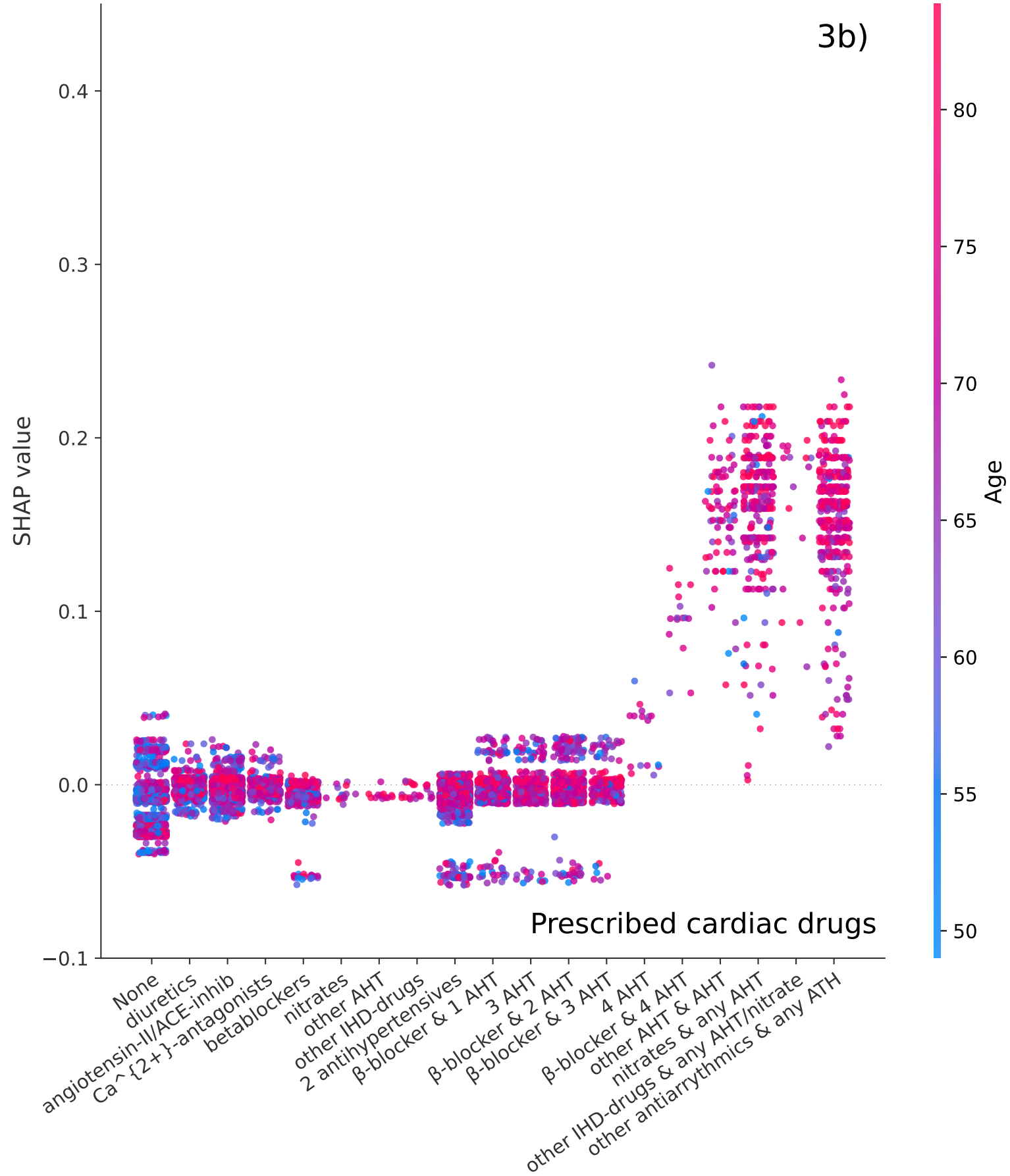
S2a) The overall importance of the 10 most important variables measured by the SHAP-values for the full machine-learning and come B (LOS >4 days or readmissions due to “medical” morbidity or LOS >4 days with no recorded morbidity).
Only the importance of prescribed anti-cholesterols and familiar disposition for venous thromboembolism differed between the m variables are summed in the bottom bar.
S2b) The SHAP-values for the full machine-learning model where values increase while negative values decrease the risk score variable with blue being lowest and red highest and each dot represents a patient.

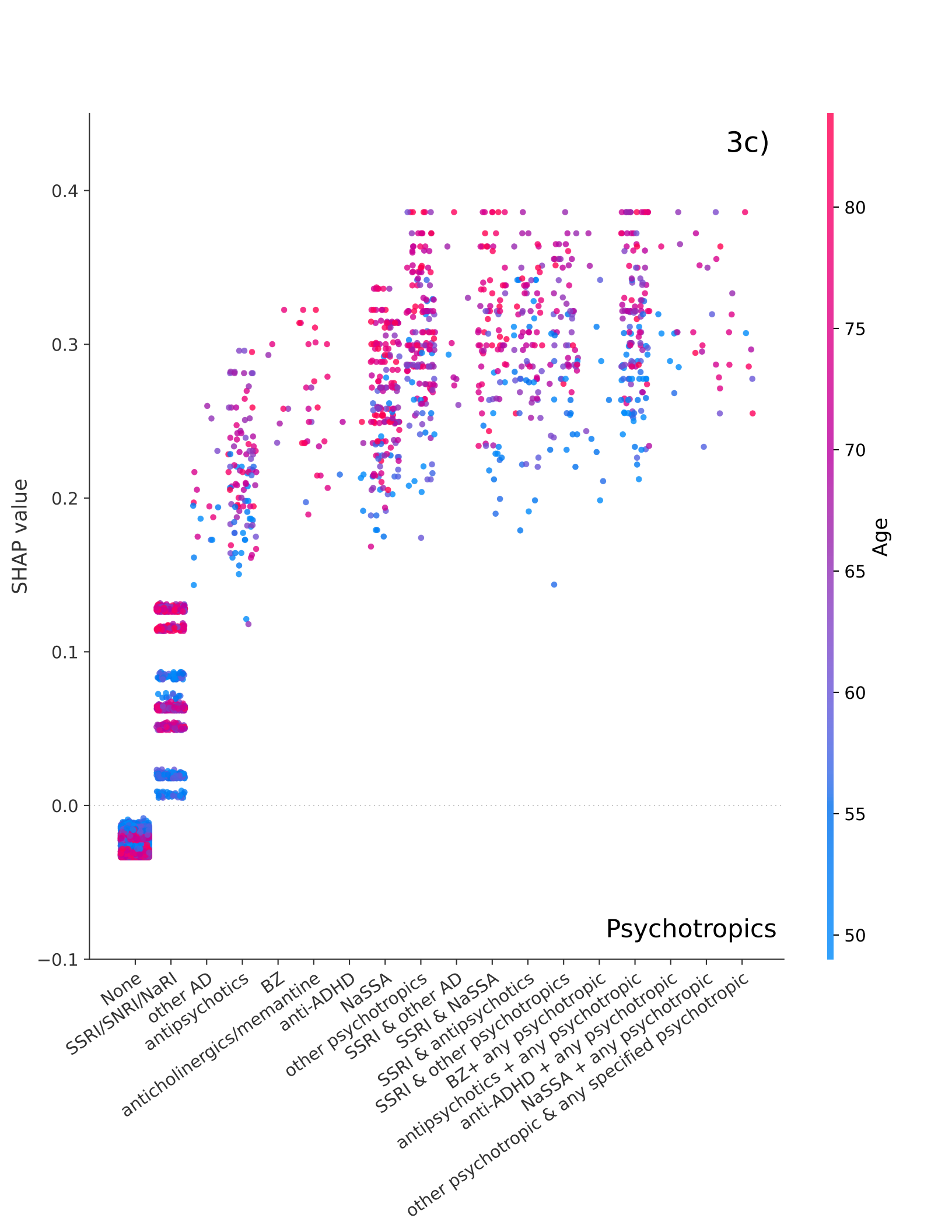
Supplemental Digital Content 4

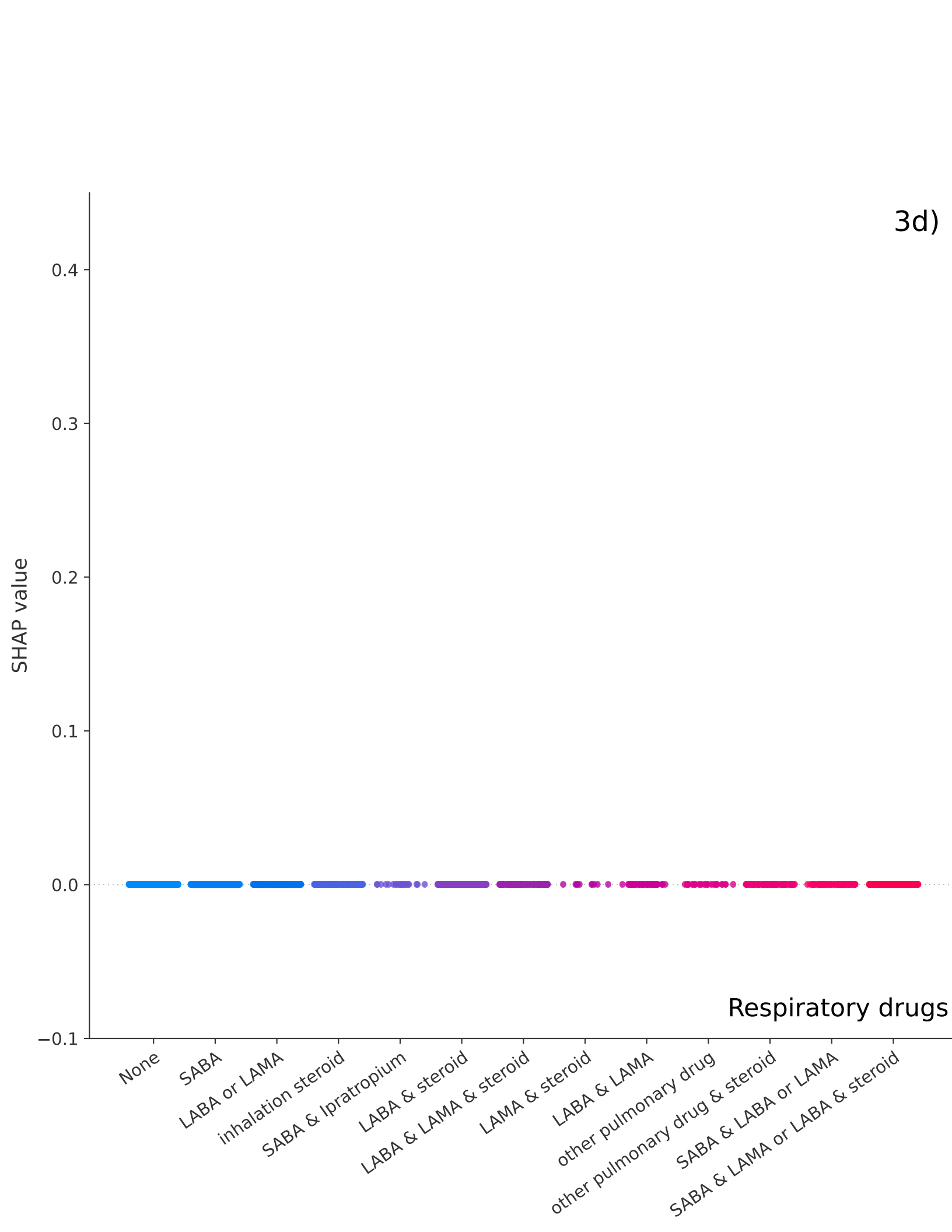
Figure S3a-d
SHAP scatter-plot on the contributions to the full machine-learning model on outcome B (LOS >4 days or readmission due to “medical” morbidity), for individual types of prescribed anticoagulants, cardiac drugs, psychotropics and respiratory drugs stratified by age.

