Addressing uncertainty and complex data structures through Bayesian and classical approaches

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

eingereicht von

Raphael Rehms

11.05.2025

Erster Berichterstatter: Prof. Dr. Helmut Küchenhoff
Zweiter Berichterstatter: Prof. Dr. Volker Schmid
Dritte Berichterstatterin: Prof. Dr. Stefanie Muff

Tag der Disputation: 30.06.2025

Acknowledgments

This thesis would not be possible without the support and trust of many wonderful and inspiring people. I owe my deepest gratitude to them. First, I want to thank Prof. Dr. Helmut Küchenhoff, Prof. Dr. Volker Schmid, and Prof. Dr. Stefanie Muff, who kindly agreed to act as reviewers for this thesis. Moreover, I would like to thank Prof. Dr. Christian Heumann and Prof. Dr. Anne-Laure Boulesteix for being part of the examination panel.
Moreover, I would like to express my deepest gratitude to...

...Helmut, who has had a more positive impact on me than he might imagine...

...Sabine, who heavily influenced my thinking and my perspective on research...

...Prof. Dr. Ulrich Mansmann, who always placed his trust in me and shared his experience and insights...

...all co-authors. They contributed significantly to all publications...

...many colleagues from the IBE and the Department of Statistics. Listing names would probably always be unfair, as so many people have been part of my life over the last few years. I am grateful for all the help, trust, shared experience, and fruitful discussions ranging from completely random topics to professional ones. I am sure that all the wonderful people know exactly that they are meant...

...my friends who really make my world a better place...

...my family. In particular...

...my mother, who gave more than a single human being could ever give...
...my wife. There are not enough words to express how much you mean to me...
...my son.

Thank you.

Summary

Answering questions based on real-world data can pose considerable challenges to analysts. It often requires the use of data that are of questionable quality, exhibit high uncertainty, and may originate from multiple sources. Such data bear a high degree of complexity in their structure with respect to the underlying data-generating process. This cumulative thesis aims to address these issues in the context of selected research areas.

The thesis is divided into two parts. The first part introduces the necessary methodology. The second part presents the four contributing articles. The methodological part provides an introduction chapter to statistical inference and probabilistic modeling by presenting Bayesian inference using Markov chain Monte Carlo as a general approach, and the generalized linear model (GLM) as a classical statistical method. Furthermore, the first part provides three chapters of methodological background in selected areas of research.

The first area to be discussed is infectious disease modeling. The focus is on time-shifting operations that can be used to combine information from multiple time series. This lays the foundation for the first two contributions, which employ a Bayesian hierarchical approach to infectious disease modeling in the context of COVID-19 data.

Next, an overview of measurement error theory is presented, followed by a discussion on how the Bayesian approach addresses these challenges. The third contribution demonstrates the flexibility of the Bayesian approach by applying it to data from the Wismut cohort, which presents considerable complexity and requires the use of multiple measurement error models.

Finally, the last discussed chapter delves into the field of federated learning and privacy-preserving methods. The fourth contribution builds on the presented methodological background to develop an algorithm that is able to validate learned classification models through a GLM-based formulation of the ROC curve.

An underlying theme of this thesis is the notion of uncertainty. In the Bayesian approach, uncertainty is encoded through the formulation of prior distributions and the overall probabilistic model, which inherently propagates and quantifies the uncertainty in a posterior distribution. The fourth contribution leverages the concept of uncertainty to preserve individual privacy by adding calibrated noise.

Zusammenfassung

Die Beantwortung von Fragestellungen anhand von realen Daten kann Analysten vor enorme Herausforderungen stellen. Oft werden Daten verwendet, deren Qualität suboptimal und mit hoher Unsicherheit verbunden ist. Darüber hinaus ist es gegebenenfalls erforderlich, Daten aus mehreren Quellen zu verwenden. Dies kann zu einer erheblichen strukturellen Komplexität im Hinblick auf den datengenerierenden Prozess führen. Die vorliegende kumulative Arbeit versucht Lösungen im Kontext ausgewählter Forschungsbereiche zu geben.
Sie gliedert sich in zwei Teile. Der erste Teil stellt den erforderlichen methodologischen Hintergrund vor, während der zweite Teil alle beigetragenen Fachartikel enthält. Im methodologischen Teil wird zunächst ein einführendes Kapitel zu statistischer Inferenz und probabilistischen Modellierungskonzepten dargelegt. Es werden bayesianische Inferenz unter Verwendung von Markov-Chain-Monte-Carlo als genereller Ansatz sowie das generalisierte lineare Modell (GLM) als klassische statistische Methode diskutiert. Im Anschluss folgen drei Kapitel, die Hintergründe zu ausgewählten Forschungsbereichen liefern, die Teil dieser Arbeit sind.
Der erste betrachtete Bereich ist die Modellierung von Infektionskrankheiten. Hierbei wird der Fokus auf Zeitverschiebungsoperationen gelegt, die genutzt werden können, um Informationen von mehreren Zeitreihen zu kombinieren. Dies bildet die Grundlage für die ersten beiden beigetragenen Fachartikel. Diese nutzen einen bayesianisch-hierarchischen Ansatz im Kontext von Daten der COVID-19-Pandemie.
Im Anschluss wird ein Überblick zu Messfehler-Methodologie gegeben und wie diese durch einen bayesianischen Ansatz berücksichtigt werden können. Der dritte und thematisch zugehörige beigetragene Fachartikel demonstriert die Flexibilität des bayesianischen Ansatzes anhand von Daten der Wismut-Kohorte – Daten, die eine beträchtliche Komplexität aufweisen und die Modellierung mehrerer Messfehler-Modelle erforderlich macht.
Das abschließend behandelte Kapitel widmet sich dem föderierten Lernen und Methoden zum Datenschutz. Dies bildet das Fundament für den vierten Fachartikel. In diesem wird ein Algorithmus entwickelt und implementiert, der Klassifikationsmodelle durch eine GLM-basierte Formulierung einer ROC-Kurve validiert.
Ein wichtiges latentes Thema dieser Arbeit ist das Konzept von Unsicherheiten. Beim bayesianischen Ansatz wird Unsicherheit durch die Formulierung von Priorverteilungen und das gesamte probabilistische Modell dargestellt welche durch die Posterioriverteilung eine direkte und angemessene Quantifizierung erlaubt. Der vierte beigetragene Fachartikel nutzt das Konzept von Unsicherheit, um den den Schutz der Privatsphäre einzelner Personen zu gewährleisten, indem kalibriertes Rauschen addiert wird.

Contents

Part I

Introduction and Background

Chapter 1

Overview

The analysis of collected data is a fundamental part of empirical research. However, data in the real world may exhibit quality limitations due to practical constraints, limited resources during collection, or evolving research requirements imposed on historically collected data sets. This means that data often carry a high degree of uncertainty and a large potential for confounding. We may also face challenges such as integrating data from multiple sources. It may not even be possible to merge data at a single location for further analysis due to legal or privacy barriers. How can these issues be addressed? This thesis tries to offer an answer for specific cases.

The underlying theme of this thesis is the notion of uncertainty. Bayesian modeling offers an intuitive way to deal with uncertainty. Through the combination of different probabilistic sub-models and prior distributions that encode a researcher's beliefs into a larger model, one can handle even very complex situations. This is exactly the case in the first three contributions, where the Bayesian approach proves to be an ideal fit.

The first two contributions address challenges in infectious disease modeling. They leverage data to model the infection dynamics of the COVID-19 pandemic. Throughout the pandemic, various structural changes, such as vaccinations, emerging variants of concern, varying testing strategies, and non-pharmaceutical interventions, affect the disease dynamics. One can model all these different influences and combine them with appropriate prior information and hierarchical modeling techniques.

In the third contribution, we exploit the flexibility of the Bayesian approach for data problems where measurement errors are prevalent. Even though the approach is generally applicable, we discuss the problem with a focus on occupational cohort studies, where one is interested in the association between an exposure and an outcome of interest. In this area of research, measurement errors are a well-known problem. The Bayesian approach provides sufficient flexibility to handle problems where the measurement error process is highly complex.

The last area to be discussed is the field of privacy-preserving federated learning. While numerous approaches exist for conducting federated learning, methods for model validation fall short. Therefore, the last contribution combines different aspects of federated learning with the well-established framework of generalized linear models (GLMs) to construct a privacy-preserving ROC analysis approach. Uncertainty plays a crucial role in this context as well. The proposed algorithm adds calibrated noise to guarantee individual privacy.

Even though the notion of uncertainty is a constant companion throughout this thesis, not all aspects are covered. In particular, confounding plays an important role in statistical modeling. However, this topic

is not explicitly discussed, and we refer to the wide body of literature.[1]

The thesis is divided into two parts. The first part is dedicated to establishing the methodological foundation, which includes an introduction to Bayesian modeling using Markov chain Monte Carlo and GLMs (Chapter 2). Subsequently, the following three chapters (Chapters 3-5) discuss the relevant theory for the various fields of research. The second part provides the contributions to this thesis, where the first two contributions (Chapters 6 and 7) are concerned with infectious disease modeling, and the next contribution (Chapter 8) focuses on measurement errors. While the first three contributions utilize a Bayesian hierarchical approach, the last contribution (Chapter 9) presents a ROC-GLM approach in the field of privacy-preserving federated learning.

[1] We refer to Greenland, Pearl, and Robins [47] as one of the most prominent introductory papers to this topic.

Chapter 2

Inference, probabilistic models and estimation

In this chapter, we discuss the general ideas and methodology that are required for all subsequent chapters and therefore also for all contributions. We consider methods that are concerned with probabilistic modeling and the estimation of such models using observable data. The subsequent chapters address specific modeling solutions for selected fields of research and specific situations.

2.1 Statistical inference

Wasserman [96] defines statistical inference as *"the process of using data to infer the distribution that generated the data."* It is not necessary that we are always interested in the full distribution. We may also just want to focus at specific parts of the distribution, e.g. certain moments or other aggregated quantities of interest. In addition to the rather classical idea of generalizing from collected data to the underlying population, Gelman, Hill, and Vehtari [40] also state two further objectives. First, the generalization from a treatment to a control group, which is a common task in causal inference. Second, the generalization from observed measurements to underlying constructs. The latter is of particular interest, as it represents a major aim for a large fraction of this thesis. Latent variables, which are inferred from uncertain and complex data settings are a central part of the first three contributions (Chapters 6, 7 and 8).
It is worth mentioning, that we are concerned with *statistical inference* and may distinguish it from its use in the context of probabilistic and generative machine learning, particularly contemporary large language models (LLMs). In this domain, the term rather refers roughly to the generation of artificial new data from new input data after a model has been learned. This could, for instance, be the generation of new tokens after a set of input tokens is provided.

