UNIVERSITY OF
COPENHAGEN

Ph.D. Thesis
by
Christian Michelsen

Biological Data Science

Ancient genomics, anesthesiology, epidemiology,
and a bit in between

Submitted: 2022-11-17

*This thesis has been submitted to the
PhD School of The Faculty of Science,
University of Copenhagen*

Supervisor	Troels C. Petersen	Niels Bohr Institute
Cosupervisor	Thorfinn S. Korneliussen	Globe Institute

Til kvinderne i mit liv

Table of 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, 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 contact tracing.

Finally, in the fourth paper I 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 at the time. 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:

Paper 1: **Christian Michelsen**[†], Mikkel W. Pedersen[†], Antonio Fernandez-Guerra, Lei Zhao, Troels C. Petersen, Thorfinn S. Korneliussen (2022). *"metaDMG: An Ancient DNA Damage Toolkit"*. Submitted to Methods in Ecology and Evolution.

Paper 2: **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"*. Accepted and in review at BMJ Open.

Paper 3: 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"*. Published in: Royal Society Open Science 9.9. issn: 2054-5703. doi: 10.1098/rsos.220018.

Paper 4: Susmita Sridar[†], Mathias S. Heltberg[†], **Christian Michelsen**[†], Judith M. Hattab, Angela Taddei (2022). *"Microscopic single molecule dynamics suggest underlying physical properties of the silencing foci"*. Paper draft.

Shared first authorship is indicated with a dagger (†) next to the name.

1 *Introduction*

The primary content of my thesis is the four papers that are included in the thesis. This chapter is meant as a brief introduction to the background needed to understand the basics of the methods used throughout the papers. As such, this chapter is not meant to be a comprehensive guide to statistics and bioinformatics used in the papers. The original research motivation supporting the funding of this Ph.D. was very multi-disciplinary and the papers included in my thesis are also clearly influenced by this.

In Section 1.1, I will shortly introduce the field of ancient genomics and the statistical methods used to identify ancient DNA will be explained. Paper I, see Chapter 2, utilize modern Bayesian methods to classify which species are ancient and which ones are not. Bayesian methods are great when possible, however, they also rely on some statistical model being defined. In the case of Paper I, the model is a Beta-Binomial distribution combined with an exponential-decay damage model.

Sometimes, however, the model is not known and the data generating process has to be inferred by other means. This is the case in Paper II, see Chapter 3, where we utilize machine learning methods to extract this information. This paper deals with estimating the individual risk scores for each patient being re-hospitalized after a knee or hip operation. Section 1.2 introduces the reader to basic classification with machine learning models.

While the former two papers are based on real life data, Paper III, see Chapter 4, concerns the development of a new agent based model for COVID-19. The model is based on the SIR model, but with a more detailed description of the disease and the transmission process. The model is used to simulate the spread of the virus in Denmark and to estimate the effect of contact tracing. The model is also used to simulate and predict the spread of the "alpha" variant of COVID-19 in Denmark. Section 1.3 introduces the reader to the basics of agent based models.

Finally, the method of Bayesian model comparison of different diffusion models is introduced in Paper IV, see Chapter 5. In particular, this paper deals with different mixture-models of independent Rayleigh-distributions and how they can be used to extract important information about the underlying diffusion processes of a polymer bridging model in cell nuclei, see Section 1.4.

1.1   *Ancient DNA and Bayesian Statistics*

Previously, the only way to study ancient animals, plants, and other species was by studying their fossils. This changed in the middle of the 1980's when the first DNA was recovered from almost 5000 years old ancient mummies, showing that it was indeed possible to extract and sequence ancient DNA (Pääbo, 1985a; Pääbo, 1985b). This discovery, along with a dozen other pushing the boundary for what is scientifically possible with ancient DNA, led to Svante Pääbo being awarded with the Nobel Prize in Physiology or Medicine in 2022 for "his discoveries concerning the genomes of extinct hominins and human evolution" (Karolinska Institutet, 2022).

The field of ancient DNA (aDNA) was drastically changed with the invention of the Polymerase Chain Reaction, PCR, method (Mullis et al., 1986) along with the Next Generation Sequencing technology which revolutionized the speed and throughput of genomic sequencing while decimating the cost (Slatko, Gardner, and Ausubel, 2018). This technological advance has lead to better understanding of human migration and the genealogical tree of modern humans including the previously unknown human (sub)species; the Denisova hominin (Krause et al., 2010).

Leaving the homocentric world view, aDNA also allows for the study of archaic animals. In recent years, the boundary of how old DNA can be sequenced has been severely pushed; in 2013 with the early Middle Pleistocene 560–780 kyr BP horse (Orlando et al., 2013) and in 2021 with the million-year-old mammoths (van der Valk et al., 2021). High-throughput sequencing not only allows for the sequencing of single genomes – like single humans, animals, or plants – but also for sequencing of entire communities of organisms, so-called metagenomics. By analysing the DNA in environmental samples, environmental DNA, one can survey the rich plant and animal assemblages of a given area and at a specific time in the past. A new paper published in Nature shows that it is now possible to perform metagenomic sequencing on environmental DNA that is 2 million years old, see Appendix A. This is a direct application of the statistical method developed in Paper I, see Chapter 2, showing that `metaDMG`, the method, can help to push the boundary of what is possible with ancient DNA.

Ancient DNA is difficult to work with since it often contains only a limited amount of biological material due to bad preservation, leading to low endogenous content with high duplication rates, making high-depth sequencing difficult (Renaud et al., 2019). Genotype likelihoods are often used to alleviate the problem of low-coverage data (Nielsen et al., 2011). In addition to this, the DNA is often highly degraded. In particular, the two prominent issues with aDNA is fragmentation

and deamination (Dabney, Meyer, and Pääbo, 2013; Peyrégne and Prüfer, 2020). Fragmentation refers to the fact that the DNA is broken into very short fragments, often with a fragment size of less than 50 bp. This leads to low-quality mapping issues and reference biases, which can somewhat be mitigated by the use variant graphs (Martiniano et al., 2020). Deamination is a process in which cytosine (C) in the single-stranded overhangs in the end of the DNA molecules is often hydrolized to uracil (U) which is then read as thymine (T) by the DNA polymerase. This particular type of postmortem damage is known as cytosine deamination, or C-to-T transitions, and is a one of the main reasons behind nucleotide misincorporations in ancient DNA (Briggs et al., 2007). Due to the short fragment sizes in ancient DNA, they will often contain overhangs with over-expressed C-to-T frequency. In the case of single-genome analysis, previous solutions have been to either remove all transitions and only keep transversions, or apply trimming at the read ends (Schubert et al., 2012). For an illustration of both fragmentation and deamination of ancient DNA, see Figure 1.
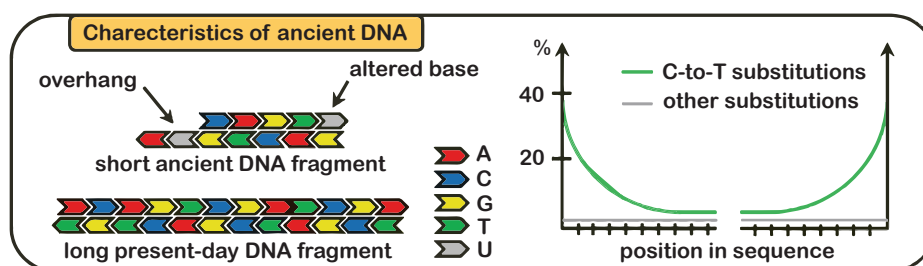
Figure 1.
Illustration of DNA damage. Ancient DNA is often highly fragmented with short reads compared to modern, present-day DNA, and can contain uracils (U). These uracils will then be misread as thymines (T) while sequencing leading to C-T nucleotide misincorporations. This is primaryly happening at the end of the reads. Modified from (Peyrégne and Prüfer, 2020).

Currently, a handfull of different methods for quantifying ancient DNA damage exists. In particular, the mapDamage 2.0 software has been the gold standard for how to measure ancient DNA damage in the field (Jónsson et al., 2013), however, it uses slow algorithms leading to unfeasible runtimes for large datasets. Newer, faster methods are being developed all of the time, such as PyDamage (Borry et al., 2021) which tackle some of mapDamage's limitations, although even faster methods suited at metagenomic analysis for large-scale datasets are still lacking.