Legend:
3a) Prescribed anticoagulants
VKA: vitamin K antagonists ASA: acetylsalicylic acid DOAC: direct oral anticoagulant ADP: Adenosine diphosphate ACE: angiotensin converting enzyme
3b) Prescribed cardiac drugs
ACE: angiotensin converting enzyme AHT: antihypertensive. Other AHT were defined as AHT different from diuretics ANG-II/ACE inhibitors or Ca2+antagonists. IHD: Ischemic heart disease
3c) Prescribed psychotropics
SSRI: Selective serotonin inhibitor SNRI: Serotonin and norepinephrine reuptake inhibitor NaRI: Norepinephrine reuptake inhibitor NaSSA: Norepinephrine and specific serotonergic antidepressants. AD: antidepressants BZ: Benzodiazepines (likely underreported due to limited general reimbursement in Denmark). ADHD: Attention-deficit/hyperactivity disorder
3d) Prescribed respiratory drugs
The model found no additional information from this variable why all values equal 0.
SABA: Short-acting beta agonist LABA: long-acting beta agonist LAMA: Long-acting muscarinic antagonist.


3a)
SHAP value
0.4
0.3
0.2
0.1
0.0
−0.1
Prescribed anticoagulants
None
VKA
Heparin+ASA
DOAC
ASA
Dipyradimol
ADP
ASA+Dipyradimol
VKA+ASA
DOAC+ASA
VKA+ADP
DOAC+ADP
VKA+Heparin
DOAC+ASA+ADP
ASA+ADP
ASA+ADP+Heparin
ASA+ADP+Dipyradimol
Age
80
75
70
65
60
55
50


3b)
SHAP value
0.4
0.3
0.2
0.1
0.0
−0.1
Prescribed cardiac drugs
None
diuretics
angiotensin-II/ACE-inhib
Ca^{2+}-antagonists
betablockers
nitrates
other AHT
other IHD-drugs
2 antihypertensives
β-blocker & 1 AHT
3 AHT
β-blocker & 2 AHT
β-blocker & 3 AHT
4 AHT
β-blocker & 4 AHT
other AHT & AHT
nitrates & any AHT
other IHD-drugs & any AHT/nitrate
other antiarrythmics & any ATH
Age
80
75
70
65
60
55
50


3c)
SHAP value
0.4
0.3
0.2
0.1
0.0
−0.1
Psychotropics
Age
80
75
70
65
60
55
50
None
SSRI/SNRI/NaRI
other AD
antipsychotics
BZ
anticholinergics/memantine
anti-ADHD
NaSSA
other psychotropics
SSRI & other AD
SSRI & NaSSA
SSRI & antipsychotics
SSRI & other psychotropics
BZ+ any psychotropic
antipsychotics + any psychotropic
anti-ADHD + any psychotropic
NaSSA + any psychotropic
other psychotropic & any specified psychotropic


3d)
SHAP value
0.4
0.3
0.2
0.1
0.0
−0.1
Respiratory drugs
None
SABA
LABA or LAMA
inhalation steroid
SABA & Ipratropium
LABA & steroid
LABA & LAMA & steroid
LAMA & steroid
LABA & LAMA
other pulmonary drug
other pulmonary drug & steroid
SABA & LABA or LAMA
SABA & LAMA or LABA & steroid

4 *Paper III: COVID-19 and Agent Based Modelling*

The following pages contain the article:

Mathias S. Heltberg, Christian Michelsen, Emil S. Martiny, Lasse E. Christensen, Mogens H. Jensen, Tariq Halasa and Troels C. Petersen (2022). "Spatial Heterogeneity Affects Predictions from Early-Curve Fitting of Pandemic Outbreaks: A Case Study Using Population Data from Denmark". In: Royal Society Open Science 9.9. issn: 2054-5703. doi: 10.1098/rsos.220018.

ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Heltberg ML, Michelsen C, Martiny ES, Christensen LE, Jensen MH, Halasa T, Petersen TC. 2022 Spatial heterogeneity affects predictions from early-curve fitting of pandemic outbreaks: a case study using population data from Denmark. *R. Soc. Open Sci.* **9**: 220018. https://doi.org/10.1098/rsos.220018

Received: 18 January 2022
Accepted: 16 August 2022

Subject Category:
Mathematics

Subject Areas:
mathematical modelling/biophysics/computational biology

Keywords:
pandemics, agent-based modelling, spatial heterogenity, fitting, COVID-19

Author for correspondence:
Mathias L. Heltberg
e-mail: heltberg@nbi.ku.dk

†These authors contributed equally.

THE ROYAL SOCIETY PUBLISHING

Spatial heterogeneity affects predictions from early-curve fitting of pandemic outbreaks: a case study using population data from Denmark

Mathias L. Heltberg[1,2,3,†], Christian Michelsen[1,†], Emil S. Martiny[1,†], Lasse Engbo Christensen[4], Mogens H. Jensen[1], Tariq Halasa[5] and Troels C. Petersen[1]

[1]Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, Copenhagen E 2100, Denmark
[2]Laboratoire de Physique, Ecole Normale Superieure, Rue Lhomond 15, Paris 07505, France
[3]Infektionsberedskab, Statens Serum Institute, Artillerivej, Copenhagen S 2300, Denmark
[4]DTU Compute, Section for Dynamical Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Anker Engelunds Vej 101A, Kongens Lyngby 2800, Denmark
[5]Animal Welfare and Disease Control, University of Copenhagen, Gronnegårdsvej 8, Frederiksberg C 1870, Denmark

MLH, 0000-0002-9699-4075; LEC, 0000-0001-5019-1931

The modelling of pandemics has become a critical aspect in modern society. Even though artificial intelligence can help the forecast, the implementation of ordinary differential equations which estimate the time development in the number of susceptible, (exposed), infected and recovered (SIR/SEIR) individuals is still important in order to understand the stage of the pandemic. These models are based on simplified assumptions which constitute approximations, but to what extent this are erroneous is not understood since many factors can affect the development. In this paper, we introduce an agent-based model including spatial clustering and heterogeneities in connectivity and infection strength. Based on Danish population data, we estimate how this impacts the early prediction of a pandemic and compare this to the long-term development. Our results show that early phase SEIR model predictions overestimate the peak number of infected and the equilibrium level by at least a factor of two. These results are robust to variations of parameters influencing connection distances and independent of the distribution of infection rates.

1. Introduction

Over the past years, the pathogen now known as SARS-CoV-2 has spread dramatically, risen in several waves, paralyzing societies, resulting in a large number of deaths and severe economic damage worldwide [1,2]. Mathematical models have estimated the reproduction number and guided the authorities in an attempt to minimize the damage caused by the virus [3–6]. Even though modern algorithms using machine learning have helped the process [7,8], the majority of models used to predict the size of the pandemic (or a rising wave of the disease) have been variants of the SIR/SEIR model. The SIR model was originally proposed in 1927, in the seminal work of Kermack and McKendrick, who successfully described the evolution of a pandemic, using a mean field approximation where all individuals are described as one population [9]. In the investigations of the SARS-CoV-2 pandemic, the mathematical models have varied in complexity including simple deterministic compartmental models [6,10], meta-population compartmental models [11–13], individual based models without including spatial specifications [4,14,15] and spatio-temporal agent-based models [16].

One aspect in the modelling is the ability to predict the infection peak height and the number of individuals who will be infected based on the early rise in the number of infected (before governmental interference). Earlier work has pointed out the importance of including heterogeneity when modelling the spread of infectious disease such as contact patterns between individuals [17], population mixing assumptions [18], heterogeneities caused by super-spreaders [15], and the spatial dependency of COVID-19 [19,20]. These mathematical models have not combined heterogeneous elements nor quantified how much the early SIR/SEIR predictions might be biased.

In this paper, we include geographical distributions based on an entire population, using population data of Denmark. When the SIR model was originally formulated, 95 years ago, data was not available to investigate the effects of geographical and demographic differences among the population, which might be one of the reasons why fundamental properties for diseases, such as the basic reproduction number (R_0), can vary significantly between different regions [21]. However, with modern collection of data, these geographical aspects might be accounted for. Our main goal of this work is therefore to investigate the importance of heterogeneities in a geographically distributed population on the spread of a pandemic. We find that the heterogeneity arising from spatial inhomogeneities causes an increase in the early stage of the pandemic, affecting the initial forecast and highlighting the importance of early intervention in order to minimize the effects of the pandemic.

1.1. Construction of the model

In order to investigate the effect of a geographically distributed population, we extracted the number of infected per commune (from the Danish Serum Institute [22]) and divided this number with the number of inhabitants in each commune to obtain the number of infected per individual in each commune. This number we then plotted against the number of inhabitants in that specific commune (extracted from statistics Denmark [23]). Doing so, we found a strong correlation between the population density and the number of infections per inhabitant as seen in figure 1*a*. This observation has been made for many other countries [24–29] and underlines the aspect of disease spreading that has been observed since ancient times; that densely populated regions often have larger pandemics than the rural areas. Note that in the very early stage of a pandemic, before the exponential growth rate is reached, micro outbreaks will guide its evolution and these events can likely take place in regions with low density [30].