2.1.1 Probabilistic modeling

The heart of statistical inference is probably the model that reflects or encodes the scientific question under consideration.[1] The model is an essential part of both Bayesian and frequentist approaches. We will discuss both ideas in more detail in the next section (section 2.1.3). Here, we focus on *probabilistic*

[1] We could, of course, consider multiple models. For instance, in a Bayesian workflow, it is common to define models iteratively by critical model checking and testing [43, 12]. Another example would be a comparison or ensemble of models. However, one could still interpret this as one large, hierarchically ordered model. Moreover, models may not be fixed forever as new findings and other factors over time will emerge leading to updated versions.

models in which we define relationships between quantities using *distributional* assumptions. This usually leads to a joint distribution over all random quantities [79]. They may define an assumed (marginal) data-generating process (DGP) that we use as a working model. With "working model", we want to highlight that the model is almost always a simplification of the actual DGP. This also aligns with the famous statement *"all models are wrong"* by Box [12]. However, the same text also highlights their importance. In particular, it is stated that *"scientists must be alert to what is importantly wrong"*. With this in mind, we may be able to answer scientific questions in an adequately way, making a working model useful. For instance, if we are interested in an unbiased effect for a specific association, we may decide on required simplifications to make a problem feasible. In the end, we are left with defining a model that we hope, it represents or offers what we require. Therefore, we could rather interpret a model as a story we assume or permit. This could range from defining a precise and possibly complex DGP, a robust simplification, or a highly non-linear black-box function.
We could therefore view this in light of the statement by Coombs [23] that *"we buy information with assumptions"*[2]. With this we mean that our findings are always conditioned on a chosen story with a subjective component, and therefore exhibit inherent uncertainty.
If we define a probabilistic relationship, we generally use the notion of distributions in which the functional form is determined by a set of parameters. Here, we restrict ourselves to models that are parameterized by a finite set, i.e., we focus on the class of parametric models. This stands in contrast to non-parametric models where a relationship cannot be described by a finite set [96]. Ghosal and van der Vaart [44] describes a parametric model as setting a very strong prior on a thin subset of all possible distributions. We could relate this again to the subjectivity of a model and therefore the inherent uncertainty about model choice. However, an in-depth discussion about the merits and demerits is not part of this work.

2.1.2 Directed acyclic graphs

It is sometimes helpful to illustrate a model as a directed acyclic graph (DAG). This is commonly done in Bayesian modeling, causal inference, and probabilistic machine learning ([83, 97, 71, 79]). They are also sometimes called "Bayesian networks" or "belief networks". It is important to note, that DAGs are not inherently Bayesian, nor is Bayesian inference a prerequisite. It is rather a tool that is used to illustrate the relationships between different quantities, such as data, parameters and latent variables. However, Murphy [79] argues that the term "belief" refers to subjective probability. This is a common interpretation in the Bayesian paradigm [97], implying the subjectivity that is inherent in modeling.

2.1.3 Two perspectives

The probabilistic model is an essential component of both concepts that are used in statistical inference, namely the frequentist (or classical) and Bayesian approach. The debate will not be discussed in full detail here. However, we will briefly outline the distinction as both are relevant to this thesis. Wasserman [96] distinguishes frequentist and Bayesian methods in Chapter 11 of his book using three postulates for each of them. We briefly state them in a direct comparison.

- While in the frequentist perspective, probabilities are considered to be objective and are represented through limiting relative frequencies, the Bayesian idea embraces a subjective view in which probabilities represent the degree of belief (or uncertainty).

[2] Coombs' interpretation was mainly concerned with psychological data. The quote is therefore actually a bit misused. However, the statement generalizes perfectly to statistical models.

- In the frequentist framework, parameters are unknown, but fixed quantities where we cannot make a probability statement about a parameter itself. A Bayesian view would allow making a direct probability statement. about a parameter.
- Frequentists focus on well-defined long run frequency properties (e.g. convergence or confidence intervals) in contrast to the Bayesian idea of inferring a full probability distribution.

Even though the postulates help to distinguish the two points of view, in practice, the distinction may not always be clear, as many ideas can be framed from either viewpoint (see for instance [55, 53]). This also means that much of the criticism apply for both approaches (for instance, the dependence on a probabilistic modeling story [46]). From a *practical* standpoint, we may even ignore the distinction as both use basically the same mathematical foundation and rather use tools from both views to solve a given problem.[3] We will not discuss this topic in more depth and use the classical distinction to introduce the relevant parts for the rest of the thesis.

2.2 Bayesian inference

The Bayesian idea dates back to the work of Thomas Bayes in 1763 [8]. Pierre Simon Laplace independently proposed the same idea (probably independently) in 1774 and developed it into the form we use today [72].

2.2.1 Bayes' rule

In the Bayesian framework, we consider a set of parameters and possible latent variables, denoted by θ where we express our *belief* using a prior distribution $p(\theta)$.[4] Our belief can also be viewed as *prior uncertainty* regarding θ [78]. Furthermore, we require observed data $\mathcal{D}$. Note that we often condition on further quantities as fixed parameters. We then update our belief about θ by forming the posterior $p(\theta \mid \mathcal{D})$ using the likelihood $p(\mathcal{D} \mid \theta)$ within Bayes' rule:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})}, \tag{2.1}$$

where the denominator is a normalizing constant, also known as the model evidence or marginal likelihood [11]. The marginal likelihood plays an important role in Bayes factors, as it serves as a measure of model fit to the data and, therefore, as a basis for comparing different models [90]. We can rewrite the denominator as

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta)\, dp(\theta) \tag{2.2}$$

to make this more clear. If we want to make predictions using new data $\mathcal{D}_{\text{new}}$, we can utilize the posterior

$$p(\mathcal{D}_{\text{new}} \mid \mathcal{D}) = \int p(\mathcal{D}_{\text{new}} \mid \theta)\, dp(\theta \mid \mathcal{D}),$$

which yields the posterior predictive distribution. The posterior predictive provides a natural uncertainty quantification that accounts for the modeled uncertainty with respect to θ and the data.

[3] It may be not fruitful to debate in terms of subjectivity and objectivity as it is often done in this context, but rather refer to attributes that can be motivated from those. We refer to Gelman and Hennig [39] for a more elaborate discussion of their role in statistical data analysis.

[4] Sometimes, the latent quantities, i.e., the parameters and latent variables, are referred to simply as latent variables [11].

Note that $p(\mathcal{D} \mid \theta)\, p(\theta)$ can be rather complex with multiple hierarchical layers with conditional distributions. We refer to the subsequent chapters for examples. The full definition in terms of a posterior allows the uncertainty expressed through probability distributions can propagate through different levels.

2.2.2 Learning the posterior

Using (2.1) implies that we can compute the normalizing constant given in (2.2) . This is only for a small subset of models possible (for instance, in the case of conjugate priors, where the posterior and prior belong to the same distributional family) [42]. If a closed-form solution cannot be found, one has to rely on numerical methods. The workhorse is probably Markov chain Monte Carlo, which will be discussed in more detail in the next section. Since the method is associated with a high computational burden, alternative approaches exist that try to alleviate the problem. Some prominent examples include variational inference [11], integrated nested Laplace approximation [88], sequential Monte Carlo [21], normalizing flows [100], and Stein variational gradient descent [69].

2.2.3 Markov chain Monte Carlo

The origin of Markov chain Monte Carlo (MCMC) dates back to the work of Metropolis et al. [74]. However, one could also pin down the origins to roughly one decade earlier, when the first Monte Carlo methods were put into action [73]. The method was later generalized by Hastings [54]. These are probably the canonical citations commonly referenced when delving into of MCMC methods. Even though the method has been optimized and has undergone a lot of changes over the past decades, the basic idea remains the same: we generate a Markov chain that converges to a stationary distribution of interest.[5] With respect to Bayesian inference, this is exactly the posterior since we only have access to the numerator of equation 2.1.

Metropolis-Hastings algorithm

In the standard Metropolis-Hastings (MH) algorithm, it is required that we can evaluate the numerator of the target distribution and that we can propose new values from a selected proposal distribution. The algorithm constructs a Markov chain $\boldsymbol{\theta} = \{\theta^{(t)} : t \in T\}$, where T is a discrete set, usually $\mathbb{Z}^+$. We use the proposal distribution conditioned on a previous state, i.e., $q(\cdot \mid \theta^{(t)})$ [6] and our (unnormalized) distribution from (2.1), i.e. a $\tilde{p}(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)\, p(\theta)$. First, we initialize the chain (e.g. randomly) at $t = 1$ with $\theta^{(1)}$. Then, we iterate T times as follows. We use the proposal distribution to obtain a new candidate state $\theta^{((t+1)^*)}$. Afterwards, we calculate

$$\alpha(\theta^{(t)}, \theta^{((t+1)^*)}) = \min\left(1, \frac{\tilde{p}(\theta^{((t+1)^*)} \mid \mathcal{D}) q(\theta^{(t)} \mid \theta^{((t+1)^*)})}{\tilde{p}(\theta^{(t)} \mid \mathcal{D}) q(\theta^{((t+1)^*)} \mid \theta^{(t)})}\right) \tag{2.3}$$

and set $\theta^{(t+1)} = \theta^{((t+1)^*)}$ with probability $\alpha(\theta^{(t)}, \theta^{((t+1)^*)})$, or retain the old state with probability $1 - \alpha(\theta^{(t)}, \theta^{((t+1)^*)})$, i.e., $\theta^{(t+1)} = \theta^{(t)}$. This is equivalent to drawing a random number from a uniform distribution over $[0, 1]$ and checking whether the drawn number is less than the ratio in (2.3). For computational reasons, we could also consider evaluating everything on the log scale. This makes the algorithm numerically stable [14], and evaluation can be accelerated since addition operations are computationally cheaper, and probability distributions such as the Gaussian no longer require the evaluation

[5] See for instance Meyn and Tweedie [75] for an in-depth discussion on Markov chains.

[6] It is actually not necessary, to have a conditional proposal distribution. It can take any form under some regularity conditions [45].

of an exponential function. Moreover, it is worth mentioning that an asymmetric proposal distribution would imply, that the proposal ratio $q(\theta^{(t)} \mid \theta^{((t+1)^*)})/q(\theta^{((t+1)^*)} \mid \theta^{(t)})$ in (2.3) reduces to one. This would reduce the MH algorithm to the Metropolis algorithm [74].
The presented procedure guarantees that the generated Markov chain converges to the posterior as its stationary distribution. This works because the MH ratio (2.3) is constructed, in such a way that the normalizing constant cancels out and only the numerator is required.
In general, the standard procedure is to discard a portion of the samples at the beginning (burnin). Furthermore, it can be useful to run an adaptive phase at the beginning to calibrate the proposal distribution for better sampling properties. In the classical algorithm, this would imply that one decreases or increases the dispersion of a simple proposal distribution, e.g., from a Gaussian proposal. However, this procedure is also used in more recent versions of the algorithm, where other features of the proposal may be tuned.[7].

Sampling quality and dimension of θ

The quality of the sampling in the MH algorithm heavily depends on the dimension of the parameter space that must be explored. If the dimension is high, this may be problematic because the curse of dimensionality impairs acceptance probabilities [10]. Even though the application of (2.3) guarantees that the generated Markov chain will eventually sample from the distribution of interest, it does not guarantee efficient sampling. In high dimensions, sampling primarily focuses on the typical set and any proposed state outside this set would lead to a very small acceptance probability. To alleviate the problem, one may use a tailored proposal distribution that generates new states with a high probability of being in the typical set. If a simple proposal distribution is used, such as a Gaussian in a random walk Metropolis, a very small standard deviation is required. If one wishes to stick with simple proposals, one can also propose a new state for a subset of the parameters or even a single element of θ. This resembles the Gibbs sampler [14], where each component of θ is updated one at a time. Even though this procedure allows sampling while preserving a certain degree of simplicity, it may lead to a high autocorrelation within the generated chain and the effective sample size is low. Therefore, many more iterations are required implying a larger computational budget. Alternatively, one might consider more complex proposal procedures. These are, for instance, algorithms that take the gradients of the posterior into account. The most popular versions are probably Metropolis-adjusted Langevin algorithm [28], hybrid Monte Carlo (also called Hamiltonian Monte Carlo) [30, 14], or the No-U-Turn Sampler [59]. While these methods offer more informed proposals, they also have limitations. For instance, they require the parameter space to be $\Theta = \mathbb{R}^p$, where p is the dimension of θ. One may use transformations or marginalization to satisfy this constraint. However, depending on the problem (and consequently, the specified models), these options are not straightforward to use. Moreover, an implementation is not a trivial task when the data structure is complex. In such cases, the standard approach of using probabilistic programming languages or inference frameworks (e.g. [15, 1, 86, 16]), which provide well-tested implementations of these algorithms, does not offer sufficient flexibility.