In Paper I, see Chapter 2, we introduce the `metaDMG` software which utilizes the C-to-T deamination pattern[1] to identify ancient DNA damage. One of the key features of this method is the beta-binomial model which allows the uncertainty to fitted independently of the mean leading to improved accuracy of the damage estimation. Since the data is based on misincorporation counts, in particular the number of C-to-T transitions, k, out of N total C's, the classical likelihood to use

[1] for the forward strand and the G-to-A deamination pattern for the reverse strand

for this type of data is a binomial distribution. The mean and variance of the binomial distribution is given by:

$$
\begin{aligned}
\mathbb{E}\left[k\right] &= Np \\
\mathbb{V}\left[k\right] &= Np(1-p),
\end{aligned}
\tag{1}
$$

where p is the probability of success (a C-to-T substitution). One of the issues, however, is that the variance of the binomial distribution is proportional to the mean. The binomial distribution is thus not flexible enough to accommodate large amounts of variance in the data, so-called overdispersion (McElreath, 2020). One way to accommodate overdispersion is to instead use a beta-binomial model. The beta-binomial model is a generalization of the binomial distribution where the variance is allowed to be flexible. Technically, the beta-binomial model assumes that p is a random variable which follows a beta distribution $p \sim \text{Beta}(\mu, \varphi)$ where the beta distribution is parameterized[2] in terms of its mean, μ, and dispersion parameter, φ, (Cepeda-Cuervo and Cifuentes-Amado, 2017). The mean and variance of this beta-binomial model is then given by:

$$
\begin{aligned}
\mathbb{E}\left[k\right] &= N\mu \\
\mathbb{V}\left[k\right] &= N\mu(1-\mu)\frac{\varphi + N}{\varphi + 1}.
\end{aligned}
\tag{2}
$$

Comparing Equation 1 and Equation 2, we see that the variance of the beta-binomial model is no longer (strictly) proportional to the mean, but instead is a function of the dispersion parameter, φ, allowing for higher variance than the binomial-only model. When $\varphi = 0$, the variance of the beta-binomial model is N times larger, and when $\varphi \to \infty$ the variance reduces to the variance of the binomial model, showing that the beta-binomial model is a generalization of the binomial model.

Equation 2 shows how we model the C-to-T damage at a specific base position in the read. We model the position-dependent damage frequency, $f(x) = k(x)\,N(x)$, see Figure 1, as a function of the distance from the end of the read, x, with an exponential decay:

$$
f(x; A, q, c) = A(1-q)^{x-1} + c.
\tag{3}
$$

Here A is the scale factor, or amplitude, q is the decay rate, and c is a constant offset, the baseline damage. Since x is discrete, this is similar to a (modified) geometric sequence starting from $x = 1$. The combination of Equation 2 and Equation 3 is illustrated in Figure 2, which shows the position-dependent, decreasing damage

frequency. The figure also shows the increase in uncertainty in the beta-binomial model compared to the binomial-only model.
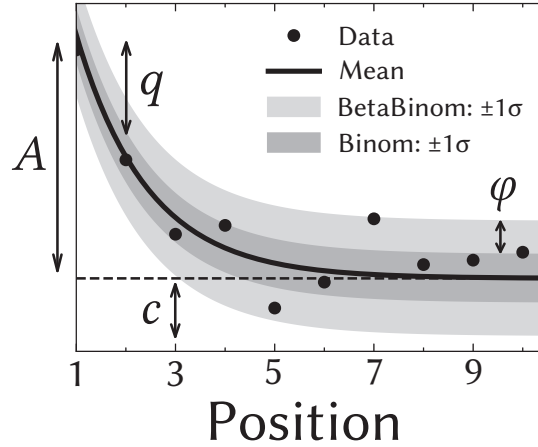

Figure 2.
Illustration of the damage model. The figure shows data points as circles and the damage frequency, $f(x)$, as a solid line. The amplitude of the damage is A, the offset is c, and the relative decrease in damage pr. position is given by q. The damage uncertainty for a binomial model is shown in dark grey and the uncertainty for a beta-binomial model in light grey.

The damage framework described above is based on the nucleotide misincorporations, i.e. the C-to-T transitions. The background for this data can be from either sequence files (fasta or fastq files) mapped to a single genome or from metagenomic data consisting of multiple mapped reads. As such, the damage framework is a general tool for estimating damage. In the metagenomic case where single DNA reads are mapped to multiple species, `metaDMG` performs a simple lowest common ancestor based on the ngsLCA algorithm (Wang et al., n.d.). This means that for each read that map to multiple reference, i.e. has multiple alignments, the taxonomic tree is traversed for each alignment until a common ancestor is found. This is the so-called lowest common ancestor (LCA). Figure 3 illustrates the LCA for a read that maps to different (sub)species. In this example, the LCA of alignment 1 and 2 is the Subspecies I while the LCA for all four alignments is the Genus X. `metaDMG` works by default with the NCBI taxanomic database but can also be used with custom databases.

Given the nucleotide misincorporations, either coming from a single-reference alignment file or after LCA in the metagenomic case, we fit eq. (2) and (3) with a Bayesian model. This is done to ensure the optimal inference of the parameters, A, q, and c, and to account for the uncertainty in the data. Bayesian inference also allows for the inclusion of domain knowledge in the form of the prior distribution by Bayes theorem. Bayes theorem is based on the law of conditional probability (Barlow, 1993) stating that the probability of two events, A and B, both happening,

Figure 3.
Illustration of the lowest
common ancestor (LCA) for
taxonomic trees. Here the
LCA of alignment 1 and 2 is
Subspecies I, while the LCA
of all four reads is Genus X.
The dots (…) refers to other
taxonomic levels, e.g. family
and order.
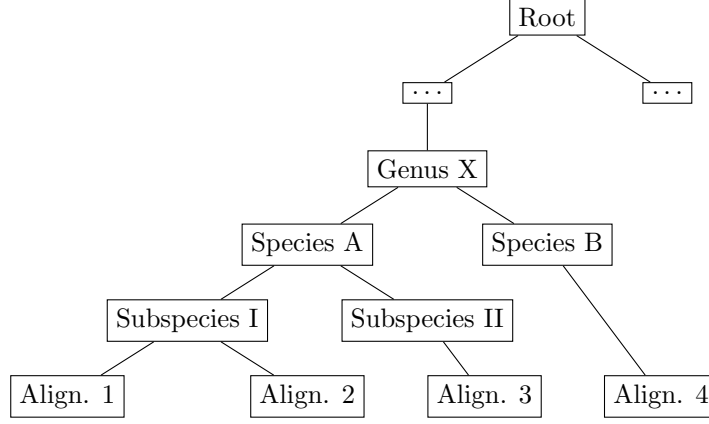
$P(A \cap B)$, is given by the probability of B, $P(B)$ times the probability of A given B, $P(A|B)$:

$$P(A \cap B) = P(B)P(A|B). \tag{4}$$

Similarly, $P(A \cap B)$ can also be expressed in terms of the probability of A:

$$P(A \cap B) = P(A)P(B|A). \tag{5}$$

Combining Equation 4 and Equation 5 and rearranging terms gives the Bayes theorem:

$$P(\theta|x) = \frac{P(\theta)P(x|\theta)}{P(x)}, \tag{6}$$

with a change of variables where x now refers to the observed data and θ the parameter(s) of the model. The first term in the numerator, $P(\theta)$, is the prior distribution and describes the probability distribution assigned to θ before observing any data. The second term is the likelihood function, $P(x|\theta)$, which is the probability of observing the data, x, given the parameter(s), θ. Together these two terms combine to a compromise between data and prior information.

The numerator, $P(x)$, also known as the evidence, can be treated as a data-related normalization factor. In the case of continuous θ, this can calculated as the marginalization of the likelihood function over θ:

$$P(x) = \int_{\theta} P(x|\theta)P(\theta) \, d\theta. \tag{7}$$

This equation, however, is often intractable to compute in the higher-dimensional case. Luckily, it can be shown that Markov Chain Monte Carlo (MCMC) sampling can approximate the posterior distribution, $P(\theta|x)$, and asymptotically converge to the correct distribution (Gelman et al., 2015).