1.2. Disease simulation

To simulate evolution of the disease, we assigned each individual (agent) to a state (predominantly initialized in state *S*) and assigned four states to the exposed phase and four states to the infectious phase, in order to achieve an Erlang distribution (which is related to the Gamma distribution) of time in each phase [31]. Once in the exposed phase, the infected agent has a rate to move into another state, where the rate is fixed based on experimental data in order to achieve a mean time in the exposed phase of approximately 4 days (table 1). Each agent in the Infectious phase can infect other agents that have a connection to this agent in the network. This definition of agents in discrete states is naturally a simplification of the real pandemic, and we stress that this mathematical model aims at describing the spread of the disease in a simple way that does not capture all aspects of the real disease. We do not believe that this impacts our main conclusions in any way, as we are aware that one should always be careful when making these kinds of simplifications. To investigate the effect of

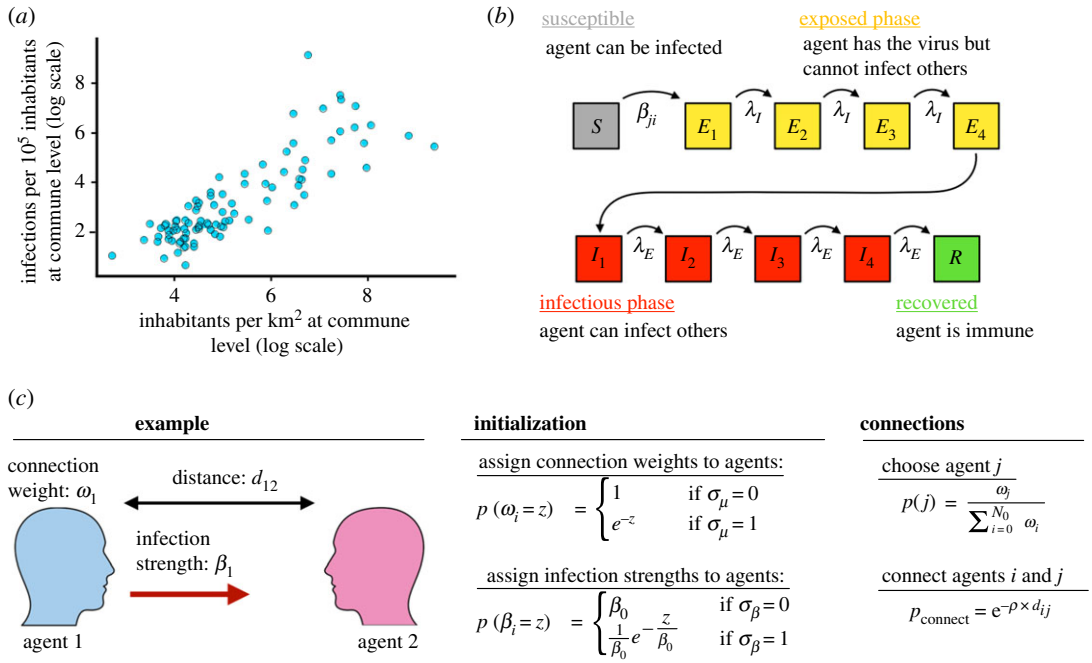


Figure 1. (*a*) Population density (*x*-axis) and the number of infections per 10^5 inhabitants (*y*-axis) for each commune in Denmark. (*b*) Illustration of the modified susceptible-exposed-infected-removed (SEIR) model used. It consists of 10 consecutive states (S, E_{1-4}, I_{1-4} and R), with transition rates governed by β, λ_E and λ_I, respectively. (*c*) Illustration of how the spatial network is generated and heterogeneities in individuals included.

infection heterogeneities, we assigned an infection strength to each connection in the network, so some agents were more infectious than others. In order to control the degree of this heteogeneity, we assigned a boolean parameter σ_β, that if switched on generated an exponential distribution in infection strengths, keeping the mean field reproduction number fixed. The reproduction number between the ABM and the SIR model is related through the parameter $\tilde{\beta} = \beta(\mu/2N_0)$. All transitions between states and infection of other individuals were done using the Gillespie algorithm [32]. This is schematized in figure 1*b*.

1.3. Network creation

In order to construct the underlying network, we created a set-up whereby two agents were chosen at random but based on their individual connectivity weight each iteration and connected with some probability based on their spatial position. To include the possibility of highly connected individuals independent of their spatial position, we assigned a boolean parameter σ_μ that, if switched on, generated an exponential distribution in weights for the individuals, keeping the mean field reproduction number fixed similar to the heterogeneity in infection strengths. To include the spatial position in the network, we introduced a parameter ρ, so the probability of connecting two chosen agents decayed exponentially with the distance between them: $p_{\text{connect}} = \mathrm{e}^{-\rho \times d_{ij}}$. In order to allow some long-distance connections we introduced another parameter $\varepsilon \in [0; 1]$, that determines the fraction of distance-independent contacts. To construct the network of spatially distributed contacts, we chose the parameters using data based on:

— The geographical location of people in Denmark (from Boligsiden [33])
— The average number of contacts per individual per day of 11 (from HOPE [34]). Given an average infectious period of 4 days, we approximate the average number of effective contacts to be $\mu = 40$
— The average commuting distance $\rho = 0.1$ km^{-1} and the fraction of long-distance commutes $\epsilon_\rho = 4\%$ (from statistics Denmark [23])

This is schematized in figure 1*c* and further described in the Methods section. All 10 parameters in this model are defined and outlined in table 1. We note that in order to keep the parameters space low, this model does not include the effects of temporal changes such as seasonality and holidays. While all agents

Table 1. Overview of the 10 parameters applied in this study, their typical value, and the ranges we have considered. The first six parameters are standard SEIR parameters, whereas the last four parameters define the heterogeneity in the model. These four parameters do not affect the SEIR model.

variable	description	value	range	units
N_0 :	population size	5.8×10^6	10^5–10^7	—
N_{init} :	number of individuals initially infected	100	1–10^4	—
μ :	average number of network contacts	40	10–100	—
β :	typical infection strength	0.01	0.001–0.1	d^{-1}
λ_E :	rate to move through $\frac{1}{4}$ of latency period	1	0.5–4	d^{-1}
λ_I :	rate to move through $\frac{1}{4}$ of infectious period	1	0.5–4	d^{-1}
σ_μ :	population clustering spread	0	0–1	—
σ_β :	interaction strength spread	0	0–1	—
ρ :	typical acceptance distance	0.1	0–0.5	km^{-1}
ϵ_ρ :	fraction of distance-independent contacts	0.04	0–1	—

have been assigned parameters to their infection network that are derived from statistics of Denmark for both employees and students, we have not divided each agent into specific occupations.

Before including heterogeneity, we compared the ABM to the corresponding SEIR model as a test, and found them to agree within 5% for all parameter configurations tested. Here, we also tested the effect of the number of individuals initially infected (see electronic supplementary material). This concludes that the SEIR and ABM model are calibrated to have the same reproduction number in the absence of heterogeneities. Next, we will introduce heterogeneities into the system, while keeping the sum of contacts and infection strengths constant, to study how this affects the evolution of the pandemic.

2. Results

2.1. Geographical distributions in a population and large variances in numbers of contacts leads to increased infection levels

Having introduced heterogeneity, the distributions of connections in this network were created automatically through the population clustering, see figure 2*a*. This naturally leads to individuals living in densely populated areas having higher number of connections. In an example simulation with 100 initially infected individuals, $N_{\text{init}} = 100$, we observed a spatial difference in areas affected by the disease (figure 2*b*), as expected. Note that we also show the effective reproduction number ($\mathcal{R}_{\text{eff}}$) as a function of time in the lower part of the inserted panel. One region reached local endemic steady state (green arrow, figure 2*b*) while other regions of similar density were highly infected (red arrow, figure 2*b*) and yet other districts were almost unaffected (grey arrow, figure 2*b*). To quantify the effect of population clustering, we compared the ABM result to the reference SEIR model of similar parameters. Generally, we observed that the epidemic developed faster with a higher infection peak I_{peak}, but also subsided quicker, leading to a lower number of infected once reaching endemic steady state, R_∞ (figure 2*c*,*d*).

In order to explore how population clustering affects the epidemic, we chose a reference value of infection rates, $\beta = 0.01$, and an alternative value of $\beta = 0.007$. In the absence of spatial dependence ($\rho = 0\ \text{km}^{-1}$), these correspond to initial reproduction numbers $\mathcal{R}_0 \approx 1.7$ and 1.1, respectively. Here, we define the reproduction number as the average number of agents each infectious agent will infect in the first part of the disease. Increasing the spatial dependence (i.e. increasing ρ) leads to a significant rise in the infection peak for the ABM, $I_{\text{peak}}^{\text{ABM}}$, compared to the (unaffected) SEIR model, $I_{\text{peak}}^{\text{SEIR}}$ for both the reference value and the alternative lower value of β (black and blue points, figure 2*e*). We introduced heterogeneity in infection strengths ($\sigma_\beta = 1$, see figure 1*b*), thus making some individuals much more infectious than others (i.e. including *super shedders*). We found no significant impact from this effect (red points in figure 2*e*). Similarly, we introduced heterogeneity in connection weights ($\sigma_\mu = 1$, see figure 1*b*), thus making some individuals much more likely to form contacts than others