2.3 Classical approach

In this section, we discuss generalized linear models (GLMs) as a representative classical approach. In a GLM, we use a class of probabilistic models in which we regress an outcome on a set of covariates. It forms the foundation for Chapter 5. GLMs represent one of the most important methods of classical

[7] This could be, for instance, the number of steps, discretization time, or the mass matrix required in proposals that rely on Hamiltonian dynamics [14, 10, 58]

statistics and are the predecessor of a vast amount of more complex regression models (see for instance Wood [99]).

Generalized linear model

The GLM was introduced by Nelder and Wedderburn [80] defining a unifying framework for different regressions problems. We consider $n \in \mathbb{N}$ observations, represented in a data set $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ where $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}$. Each y_i is a realization of a random variable Y_i that belongs to the exponential family. Here, we consider θ as the parameters of interest, which usually represents $p+1$ regression coefficients. We use a vector of $p \in \mathbb{N}$ covariates (or features), often including an intercept; i.e., the i-th covariate vector is given by $\boldsymbol{x}_i = (1, x_{i,1}, \ldots, x_{i,p})$, where the first element is reserved for the intercept. The expected value of the dependent variable for the i-th observation is given by $\mu_i = E(Y_i)$ and is defined via a monotonic function h applied to a linear predictor:

$$\begin{aligned} \mu_i &= h(\eta_i) \\ \eta_i &= \boldsymbol{x}_i^T \boldsymbol{\beta}. \end{aligned}$$

We are mainly interested, to estimate the vector of coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$. For the fourth contribution, we consider a binary outcome, for which we model the probability of observing $Y_i = 1$ using the cumulative distribution function (CDF) of a Gaussian distribution Φ:

$$P(Y_i = 1) = \Phi(\eta_i)$$

Note that the canonical link would be, $g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$ while the probit model uses the probit function $g(\mu_i) = \Phi^{-1}(\mu_i)$, i.e., the inverse CDF of a standard Gaussian. One can motivate the probit model using a latent Gaussian variable with a threshold as is done in the contribution.
We write the log-likelihood of the model, parameterized by θ as a function of the data $\mathcal{D}$ as

$$\ell_\theta(\mathcal{D}) = \log p(\mathcal{D} \mid \theta) = \sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{x}_i, \theta).$$

$p(\mathcal{D} \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{x}_i, \theta)$ implies a conditional independence on the covariates $\boldsymbol{x}_i$ that are considered to be fixed.[8] Therefore, the log-likelihood is a function of θ.

Model fitting

The parameters θ of a GLM can be estimated using a gradient-based optimization method. Here, we consider the Fisher-scoring algorithm, which is also employed in the corresponding contribution. For that, we iteratively optimize the parameter estimate $\hat{\theta}_s$, where the subscript s is used to denote the s-th iteration. The update is performed by calculating

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \mathcal{I}^{-1}(\hat{\theta}_s)\mathcal{V}(\hat{\theta}_s)$$

[8] In the Bayesian framework we stay a bit more generic compared to the GLM case, where we do not define the data $\mathcal{D}$ explicitly like in the GLM. In the view of a GLM, we consider the data to be used in a "regress Y on X relationship" logic. We do this deliberately, to convey the generality of the presented Bayesian approach while we focus on a classical regression problem in this part. For instance, in the first two contributions, we consider up to four different outcomes that are used to inform a latent series.

where we use the first- and second-order gradient information, i.e., the score function and observed Fisher information, given by

$$\mathcal{V}(\hat{\theta}_s) = \left[\frac{\partial \ell_\theta(y, x)}{\partial \theta}\right]_{\theta=\hat{\theta}_s},$$
$$\mathcal{I}(\hat{\theta}_s) = \left[\frac{\partial \mathcal{V}(\theta)}{\partial \theta}\right]_{\theta=\hat{\theta}_s}.$$

2.4 Discussion and link to contributions

In this chapter, we discussed several aspects of statistical inference, covering the Bayesian approach in general and GLMs as a representative method within the frequentist framework. In the Bayesian framework, a common argument is its great flexibility. One of the main reasons for this flexibility is probably the generality of MCMC, which, at least in theory, works for a very wide range of problems. Therefore, the Bayesian approach often focuses a lot on the modeling part itself. This flexibility makes it ideally suited for the first three contributions. One can construct the joint posterior by defining small sub-models for specific components and concatenating them via conditional independence. Nonetheless, this flexibility is not a "silver bullet". Inferring a posterior distribution often implies a high computational burden. Moreover, the high complexity of the problems requires a custom implementation and considerable effort has to be put into the implementation to guarantee reliable software that minimizes errors, numerical stability and modularity for further development. This is exactly the case for the contributed articles. All of them employ a MCMC sampling scheme that is, at its core, always a MH algorithm, but is heavily customized to fit the problems.

Even though the Bayesian approach embraces subjectivity, it still require scrutiny. The flexibility carries the risk of becoming unnecessarily complex, and one has to weigh accuracy and robustness carefully. Additionally, probabilistic models can be sensitive to prior distributions or model assumptions (see, for instance Gelman and Yao [41]). One must investigate the quality of models, for example through simulation studies and sensitivity analyses. A Bayesian approach does not directly provide any guarantees on long-run performance; therefore, the frequentist properties of a developed model should also be investigated, e.g., using simulation studies.

On the other hand, frequentist models may offer a simple and direct approach for many problems. Models like the GLM are well understood and supported by a rich body of theory and literature. One can rely on a fixed model class that is often proven to perform well. However, due to this inflexibility, a researcher would possibly opt for smaller samples where stricter assumptions can be defended, or even make simplifying assumptions about the model itself. This bears the risk that scientific questions are made to fit the tools rather than choosing the most tailored solution to answer the question of major interest. In situations such as contribution four, the problem naturally lends itself to the application of these classical methods, which allow for a robust and fast solution.

Chapter 3

Modeling of infectious diseases

The mathematical modeling of infectious diseases has a long tradition, dating back centuries [26, 56]. The first highly recognized work was published by Daniel Bernoulli in 1760 which formulated an infection process as a deterministic differential equation [3]. However, as progressive this work has been at this time, the modeling approach was put to sleep and was awakened from its slumber many years later in the 20th century [57, 52]. Since then, the field of mathematical modeling of infectious diseases has come a long way with landmarks like the first edition of the book of N. T. J. Bailey in 1975 [4, 57]. With the outbreak of the COVID-19 pandemic [104], that is arguably one of the most influential events in the past years, resulting in millions of deaths [20, 29], the modeling of infectious diseases gained even more traction leading to a plethora of scientific publications featuring a high number of modeling approaches (see for instance [20, 25]).
Institutions and scientists published data related to the pandemic (see for instance [29, 24, 51, 49]). These sources made it possible for scientists around the world to leverage data from many countries to gain insights and understand characteristics of the infection dynamics. With the progression of the pandemic, the challenges became more complex. New variants led to a faster spread of the virus and affected other parts of the dynamics, such as the severity of an infection. The development and distribution of vaccines had opposite effects. Social and political actions also changed as more information became available and infrastructure was adapted.
In this chapter, we discuss the relevant methods and selected modeling approaches that form the foundation for the first two contributions (Chapters 6 and 7).

3.1 Overview

To infer infection dynamics, we must rely on available data. Officially reported cases do not represent the true number of infections, which therefore has to be treated as a latent variable in a probabilistic modeling approach. One can, of course, make simplifying assumptions that allow official data to be used directly. However, this would either limit a potential analysis to a very short observation window or one would risk biases in the estimation. When the COVID-19 pandemic hit Europe, the most relevant data series were reported cases, collected through testing, reported number of deaths, or numbers that are related to hospitals like hospital admissions or occupations of intensive-care units. These sources bear a lot of uncertainty with respect to location and time. For instance, reporting or testing policies changed over time and varied between countries making it non-trivial to use them in a model. The next section aims to introduce a method, that can be used to combine the different data sources to infer the actual number of infections.

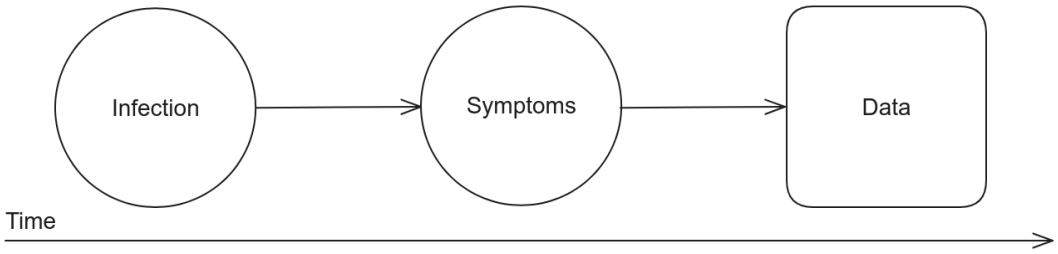


Figure 3.1: Simplified DGP for an infected individual. After an infection, a person will develop symptoms. Afterwards the person will possibly contribute to official data.

In the subsequent section, we will discuss infection dynamics and the role of the reproduction number. While SIR-based models are probably the most recognized approach in standard infectious disease modeling, other methods exist that are closely related. In this section, we consider a generalized version: the renewal equation that characterizes the infection process of a communicable disease.

3.2 Multiple sources of information

All the reported quantities mentioned above (cases, deaths, hospital) are not flawless and come with their own issues that can lead to contradicting information. However, we want to use as much information as possible. How can this problem be addressed? Contribution I attempts to give an answer. The goal is to find a way that allows us to include all time series. This can be done by incorporating the unique problems of each series into sub-models with its own likelihood and combine them to infer a latent variable. Since all series are count data by definition, that may exhibit overdispersion, one can use a negative binomial distribution parameterized with its mean and a dispersion parameter. Therefore, the general form to model the observed counts on day $t \in \{1, \dots, T\}$ can be expressed as a negative binomial (NB) distribution. We consider a random variable $X_{t,m}$, observed at time points $t \in \{1, \dots, T\}$ in location $m \in \{1, \dots M\}$, which is assumed to be

$$X_{t,m} \sim NB(\mu^X_{t,m}, \phi^X). \tag{3.1}$$

Here, we parameterize the NB distribution not in the standard form, but in terms of a location parameter $\mu^X_{t,m}$ and an overdispersion parameter ϕ^X.[1] The variance is then given by $\mathbb{V}(X) = \mu^X_{t,m} + \frac{{\mu^X_{t,m}}^2}{\phi^X}$. One can make two important observations from this chosen parametrization. 1) For $\phi^X \to \infty$, the variance of $X_{t,m}$ becomes equal to its mean and the distribution reduces to a Poisson distribution without overdispersion. 2) Like in the Poisson case, the variance of the random variable depends on the location parameter. Hence, a small value of $\mu^X_{t,m}$ will result in a small variance of $X_{t,m}$ and a large value will lead to a large variance. Note that this is always relative to a given value of ϕ, that also affects the variance. Using a Bayesian framework, we can set an adequate prior on ϕ^X to control this variance.