Traditionally MCMC methods such as Metropolis Hastings (MH) or Gibbs sampling have been used for Bayesian inference, however, these methods are often slow and require a lot of tuning. In the last decades, a new class of MCMC methods have been developed, namely Hamiltonian Monte Carlo (HMC) methods. While traditional MH uses a Gaussian random walk, HMC is a gradient-based MCMC method that uses Hamiltonian dynamics to guide the sampling. This makes HMC more efficient than traditional MCMC methods and allows for sampling from high-dimensional distributions (Betancourt, 2018; Neal, 2011). A particularly efficient variant of HMC is the No-U-Turn Sampler (NUTS). NUTS is a variant of HMC that automatically tunes the step size and number of steps to take in the Hamiltonian dynamics (Homan and Gelman, 2014).

Most statistical domain-specific languages (DSL) such as Stan (Carpenter et al., 2017), Pyro (Bingham et al., 2019), NumPyro (Phan, Pradhan, and Jankowiak, 2019) or Turing.jl (Ge, Xu, and Ghahramani, 2018), implement HMC and in particular the NUTS algorithm. Since `metaDMG` is implemented in Python, we use NumPyro for the Bayesian inference of the damage model as it is easy to implement and computationally efficient since it which uses JAX (Bradbury et al., 2018) under the hood for automatic differentiation and just-in-time (JIT) compilation.

Even though NumPyro is fast and `metaDMG` is efficiently implemented, the Bayesian inference of the damage model is still computationally expensive. Thus, we have decided to also include a faster, approximate method of Bayesian inference: the maximum a posteriori (MAP) estimate. The MAP estimate is the point estimate of the posterior distribution that maximizes the posterior probability density function, i.e. the posterior mode:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} P(\theta|x) = \arg\max_{\theta} P(\theta)P(x|\theta), \qquad (8)$$

where the second equality is due to the evidence being independent of θ. Since this is a point estimate, $\hat{\theta}_{\text{MAP}}$ does not fully explain the full posterior, however, it is often a good approximation*. Comparing $\hat{\theta}_{\text{MAP}}$ to the maximum likelihood estimate (MLE):

*Especially when the posterior is unimodal, which it generally is in the case of `metaDMG`.

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} P(x|\theta), \qquad (9)$$

the MAP estimate can be seen as a regularized version of the MLE estimate (Murphy, 2012). To further optimize the computational efficacy of the MAP estimation in `metaDMG`, we JIT compile the MAP estimation function using Numba (Lam, Pitrou, and Seibert, 2015) and mathematically optimize the function with iMinuit (Dembinski et al., 2021).

1.2   *Anestesiology – a Machine Learning Approach*

This section explains the technical background behind Paper II, see Chapter 3. This study investigates the potential advantages of using a modern machine-learning model compared to classical logistic regression to predict the risk of patients being re-hospitalized after fast-track hip and knee replacements. In particular, the patients were grouped into two groups where the "risk-patients" all stayed in the hospital for more than four days after the operation or were readmissioned to the hospital within 90 days after the operation. As such, this is a binary classification problem where the patient's risk-score is predicted based on historical data.

Most classification and regression problems fall under the same machine learning (ML) branch called supervised learning. In supervised learning, the goal is to find the optimal hypothesis h^* in the hypothesis set $\mathcal{H}$ that matches the unknown, "true" data-generating function $f : \mathcal{X} \to \mathcal{Y}$ as good as possible, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. Assuming that we have access to realizations of f, the so-called training data $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, we can use a learning algorithm $\mathcal{A}$ combined with the training data to estimate h^* (Abu-Mostafa, Magdon-Ismail, and Lin, 2012). Here N refers to the number of training samples and $\mathbf{x}_i$ is the ith observation with the true label y_i. This process is illustrated in Figure 4.
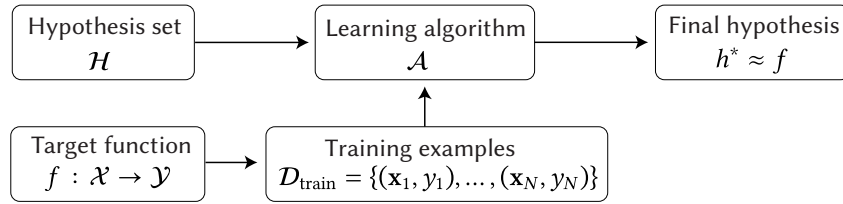
Figure 4.
Illustration of how to learn from data in a supervised learning setting. Adapted from (Abu-Mostafa, Magdon-Ismail, and Lin, 2012).

Both the logistic regression (LR) and ML model can be viewed through the lens of Figure 4, just with $|\mathcal{H}_{\text{LR}}| \ll |\mathcal{H}_{\text{ML}}|$, i.e. the machine learning model is a lot more complex than the logistic regression model[3]. To predict the performance of h^* on new, unseen data, the naive method would be to train on all of the data and evaluate on the same, however, this would have a high risk of overfitting the data and thus biasing the predicted performance[4] (Abu-Mostafa, Magdon-Ismail, and Lin, 2012).

3 And the hypothesis space thus is significantly larger.

4 Especially for high cardinality hypothesis sets.

To avoid this and get more accurate estimates of the performance of h^*, we use a technique called cross-validation (CV). In the simplest way, this can be done by splitting the data into two sets, one called the training and one called the validation set, and then only train on the training set. Afterwards the trained model can be evaluated on the validation set without biasing the performance estimate. This process can further be refined by splitting the data into K folds and then repeating the process K times, where each fold is used as the validation set once. This is called K-fold cross-validation and is illustrated in Figure 5a (Murphy, 2012; Hastie, Tibshirani, and Friedman, 2016). K-fold cross validation works well in many cases, yet in the case of temporal data, it also risks introducing bias in the performance estimates, since, in the different folds, it, effectively, is allowed to "look into the future". The most extreme case of this is shown in the bottom of Figure 5a where the model trains on all future and present data and is then evaluated only on past data. In many time dependent datasets, this is undesirable. Instead, we use a technique called temporal cross validation (Tashman, 2000), see Figure 5b, which circumvents this problem by only allowing the model to train on past data and evaluate on future data. As the patient data is time dependent[5], this is the technique we use in Paper II.

[5] The fraction of rehospitalizations decreased over time due to surgerilogical improvements.

(a) K-fold cross validation   (b) Temporal cross validation
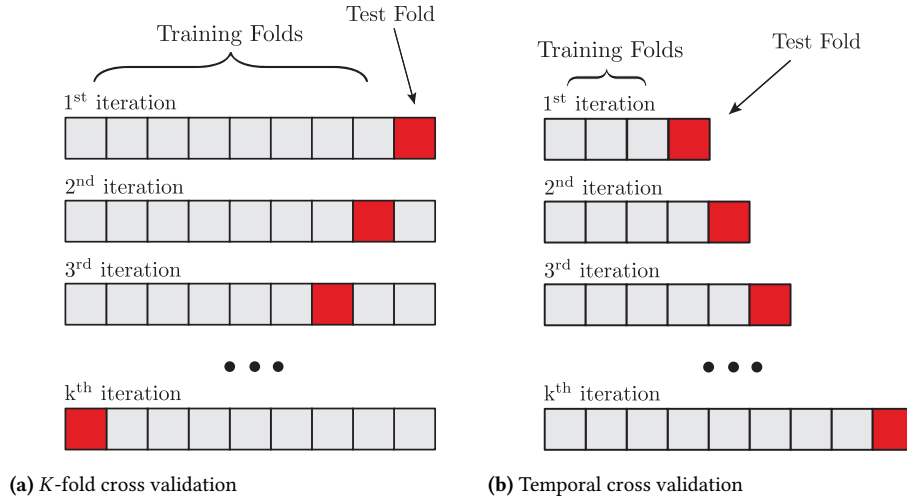
Figure 5.
Two types of cross validation: K-fold cross validation, and temporal cross validation. Both figures from Michelsen, 2020.