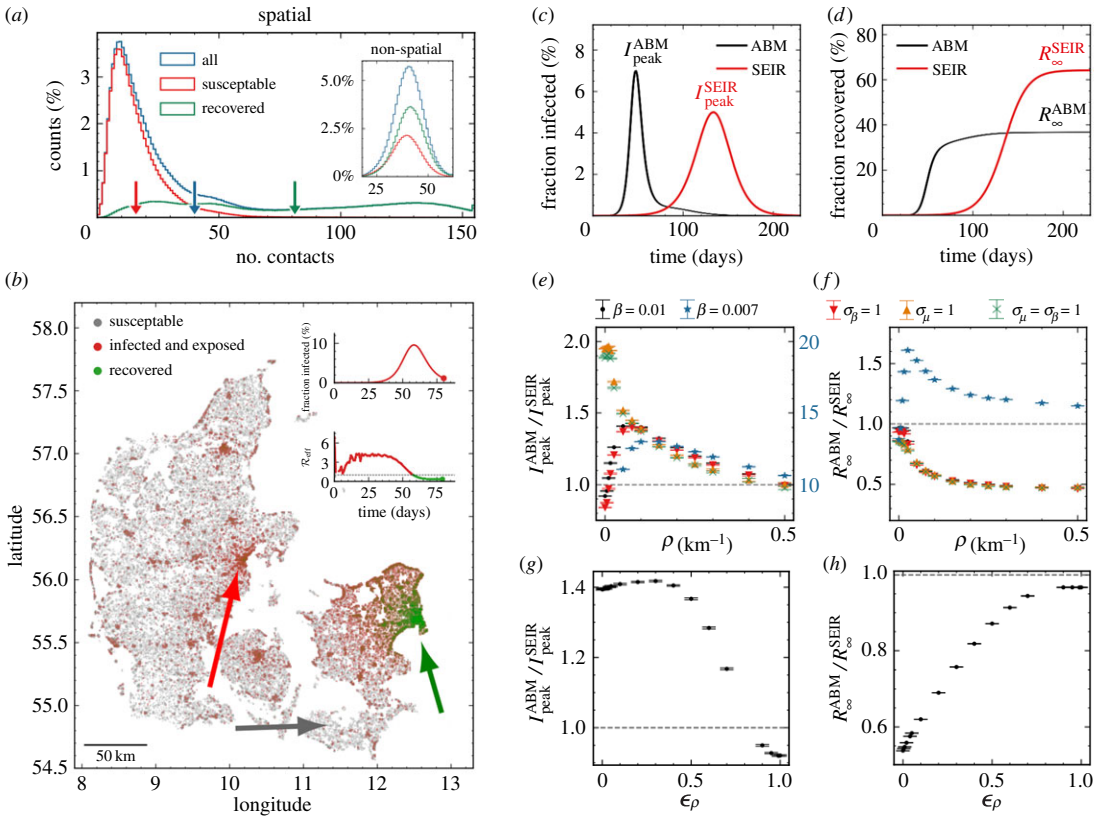


Figure 2. (*a*) Histograms showing the number of susceptible (red) and recovered (green) individuals at the end of an epidemic with $\rho = 0.1$ km^{-1}. The distribution before the epidemic is shown in blue. The arrows show the mean of each distribution. The inset shows the same for $\rho = 0$ km^{-1}. (*b*) Visualization of the spatial position of individuals during the infection and which state they are in. Green arrow: largest city in Denmark (Copenhagen): mostly recovered. Red arrow: Second largest city in Denmark (Aarhus): mostly infected. Grey arrow: low-population area: mostly susceptible (i.e. have not been infected). (*c*) Number of infected individuals as a function of time. Data shown for the spatially distributed network ($\rho = 0.1$ km^{-1}). Simulation was repeated 10 times. (*d*) Cumulative sum of individuals who have had the disease as a function of time (with $\rho = 0.1$ km^{-1}). (*e*) Relative difference in maximal number of infected, I_{peak}, between deterministic (SEIR) and ABM as a function of ρ, and shown for different parameters. Note the data for $\beta = 0.007$ are shown in blue with a factor 10 scaling (right *y*-axis). (*f*) Relative difference in total number of infected at the end of the epidemic, R_∞, between deterministic (SEIR) and ABM as a function of ρ. Colours similar to (*e*). (*g*) Same as (*e*), but as a function of ϵ_ρ. (*h*) Same as (*f*), but as a function of ϵ_ρ.

(i.e. including *super connecters*). This leads to a significant effect for $\rho = 0$ km^{-1}, which converges towards the other curves for $\rho > 0.1$ km^{-1} (orange (only super connecters) and green (super connecters and super shedders) points in figure 2*e*). The total number of individuals that have been in the infectious state, when there are not enough susceptible agents for the disease to keep infecting new individuals, is termed R_∞, and this converged towards half of the SEIR model prediction as a function of ρ except for $\beta = 0.007$ where the endemic steady state level is larger than the one obtained by the SEIR model (figure 2*f*). We note that in reality, individuals can lose immunity and therefore new waves can emerge. But for a completely susceptible population, R_∞ describes the fraction of the population that will get the disease during a specific wave. Fixing $\rho = 0.1$ km^{-1} and increasing the fraction of distance-independent contacts, ϵ_ρ, we found that $I_{\text{peak}}^{\text{ABM}}$ is almost unaffected for $\epsilon_\rho < 0.5$ (figure 2*g*), while R_∞^{ABM} increases linearly towards the SIER model R_∞^{SEIR}, as expected (figure 2*h*).

2.2. Fitting early infection curves leads to significant bias in estimating the size of the pandemic

Next, we consider how these heterogeneities bias the traditional SEIR model predictions, especially the predictions based on fits to the number of infected (i.e. the curve to be flattened) in the beginning of the epidemic (see Methods). Without spatial dependence, the predicted curves fitted the number of infected

individuals very well (figure 3*a*). Introducing spatial dependence ($\rho = 0.1\ \mathrm{km}^{-1}$) leads to a severe overestimation of the epidemic based on the number of early infection cases (figure 3*b*). This result can be interpreted by the fact that in societies where population density and thus individual contact number varies significantly, the early phase will be driven by people with many contacts (*super connecters*). This typically happens in cities where the population density is high. Increasing the spatial dependence ρ, we found that the SEIR model predictions overestimated the infection peak height I_{peak} and the total number of infected R_∞ significantly even for very small spatial heterogeneities (figure 3*c*, *d*). We observed this general trend for all tested combinations of parameters and heterogeneities. In particular, we found that if long-distance connections ϵ_ρ are below 10%, the bias in the estimated infection peak height, I_{peak}, was constant within statistical uncertainty (figure 3*e*). For the total number of infected, R_∞, we observed an almost linear regression to the SEIR model as ϵ_ρ approaches one. However, even when $\epsilon_\rho = 0.25$, the prediction bias was still a factor of two (figure 3*f*). We concluded from these curves a general trend; if one fits an SEIR model to infection numbers during the beginning of an epidemic, and use these estimates to predict the characteristics of the epidemic at a national level, one overestimates the number of infected by at least a factor of two.

3. Discussion

In summary, this work outlines that the degree of population clustering in Denmark creates a discrepancy between the early predictions made by the SEIR models and the underlying agent-based interactions. It results in a significant overestimation of the impact of the disease, both in terms of maximal number of simultaneously infected (by a factor of 3) and the endemic steady state level (by a factor of 2.5). Such discrepancies have been observed for earlier pandemics, for instance, the 1918 Spanish flu, where the predicted number of individuals that would get the disease within a season turned out to be higher than the actual outcome [35]. The present results can be an important element in explaining these mismatches, even though other elements, such as for instance social distancing and the population behaviour, play a vital part. When facing a rising pandemic, societies are faced with the task of laying out strategies to minimize the consequences, including the importance of *flattening the curve*. While this is truly crucial to avoid overpopulated hospitals, the understanding of the pandemic should be taken seriously enough that we might specify to a higher degree of certainty which curve to be flattened. Our results highlight an important element in the prediction of infection levels and quantify the effect of density heteogeneities. We are aware that these results are not directly applicable to the pandemic of SARS-CoV-2 as a whole, since numerous mutations have increased the infection rates compared to the early estimates and created a strong heterogeneity in the infection worldwide. Furthermore, the actual evolution of the pandemic was highly affected by the different governmental interventions, that are not included in this work. However, this study emphasizes the abnormally large reproduction rates in the beginning of a pandemic, due to individuals with more connections than the rest of the population and attempts to quantify this bias, when countries should estimate the severity of a disease based on the data collected in the early phase. This also underlines the benefits by making lockdowns early in the pandemic, when a population is highly susceptible (for instance to a new mutation) and therefore can be driven by *super connectors*. Since people living in city-clusters are more likely to have many contacts, or infection events, they are on average more likely to be affected in the early stage of the pandemic (if they do not implement social distancing). By removing contacts from these individuals, through some level of interaction in order to reduce the number of social contacts, one can avoid the worst peak while affecting the lowest number of people. While our work describes some fundamental aspects of the disease spreading, this model does not consider asymptomatic individuals, which has been an important aspect of the SARS-CoV-2 pandemic [36,37]. Effectively, asymptotic individuals would correspond to a very heterogenous distribution of time the agents spend in the infectious state. While agents with symptoms would predominantly isolate themselves and thereby significantly reduce their ability to infect other agents, asymptomatic agents would have a long time in the infectious state, thereby infecting more individuals. In this work, we have not considered the observation that individuals lose their immunity to SARS-CoV-2 which was first studied in the Brazilian city of Manaus. For this model, the temporal decline of immunity would lead to more pandemic 'waves', but for a fixed disease transmissibility this would not alter the maximal height of the peak number of infected, since this occurred for all the initially susceptible population. Finally, we note that this work does not include a vast range of divisions for the population, including age, socio-economic status etc. We have not included this directly, since we wanted to estimate as cleanly as possible how the heterogeneity in the

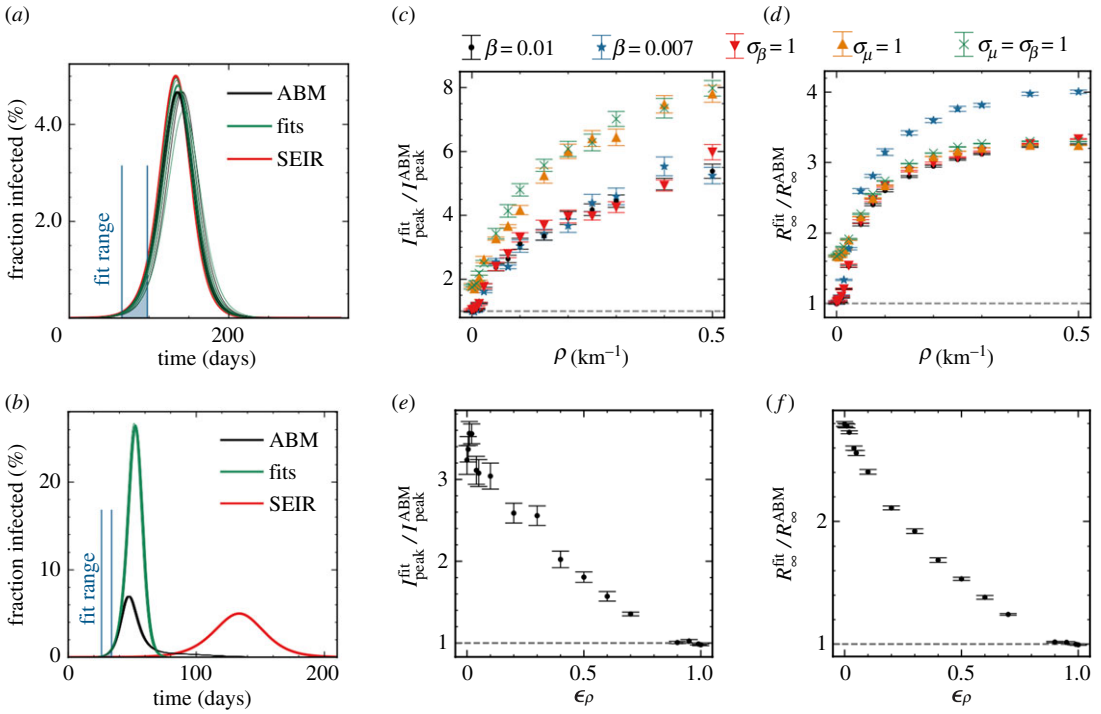


Figure 3. (*a*) Number of infected individuals for the ABM in black, the SEIR model in red and the SEIR fits to the ABM data in green. Blue lines show the interval where parameters are fitted (also shown below the curves). Here, $\rho = 0$ km^{-1}. (*b*) Same as (*a*) but with population clustering ($\rho = 0.1$ km^{-1}). (*c*) Relative difference in maximal number of infected, I_{peak}, between the fit and the ABM for different values of ρ. Simulations repeated 10 times for each data-point. (*d*) Relative difference in total number of infected at the end of the epidemic, R_∞, between the fit and the ABM for different values of ρ. (*e*) Same as (*c*), but as a function of ϵ_ρ. (*f*) Same as (*d*), but as a function of ϵ_ρ.