We aim to infer the true and unobserved number of actually infected individuals, denoted as $I_{t,m} \in \mathbb{N}_0$ from observed data. For that, we can deduce an intuitive DGP that represents the connection between $I_{t,m}$ and the data. Given that an individual gets infected, incubation period passes until symptoms are developed. After a disease becomes recognizable through symptoms, it may lead to different outcomes. An individual has a certain probability of being tested, possibly receiving a positive test result, and,

[1] Therefor, the density is given by $p(x \mid \mu_{t,m}, \phi) = \frac{\Gamma(\phi+x)}{x!\Gamma(\phi)} \left(\frac{\mu_{t,m}}{\mu_{t,m}+\phi}\right)^x \left(\frac{\phi}{\mu_{t,m}+\phi}\right)^\phi$.

consequently, being reported as a case. The person may also develop severe symptoms such that an admission to a hospital is necessary. In some cases, an infection can be fatal. Figure 3.1 shows this process as a simplified image over time. Note that we denote the last node as "Data". With this, we mean that an individual may contribute to reported data. This could be in the form of a case, in hospital statistics, or in fatality counts. A single case can, therefore, also contribute to multiple time series. The DGP assumes that symptoms will be developed after an infection, and that a contribution to the data can only occur after symptom onset. Only a fraction of infected individuals will experience an event that contributes to the data. Moreover, the time until an event occurs varies among infected individuals. This implies a distribution within the time domain: Some individuals will experience an earlier event, while others will experience later events. When considering an aggregated level (e.g., a country) and staying in discrete time, we can express this varying time shift for a share of the infected individuals as

$$\mu_{t,m}^{X'} = \pi^{X'} \sum_{u \leq t} X_{u,m} \left(F_{X'}(t-u+1) - F_{X'}(t-u) \right), \tag{3.2}$$

where $F_{X'}$ is a continuous distribution that represents the time until an event occurs, e.g., the time until an individual develops symptoms after an infection. The expected number of all events $\mu_{t,m}^{X'}$ at time t is therefore a function of all counts that occurred in the past $X_{u,m}, u \leq t$. The term $(F_{X'}(t-u+1) - F_{X'}(t-u))$ can be seen as a discretization, that represents the whole probability mass between two consecutive days $(t-u)$ and $(t-u+1)$. Thus, (3.2) states that we obtain the expected value $\mu_{t,m}^{X'}$ for a random variable $X'_{t,m}$ at time t in location m as the sum of all past values $X_{u,m}, \forall u \leq t$, weighted by their probability that an event is occurring after u and $u+1$ days. This sum in (3.2) is then multiplied by a factor $\pi^{X'}$ to reflect a possible fraction or multiple of the value. To illustrate, consider a simplified example: Suppose that the distribution $F_{X'}$ has all its probability mass concentrated on one specific day (e.g., a one-day lag) and that $\pi^{X'} = 1$. In this case, the expectation of $\mu_{t,m}^{X'}$ would simply be the value of $X_{t-1,m}$ and the process would resemble a random walk with negative binomial innovations (see e.g., [31] for further information on time series modeling). As a second example, consider a distribution that has a uniform probability mass over three days, such that every day has a mass of $\frac{1}{3}$, and let $\pi^{X'} = 0.1$. This would shift and "smooth" the past values.[2] $\pi^{Y} = 0.1$ scales this down such that the event occurs only in 10% of the cases. As abstract as this concept is so far, it can be directly used to formulate a relationship between the unobserved infections during an outbreak and the officially reported time series. The concept is used in both contributions of this thesis and was originally borrowed from Flaxman et al. [36], where it is used to model a DGP for fatality counts. The corresponding time shifting distribution, therefore, models the time between an infection and the time until death. Contribution I of this thesis generalizes the concept to utilize multiple considered time series: In a first step account for the incubation period until infected individuals would develop symptoms. This is considered to be a deterministic shift and we do not treat the number of symptomatic cases as stochastic. This would cause identification problems because we would have two latent, unobserved variables with a direct dependency. Moreover, not all infected cases develop symptoms. In this case, we would rather refer to them as pseudo-symptomatic cases.
Using the time-shifted variable of counted cases, we can follow the assumed DGP and consider the potential contribution to the reported data. For instance, an individual may get tested and subsequently reported as an actual case. We can use the logic from equation (3.3) to model this process. This requires

[2] Note that this is not an actual smoother such as a 3-day moving average. Equation (3.2) can be seen as a partial convolution since it only considers the *past* values at $u \leq t$ [13]. However, one could represent it as a standard convolution by discarding certain values. Computationally, this can be faster as the fast Fourier transform can be utilized to reduce the convolution operation to a multiplication operation in the Fourier domain [27] and may be optimized in high-level programming languages [81].

a time-shifting distribution that reflects the time between symptom onset and the actual contribution to the data.

Accounting for uncertainties in reported data

The mapping to subsequent events from infections can be highly dependent on the current regime in a pandemic. For instance, in phases of high incidence within a certain country, official testing capacities were not sufficient. This implies that only a small fraction of actual cases was reported as such while this may not play a crucial role in phases of low incidence. Moreover, official testing policies also have changed over time, and other location-specific, seasonal effects (such as weekend effects) may impaired reporting further.
The time-shifting distribution and scaling factor $\pi^{X'}$ should therefore reflect local and time-specific features. Hence, we would like to adapt both quantities accordingly, such that they can represent individual characteristics. Making them part of a probabilistic model gives us the flexibility to account for these particular uncertainties in the data. However, one has to be cautious as high flexibility gives rise to overfitting and identification problems.
Besides reported cases, we can use the same procedure to connect the true cases to the other series, such as reported deaths or hospital data. Within these series, we may consider that new dynamics, such as improved severity due to new variants or vaccination coverage in the population, affect the share of individuals who die, are admitted to a hospital, or remain in intensive-care units. This, again, implies a flexible representation of the time-shifting distribution and scaling factor. Combining all available series while accounting for the mentioned uncertainties may offer a way to infer a meaningful picture of the actual infections.

Contributions I and II heavily rely on this concept, in which we define all the connections to the data via time-shifting distributions as sub-models of the likelihood. Within a Bayesian approach we attempt to estimate parts of the characteristics, that are required, to obtain a meaningful $F_{X'}$ and $\pi_{X'}$.

3.3 Modeling the transmission process

After we have modeled the connection between the latent variable representing infected individuals, we can use this relationship to infer features of the disease transmission process. This can be approached from different angles. One prominent approach is the simulation of a full, possibly stochastic, pseudo- or meta-population (see, for instance, [34, 35]). These models require strong assumptions about a population and its behavior that may have a significant impact on the conclusions. However, they provide high flexibility and one can examine counterfactual outcomes by making small modifications. The transmission can also be modeled via network approaches that can be combined with other ideas [82, 19, 37]. The last approach to be mentioned here, which is arguably the most popular, is the method of compartment models, often also referred as SIR models [101, 13]. Even though the term "SIR" imply the basic form, it generally refers to much more complex models with many compartments compared to the three standard ones (susceptible, infectious and recovered).

Renewal equation

The compartment modeling approach is closely connected with the utilized approach in the contributions of this thesis, namely, the renewal equation. Although the motivation stems from the Euler–Lotka equation that is originally used by demographers, ecologists and evolutionary biologists, the renewal equation can

also be used to model the transmission of an infectious disease where it provides an intuitive interpretation [13]. The connection to SIR models stems from branching process theory, where the difference is the higher flexibility of renewal equations through a more flexible generation time distribution. Also known as the serial interval, it defines the duration of a disease transmission, i.e., the time an infected individual transmits the disease to a secondary susceptible [95, 36]. The renewal equation can be formulated in continuous time. However, since the contributed articles consider a equidistant discrete time grid, we write it as a discrete sum:

$$\mu_{t,m} = R_{t,m} \sum_{u<t} I_{u,m} \left(F_\gamma(t - u + 1) - F_\gamma(t - u) \right), \tag{3.3}$$

where the expected number of current infections $\mu_{t,m}$ depends on past infections $I_{u,m}, u \leq t$, the generation time (or serial interval) distribution F_γ, and the instantaneous reproduction number $R_{t,m}$. We can interpret the expected number of current infections as a weighted sum of past infections that is scaled by the reproduction number. It is immediately recognizable that equation (3.3) is a special case of the time-shifting procedure in (3.2), where we have an autoregressive character and an intuitive interpretation of π^Y as the reproduction number $R_{t,m}$. The generation time distribution F_γ quantifies the probability that an infected individual transmits the disease to a secondary susceptible within a given time interval.

Reproduction number

$R_{t,m}$ is a dynamic quantity with changes over time and can depend on both the population and its location. Different reproduction numbers can be defined, where all of them share the same core interpretation as the expected number of secondary infections caused by a single infected individual. In the following, we discuss the two reproduction numbers, that are utilized in Contributions I and II. The *basic* reproduction number R_0 is defined as before, but with the assumption to be in a population of fully susceptible individuals [95, 13]. We can view R_0 as a threshold quantity that determines whether a disease can invade and persist in a new host population [91]. For values less than one, infections would causing the disease to eventually die out. For values greater than one, the disease has a positive growth rate and can invade the population.
The *instantaneous* reproduction number R_t is a property of the disease dynamics at a specific time t. It is the expected number of secondary infections, assuming that the conditions at time t remain unchanged. This means that for an arbitrary time point, it can be modeled as a function of the basic reproduction number and other circumstances that are prevalent at this time. These conditions may include a smaller pool of susceptible individuals (e.g, immunization through infection or vaccination), changed contact patterns (e.g. though changes in social behavior), or a drift in the contagiousness through mutations of the virus.

3.4 Discussion and link to contributions

The presented concept of inferring a latent series from observed data to learn disease dynamics implies that the population (with respect to a location m) is homogeneous and that individual characteristics can be modeled by independent noise processes. To relax this rather strict assumption, it would be necessary to use fine-grained data that provide this information. Alternatively, one could impose assumptions about the population as it is done in a pseudo-population approach, suggesting that the boundaries between different approaches are not strict.
As discussed in Chapter 2, the strength of the Bayesian approach lies in its high flexibility. An assumed DGP can be defined through the combination of sub-models and prior information. This makes it a perfect fit for modeling infectious diseases. It gives the freedom to model all relevant parts of the DGP under

a complex data situations with high uncertainty. The probabilistic formulation allows the propagation of this uncertainty in the model to obtain reasonable posterior estimates for quantities of interest. The use of data from different locations also allows the formulation of hierarchical levels that can be utilized in in a probabilistic model. This provides a direct approach for sharing information between locations while allowing a certain degree of flexibility.

Contribution I (Section 6) utilizes the presented concepts to demonstrate how multiple time series can be integrated to infer the effect of NPIs during the COVID-19 pandemic by leveraging data from 20 European countries. Through the Bayesian formulation, it is straightforward to model quantities of interest like NPIs or the reproduction number.

Contribution II (Section 7) builds on the methods developed in the first contribution. The idea stems from the observation that the data structure at the European level - consisting of daily counts for every county - is fundamentally the same as that at the German level when using federal states as the location unit. Even though, the data structure can be framed in the same way, the collected data itself differs for some of the time series, implying the necessity to adapt the developed method.