The training of the learning model $\mathcal{A}$ itself is model-dependent and will not be covered in this thesis, see (Michelsen, 2020) for a more detailed description of the training process. This is not the only way to optimize the performance of $\mathcal{A}$, albeit it is the primary one. In addition to the internal parameters of the model, some parameters are external to the model in the sense that they are not optimized by the model itself, but rather by the user. These are called hyperparameters and are often optimized using a technique called hyperparameter optimization (HPO).
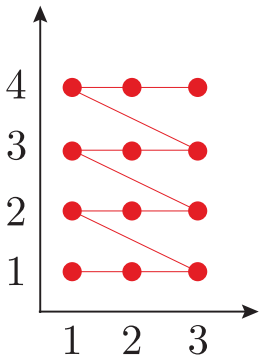
Figure 6.
Illustration of grid search.
Figure from Michelsen,
2020.

6 As such, grid search suffers from the curse of dimensionality.
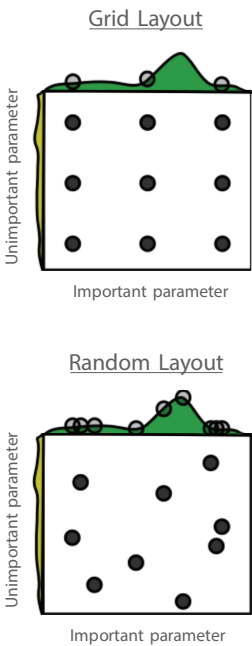
Figure 7.
Illustration comparing grid
search to random seach.
The height of green surve
is the score-function which
has to be optimized. Figure
adapted from Bergstra and
Bengio, 2012.

In the case of logistic regression, the number of variables to include would be an example of a hyperparameter; in the case of a decision tree model, the depth of the tree. Hyperparameter optimization can be performed in many ways, where the classical one is through grid search, see Figure 6.

In grid search all combinations of the hyperparameters (the cartesian product) are tried and the best combination is chosen. This is a simple and intuitive approach, however, it scales exponentially, i.e. very poorly, with the number of hyperparameters[6]. In addition to this, it depends on the user-defined grid, which might not be optimal. To circumvent this, a technique called random search (RS) was developed (Bergstra and Bengio, 2012). Random search is a randomized version of grid search, where the hyperparameters are sampled randomly from a distribution. This allows for a more efficient sampling of the hyperparameter space, see Figure 7, and further lets the user decide on the number of iterations beforehand.

The disadvantage of random search is that all draws are fully independent. While this allows for easy parallelisation of the algorithm, this also means that each new sample might be infinitesimal close in the hyperparameter space to a previous sample with bad performance, which with high probability will thus also have a high loss. An approach that does take the history of the previous samples' performance into consideration is Bayesian optimization (Brochu, Cora, and de Freitas, 2010). In Bayesian optimization each successive hyperparameter is chosen based on an acquisition function which optimizes the expected improvement in the performance of the model. This is illustrated in Figure 8. This leaves the user with the task of choosing between "exploitation" and "exploration" of the hyperparameter space in the definition of the acquisition function, yet most implementations of bayesian optimization have decent default settings.

We use the Python package Optuna (Akiba et al., 2019) for HPO in Paper IV due to its ease of use and its support for Bayesian optimization. In particular, we use the Tree-structured Parzen Estimator algorithm for the Bayesian optimization and a median stopping rule to minimize optimization time (Bergstra, Bardenet, et al., 2011). This allowed for a good comporise between optimization time and performance.

While model performance is often paramount, in some fields – such as medicine – being able to explain the model's predictions is almost as important. This is especially true in the case of medical decision support systems, where the model is used to make decisions about the patient's treatment. Model explainability helps to build trust in the model for both the patient and the medical staff alike.

In Paper II, we employ the SHapley Additive exPlanations (SHAP) values which provide estimates on which variables contribute most to the risk score predictions
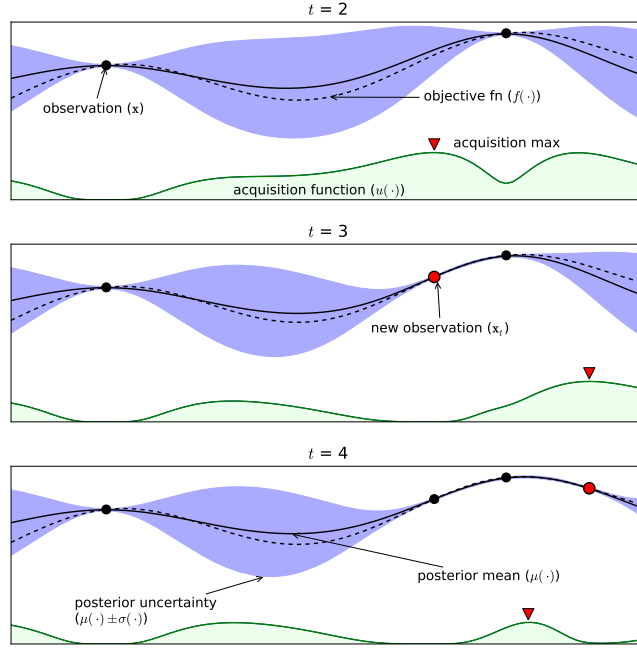
Figure 8.
Illustration of the learning process of Bayesian optimization. The previous observations are shown as black dots and the true objective function is shown as a dashed black line. This line is fitted with Gaussian processes which is shown as the solid line with its uncertainty in purple. The acquisition function is shown in green and its maximum decides what the next iteration of the hyperparameter value(s) should be (Michelsen, 2020).

(Scott M Lundberg and Lee, 2017; Scott M. Lundberg, Erion, et al., 2020). SHAP values allow for not only a global explanation of the model, i.e. which features are most important overall, but also a local explanation, i.e. why a single patient was predicted to be at risk of being re-hospitalized. It has previously been shown that the interaction between SHAP values and medical doctors can improve the performance of anaesthesiologists (Scott M. Lundberg, Nair, et al., 2018).

While the aim of Paper II is to show how modern machine learning techniques can be used to improve the risk prediction process, the usefulness of the SHAP values in a medical context is demonstrated in the paper in Appendix B. The paper uses the SHAP values to compare the preoperative haemoglobin level in the patient with the risk-score, stratified by sex and operation type (knee vs hip replacement). Currently, the WHO guidelines for the haemoglobin levels are gender specific (Anaemias and Organization, 1968), however, this study finds no significant gender difference and a haemoglobin threshold close to the WHO suggestions for men.

1.3   *COVID-19 and Agent Based Models*

In early 2020, a contagious disease called COVID-19 started to spread in Europe, including Denmark. With new infections showing up faster and faster, governments started to implement different measures to limit the spread of the deadly disease, including lockdowns, travel restrictions, and social distancing, measures

not previously seen in peacetimes since the Spanish flu in 1918. This was the background for the work that we did in 2020 which became the basis for Paper III, see Chapter 4. This paper deals with the development of a new agent based model for COVID-19 in Denmark in collaboration with Statens Serum Institut (SSI), the Danish CDC.

Historically, most mathematical models of infectious diseases were variations of the SIR model which describe the evolution of a pandemic by approximating all individuals as one population (Kermack, McKendrick, and Walker, 1927). As one of the simplest compartmental models, the susceptible-infectious-recovered (SIR) model is based on a system of three non-linear differential equations that describe the transition between each state, or compartment, of the model (Kröger and Schlickeiser, 2020). Initially the entire population is susceptible until time $t = 0$ at which some individuals become not only infected, but also infectious, allowing the disease to spread. After having been infectious, the individuals recovers and becomes immune to the disease and are stops being infectious. Several variations of the SIR model exist, including the SIS model, where the recovered individuals become susceptible again (Hethcote, 1989). Another variation is the SEIR model, which includes an exposed state, where individuals are infected but not yet infectious, which is the basis for the model used in Paper III.

SIR-like models suffer from several shortcomings, including the assumptions that the population is homogeneous and that the disease is transmitted at a constant rate. In reality, neither the population nor the transmission rates are homogenous. These are some of the reasons why we chose to use an agent based model (ABM). Agent based models simulate individual agents in a population that can have complex interactions patterns, e.g. based on their geography (Wilensky and Rand, 2015).