contact pattern, arising from a geographically distributed population, could affect the evolution of a disease. We are aware that for instance the distribution of age has an enormous impact on the health risk and that this risk is vital in the prediction of hospitalizations in modern society. However, our aim was to understand the bias in the prediction of a disease, based on the data that comes during the early periods of a disease, independently of the mortality of this disease. Mathematical predictions of disease progression have been heavily criticized [38,39] and it is important to improve the theoretical foundations of the mathematical descriptions, in order to increase the confidence in the predictions. Our work highlights the importance of estimating the spatial clustering and connectivity skewness in the population in order to correct the predictions based on SEIR models, by quantifying their biases from not including spatial clustering. We hope that this work could serve as an input to the modelling and prediction of future pandemics and the importance of avoiding super-spreaders in high-density areas.

3.1. Methods

3.1.1. Construction of spatial network

We initialized N_0 agents on a network generating a total of $\mu \times N_0$ links between two agents, with an assigned interaction strength β_{ij} for each link. The average contact number, μ, was fixed to 20, based on results from the Danish HOPE project, gathering data on population behaviour since April 2020 [34]. In order to include a realistic, geographical distribution of the population, we randomly selected agent locations from a two-dimensional kernel density estimate we had generated based on housing sales in Denmark 2007–2019 (data given with permission from Boligsiden, [33]). We note that in this distribution, we do not take specific geographical elements such as roads or environment into account (which has been previously studied for other diseases [40]) as we assume that this effect is small in a country like Denmark, where all parts are connected and natural obstacles such as mountains and rivers are not present. To connect the agents, we used a hit and miss method, where two random agents are first suggested and then connected with probability, $p(d) = e^{-\rho \cdot d_{ij}}$. Here, d_{ij} is the distance between agents and

ρ is a parameter with units of inverse distance. We choose $\rho = 0.1\ \mathrm{km}^{-1}$ (i.e. 10 km) which is the average distance travelled by labour force (statistics Denmark [23]). To allow some long-distance interactions, we introduced a parameter $\epsilon_\rho = 4\%$ representing the fraction of distance-independent connections. This value we based on the fraction of workers travelling longer than 50 km to work (statistics Denmark [23]).

3.1.2. Fits and predictions

We defined an early phase to be the period of time when between 0.1% and 1% of the population were infected (blue lines figure 3*a*). We then fitted β and a time delay, τ, to the SEIR model with a χ^2-fit (assuming Poissonian statistics) and kept λ_E and λ_I fixed to the true numbers (used in the simulation). The initial number of infected, N_{init}, was also fixed to the true numbers. The fit parameters were then inserted into the SEIR model, and $I_{\mathrm{peak}}^{\mathrm{fit}}$ and R_∞^{fit} were extracted from the fitted model and compared to the $I_{\mathrm{peak}}^{\mathrm{ABM}}$ and R_∞^{ABM} from the ABM simulation.

Data accessibility. Data and relevant code for this research work are stored in GitHub: www.github.com/ChristianMichelsen/NetworkSIR and have been archived within the Zenodo repository: https://zenodo.org/badge/latestdoi/258223118.

Authors' contributions. C.M.: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; E.S.M.: investigation, software, validation, visualization, writing—review and editing; L.E.C.: supervision, validation, writing—review and editing; T.C.P.: conceptualization, investigation, methodology, project administration, software, supervision, validation, visualization, writing—original draft, writing—review and editing; M.L.H.: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; M.H.J.: formal analysis, investigation, supervision, validation, writing—review and editing; T.H.: conceptualization, investigation, supervision, validation, visualization, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare that we have no competing interests.

Funding. M.L.H. acknowledges the Carlsberg Foundation grant no. CF20-0621 and the Lundbeck Foundation grant no. R347-2020-2250. E.S.M. and M.H.J. acknowledge support from the Independent Research Fund Denmark grant no. 9040-00116B and Danish National Research Foundation through StemPhys Center of Excellence, grant no. DNRF116.

Acknowledgements. The authors are grateful to the Danish expert group of SARS-CoV-2 modelling led by Statens Serum Institute, especially Robert L. Skov, Kåre Mølbak, Camilla Holten Møller, Viggo Andreasen, Kaare Græsbøl, Theis Lange, Carsten Kirkeby, Frederik P. Lyngse, Matt Denwood, Jonas Juul, Sune Lehman, Uffe Thygesen and Laust Hvas Mortensen. Furthermore, we thank Kim Sneppen for valuable discussions.

References

1. Chinazzi M *et al.* 2020 The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID–19) outbreak. *Science* **368**, 395–400. (doi:10.1126/science.aba9757)
2. WHO: see www.who.int/news-room/detail/27-04-2020-who-timeline-covid-19 (accessed 29 September 2020).
3. Anderson RM, Heesterbeek H, Klinkenberg D, Hollingsworth TD. 2020 How will country-based mitigation measures influence the course of the COVID–19 epidemic? *Lancet* **395**, 931–934. (doi:10.1016/S0140-6736(20)30567-5)
4. Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, Flasche S. 2020 Feasibility of controlling COVID–19 outbreaks by isolation of cases and contacts. *Lancet Global Health* **8**, e488–e496. (doi:10.1016/S2214-109X(20)30074-7)
5. Keeling MJ, Hollingsworth TD, Read JM. 2020 Efficacy of contact tracing for the containment of the 2019 novel coronavirus (COVID-19). *J. Epidemiol. Commun. Health* **74**, 861–866. (doi:10.1101/2020.02.14.20023036)
6. Kuniya T. 2020 Prediction of the epidemic peak of coronavirus disease in Japan, 2020. *J. Clin. Med.* **9**, 789. (doi:10.3390/jcm9030789)
7. Ghafouri-Fard S, Mohammad-Rahimi H, Motie P, Minabi MA, Taheri M, Nateghinia S. 2021 Application of machine learning in the prediction of COVID-19 daily new cases: a scoping review. *Heliyon* **7**, e08143. (doi:10.1016/j.heliyon.2021.e08143)
8. Fokas AS, Dikaios N, Kastis GA. 2020 Mathematical models and deep learning for predicting the number of individuals reported to be infected with SARS-CoV-2. *J. R. Soc. Interface* **17**, 20200494. (doi:10.1098/rsif.2020.0494)
9. Kermack WO, McKendrick AG. 1927 A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A* **115**, 700–721. (doi:10.1098/rspa.1927.0118)
10. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J. 2020 Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493. (doi:10.1126/science.abb3221)
11. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, Abbott S. 2020 The effect of control strategies to reduce social mixing on outcomes of the COVID–19 epidemic in Wuhan, China: a modelling study. *Lancet Public Health* **5**, e261–e270. (doi:10.1016/S2468-2667(20)30073-6)
12. van Bunnik BA, Morgan AL, Bessell P, Calder-Gerver G, Zhang F, Haynes S, Lepper HC. 2020 Segmentation and shielding of the most vulnerable members of the population as elements of an exit strategy from COVID–19 lockdown. *medRxiv* (doi:10.1101/2020.05.04.20090597).
13. Danon L, Brooks-Pollock E, Bailey M, Keeling MJ. 2020 A spatial model of COVID–19 transmission in England and Wales: early spread and peak timing. medRxiv (doi:10.1101/2020.02.12.20022566).
14. Chang SL, Harding N, Zachreson C, Cliff OM, Prokopenko M. 2020 Modelling transmission and control of the COVID-19 pandemic in Australia. *Nat. Commun.* **11**, 1–13. (doi:10.1038/s41467-020-19393-6)
15. Sneppen K, Nielsen BF, Taylor RJ, Simonsen L. 2021 Overdispersion in COVID-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control. *Proc. Natl Acad. Sci. USA* **118**, e2016623118. (doi:10.1073/pnas.2016623118)
16. Milne GJ, Xie S. 2020 The effectiveness of social distancing in mitigating COVID–19 spread: a

modelling analysis. *medRxiv* (doi:10.1101/2020.03.20.20040055).