Chapter 4

Measurement error

This chapter aims to provide the methodological background on measurement error (ME). We discuss the topic in the context of occupational cohort studies.
In general, a ME can vaguely be described as a discrepancy between a true value of a quantity and the measurement of it. With respect to probabilistic models, a ME can be defined as *not observing* the true value of a random variable X, but an error-prone version, denoted Z. With respect to epidemiological studies we only consider in the following measurement errors in the *exposure*, e.g., where one is interested in the association between a dependent variable Y and and an exposure X. Moreover, we consider only errors on variables, measured on a continuous scale. Hence, categorical errors and errors in the dependent variable are beyond the scope of this text.

4.1 Types of measurement error

A ME can be categorized based on several characteristics. These characteristics are generally defined in the *ME model* [65]. It is important to note that measurement error can always be categorized according to additional characteristics. However, the context often implies some of these features without stating them explicitly. To illustrate, consider the standard introduction to ME, which motivates the concept using an unshared, additive, classical ME that is therefore implicitly assumed to be non-differential. For instance, see Yi [102], Carroll et al. [17], and Gustafson [50], where all of these sources use exactly this type of error as a introductory example in the context of a linear regression. The reason is probably that it is easy to understand and one can directly show, the consequence of the error (see also Section 4.2). In the following we distinguish MEs based on characteristics that are important for Contribution III (Chapter 8).

Classical Error

As stated earlier, the classical ME is the standard model used to introduce the concept [22]. The standard formulation is given by

$$Z = X + U_C, \quad \text{where } X \perp\!\!\!\perp U_C,$$

where we consider Z, X and U_C to be real-valued random variables. Z represents the observed version of the true variable X, perturbed by an error U_C. In the most prominent error model, it is assumed that the measurement error follows a zero-mean Gaussian distribution, i.e., $U_C \sim N(0, \sigma_C^2)$. Therefore, the assumption in this model is that the *observed* variable deviates randomly from the true value without any

systematic pattern. However, the most essential feature of a classical ME is the assumed independence of X and U_C.

Berkson Error

In contrast to the classical error, a Berkson error assumes that the *true* value deviates from a known value [9]. Mathematically, the ME model can be written as

$$X = Z + U_B, \quad \text{where } Z \perp\!\!\!\perp U_B.$$

One may again consider a Gaussian measurement error, $U_B \sim N(0, \sigma_B^2)$. In this case, the error is independent of the observed variable Z. This type of error is prominent in occupational epidemiology [65]. To give a typical example in this context, we consider groups $j = 1, \dots, J$ of workers where each group j is assigned a measured or estimated amount of exposure Z_j. Hence, each individual within group j gets assigned exactly the same value, e.g. the mean (or level) exposure measured through an ambient measurement at a workplace j. However, an individual i in group j is assumed to deviate from this mean value. Therefore the actual exposure for an individual can be written as $X_{i,j}$ for all individuals $i = 1, \dots, n_j$.

Combination of classical and Berkson error

One can consider a combination of classical and Berkson error models [60]. To illustrate, consider the previous example. However, the mean exposure for group j is not measured exactly, but rather is measured with error now. This is a common case in occupational cohort studies. For instance, this occurs when exposures are measured through a job-exposure matrix (JEM). Mathematically, this can be written as

$$\begin{aligned} Z_j &= X_j + U_{C_j} \\ X'_{i,j} &= Z_j + U_{B_i}, \end{aligned} \tag{4.1}$$

where we still use Z_j to denote the measured exposure level for each group $j = 1, \dots, J$, which is an error-prone version of X_j. Furthermore, we allow a deviation for each single worker $i = 1, \dots, n_j$ *within* the group j, leading to the individual exposure $X'_{i,j}$ that is not observable. Note that this model only has the measured version of the mean exposure level as an observed quantity.

Shared and unshared errors

(4.1) implicitly suggests a shared classical error: The error U_{C_j} affects all individuals in group j exactly in the same manner, since everyone is subject to the same mean ambient measurement. They *share* the same error. When looking at the Berkson error formulation, the example implies an *unshared* Berkson error component, as every individual experiences an individual-specific deviation U_{B_i}. This also implies a different dimension of the latent exposure: The mean ambient exposure is measured for J groups and therefore we could define it as an error $U_C \in \mathbb{R}^J$.[1] The dimension of the actual individual exposure, which is also influenced by the Berkson error, is defined by adding the number of all individuals in all groups, i.e. the full Berkson error U_B has dimension n with $n = \sum_{j=1}^{J} n_j$ as the total number of individuals in the cohort or data set. In this example, the Berkson error is unshared among workers. However, one can also consider a shared Berkson error, where a deviation from an overall mean level of exposure represents

[1] The missing subscript on U_C implies a larger dimension, than for U_{C_j} as U_C denotes the vector of all mean exposure values, i.e. $U_C = (U_{C_1}, \dots, U_{C_J})^T$.

different workplaces [61]. For instance, all individuals in a workroom may share the same Berkson error if the overall level is taken over the whole building and we want to model the exposures on a room level. The assumed model for shared and unshared error structures has a significant impact on computation because it heavily affects the dimension of the implied latent variable (see also Chapter 2).

Additive and multiplicative errors

The equations from before are always written using an *additive* error. However, although the additive structure is more popular, one may also consider a *multiplicative* error model, wich is more prominent in occupational and environmental epidemiology [61, 2]. In this case, the dependence assumptions between errors and observed variables remain the same as in the previous cases. Instead of a Gaussian distribution as the most popular choice for additive errors, we can use a log-Gaussian distribution for multiplicative errors. For instance, the classical error model would be written as $Z = X \cdot U_C$, where we assume that $\log(U_C) \sim N\left(-\frac{\sigma_C^2}{2}, \sigma_C^2\right)$. In this case, the location parameter of the underlying Gaussian distribution is set to $-\frac{\sigma_C^2}{2}$. Since the mean of a log-Gaussian distribution is defined as $\exp(\mu + \frac{\sigma^2}{2})$, setting $\mu = -\frac{\sigma_C^2}{2}$ ensures that the mean equals one. This is analogous to the definition of a zero-mean Gaussian for an additive error, as the error does not change the value of the error-prone variable in expectation.

More on types of errors

So far, the description of different error types represents the relevant standard theory that is required for Contribution III (Chapter 8). However, some other concepts play a minor but relevant role. The first one to be mentioned is the distinction between differential and non-differential ME. If the outcome (dependent variable) is conditionally independent of the error given the true non-error-prone covariate X, the error is called non-differential. This means that the error has no influence on the outcome and is stochastically independent [103]. On the other hand, if the ME is not independent of the outcome, the ME is considered differential. In Contribution III, we assume a non-differential ME.
Other characteristics worth mentioning include the differentiation between ME on the outcome and/or on covariates [92], and ME with a different sample space, e.g., a categorical ME that can be interpreted as misclassification [65, 50].

4.2 Accounting for measurement error

We consider statistical problems in which we are interested in the association between an outcome and covariates (exposure), with the strength of this association parameterized by θ_d, possibly as a part of a larger parameter set θ_D (see disease model in Section 4.2.2). Inference on error-prone data generally leads to biased results [17, 103, 50]. For trivial cases, such as linear regression, one can directly derive a formula for the resulting attenuation of the coefficient. In more complex cases involving a non-linear link, interactions, or more complex dependence structures, the behavior is not always clear and may even lead to reverse attenuation [17, 62, 77, 94, 2]. Consequently, solutions to account for ME in data is inevitable.

4.2.1 Approaches to account for measurement error

Many different approaches have been proposed over the past decades. Here, we focus on the Bayesian approach used in Contribution III. We refer to Keogh et al. [65] and Shaw et al. [89] for an overview of alternative ideas.

4.2.2 Bayesian approach to account for measurement error

In the Bayesian approach, we model the ME as part of the DGP. This is usually a probabilistic model that defines the relationship between an outcome of interest and the exposure and potential other potential covariates. To model the DGP, we require at least three sub-models [87].[2] These are the disease model, the measurement model and the exposure model. We connect them through conditional independence assumptions to one large model. We will discuss each of them, along with a brief discussion of prior distributions.

Disease model

The disease model defines the association between an outcome Y_i and the covariate of interest X_i for individuals $i = 1, \dots, n$, where the association is quantified by the parameter θ_d. For simplicity, we stick to a single covariate. We write this association as $p(y_i \mid x_i, \theta_D)$. θ_D is the collection of *all* parameters of the disease model. This collection contains at least θ_d. Other parameters may represent additional characteristics like dispersion parameters or additional hazard parameters.

Measurement model

The measurement model describes the relationship between the observed, error-prone version Z_i of the variable X_i, that is $p(z_i \mid x_i, \theta_M)$. For instance, in the most prominent example of an unshared classical additive error, which is a zero-mean Gaussian, the model is defined as stated earlier in Section 4.1 with only one parameter, $\theta_M = \sigma_C^2$. However, the model can be arbitrarily complex, as long as it is identifiable (e.g., through adequate prior assumptions).

Exposure model

The exposure model is only relevant for classical errors. It defines the distributional assumption of the latent variable X_i, where the model parameters are denoted by θ_E and therefore $p(x_i \mid \theta_E)$.

Prior distributions

In the Bayesian approach, we define prior distributions on all parameters of the three models, i.e., $p(\theta_D), p(\theta_M)$, and $p(\theta_E)$. As discussed in Chapter 2, we can deliberately express our beliefs or uncertainty about the parameters here. A reasonable approach would be to impose additional hierarchical structures that define prior distributions on the parameters of the prior distributions (sometimes called hyperpriors).

4.2.3 Inference

Given all models, they can be combined into an unnormalized posterior distribution:

$$p(\theta_D, \theta_M, \theta_E, x \mid y, z) \propto p(\theta_D)p(\theta_M)p(\theta_E) \prod_{i=1}^{N} p(y_i \mid x_i, \theta_D)p(z_i \mid x_i, \theta_M)p(x_i \mid \theta_E) \tag{4.2}$$

[2] This idea is equivalent to the likelihood-based approach that also defines the same sub-models. The main differences lie, as outlined in chapter 2, in the treatment of parameters and the estimation process.

As for the most Bayesian problems, it is only in very rare cases possible to find a closed form solution and one has to opt for numerical solutions as described in chapter 2. The true exposure can be be a high-dimensional latent variable. However, depending on the chosen algorithm and measurement model, it may be possible to utilize the error structure to mitigate the computational burden a bit by implementing custom computing routines.

4.3 Discussion and link to contributions

The probabilistic model in (4.2) is written in an abstract form, without stating exact relationships. While the disease model is, in many cases, straightforward (e.g., a regression or survival problem), the measurement error model can get rather complex. In many cases, the model is just a simple unshared. However, if the measurement process is complex, the measurement model will also be complex. This is exactly the case of contribution III (Chapter 8), which uses data from the Wismut cohort [66]. The measurement model is a combination of many different shared Berkson and classical error components. Further, we deal not only with one, but multiple measurement models. The reason for the involved error structure is the complex data situation arising from multiple different approaches that were used to obtain exposure measurements over different periods and locations. Even though the data can be organized into a single table in long format, it is inherently complex. Therefore, the Bayesian approach is perfectly tailored to the problem, as it allows for the definition of custom sub-models for each part of the DGP. Contribution III also demonstrates how a custom MCMC sampler can be used to exploit the prevalent error structure.