In particular, we implemented a continuous-time, stochastic, spatial ABM using the Gillespie algorithm, a stochastic simulation algorithm (Gillespie, 1977). The model is JIT compiled with Numba (Lam, Pitrou, and Seibert, 2015) to speed up the simulation, allowing simulating the Danish population of 5.8 million people in a couple of hours instead of days. The model allows for the individual tuning of the three main effects; A) heterogeneities in the infection strength[7], B) number of connections[8], C) and the spatial clustering of the agents. In the absence of any of these effects, we find that the ABM's predictions matches the SIR model's predictions within ±5%. Once we allowed for spatial clustering, we found that the epidemic developed faster and with a higher infection peak compared to the SIR model, but that the total number of infected in the end of the epidemic was lower.

In real-life scenarios, one does not have the opportunity to let the epidemic run loose and afterwards evaluate the strength of the epidemic; the goal is to predict the

7 allowing *super-shedders*

8 allowing *super-connecters*

intensity in the very beginning of the epidemic and implement lockdown-related measured based on this estimate. In the second part of Paper III, we show that once spatial clustering is introduced, fitting standard SEIR-models to infection numbers from the first few days of the epidemic, predictions are overestimated by a factor of two. The results are a significant over-estimation of the impact of the epidemic. Since the population is highly susceptible in the beginning of an epidemic, this also highlights the benefits of early lockdowns to reduce the effect of the super connectors.

The developed ABM was further used by SSI to estimate the effect of contact tracing related to COVID-19 in Denmark, see Appendix C. It was further used to estimate spread of the "alpha" variant of COVID-19 (B.1.1.7) in Denmark, see Appendix D. Based on data available January 2nd 2021, the model predicted that the "alpha" variant would be the dominant variant in Denmark February 10–20, 2021. It became the dominant variant in Week 7: February 15–21, 2021 (Bager et al., 2021).

1.4   *Diffusion Models and Bayesian Model Comparison*

asdasdas

Bibliography

Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin (2012). *Learning From Data*. S.l.: AMLBook. 213 pp. ISBN: 978-1-60049-006-4.

Akiba, Takuya et al. (2019). "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. New York, NY, USA: Association for Computing Machinery, pp. 2623–2631. ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330701.

Anaemias, WHO Scientific Group on Nutritional and World Health Organization (1968). *Nutritional Anaemias : Report of a WHO Scientific Group [Meeting Held in Geneva from 13 to 17 March 1967]*.

Bager, Peter et al. (2021). "Risk of Hospitalisation Associated with Infection with SARS-CoV-2 Lineage B.1.1.7 in Denmark: An Observational Cohort Study". In: *The Lancet Infectious Diseases* 21.11, pp. 1507–1517. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(21)00290-5.

Barlow, R. J. (1993). *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*. Chichester, England ; New York: Wiley. 222 pp. ISBN: 978-0-471-92295-7.

Bergstra, James, Rémi Bardenet, et al. (2011). "Algorithms for Hyper-Parameter Optimization". In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc.

Bergstra, James and Yoshua Bengio (2012). "Random Search for Hyper-Parameter Optimization". In: *Journal of Machine Learning Research* 13.10, pp. 281–305.

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

Bingham, Eli et al. (2019). "Pyro: Deep Universal Probabilistic Programming". In: *Journal of Machine Learning Research* 20, 28:1–28:6.

Borry, Maxime et al. (2021). "PyDamage: Automated Ancient Damage Identification and Estimation for Contigs in Ancient DNA de Novo Assembly". In: *PeerJ* 9, e11845. ISSN: 2167-8359. DOI: 10.7717/peerj.11845.

Bradbury, James et al. (2018). *JAX: Composable Transformations of Python NumPy Programs*. Version 0.2.5.

Briggs, Adrian W. et al. (2007). "Patterns of Damage in Genomic DNA Sequences from a Neandertal". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.37, pp. 14616–14621. ISSN: 0027-8424. DOI: 10.1073/pnas.0704665104. pmid: 17715061.

Brochu, Eric, Vlad M. Cora, and Nando de Freitas (2010). *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*. DOI: 10.48550/arXiv.1012.2599. arXiv: 1012.2599 [cs].

Carpenter, Bob et al. (2017). "Stan: A Probabilistic Programming Language". In: *Journal of statistical software* 76.1.

Cepeda-Cuervo, Edilberto and MARíA VICTORIA Cifuentes-Amado (2017). "Double Generalized Beta-Binomial and Negative Binomial Regression Models". In: *Revista Colombiana de Estadística* 40.1, 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. pmid: 23729639.

Dembinski, Hans et al. (2021). *Scikit-Hep/Iminuit: V2.8.2*. Version v2.8.2. Zenodo. DOI: 10.5281/ZENODO.3949207.

Ge, Hong, Kai Xu, and Zoubin Ghahramani (2018). "Turing: A Language for Flexible Probabilistic Inference". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1682–1690.

Gelman, Andrew et al. (2015). *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC. 675 pp. ISBN: 978-0-429-11307-9. DOI: 10.1201/b16018.

Gillespie, Daniel T. (1977). "Exact Stochastic Simulation of Coupled Chemical Reactions". In: *The Journal of Physical Chemistry* 81.25, pp. 2340–2361. ISSN: 0022-3654. DOI: 10.1021/j100540a008.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2016). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7.

Hethcote, Herbert W. (1989). "Three Basic Epidemiological Models". In: *Applied Mathematical Ecology*. Ed. by Simon A. Levin, Thomas G. Hallam, and Louis J. Gross. Biomathematics. Berlin, Heidelberg: Springer, pp. 119–144. ISBN: 978-3-642-61317-3. DOI: 10.1007/978-3-642-61317-3_5.

Homan, Matthew D. and Andrew Gelman (2014). "The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *The Journal of Machine Learning Research* 15.1, pp. 1593–1623. ISSN: 1532-4435.

Jónsson, Hákon et al. (2013). "mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters". In: *Bioinformatics* 29.13, pp. 1682–1684. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btt193.

Karolinska Institutet, The Nobel Assembly at (2022). *The Nobel Prize in Physiology or Medicine 2022*. NobelPrize.org. URL: https://www.nobelprize.org/prizes/medicine/2022/press-release/ (visited on 2022).

Kermack, William Ogilvy, A. G. McKendrick, and Gilbert Thomas Walker (1927). "A Contribution to the Mathematical Theory of Epidemics". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 115.772, pp. 700–721. DOI: 10.1098/rspa.1927.0118.

Krause, Johannes et al. (2010). "The Complete Mitochondrial DNA Genome of an Unknown Hominin from Southern Siberia". In: *Nature* 464.7290 (7290), pp. 894–897. ISSN: 1476-4687. DOI: 10.1038/nature08976.

Kröger, M and R Schlickeiser (2020). "Analytical Solution of the SIR-model for the Temporal Evolution of Epidemics. Part A: Time-Independent Reproduction Factor". In: *Journal of Physics A: Mathematical and Theoretical* 53.50, p. 505601. ISSN: 1751-8113, 1751-8121. DOI: 10.1088/1751-8121/abc65d.

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.

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.

Lundberg, Scott M., Gabriel Erion, et al. (2020). "From Local Explanations to Global Understanding with Explainable AI for Trees". In: *Nature Machine Intelligence* 2.1 (1), pp. 56–67. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0138-9.

Lundberg, Scott M., Bala Nair, et al. (2018). "Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia during Surgery". In: *Nature Biomedical Engineering* 2.10 (10), pp. 749–760. ISSN: 2157-846X. DOI: 10.1038/s41551-018-0304-0.

Martiniano, Rui et al. (2020). "Removing Reference Bias and Improving Indel Calling in Ancient DNA Data Analysis by Mapping to a Sequence Variation Graph". In: *Genome Biology* 21.1, p. 250. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02160-7.

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.

Michelsen, Christian (2020). "A Physicist's Approach to Machine Learning – Understanding the Basic Bricks". University of Copenhagen.