17. Bansal S, Grenfell BT, Meyers LA. 2007 When individual behaviour matters: homogeneous and network models in epidemiology. *J. R. Soc. Interface* **4**, 879–891. (doi:10.1098/rsif.2007.1100)
18. Kong L, Wang J, Han W, Cao Z. 2016 Modeling heterogeneity in direct infectious disease transmission in a compartmental model. *Int. J. Environ. Res. Public Health* **13**, 253. (doi:10.3390/ijerph13030253)
19. Kang D, Choi H, Kim JH, Choi J. 2020 Spatial epidemic dynamics of the COVID–19 outbreak in China. *Int. J. Infect. Dis.* **94**, 96–102. (doi:10.1016/j.ijid.2020.03.076)
20. Giuliani D, Dickson MM, Espa G, Santi F. 2020 Modelling and predicting the spatio-temporal spread of COVID-19 in Italy. *BMC Infect. Dis.* **20**, 1–10. (doi:10.1186/s12879-020-05415-7)
21. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. 2019 Complexity of the basic reproduction number (R0). *Emerg. Infect. Dis.* **25**, 1–4. (doi:10.3201/eid2501.171901)
22. Danish Serum Institute: www.ssi.dk (accessed 29 September 2020).
23. Statistics Denmark: www.statistikbanken.dk (accessed 29 September 2020).
24. Wong DW, Li Y. 2020 Spreading of COVID-19: density matters. *PLoS ONE* **15**, e0242398. (doi:10.1371/journal.pone.0242398)
25. Ganasegeran K, Jamil MFA, Ch'ng ASH, Looi I, Peariasamy KM. 2021 Influence of population density for COVID-19 spread in Malaysia: an ecological study. *Int. J. Environ. Res. Public Health* **18**, 9866. (doi:10.3390/ijerph18189866)
26. Kodera S, Rashed EA, Hirata A. 2020 Correlation between COVID-19 morbidity and mortality rates in Japan and local population density, temperature, and absolute humidity. *Int. J. Environ. Res. Public Health* **17**, 5477. (doi:10.3390/ijerph17155477)
27. Bhadra A, Mukherjee A, Sarkar K. 2021 Impact of population density on COVID-19 infected and mortality rate in India. *Model. Earth Syst. Environ.* **7**, 623–629. (doi:10.1007/s40808-020-00984-7)
28. Chen K, Li Z. 2020 The spread rate of SARS-CoV-2 is strongly associated with population density. *J. Travel Med.* **27**, taaa186. (doi:10.1093/jtm/taaa186)
29. Martins-Filho PR. 2021 Relationship between population density and COVID-19 incidence and mortality estimates: a county-level analysis. *J. Infect. Public Health* **14**, 1087–1088. (doi:10.1016/j.jiph.2021.06.018)
30. Hittner JB, Fasina FO, Hoogesteijn AL, Piccinini R, Maciorowski D, Kempaiah P, Rivas AL. 2021 Testing-related and geo-demographic indicators strongly predict COVID-19 deaths in the united states during March of 2020. *Biomed. Environ. Sci.* **34**, 734–738.
31. Huang S, Li J, Dai C, Tie Z, Xu J, Xiong X, Lu C. 2021 Incubation period of coronavirus disease 2019: new implications for intervention and control. *Int. J. Environ. Health Res.* **32**, 1707–1715. (doi:10.1080/09603123.2021.1905781)
32. Gillespie DT. 1977 Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361. (doi:10.1021/j100540a008)
33. Boligsiden: www.boligsiden.dk (accessed 29 September 2020).
34. HOPE project: www.hope-project.dk (accessed 29 September 2020).
35. Andreasen V, Viboud C, Simonsen L. 2008 Epidemiologic characterization of the 1918 influenza pandemic summer wave in Copenhagen: implications for pandemic control strategies. *J. Infect. Dis.* **197**, 270–278. (doi:10.1086/524065)
36. Arcede JP, Caga-Anan RL, Mentuda CQ, Mammeri Y. 2020 Accounting for symptomatic and asymptomatic in a SEIR-type model of COVID-19. *Math. Model. Nat. Phenomena* **15**, 34. (doi:10.1051/mmnp/2020021)
37. Guan J, Zhao Y, Wei Y, Shen S, You D, Zhang R, Chen F. 2022 Transmission dynamics model and the coronavirus disease 2019 epidemic: applications and challenges. *Med. Rev.* **2**, 89–109. (doi:10.1515/mr-2021-0022)
38. Holmdahl I, Buckee C. 2020 Wrong but useful—what COVID–19 epidemiologic models can and cannot tell us. *N. Engl. J. Med.* **383**, 303–305. (doi:10.1056/NEJMp2016822)
39. Wynants L, Van Calster B, Bonten MM, Collins GS, Debray TP, De Vos M, Schuit E. 2020 Prediction models for diagnosis and prognosis of COVID–19 infection: systematic review and critical appraisal. *BMJ* **369**, m1328. (doi:10.1136/bmj.m1328)
40. Rivas AL, Fasina FO, Hoogesteyn AL, Konah SN, Febles JL, Perkins DJ, Smith SD. 2012 Connecting network properties of rapidly disseminating epizoonotics. *PLoS ONE* **7**, e39778. (doi:10.1371/journal.pone.0039778)

5 *Paper IV: Bayesian Inference and Diffusion*

The following pages contain the article:

Susmita Sridar, Mathias S. Heltberg, Christian S. 6 Michelsen Judith M. Hattab, Angela Taddei (2022). “Microscopic single molecule dynamics suggest underlying physical properties of the silencing foci”.

Microscopic single molecule dynamics suggest underlying physical properties of the silencing foci

Susmita Sridar [1 †] ✉, **Mathias Spliid Heltberg** [2 †] ✉, **Christian Stentoft Michelsen** [2 †], **Judith Mine Hattab** [1], **Angela Taddei** [1]

[1]Institut Curie, PSL University, Sorbonne Universite, CNRS, Nuclear Dynamics, Paris, France; [2]Niels Bohr Institute, University of Copenhagen

✉ **For correspondence:** mathias.heltberg@nbi.ku.dk (MH)

[†]Authors contributed equally.

Present address: Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark

Data availability: Data availability is available on Zenodo or the Github repository.

Funding: This work was supported by XXX Foundation. The funders had no role in the decision to publish.

Competing interests: The author declare no competing interests.

Abstract

In order to obtain fine-tuned regulation of protein production while maintaining cell integrity, it is of fundamental importance to living organisms to express a specific subset of the genes abailable in the genome. One way to achieve this is through the formation of subcompartments in the nucleus, known as foci, that can form at various locations on the DNA fibers and repress the transcriptional activity of all genes covered. In this work we investigate the physical nature of such foci, by applying single molecule microscopy in living cells. Here we study the motion of the protein SIR3. By combining various statistical methods, and combining a frequentist with a bayesian approach, we extract the diffusion properties for motion in a repair foci. In order to obtain useful information based on this, we derive similar measures for the foci itself, the motion of SIR3 outside the foci and other mutants of the cell. We reveal that the behaviour inside a repair foci is highly immobile and we compare this to theoretical expressions. Based on this we hypothesize that the repair foci is probably not a result of a second order liquid-liquid phase separation but rather a so-called Polymer Bridgng Model with numerous binding sites.

1 | INTRODUCTION

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris. Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

2 | METHODS & MATERIALS

2.1 | Diffusion model

For each of the different types of data (XXX), we load in the cells and group them by cell number and ID. For each group we compute the distance Δr between the subsequent observations $\vec{x}_i$:

$$\Delta r_i = \|\vec{x}_{i+1} - \vec{x}_i\|. \tag{1}$$

E.g., for Wild Type 1, we find 914 groups across 43 different cells, leading to a total of $N = 10.025$ distances. We model the diffusion distances with a Rayleigh likelihood, where the Rayleigh distribution is given by:

$$\text{Rayleigh}(x;\sigma) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}, \quad x > 0. \tag{2}$$

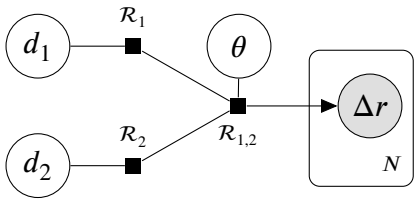


Figure 1. A graphical representation of the Bayesian model case of two diffusion components using the directed factor graph notation (Dietz, 2022). Here d_1 is the diffusion coefficient, $\mathcal{R}_1$ is the d-parameterized Rayleigh distribution and $\mathcal{R}_{1,2}$ is the mixture model of the Rayleigh distributions with a θ prior.

In this study, we parameterize the Rayleigh distribution in terms of the diffusion coefficient d, which is related to the scale parameter σ in eq. (2), through the XXX parameter, τ:

$$\sigma = \sqrt{2d\tau}, \tag{3}$$

with $\tau = 0.02$ in the current study. In the simplest form, where we assume only a single diffusion coefficient, d, the Bayesian model for this process is:

$$\begin{aligned}
&[d \text{ prior}] && d \sim \text{Exponential}(0.1) \\
&[\text{transformation}] && \sigma = \sqrt{2d\tau} \\
&[\text{likelihood}] && \Delta r_i \sim \text{Rayleigh}(\sigma).
\end{aligned} \tag{4}$$

A more realistic diffusion model include more than a single diffusion coefficient. Figure 1 shows this for the two-component case in directed factor graph notation (Dietz, 2022). In particular, the figure shows the combination of the $K = 2$ diffusion coefficients d_k through a mixture model $\mathcal{R}_{1,2}$ of the two d-parameterized Rayleigh distributions $\mathcal{R}_k$ with a θ-prior. We model each of the distances as independent, indicated by the N-replications plate. In equations, the figure is similar to:

$$\begin{aligned}
&[d_1 \text{ prior}] && d_1 \sim \text{Exponential}(0.1) \\
&[d_2 \text{ prior (ordered)}] && d_2 \sim \text{Exponential}(0.1), \quad d_1 < d_2 \\
&[\bar{\theta} \text{ prior}] && \theta_1 \sim \text{Uniform}(0,1), \quad \bar{\theta} = \left[\theta_1, 1-\theta_1\right] \\
&[\text{mixture model}] && \mathcal{R}_{1,2}(d_1, d_2, \bar{\theta}) = \text{MixtureModel}\left(\left[\mathcal{R}(d_1), \mathcal{R}(d_2)\right], \bar{\theta}\right) \\
&[\text{likelihood}] && \Delta r_i \sim \mathcal{R}_{1,2}(d_1, d_2, \bar{\theta}).
\end{aligned} \tag{5}$$

2.2 | Model comparison

We can generalize the $K = 2$ diffusion model to higher values of K by having $d_1, \ldots, d_K$ (ordered such that $d_1 < d_k < d_K$ to prevent the classical label-switching problem in the case of mixture mod-

els (McLachlan and Peel, 2004)) diffusion coefficients and letting the mixture model's $\bar{\theta}$-prior be a random variable from a flat Dirichlet distribution (such that $\sum_k \theta_k = 1$). We find that including up to three diffusion coefficients yields appropriate results. To compare the three models of different complexity, we compute the Widely Applicable Information Criterion (WAIC) (Watanabe, 2010) which is a generalized version of the Akaike information criterion (AIC) useful for Bayesian model comparison (Gelman, Hwang, and Vehtari, 2014). In short, the WAIC is an approximation of the out-of-sample performance of the model and consists of two terms, the log-pointwise-predictive-density, lppd, and the effective number of parameters p_{WAIC}:

$$\text{WAIC} = -2\left(\text{lppd} - p_{\text{WAIC}}\right). \tag{6}$$

The lppd is the Bayesian version of the accuracy of the model and p_{WAIC} is a penalty term related to the risk of over-fitting; complex models (usually) have higher values of p_{WAIC} than simple models, (McElreath, 2020). The minus 2 factor is just a scaling included for historical reasons leading to low WAICs being better. Given two models, A and B, we compute both the individual WAIC values, W_A and W_B, their standard deviations, σ_{W_A} and σ_{W_B}, their difference, $\Delta_{A,B}$, and the standard error of their difference, $\sigma_{\Delta_{A,B}}$.