Chapter 5

Validation of models on distributed data

The performance evaluation of binary classification algorithms is an essential task in many scientific disciplines. In particular, a medical context often demands critical decisions affecting patients directly. Therefore, it is important to have reliable algorithms and predictors. Leveraging a large amount of data to obtain trustworthy results through the validation of prediction methods is therefore necessary. However, data are often distributed across different locations, e.g., different hospitals. The common approach of centralizing data at one location is often not possible. One of the major reasons is the need to protect individual privacy. We, therefore, face two challenges. 1) The algorithms must be designed in such a way that they can work with data that is distributed across different locations. 2) Algorithms must guarantee the protection of individual privacy.
In the following, we introduce the building blocks for Contribution IV (Chapter 9). The idea relies on a classical statistical method, namely the GLM, which was introduced in Chapter 2.
We consider a binary classification problem where we use a function $f : \mathbb{R}^p \to \mathbb{R}$ that maps a p-dimensional vector of covariates or features $\boldsymbol{x} \in \mathbb{R}^p$ to a score.[1] The vector is a realization of the associated random variable X. We estimate or learn this function using a statistical model on (training) data, obtaining $\hat{f}$. Furthermore, we have access to a sample of observational data $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, where y_i is binary and represents the i-th realization of the binary variable Y. Note that $\mathcal{D}$ is not the data that was used to train $\hat{f}$, but rather the data that we use to *validate* a learned classifier. To do this, we define a discrimination threshold $c \in \mathbb{R}$ within an indicator function $\mathbb{1}_{(c,\infty)}(\hat{f}(\boldsymbol{x}))$. c is used to define a binary classifier derived from $\hat{f}$ based on this threshold.

5.1 ROC curve

The receiver operating characteristic (ROC) curve is a method to evaluate the performance of a binary classification model as a function of the threshold c. ROC analysis is a well-established method with a long history in signal detection, the evaluation of diagnostic systems, and medical decision making. It was also adopted by the machine learning community decades ago [33]. If we plot the true positive rate[2] (TPR) and false positive rate[3] (FPR) of $\hat{f}$ for varying values of c, we obtain the ROC curve. The TPR (also known as sensitivity) is the proportion of actual positives correctly identified, while the FPR

[1] The function may also map $f : \mathbb{R}^p \to [0, 1]$, where an algorithm either directly models probabilities or uses a link function to squash values between 0 and 1. However, in popular cases like logistic regression, these transformations do not affect the result, as the ROC curve is invariant to monotonic transformations [85].

[2] Also called true positive fraction.

[3] Also called false positive fraction.

(which equals 1–specificity) is the proportion of actual negatives incorrectly identified as positive. With "positive" we mean that the value of the variable $Y = 1$, whereas "negative" would imply $Y = 0$. We define the TPR and FPR as functions of the threshold c (conditioned on Y):

$$\begin{aligned} TPR(c) =& P(\hat{f}(X) \geq c \mid Y = 1) \\ FPR(c) =& P(\hat{f}(X) \geq c \mid Y = 0). \end{aligned} \tag{5.1}$$

The ROC curve (or space) is the set of all pairs of FPR and TPR values derived from all possible thresholds c:

$$\text{ROC} = \{(FPR(c), TPR(c)) : c \in \mathbb{R}\}. \tag{5.2}$$

We can rewrite the ROC curve as function of a variable r: $\text{ROC}(r) = S_1(S_0^{-1}(r))$, where $S_0(c) = FPR(c)$ and $S_1(c) = TPR(c)$ are called the survivor functions of Y for the negatives (non-diseased population) and positives (diseased population). We rewrite this as

$$\begin{aligned} \text{ROC}(r) =& S_1(S_0^{-1}(r)) \\ =& P(\hat{f}(X) \geq S_0^{-1}(r) \mid Y = 1) \\ =& P(S_0(\hat{f}(X \leq r) \mid Y = 1), \end{aligned}$$

where $S_0^{-1}(r) = c$. The ROC curve can therefore be expressed in terms of *placement values*, i.e. the positives (diseased) with respect to the survivor function of the negatives (non-diseased).

5.1.1 AUC

The area under the curve (AUC) is perhaps the most common measure used to summarize a ROC curve as a single value and is defined as

$$\text{AUC} = \int_0^1 \text{ROC}(r)\ dr.$$

If $\text{AUC} = 1$ the learned function $\hat{f}$ would perfectly discriminate with respect to the validation data. For $\text{AUC} = 0.5$, the classifier is useless because it performs no better than random guessing. If $\text{AUC} < 0.5$, it would be preferable to simply flip the output of the classifier.

5.1.2 Estimation of the ROC

The ROC curve is not directly observable and we must estimate it from our validation data $\mathcal{D}$. The most popular method is an empirical, non-parametric estimate. However, we adopt an alternative formulation based on a GLM. For completeness, we briefly state the standard method, which is used as a comparison measure in Contribution IV. The standard method estimates $\widehat{TPR}(c) = n_1^{-1} \sum_{i=1}^{n_1} \mathbb{1}_{(c,\infty)}(\hat{f}(\boldsymbol{x}_i^{(1)})$ and $\widehat{FPR}(c) = n_0^{-1} \sum_{l=1}^{n_0} \mathbb{1}_{(c,\infty)}(\hat{f}(\boldsymbol{x}_l^{(0)})$, where we use $\boldsymbol{x}_i^{(1)}$ and $\boldsymbol{x}_i^{(0)}$ to denote the i-th or l-th vector *within* the subgroups of positives and negatives respectively, i.e., $y_i = 0$ and $y_i = 1$, with n_1 and n_0 representing the number of data points in each group. These formulas are merely the empirical versions of (5.1).

ROC-GLM

The ROC-GLM is an alternative approach that yields a smooth ROC curve. The original idea was proposed in Pepe [84]. The approach formulates the ROC curve as a regression problem $\text{ROC}_g(r \mid \boldsymbol{\gamma}) = g(h(r)\boldsymbol{\gamma})$, where we use a link function $g : \mathbb{R} \to [0, 1], \eta \mapsto g(\eta)$, regression coefficients $\boldsymbol{\gamma} \in \mathbb{R}^l$ and covariates defined though a function $h : \mathbb{R} \to \mathbb{R}^l, r \mapsto \boldsymbol{h}(r) = (h_1(r), \ldots, h_l(r))^T$. Note that h in this context is

not the same function as defined in Chapter 2 to model the response in a GLM, but a transformation of r, and g is used to model the response (unlike the link function g in Chapter 2).
We can then fit a ROC-GLM by constructing a pseudo-data set $\mathcal{D}_{\text{ROC-GLM}} = \{(\boldsymbol{h}(r_j), u_{ij}) \mid i = 1, \ldots, n_1, j = 1, \ldots, n_R\}$ where we have $h(r_j)$ as covariates and u_{ij} as target variable. The construction is done using thresholds $R = \{r_1, \ldots r_{n_R}\}$, where we obtain the targets as $u_{ij} = \mathbb{1}_{(-\infty, r_j)}(\hat{S}_0(f(\boldsymbol{x}_i^{(1)}))$. Further, we can define as regression coefficients $\boldsymbol{\gamma}$ the function $\boldsymbol{h}(r) = (h_1(r) = 1, h_2(r) = \Phi^{-1}(r)$. Using this exact definition, we would directly obtain the bi-normal ROC model when we use a probit link, i.e. $\text{ROC}_g(r \mid \boldsymbol{\gamma}) = \Phi\left(\gamma_1 + \gamma_2 \Phi^{-1}(r)\right)$. The model can then be estimated as described in Chapter 2 using the Fisher-scoring algorithm.

5.2 Statistical disclosure control and differential privacy

Here, we state the definition of Castro [18]: *"Statistical disclosure control (SDC) comprises the set of methods for preserving individual and confidential information when releasing data"*. SDC plays a crucial role in applications where the privacy of individual data is important. A technique, can be considered part of SDC, is *differential privacy* (DP). However, DP is not often discussed in the context of SDC. In the following, we will discuss one basic method of both SDC and DP as both are essential parts of Contribution IV

5.2.1 Secure aggregation

The term "secure aggregation" is not always explicitly mentioned in SDC sources but discussed as an implicit method to preserve privacy in statistical outputs (see, for instance, Hundepool et al. [63] and Griffiths et al. [48]). This is done in a a straightforward way: Aggregated data can be considered as safe as long as the share of the individual's contribution is not too large. This logic is also often used in magnitude and frequency tables [18]. A closely related concept is k-anonymity. However, the two concepts are not identical: k-anonymity is rather a property of a microdata set (or their quasi-identifiers), while secure aggregation is a property of a function applied to the data. We may define the output of an aggregation function $a : \mathbb{R}^d \mapsto \mathbb{R}, \boldsymbol{v} \to a(\boldsymbol{v})$ to be secure if d is greater than a chosen threshold. For instance, an empirical mean or a frequency table can only be used, if a sufficient number of data points contribute to the aggregated result(s).

5.2.2 Differential privacy

Differential privacy (DP)[32] is a mathematical framework that provides quantifiable privacy guarantees. This quantifiability stands in contrast to other approaches where no formal guarantees can generally be given. A non-technical introduction to DP can be found in Wood et al. [98]. DP provides mathematical guarantees by assuming a worst-case scenario as a threat model where an adversary has *nearly all* information except a tiny part that the attacker wants to learn. To formalize this, we consider an algorithm or query $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$ that takes input from domain $\mathcal{X}$ and maps to domain $\mathcal{Y}$. The input domain $\mathcal{X}$ could be, for instance, $\mathbb{R}^d$ or $\mathbb{N}^d$, and the target domain $\mathcal{Y}$ could be $\mathbb{R}$ or $\mathbb{N}$, for some $d \in \mathbb{N}$. The algorithm $\mathcal{M}$ is considered (ε, δ)-DP if the following property holds:

$$P(\mathcal{M}(\boldsymbol{x}) \in R) \leq \exp(\varepsilon) P(\mathcal{M}(\boldsymbol{x}') \in R) + \delta, \tag{5.3}$$

with $\varepsilon \in \mathbb{R}_0^+, \delta \in [0, 1]$ and where the two inputs $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^n$ are considered to be adjacent. Adjacency in the context of DP generally means that two data sets differ by exactly one record or entry. This means

that either one value is added, deleted or changed. Qualitatively, this implies that a single individual's entry in a data set has only limited impact on the output of any analysis. Hence, an individual's contribution to the output of $\mathcal{M}$ cannot be too large. The privacy parameters ε and δ, along with the *sensitivity* of an algorithm, determine the utility of a DP algorithm. Using (5.3), we can control the privacy level ε. As $\varepsilon \to 0$, the privacy loss approaches zero while the utility would also diminish. δ can be seen as "slack" value. We can interpret $(1-\delta)$ as the probability that ε-DP holds, also referred to "pure DP". ε-DP is the special case where $\delta = 0$, i.e. $(\varepsilon, 0)$-DP [68]. The most prominent algorithm currently is probably the Gaussian mechanism, where $\mathcal{X} = \mathbb{R}^d$. This mechanism *requires* (ε, δ)-DP with $\delta > 0$ [5, 32]. This means it cannot work with pure DP. The Gaussian mechanism adds calibrated zero-mean Gaussian noise to a deterministic algorithm. The Gaussian mechanism relies, like all DP algorithms, on randomness. With respect to the Gaussian mechanism, a small ε implies a larger standard deviation for the added Gaussian noise.
This leads to a utility loss, as the true output is obfuscated, but a higher privacy is guaranteed. As DP is fundamentally probabilistic, we can frame it as deliberately introducing uncertainty in an algorithm's output to obtain a "safe" version. This implies that one must carefully weigh the two conflicting goals of utility and privacy. This gives rise to the major critique of DP: The unintuitive interpretation of ε compared to concepts like k-anonymity [93]. Even though DP provides a quantifiable degree of privacy through ε, it remains difficult to interpret. Various interpretation methods have been proposed. For instance, a Bayesian interpretations looks at the update from a prior to a posterior belief about $\boldsymbol{x}$ and $\boldsymbol{x}'$ based on the output of a DP algorithm (see e.g. Mironov [76] and Lee and Clifton [67]). An alternative interpretation embeds DP in a classical statistical hypothesis-testing framework where an adversary formulates a hypothesis about the two neighboring data sets [97, 64, 6]. Another major critique of DP concerns the resulting utility of DP algorithms: The considered attack scenario is rigorously pessimistic (worst-case), leading to an extensive obfuscation of outputs and therefore a very large utility loss [7]. More recent work argue, that higher values of $\varepsilon \geq 7$ would be sufficient, when considering a less pessimistic attack scenario or data properties (see for instance Ziller et al. [105] and Lowy et al. [70]).