Mullis, K. et al. (1986). "Specific Enzymatic Amplification of DNA in Vitro: The Polymerase Chain Reaction". In: *Cold Spring Harbor Symposia on Quantitative Biology* 51 Pt 1, pp. 263–273. ISSN: 0091-7451. DOI: 10.1101/sqb.1986.051.01.032. pmid: 3472723.

Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective.* The MIT Press. ISBN: 0-262-01802-0.

Neal, Radford M. (2011). *MCMC Using Hamiltonian Dynamics.* Routledge Handbooks Online. ISBN: 978-1-4200-7941-8 978-1-4200-7942-5. DOI: 10.1201/b10905-7.

Nielsen, Rasmus et al. (2011). "Genotype and SNP Calling from Next-Generation Sequencing Data". In: *Nature reviews. Genetics* 12.6, pp. 443–451. ISSN: 1471-0056. DOI: 10.1038/nrg2986. pmid: 21587300.

Orlando, Ludovic et al. (2013). "Recalibrating Equus Evolution Using the Genome Sequence of an Early Middle Pleistocene Horse". In: *Nature* 499.7456 (7456), pp. 74–78. ISSN: 1476-4687. DOI: 10.1038/nature12323.

Pääbo, Svante (1985a). "Molecular Cloning of Ancient Egyptian Mummy DNA". In: *Nature* 314.6012 (6012), pp. 644–645. ISSN: 1476-4687. DOI: 10.1038/314644a0.

— (1985b). "Preservation of DNA in Ancient Egyptian Mummies". In: *Journal of Archaeological Science* 12.6, pp. 411–417. ISSN: 0305-4403. DOI: 10.1016/0305-4403(85)90002-0.

Peyrégne, Stéphane and Kay Prüfer (2020). "Present-Day DNA Contamination in Ancient DNA Datasets". In: *BioEssays* 42.9, p. 2000081. ISSN: 1521-1878. DOI: 10.1002/bies.202000081.

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

Renaud, Gabriel et al. (2019). "Authentication and Assessment of Contamination in Ancient DNA". In: *Ancient DNA: Methods and Protocols*. Ed. by Beth Shapiro et al. Methods in Molecular Biology. New York, NY: Springer, pp. 163–194. ISBN: 978-1-4939-9176-1. DOI: 10.1007/978-1-4939-9176-1_17.

Schubert, Mikkel et al. (2012). "Improving Ancient DNA Read Mapping against Modern Reference Genomes". In: *BMC Genomics* 13, p. 178. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-178. pmid: 22574660.

Slatko, Barton E., Andrew F. Gardner, and Frederick M. Ausubel (2018). "Overview of Next Generation Sequencing Technologies". In: *Current protocols in molecular biology* 122.1, e59. ISSN: 1934-3639. DOI: 10.1002/cpmb.59. pmid: 29851291.

Tashman, Leonard J. (2000). "Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review". In: *International Journal of Forecasting* 16.4, pp. 437–450. ISSN: 0169-2070.

Van der Valk, Tom et al. (2021). "Million-Year-Old DNA Sheds Light on the Genomic History of Mammoths". In: *Nature* 591.7849 (7849), pp. 265–269. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03224-9.

Wang, Yucheng et al. (n.d.). "ngsLCA—A Toolkit for Fast and Flexible Lowest Common Ancestor Inference and Taxonomic Profiling of Metagenomic Data". In: *Methods in Ecology and Evolution* n/a.n/a (). ISSN: 2041-210X. DOI: 10.1111/2041-210X.14006.

Wilensky, Uri and William Rand (2015). *An Introduction to Agent-Based Modeling*. The MIT Press. ISBN: 978-0-262-73189-8. JSTOR: j.ctt17kk851.

2   *Paper I*

The following pages contain the paper:

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

3 *Paper II*

The following pages contain the paper:

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.".

4 *Paper III*

The following pages contain the paper:

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.

5 *Paper IV*

The following pages contain the paper:

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

APPENDIX

A   *Kap København*

The following pages contain the paper published in Nature 2022:

> Kurt H. Kjær, Mikkel W. Pedersen, Bianca De Sanctis, Binia De Cahsan, Thorfinn S. Korneliussen, **Christian Michelsen**, Karina K. Sand, Stanislav Jelavić, Anthony H. Ruter, Astrid M. Z. Bonde, Kristian K. Kjeldsen, Alexey S. Tesakov, Ian Snowball, John C. Gosse, Inger G. Alsos, Yucheng Wang, Christoph Dockter, Magnus Rasmussen, Morten E. Jørgensen, Birgitte Skadhauge, Ana Prohaska, Jeppe Å. Kristensen, Morten Bjerager, Morten E. Allentoft, Eric Coissac, PhyloNorway Consortium, Alexandra Rouillard, Alexandra Simakova, Antonio Fernandez-Guerra, Chris Bowler, Marc Macias-Fauria, Lasse Vinner, John J. Welch, Alan J. Hidy, Martin Sikora, Matthew J. Collins, Richard Durbin, Nicolaj K. Larsen & Eske Willerslev, *"A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA"* (Published in Nature, 2022, doi: 10.1038/s41586-022-05453-y).

The paper use the `metaDMG` tool to identify ancient species and classify the amount of ancient damage in these species. This shows, that modern modern statistical methods combined with excellent work in the ancient DNA labs can provide new insights into the past – even on data that are more than two millions years old.

XXX

B *Explainable ML and Anaemia*

The following pages contain the draft paper:

> Christoffer C. Jorgensen, **Christian Michelsen**, Troels C. Petersen, Henrik Kehlet (2022), *"Gender-specific haemoglobin thresholds in relation to preoperative risk assessment in fast-track total hip and knee arthroplasty"*.

Based on the same data as used on Paper II, the paper uses the SHAP curves to understand the machine learning model. In particular, it compares the preoperative haemoglobin level in the patient with the risk-score for being resubmitted to the hospital within 30 days after the operation, stratified by sex and operation type (knee vs hip replacement).

1

Type of article: Science Letter

Submitting author: Christoffer C Jørgensen

Department of Anaesthesia

Hospital of Northern Zeeland - Hillerød

Dyrehavevej 29, 3400 Hillerød, Denmark

Gender-specific haemoglobin thresholds in relation to preoperative risk assessment in fast-track total hip and knee arthroplasty.

C. C. Jorgensen,[1] C. Michelsen[2], T. C. Petersen,[3] H. Kehlet5[4]

1 Anaesthetist,

Department of Anaesthesia, Hospital of Northern Zeeland, Hillerød, Denmark

The Centre for Fast-track Hip and Knee Replacement, Copenhagen, Denmark

2, PhD-student 3, Ass. Professor,

The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

4 Professor,

Section for Surgical Pathophysiology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark

The Centre for Fast-track Hip and Knee Replacement, Copenhagen, Denmark

Correspondence to:

Dr. C.C Jørgensen

Email: christoffer.calov.joergensen@regionh.dk

Short title:

Preoperative anaemia and recovery in fast-track THA and TKA

Within the last decade there has been an increasing focus on anaemia, iron deficiency and transfusion strategies leading to the concept of "patient blood management" (PBM), aiming at reducing the need for blood transfusions by preoperative optimisation of haemoglobin (Hb) and iron-status and use of intra- and postoperative restrictive transfusion protocols [1].

When diagnosing preoperative anaemia most practitioners have adhered to the WHO guidelines which were developed in 1968 and use gender specific criteria of a Hb of < 130 g.l^{-1} for men and < 12 g.l^{-1} for women [2]. However, these thresholds are based on studies with less sophisticated laboratory end epidemiological techniques than presently available and are consequently under current revision [3].

Furthermore, it has been argued that the WHO definitions of anaemia may not apply to surgical patients, as the relative blood-loss is larger in women, potentially leading to increased risk of allogenic blood transfusions and morbidity when using a gender specific lower preoperative anaemia threshold [4-6].

In total hip (THA) and knee arthroplasty (TKA) it is internationally acknowledged that preoperative iron deficiency anaemia should be corrected by treatment with intravenous (i.v.) iron [7]. However, detailed knowledge of the Hb threshold to increase the risk of postoperative morbidity, indications for treatment and whether it differs in men and women is sparse. The aim of this secondary analysis was to investigate the influence of preoperative Hb level in a comprehensive machine-learning model aimed at identifying patients at "high-risk" of medical complications leading to either a length of hospital stay of >4 days or 30-days readmission after an established fast-track THA and TKA [8]. While the primary study focused on comparing potential benefits of an overall machine-learning model in preoperative risk-prediction [9], this secondary analysis focus specifically on the influence of preoperative Hb level per se and potential differences according to gender and age.