2.3 | MSD and energy

After choosing the optimal model, we extract the slow diffusion coefficient from the model, d_{slow}, and use this to compute the mean squared displacement (MSD) for the groups with a mean diffusion $D = \langle \frac{\Delta r^2}{4\tau} \rangle$ being slow, where slow is defined as $D < d_{\text{slow}} + 3\sigma_{\text{slow}}$. From the MSD, we can either infer the full XXX (Mathias) model, based on XXX equation:

$$4\sigma^2 + R_\infty^2\left(1 - \exp\left(-\frac{4dx}{R_\infty^2}\right)\right) \tag{7}$$

or simply approximate the `DCon2_WT1` (XXX) with half of the slope of the first three data points of the MSD (Mathias, why half?).

We can compute the energy, U, in two different ways; U_{left} and U_{right}. The first method is based on a geometric calculation depending on the fraction of the slow diffusion coefficient from the Wild

Type 1 calculation, $\theta_{\text{slow}}^{\text{WT}_1} \equiv \theta_1^{\text{WT}_1}$:

$$
\begin{aligned}
V_{\text{cap}} &= \frac{\pi h^2}{3(3r_0 - h)} \\
V_0 &= \frac{4\pi}{3 - 2V_{\text{cap}}} \\
V_F &= \frac{8V_0}{4\pi/3} \frac{4\pi}{3R_R^3} \\
U_{\text{left}} &= -\log\left(\theta_{\text{slow}} \frac{V_0 - V_F}{(1 - \theta_{\text{slow}}^{\text{WT}_1})V_F}\right),
\end{aligned}
\tag{8}
$$

where $r_0 = 1.0$, $h = 0.85$, and $R_R = 0.13$.

The other energy, U_{right}, can be calculated from the value of `DCon2_WT1` (half slope) from Wild Type 1, the `Db_focus` (half slope) from the Focus files, and the fast diffusion coefficient from the delta files: $\theta_{\text{fast}}^{\text{delta}} \equiv \theta_2^{\text{delta}}$:

$$
U_{\text{left}} = \log\left(\frac{\texttt{DCon2\_WT1} - \texttt{Db\_focus}}{\theta_{\text{fast}}^{\text{delta}} - \texttt{Db\_focus}}\right). \tag{9}
$$

2.4 | Implementation

The data analysis has been carried out in Julia (Bezanson et al., 2017) and the Bayesian models are computed using the Turing.jl package (Ge, Xu, and Ghahramani, 2018). We use Hamiltonian Monte Carlo sampling (Betancourt, 2018) with the NUTS algorithm (Hoffman and Gelman, 2011). In particular, each Bayesian model have been run with 4 chains, each chain 1000 iterations long after discarding the initial 1000 samples ("warm up").

3 | RESULTS

3.1 | Dynamics of SIR3 reveals two dominating populations of the motion

We started out by investigating the mobility of individual SIR3 molecules in vivo. Here we typically have 5-8 repair foci. To image SIR3 without changing its normal level, we generated haploid cells expressing the endogenous SIR3 fused to Halo (Figure 1A). Before we wanted to visualize the cells on a PALM microscope (see Materials and methods), we incubated the exponentially growing cells with fluorescent and fluorogenic JF647, a dye emitting light only once bound to SIR3. We were very used a low concentration of JF646 allowing for the observation of individual molecules (Ranjan et al., 2020; Figure 1— figure supplement 2). Rad52-Halo bound to JF646 (Rad52-Halo/JF646) were visualized at 20 ms time intervals (50 Hz) in 2-dimensions during 1000 frames until no signal was visible.

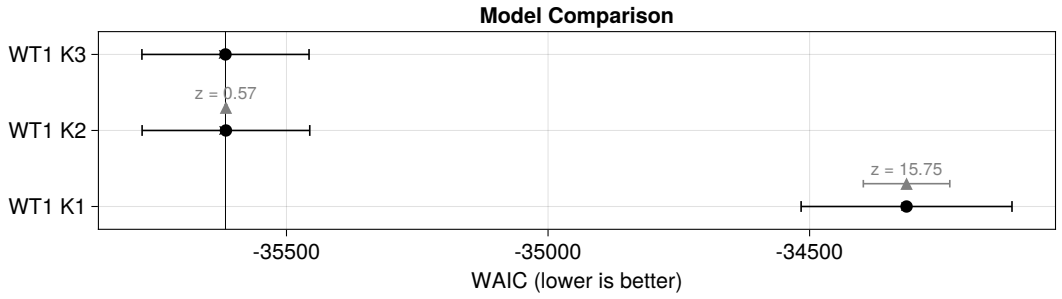


Figure 2. Comparison between diffusion models with $K = 1$, $K = 2$, or $K = 3$ diffusion coefficients for the Wild Type 1 data (WT1). The x-axis shows the WAIC score, where lower values indicate higher-performing models. The WAIC-score for each model is shown in black along with its uncertainty. The difference in WAIC-scores between the model and the best performing model (WT1 K3) is shown in grey with z being the number of standard deviations between them.

3.2 | Bayesian Analysis

Comparing the three diffusion models with 1, 2, or 3 diffusion coefficients, respectively, we find that the model with only a single diffusion component is simply not advanced enough to fully explain the data, see Figure 2. This figure shows that, even though the 3 component model is the best-performing of the models, when judging by the number of standard deviations, z, that the best model's WAIC is higher than the second best model's WAIC, it is statistically non-significant ($z < 2$). Since the performance of both the 2 and 3 component models are indistinguishable, we follow Occam's razor and continue with the former model.

Bayesian models allow for far greater flexibility than traditional frequentist models, including internal validation checks and diagnostic criteria to make sure that the model has not converged. In particular, we made sure that all $\hat{R}$-values were less than 1.01. To fully validate the $K = 2$ model, we show the traceplots and posterior distributions for the different parameters in Figure 3. The left part of the figure shows the parameter estimate as a function of MCMC iteration, i.e. traceplot, which, for correctly sampled chains, should resemble a fuzzy caterpillar (and not a skyline which would indicate bad mixing) (Roy, 2020). We find that the slow diffusion coefficient for WT1 data is: $\theta_{\text{slow}}^{\text{WT}_1} = 0.0417 \pm 0.0014\,\text{XXXunit}$.

XXX Energy computation shows that, see Figure 4.

Final results:

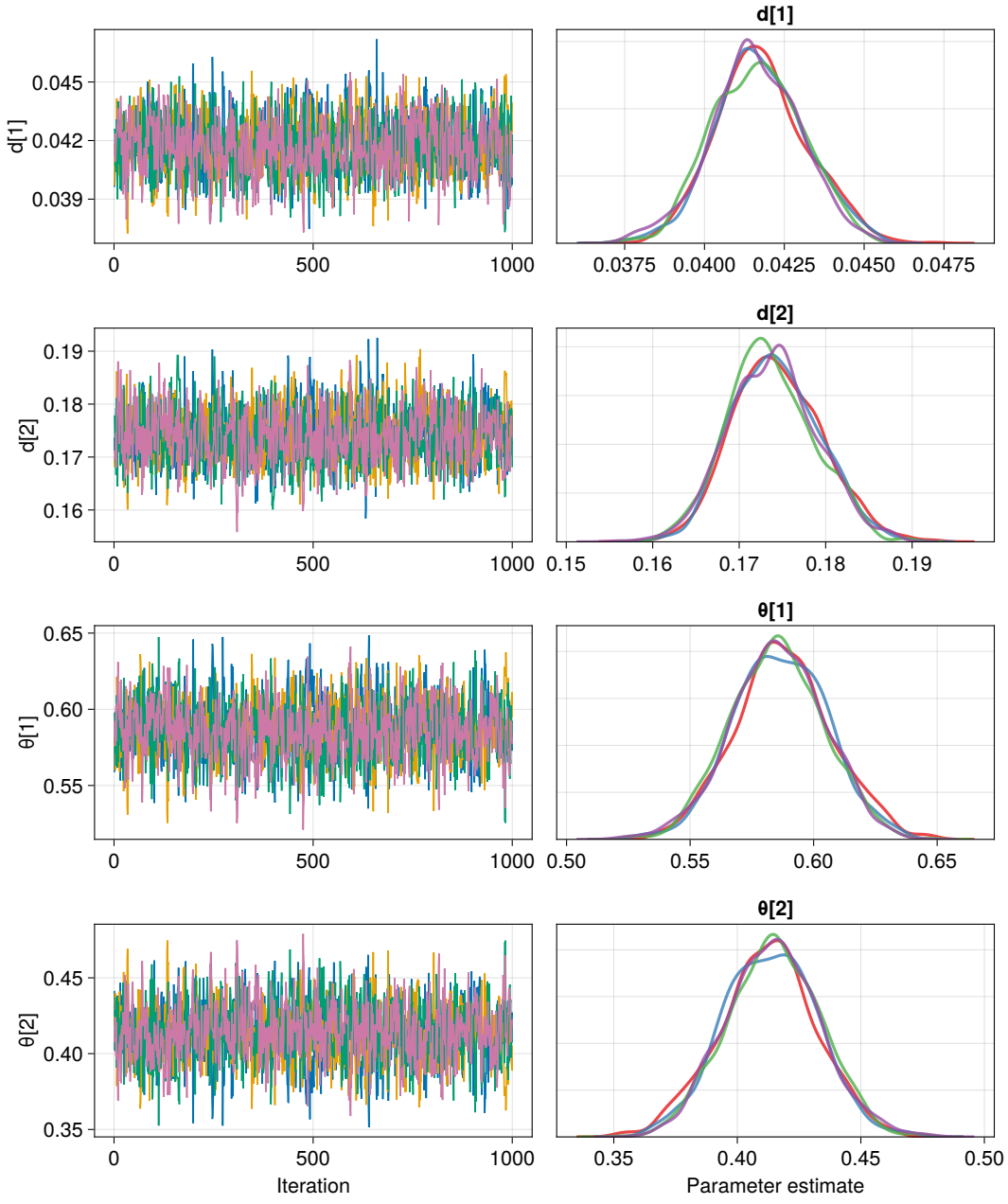


Figure 3. Results of the $K = 2$ diffusion model. Left) Traceplots. Right) Density plots.