5.3 Learning on distributed data

We consider data, that are split across different locations, e.g., different servers. Moreover, we assume that the data follow the same structure, i.e., they have identical columns and data types, leading to similar design matrices within a statistical model. Hence, the data are *horizontally* split and do not overlap. The full data set is therefore th eunion of the local data sets:

$$\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_{(k)}, n = n_{(1)} + \cdots + n_{(K)},$$

where each local data set is defined as before with an additional subscript $(k), k = 1 = 1, \ldots, K$ to define it as the data set at the k-th location: $\mathcal{D}_{(k)} = \{(\boldsymbol{x}_{(k),1}, y_{(k),1}), \ldots, (\boldsymbol{x}_{(k),n_{(k)}}, y_{(k),n_{(k)}})\}$.
The field of distributed learning is quite broad with numerous applications in machine learning. The most commonly employed strategy is relatively intuitive: An algorithm is applied to the local data $\mathcal{D}_{(k)}$ (all locations, or just a subset). Afterwards, the local results (e.g. parameter estimates) are sent to an aggregator, that combines them to obtain intermediate global parameters. This procedure is repeated until a convergence criterion is met.

5.4 Discussion and link to contributions

The above described techniques are the building blocks of the final contribution to this thesis. We can combine the previously described methods into one algorithm, that can be used to privately validate a learned classifier via a ROC curve. However, through the adaption of a GLM formulation and the use of DP, one may obtain suboptimal results.
The standard procedure in DP is to generate stochastic noise on top of a, possibly deterministic, algorithm. Thus, DP can also be seen as deliberately inducing *measurement error* to preserve privacy (see also Chapter 4). Much of the ME theory could therefore be adapted to the field of DP.
Even if a new method is defined, this does not mean it can be used directly in practice without accompanying software. Contribution IV (Chapter 9) provides an implementation in DataSHIELD [38].

Bibliography

[1] Oriol Abril-Pla et al. "PyMC: A Modern, and Comprehensive Probabilistic Programming Framework in Python". In: *PeerJ Computer Science* 9 (2023), e1516. ISSN: 2376-5992.

[2] Ben G Armstrong. "Effect of Measurement Error on Epidemiological Studies of Environmental and Occupational Exposures". In: 55.10 (1998), pp. 651–656. DOI: 10.1136/oem.55.10.651.

[3] Nicolas Bacaër. "Daniel Bernoulli, d'Alembert and the Inoculation of Smallpox (1760)". In: *A Short History of Mathematical Population Dynamics*. Springer, 2011, pp. 21–30.

[4] Norman TJ Bailey. *The Mathematical Theory of Infectious Diseases and Its Applications*. 2nd ed. 1975.

[5] Borja Balle and Yu-Xiang Wang. *Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising*. June 2018. arXiv: 1805.06530 [cs, stat].

[6] Borja Balle et al. *Hypothesis Testing Interpretations and Renyi Differential Privacy*. Oct. 2019. DOI: 10.48550/arXiv.1905.09982. arXiv: 1905.09982 [cs].

[7] Jane Bambauer, Krishnamurty Muralidhar, and Rathindra Sarathy. "Fool's Gold: An Illustrated Critique of Differential Privacy". In: *Vand. J. Ent. & Tech. L* 16 ().

[8] Thomas Bayes. "LII. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S". In: *Philosophical Transactions of the Royal Society of London* 53 (1763), pp. 370–418. DOI: 10.1098/rstl.1763.0053.

[9] Joseph Berkson. "Are There Two Regressions?" In: *Journal of the American Statistical Association* 45.250 (1950), pp. 164–80. DOI: 10.2307/2280676.

[10] Michael Betancourt. *A Conceptual Introduction to Hamiltonian Monte Carlo*. July 2018. arXiv: 1701.02434 [stat].

[11] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2017.1285773.

[12] George EP Box. "Science and Statistics". In: *Journal of the American Statistical Association* 71.356 (1976), pp. 791–799. ISSN: 0162-1459.

[13] Fred Brauer, Carlos Castillo-Chavez, and Zhilan Feng. *Mathematical Models in Epidemiology*. Vol. 69. Texts in Applied Mathematics. New York, NY: Springer New York, 2019. ISBN: 978-1-4939-9826-5 978-1-4939-9828-9. DOI: 10.1007/978-1-4939-9828-9.

[14] Steve Brooks et al. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011. ISBN: 1-4200-7942-5.

[15] Alberto Cabezas et al. "Blackjax: Composable Bayesian Inference in Jax". In: *arXiv preprint arXiv:2402.10797* (2024). arXiv: 2402.10797.

[16] Bob Carpenter et al. "Stan: A Probabilistic Programming Language". In: *Journal of statistical software* 76 (2017), pp. 1–32. ISSN: 1548-7660.

[17] Raymond J Carroll et al. *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC, 2006.

[18] Jordi Castro. "Thirty Years of Optimization-Based SDC Meth- Ods for Tabular Data". In: *Transactions on data privacy* 16 (2023), pp. 3–13.

[19] Serina Chang et al. "Mobility Network Models of COVID-19 Explain Inequities and Inform Reopening". In: *Nature* 589.7840 (Jan. 2021), pp. 82–87. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-2923-3.

[20] Aviral Chharia et al. "Accuracy of US CDC COVID-19 Forecasting Models". In: *Frontiers in Public Health* 12 (June 2024), p. 1359368. ISSN: 2296-2565. DOI: 10.3389/fpubh.2024.1359368.

[21] Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Cham: Springer International Publishing, 2020. ISBN: 978-3-030-47844-5 978-3-030-47845-2. DOI: 10.1007/978-3-030-47845-2.

[22] W. G. Cochran. "Errors of Measurement in Statistics". In: *Technometrics* 10.4 (Nov. 1968), p. 637. ISSN: 00401706. DOI: 10.2307/1267450. JSTOR: 1267450.

[23] Clyde H Coombs. "A Theory of Data." In: *Psychological review* 67.3 (1960), p. 143. ISSN: 1939-1471.

[24] *COVID-19*. https://www.ecdc.europa.eu/en/covid-19. June 2023.

[25] Estee Y. Cramer et al. "Evaluation of Individual and Ensemble Probabilistic Forecasts of COVID-19 Mortality in the United States". In: *Proceedings of the National Academy of Sciences* 119.15 (Apr. 2022), e2113561119. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2113561119.

[26] Daryl J. Daley and J. M. Gani. *Epidemic Modelling: An Introduction*. Cambridge Studies in Mathematical Biology 15. Cambridge ; New York: Cambridge University Press, 1999. ISBN: 978-0-521-64079-4.

[27] Steven B. Damelin and Willard Miller Jr. *The Mathematics of Signal Processing*. 1st ed. Cambridge University Press, Dec. 2011. ISBN: 978-1-107-01322-3 978-1-107-60104-8 978-1-139-00389-6. DOI: 10.1017/CBO9781139003896.

[28] "Discussion of the Paper by Grenander and Miller". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 56.4 (Jan. 1994), pp. 581–603. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1994.tb02001.x.

[29] Ensheng Dong, Hongru Du, and Lauren Gardner. "An Interactive Web-Based Dashboard to Track COVID-19 in Real Time". In: *The Lancet Infectious Diseases* 20.5 (May 2020), pp. 533–534. ISSN: 14733099. DOI: 10.1016/S1473-3099(20)30120-1.

[30] Simon Duane et al. "Hybrid Monte Carlo". In: *Physics Letters B* 195.2 (Sept. 1987), pp. 216–222. ISSN: 0370-2693. DOI: 10.1016/0370-2693(87)91197-X.

[31] J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. 2nd ed. Oxford Statistical Science Series 38. Oxford: Oxford University Press, 2012. ISBN: 978-0-19-964117-8.

[32] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4 (2013), pp. 211–407. ISSN: 1551-305X, 1551-3068. DOI: 10.1561/0400000042.

[33] Tom Fawcett. "An Introduction to ROC Analysis". In: *Pattern Recognition Letters* 27.8 (June 2006), pp. 861–874. ISSN: 01678655. DOI: 10.1016/j.patrec.2005.10.010.

[34] N Ferguson et al. *Report 9: Impact of Non-Pharmaceutical Interventions (NPIs) to Reduce COVID19 Mortality and Healthcare Demand*. Tech. rep. Imperial College London, Mar. 2020. DOI: 10.25561/77482.

[35] Neil M. Ferguson et al. "Strategies for Containing an Emerging Influenza Pandemic in Southeast Asia". In: *Nature* 437.7056 (Sept. 2005), pp. 209–214. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature04017.

[36] Seth Flaxman et al. "Estimating the Effects of Non-Pharmaceutical Interventions on COVID-19 in Europe". In: *Nature* 584.7820 (Aug. 2020), pp. 257–261. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-2405-7.

[37] Cornelius Fritz, Emilio Dorigatti, and David Rügamer. "Combining Graph Neural Networks and Spatio-Temporal Disease Models to Improve the Prediction of Weekly COVID-19 Cases in Germany". In: *Scientific Reports* 12.1 (Mar. 2022), p. 3930. ISSN: 2045-2322. DOI: 10.1038/s41598-022-07757-5.

[38] Amadou Gaye et al. "DataSHIELD: Taking the Analysis to the Data, Not the Data to the Analysis". In: *International Journal of Epidemiology* 43.6 (Dec. 2014), pp. 1929–1944. ISSN: 1464-3685, 0300-5771. DOI: 10.1093/ije/dyu188.

[39] Andrew Gelman and Christian Hennig. "Beyond Subjective and Objective in Statistics". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 180.4 (Oct. 2017), pp. 967–1033. ISSN: 0964-1998, 1467-985X. DOI: 10.1111/rssa.12276.

[40] Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and Other Stories*. Cambridge University Press, 2021. ISBN: 1-107-02398-X.

[41] Andrew Gelman and Yuling Yao. "Holes in Bayesian Statistics". In: *Journal of Physics G: Nuclear and Particle Physics* 48.1 (Jan. 2021), p. 014002. ISSN: 0954-3899, 1361-6471. DOI: 10.1088/1361-6471/abc3a5. arXiv: 2002.06467 [math, stat].

[42] Andrew Gelman et al. *Bayesian Data Analysis Third Edition (with Errors Fixed as of 15 February 2021)*. Third edition. Chapman and Hall/CRC, 2013.

[43] Andrew Gelman et al. "Bayesian Workflow". In: *arXiv preprint arXiv:2011.01808* (2020). arXiv: 2011.01808.

[44] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. 1st ed. Cambridge University Press, June 2017. ISBN: 978-0-521-87826-5 978-1-139-02983-4. DOI: 10.1017/9781139029834.

[45] Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov Chain Monte Carlo in Practice*. CRC press, 1995. ISBN: 1-4822-1497-0.