3

We used a well-defined cohort of elective fast-track THA and TKA patients and

evaluated the effect of preoperative Hb-level on the machine-learning model by

SHAP-analysis which evaluates the individual effect of the variables included in a

machine-learning model [10]. Furthermore, we assessed the distribution of Hb-levels

and increases in risk-profile according to gender and age.

From January 2017 to August 2017, we included 3913 patients with a median length

of stay of 1 day.   Mean preoperative Hb was 154.8 (SD:15.12) but lower in women

(149.4 vs. 162 g.l[1]: p<0.001) and there were 30.5% of women vs. 12.0% of men with

a Hb of <130 g.l[1] (p<0.001). SHAP-analysis demonstrated an immediate steep

increase in the risk-score for medical complications with a preoperative Hb below

147.6 g.l[1], and irrespective of gender and age (figure 1). Finally, the median SHAP-

value of Hb-level was 0.35 (IQR:) in the patients with a Hb-level below 147.6 g.l[1]

These results remained consistent regardless of analysing THA and TKA separately

(online Supporting Information Figure S1a+b).

Our analysis demonstrates that in a comprehensive machine-learning risk-model, the

preoperative Hb threshold was the same in men and women for an increased risk of

prolonged length of stay or readmissions due to medical issues after fast-track THA

and TKA. The threshold value of 147.6 g.l[1] is remarkably close to the 130 g.l[1]

suggested for men in the current WHO guideline. Thus, the results of our study

support the current WHO threshold for anaemia in men, but importantly also for

removing gender specific Hb criteria for preoperative anaemia in women, at least in

elective THA and TKA. Furthermore, the influence of preoperative Hb level < 147.6

g.l[1] was consistent regardless of age, supporting that the removal of gender specific

criteria should apply to all patients. Finally, the effect of Hb level on the accumulated

risk-score was clinically meaningful. Thus, figure 1, illustrates that preoperative Hb

level contributed with SHAP-values of approximately 0.4 in patients with a Hb of

<147.6 g.l[1]. This corresponds with about 50% increased odds of being a high-risk patient. In contrast, in those with Hb-levels >147.6 g.l[1] the odds of being high-risk patients decreased with about 15%.

That gender specific Hb criteria may be inappropriate and need further consideration, has also been demonstrated in cardiac surgery, where women with a preoperative Hb of 120-129 g.l[1] received more blood transfusions and had increased length of hospital stay compared to those with a Hb of >129 g.l[1] [11]. That women with a preoperative Hb level of < 130 g.l[1] may potentially benefit from iron-treatment prior to surgery was illustrated by a large study investigating preoperative Hb levels and iron deficiency in major elective surgery and finding similar incidence of iron deficiency in women with Hb < 130 g.l[1] and < 120 g.l[1] [12]. Our study has some limitations, including lack of information on perioperative blood-transfusions and potential use of preoperative i.v. iron. However, preoperative optimisation with i.v. iron was not standard in the participating departments, and even if some of the outcomes was due to transfusion-related morbidity it would not change the finding of similar SHAP-curves between men and women. Study strengths include well-established fast-track protocols, detailed data on comorbidity and patient outcomes, a complete follow-up, and use of a sophisticated machine-learning model.

In conclusion, from a machine-learning model in fast-track THA and TKA, a Hb threshold of 146.7 g.l[1] was found to increase risk of impaired recovery, regardless of gender or age, thus calling for re-evaluation of preoperative anaemia risk criteria in the elective surgical setting.

5

Competing Interests

Acknowledgements

References

1. Goodnough LT, Shander A Patient blood management. *Anesthesiology* 2012; **116:** 1367-76.
2. Nutritional anaemias. Report of a WHO scientific group. *World Health Organ Tech.Rep.Ser.* 1968; **405:** 5-37.
3. Pasricha SR, Colman K, Centeno-Tablante E, Garcia-Casal MN, Peña-Rosas JP Revisiting WHO haemoglobin thresholds to define anaemia in clinical medicine and public health. *Lancet Haematol* 2018; **5:** e60-e2.
4. Munoz M, Gomez-Ramirez S, Kozek-Langeneker S, et al. 'Fit to fly': overcoming barriers to preoperative haemoglobin optimization in surgical patients. *Br.J Anaesth.* 2015; **115:** 15-24.
5. Butcher A, Richards T, Stanworth SJ, Klein AA Diagnostic criteria for pre-operative anaemia-time to end sex discrimination. *Anaesthesia* 2017; **72:** 811-4.
6. Gombotz H, Rehak PH, Shander A, Hofmann A The second Austrian benchmark study for blood use in elective surgery: results and practice change. *Transfusion* 2014; **54:** 2646-57.
7. Gómez-Ramírez S, Maldonado-Ruiz M, Campos-Garrigues A, Herrera A, Muñoz M Short-term perioperative iron in major orthopedic surgery: state of the art. *Vox Sang* 2019; **114:** 3-16.
8. 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.
9. Michelsen C, Jorgensen C, Heltberg M, et al. Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty - a machine-learning based approach. In revision, 2022.
10. Lundberg SM, Erion G, Chen H, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* 2020; **2:** 56-67.
11. Blaudszun G, Munting KE, Butchart A, Gerrard C, Klein AA The association between borderline pre-operative anaemia in women and outcomes after cardiac surgery: a cohort study. *Anaesthesia* 2018; **73:** 572-8.
12. Muñoz M, Laso-Morales MJ, Gómez-Ramírez S, Cadellas M, Núñez-Matas MJ, García-Erce JA Pre-operative haemoglobin levels and iron status in a large multicentre cohort of patients undergoing major elective surgery. *Anaesthesia* 2017; **72:** 826-34.

7

Figure legend:

Figure 1a+b
SHapley Additive exPlanations (SHAP) curves for preoperative haemoglobin level in in relation to preoperative risk-stratification according to the machine-learning algorithm. Each dot indicates a patient with the colour indicating age (increasing from blue to red). Increasing SHAP values indicate increasing risk-score and decreasing values a decreased risk-score. The cut-off for going from a negative to a positive SHAP-value is indicated by the dotted line at a preoperative Hb level of 147.6 g.l[1]