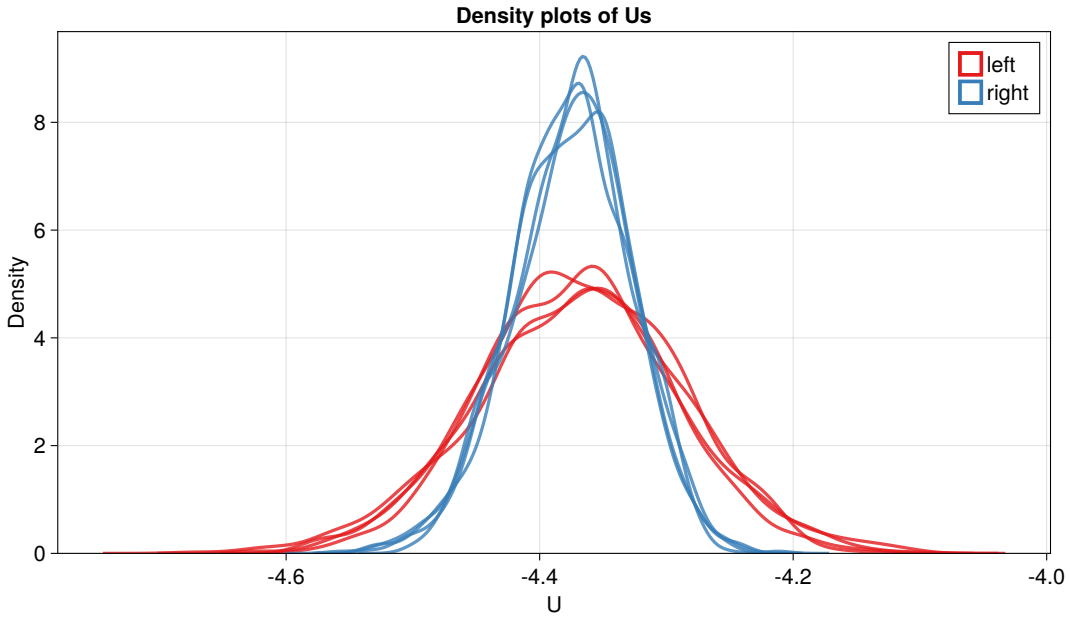


Figure 4. Density plot of the energy, U, using either the left or the right computation approach (XXX Mathias). The energy computed using the left computation is shown in red, and in blue with the right computation. The four different MCMC chains (for each approach) are shown as individual lines.

$$
\begin{aligned}
\texttt{DCon1\_WT1} &= 0.010\,29 \pm 0.001\,83 \\
\texttt{DCon2\_WT1} &= 0.017\,334\,3 \pm 0.000\,000\,4 \\
\texttt{Db\_focus} &= 0.006\,629\,0 \pm 0.000\,000\,1 \\
U_{\text{left}} &= -4.373 \pm 0.078 \\
U_{\text{right}} &= -4.375 \pm 0.047 \\
\texttt{Din\_hyper\_WT} &= 0.031\,51 \pm 0.002\,71 \\
\texttt{DCon1\_hyper\_WT} &= 0.008\,14 \pm 0.002\,34 \\
\texttt{DCon2\_hyper\_WT} &= 0.015\,391\,1 \pm 0.000\,001\,8 \\
\texttt{Db\_hyper\_focus} &= 0.001\,315\,7 \\
U_{\text{left-hyper}} &= -4.364 \pm 0.204 \\
U_{\text{right-hyper}} &= -4.108 \pm 0.047
\end{aligned}
\tag{10}
$$

4 | DISCUSSION

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

4.1 | Acknowledgment

Acknowledgements here

4.2 | Data availability

Source code is hosted at GitHub: https://github.com/ChristianMichelsen/diffusion.

REFERENCES

Betancourt, Michael (2018). "A Conceptual Introduction to Hamiltonian Monte Carlo". In: *arXiv:1701.02434 [stat]*. arXiv: 1701.02434.

Bezanson, Jeff et al. (2017). "Julia: A fresh approach to numerical computing". In: *SIAM review* 59.1. Publisher: SIAM, pp. 65–98. URL: https://julialang.org/.

Dietz, Laura (2022). "Directed factor graph notation for generative models". In.

Ge, Hong, Kai Xu, and Zoubin Ghahramani (2018). "Turing: A Language for Flexible Probabilistic Inference". en. In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. ISSN: 2640-3498. PMLR, pp. 1682–1690. URL: https://proceedings.mlr.press/v84/ge18b.html (visited on 2022).

Gelman, Andrew, Jessica Hwang, and Aki Vehtari (2014). "Understanding predictive information criteria for Bayesian models". en. In: *Statistics and Computing* 24.6, pp. 997–1016. ISSN: 1573-1375. DOI: 10.1007/s11222-013-9416-2. URL: https://doi.org/10.1007/s11222-013-9416-2 (visited on 2022).

Hoffman, Matthew D. and Andrew Gelman (2011). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *arXiv:1111.4246 [cs, stat]*. arXiv: 1111.4246.

McElreath, Richard (2020). *Statistical rethinking: a Bayesian course with examples in R and Stan*. 2nd ed. CRC texts in statistical science. Boca Raton: Taylor and Francis, CRC Press. ISBN: 978-0-367-13991-9.

McLachlan, Geoffrey J. and David Peel (2004). *Finite Mixture Models*. en. Google-Books-ID: c2_fAox0DQoC. John Wiley & Sons. ISBN: 978-0-471-65406-3.

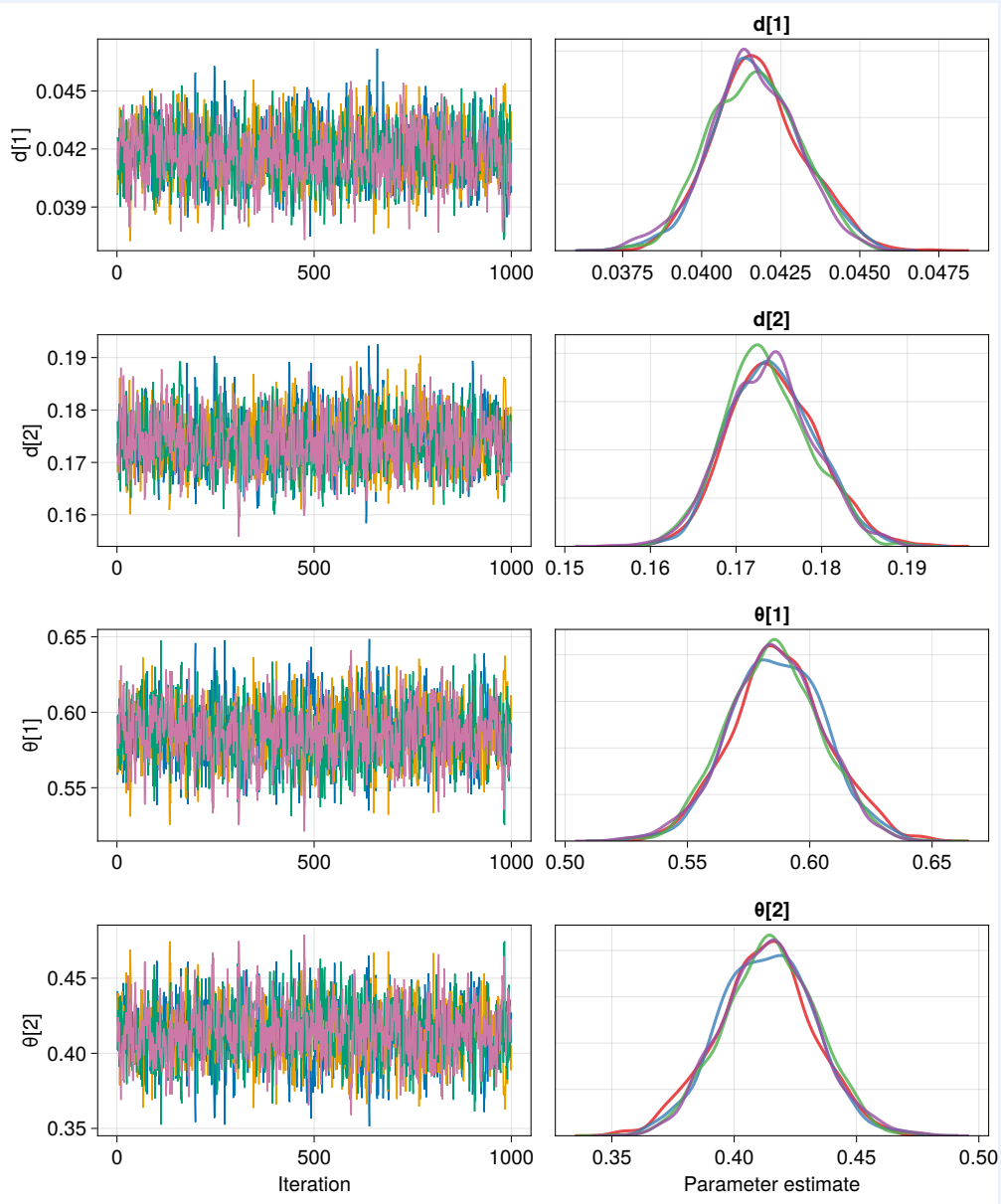
Roy, Vivekananda (2020). "Convergence Diagnostics for Markov Chain Monte Carlo". In: *Annual Review of Statistics and Its Application* 7.1. _eprint: https://doi.org/10.1146/annurev-statistics-031219-041300, pp. 387–412. DOI: 10.1146/annurev-statistics-031219-041300. URL: https://doi.org/10.1146/annurev-statistics-031219-041300 (visited on 2022).

Watanabe, Sumio (2010). "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory". In: *Journal of Machine Learning Research* 11.116, pp. 3571–3594. ISSN: 1533-7928.

Appendix 1

A | APPENDIX FIGURE 1

Here an example of an appendix figure.



Appendix 1—figure 1. Blablabla.

This document was typeset using LaTeX and the `tufte-style-thesis` class.
The style is heavily inspired by the works of Edward R. Tufte and Robert Bringhurst.
This is available on here:

`https://github.com/sylvain-kern/tufte-style-thesis/`.

Feel free to contribute!