Supplemental material
Figure 1a+b
SHapley Additive exPlanations (SHAP) curves for preoperative haemoglobin level in in relation to preoperative risk-stratification according to the machine-learning algorithm for total hip (1a) and total knee arthroplasty (1b), respectively. Each dot indicates a patient with the colour indicating age (increasing from blue to red). Increasing SHAP values indicate increasing risk-score and decreasing values a decreased risk-score. The cut-off for going from a negative to a positive SHAP-value is indicated by the dotted line at a preoperative Hb level of 147.6 g.l[1]
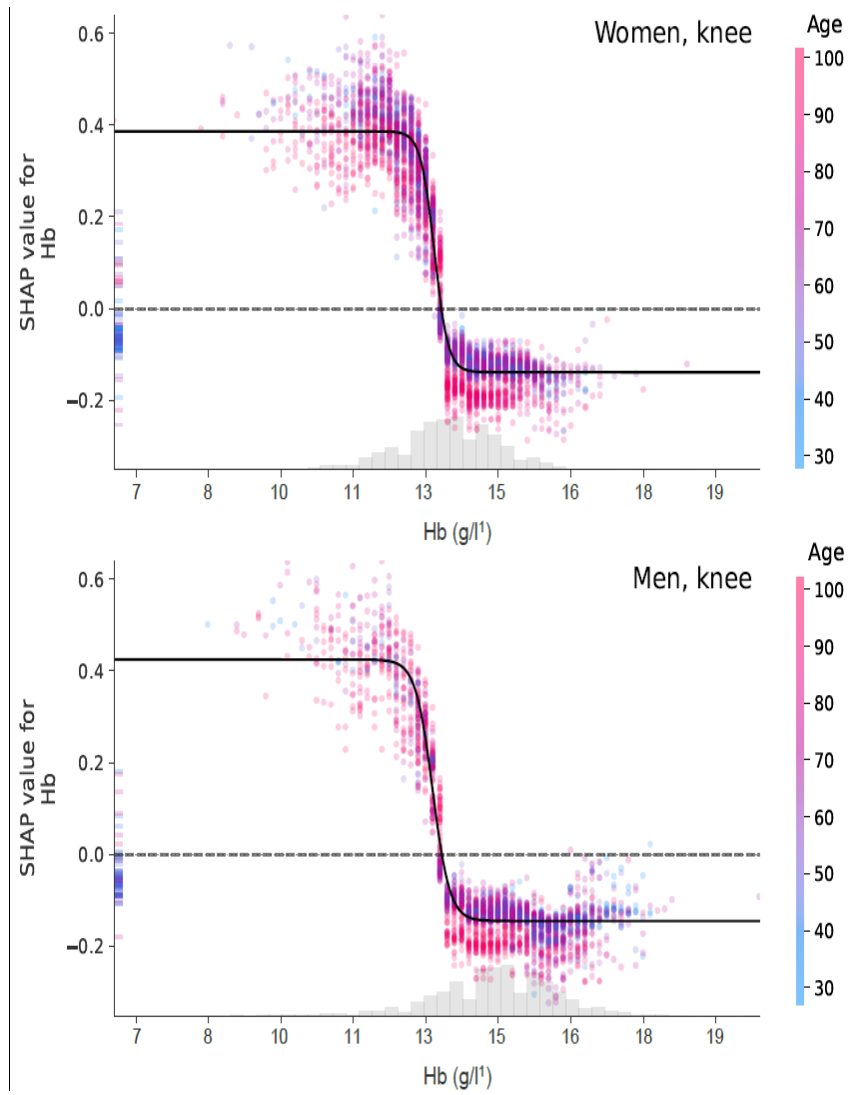
Figure 1 a+b

9
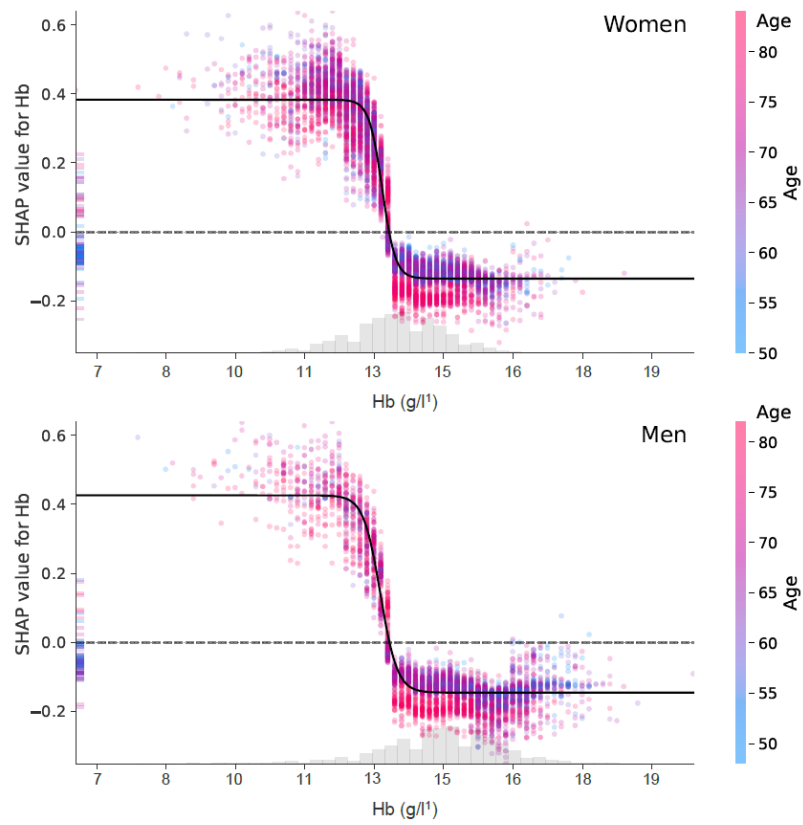
Supplemental figure 1a
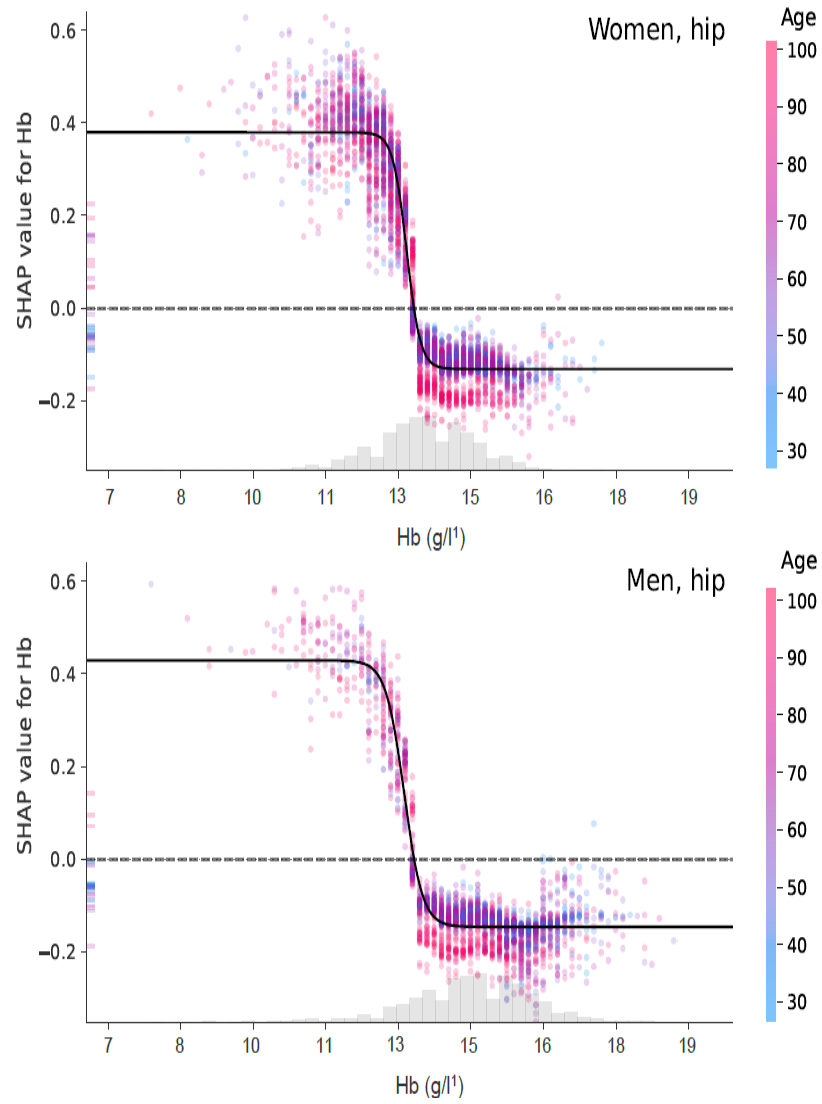
Supplemental figure 1b

C *SSI Ekspertrapport*

The following pages contain the report from Statens Serum Institut, the Danish CDC:

> Ekspertgruppen for matematisk modellering, *"Ekspertrapport af den 10. december 2020 – Effekten af kontaktopsporing"* (Statens Serum Institut, 2020).

The report is from December 10 2020 and is a summary on the effect of contact tracing related to COVID-19 in Denmark. The report is in Danish and is based on two agent based models, one from DTU and our model from NBI.

XXX

D SSI Notat

The following pages contain the report from Statens Serum Institut, the Danish CDC:

> Ekspertgruppen for matematisk modellering, *"Scenarier for udviklingen i den engelske virusvariant af SARS-COV-2 (cluster B.1.1.7)"* (Statens Serum Institut, 2021).

The report is from January 2 2021 and is a summary of the estimated spread of the "alpha" variant of COVID-19 (B.1.1.7) in Denmark. The report is in Danish and is based on two models, one from DTU and our agent based model from NBI.

XXX