[46] Sander Greenland. "Bayesian Perspectives for Epidemiological Research: I. Foundations and Basic Methods". In: *International Journal of Epidemiology* 35.3 (June 2006), pp. 765–775. ISSN: 1464-3685, 0300-5771. DOI: 10.1093/ije/dyi312.

[47] Sander Greenland, Judea Pearl, and James M. Robins. "Confounding and Collapsibility in Causal Inference". In: *Statistical Science* 14.1 (Feb. 1999). ISSN: 0883-4237. DOI: 10.1214/ss/1009211805.

[48] Emily Griffiths et al. "Handbook on Statistical Disclosure Control for Outputs". In: *Safe Data Access Professionals Working Group* (2019).

[49] Emanuele Guidotti and David Ardia. "COVID-19 Data Hub". In: *Journal of Open Source Software* 5.51 (July 2020), p. 2376. ISSN: 2475-9066. DOI: 10.21105/joss.02376.

[50] Paul Gustafson. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Interdisciplinary Statistics Series 13. Boca Raton, Fla. [u.a]: Chapman & Hall/CRC, 2004. ISBN: 978-1-58488-335-7.

[51] Thomas Hale et al. "A Global Panel Database of Pandemic Policies (Oxford COVID-19 Government Response Tracker)". In: *Nature Human Behaviour* 5.4 (Mar. 2021), pp. 529–538. ISSN: 2397-3374. DOI: 10.1038/s41562-021-01079-8.

[52] William Heaton Hamer. *The Milroy Lectures on Epidemic Diseases in England*. 1906.

[53] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York, 2009. ISBN: 978-0-387-84857-0 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7.

[54] W Keith Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: (1970). ISSN: 1464-3510.

[55] Leonhard Held and Daniel Sabanés Bové. *Applied Statistical Inference: Likelihood and Bayes*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. ISBN: 978-3-642-37886-7 978-3-642-37887-4. DOI: 10.1007/978-3-642-37887-4.

[56] Leonhard Held et al. *Handbook of Infectious Disease Data Analysis*. CRC Press, 2019.

[57] Herbert W. Hethcote. "The Mathematics of Infectious Diseases". In: *SIAM Review* 42.4 (Jan. 2000), pp. 599–653. ISSN: 0036-1445, 1095-7200. DOI: 10.1137/S0036144500371907.

[58] Matthew D. Hoffman and Andrew Gelman. *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*. Nov. 2011. DOI: 10.48550/arXiv.1111.4246. arXiv: 1111.4246 [stat].

[59] Matthew D. Hoffman and Andrew Gelman. *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*. Nov. 2011. DOI: 10.48550/arXiv.1111.4246. arXiv: 1111.4246 [stat].

[60] Sabine Hoffmann et al. "Accounting for Berkson and Classical Measurement Error in Radon Exposure Using a Bayesian Structural Approach in the Analysis of Lung Cancer Mortality in the French Cohort of Uranium Miners". In: *Radiation Research* 187.2 (Jan. 2017), p. 196. ISSN: 0033-7587. DOI: 10.1667/RR14467.1.

[61] Sabine Hoffmann et al. "Shared and Unshared Exposure Measurement Error in Occupational Cohort Studies and Their Effects on Statistical Inference in Proportional Hazards Models". In: *PLOS ONE* 13.2 (Feb. 2018). Ed. by Jane Hoppin, e0190792. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0190792.

[62] Sabine Hoffmann et al. "Shared and Unshared Exposure Measurement Error in Occupational Cohort Studies and Their Effects on Statistical Inference in Proportional Hazards Models". In: *PLOS ONE* 13.2 (Feb. 2018). Ed. by Jane Hoppin, e0190792. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0190792.

[63] Anco Hundepool et al. "Handbook on Statistical Disclosure Control". In: *ESSnet on Statistical Disclosure Control* (2010), pp. 2023–12.

[64] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. "The Composition Theorem for Differential Privacy". In: *IEEE Transactions on Information Theory* 63.6 (June 2017), pp. 4037–4049. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/TIT.2017.2685505.

[65] Ruth H. Keogh et al. "STRATOS Guidance Document on Measurement Error and Misclassification of Variables in Observational Epidemiology: Part 1—Basic Theory and Simple Methods of Adjustment". In: *Statistics in Medicine* 39.16 (July 2020), pp. 2197–2231. ISSN: 0277-6715, 1097-0258. DOI: 10.1002/sim.8532.

[66] M. Kreuzer et al. "Cohort Profile: The German Uranium Miners Cohort Study (WISMUT Cohort), 1946-2003". In: *International Journal of Epidemiology* 39.4 (Aug. 2010), pp. 980–987. ISSN: 0300-5771, 1464-3685. DOI: 10.1093/ije/dyp216.

[67] Jaewoo Lee and Chris Clifton. "How Much Is Enough? Choosing ε for Differential Privacy". In: *Information Security: 14th International Conference, ISC 2011, Xi'an, China, October 26-29, 2011. Proceedings 14*. Springer, 2011, pp. 325–340. ISBN: 3-642-24860-8.

[68] Yehuda Lindell, ed. *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*. Information Security and Cryptography. Cham: Springer International Publishing, 2017. ISBN: 978-3-319-57047-1 978-3-319-57048-8. DOI: 10.1007/978-3-319-57048-8.

[69] Qiang Liu and Dilin Wang. "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm". In: *Advances in neural information processing systems* 29 (2016).

[70] Andrew Lowy et al. *Why Does Differential Privacy with Large Epsilon Defend Against Practical Membership Inference Attacks?* Feb. 2024. arXiv: 2402.09540 [cs].

[71] Richard McElreath. *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, 2020. ISBN: 978-0-367-13991-9.

[72] Sharon Bertsch Mcgrayne. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press, 2011. ISBN: 978-0-300-16969-0. JSTOR: j.ctt1np76s.

[73] Nicholas Metropolis. "The Beginning of the Monte Carlo Method". In: *Los Alamos Science Special Issue* 15 (1987), pp. 125–130.

[74] Nicholas Metropolis et al. "Equation of State Calculations by Fast Computing Machines". In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092. ISSN: 0021-9606.

[75] Sean P Meyn and Richard L Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012. ISBN: 1-4471-3267-X.

[76] Ilya Mironov. "Renyi Differential Privacy". In: *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. Aug. 2017, pp. 263–275. DOI: 10.1109/CSF.2017.11. arXiv: 1702.07476 [cs].

[77] Stefanie Muff and Lukas F. Keller. "Reverse Attenuation in Interaction Terms Due to Covariate Measurement Error". In: *Biometrical Journal* 57.6 (Nov. 2015), pp. 1068–1083. ISSN: 0323-3847, 1521-4036. DOI: 10.1002/bimj.201400157.

[78] Stefanie Muff et al. "Bayesian Analysis of Measurement Error Models Using Integrated Nested Laplace Approximations". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 64.2 (Feb. 2015), pp. 231–252. ISSN: 0035-9254, 1467-9876. DOI: 10.1111/rssc.12069.

[79] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts: The MIT Press, 2023. ISBN: 978-0-262-37599-3 978-0-262-37600-6.

[80] John Ashworth Nelder and Robert WM Wedderburn. "Generalized Linear Models". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 135.3 (1972), pp. 370–384.

[81] *Numpy.Convolve*. https://numpy.org/doc/stable/reference/generated/numpy.convolve.html. Jan. 2025.

[82] George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. *Transfer Graph Neural Networks for Pandemic Forecasting*. Apr. 2021. DOI: 10.48550/arXiv.2009.08388. arXiv: 2009.08388 [cs].

[83] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 1. Aufl. s.l.: Elsevier Reference Monographs, 2014. ISBN: 978-1-55860-479-7.

[84] Margaret Sullivan Pepe. "An Interpretation for the ROC Curve and Inference Using GLM Procedures". In: *Biometrics* 56.2 (June 2000), pp. 352–359. ISSN: 0006341X. DOI: 10.1111/j.0006-341X.2000.00352.x.

[85] Margaret Sullivan Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series 28. Oxford ; New York: Oxford University Press, 2003. ISBN: 978-0-19-850984-4.

[86] Du Phan, Neeraj Pradhan, and Martin Jankowiak. "Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro". In: (Dec. 2019). DOI: 10.48550/arXiv.1912.11554. arXiv: 1912.11554 [stat].

[87] Sylvia Richardson and Walter R. Gilks. "A Bayesian Approach to Measurement Error Problems in Epidemiology Using Conditional Independence Models". In: *American Journal of Epidemiology* 138.6 (Sept. 1993), pp. 430–442. ISSN: 1476-6256, 0002-9262. DOI: 10.1093/oxfordjournals.aje.a116875.

[88] Håvard Rue, Sara Martino, and Nicolas Chopin. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 71.2 (2009), pp. 319–392. JSTOR: 40247579.

[89] Pamela A. Shaw et al. "STRATOS Guidance Document on Measurement Error and Misclassification of Variables in Observational Epidemiology: Part 2—More Complex Methods of Adjustment and Advanced Topics". In: *Statistics in Medicine* 39.16 (July 2020), pp. 2232–2263. ISSN: 0277-6715, 1097-0258. DOI: 10.1002/sim.8531.

[90] Angelika M. Stefan et al. "A Tutorial on Bayes Factor Design Analysis Using an Informed Prior". In: *Behavior Research Methods* 51.3 (June 2019), pp. 1042–1058. ISSN: 1554-3528. DOI: 10.3758/s13428-018-01189-8.

[91] Pauline van den Driessche. "Reproduction Numbers of Infectious Disease Models". In: *Infectious Disease Modelling* 2.3 (Aug. 2017), pp. 288–303. ISSN: 24680427. DOI: 10.1016/j.idm.2017.06.002.

[92] Tyler J VanderWeele and Miguel A. Hernan. "Results on Differential and Dependent Measurement Error of the Exposure and the Outcome Using Signed Directed Acyclic Graphs". In: *American Journal of Epidemiology* 175.12 (June 2012), pp. 1303–1310. ISSN: 0002-9262, 1476-6256. DOI: 10.1093/aje/kwr458.

[93] Saskia Nuñez von Voigt, Luise Mehner, and Florian Tschorsch. *From Theory to Comprehension: A Comparative Study of Differential Privacy and k-Anonymity*. Apr. 2024. DOI: 10.48550/arXiv.2404.04006. arXiv: 2404.04006 [cs].

[94] Sholom Wacholder. "When Measurement Errors Correlate with Truth: Surprising Effects of Nondifferential Misclassification". In: *Epidemiology* 6.2 (1995), pp. 157–161.

[95] J Wallinga and M Lipsitch. "How Generation Intervals Shape the Relationship between Growth Rates and Reproductive Numbers". In: *Proceedings of the Royal Society B: Biological Sciences* 274.1609 (Feb. 2007), pp. 599–604. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2006.3754.

[96] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. New York, NY: Springer New York, 2004. ISBN: 978-1-4419-2322-6 978-0-387-21736-9. DOI: 10.1007/978-0-387-21736-9.

[97] Larry Wasserman and Shuheng Zhou. "A Statistical Framework for Differential Privacy". In: *Journal of the American Statistical Association* 105.489 (Mar. 2010), pp. 375–389. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/jasa.2009.tm08651.

[98] Alexandra Wood et al. "Differential Privacy: A Primer for a Non-Technical Audience". In: *SSRN Electronic Journal* (2018). ISSN: 1556-5068. DOI: 10.2139/ssrn.3338027.

[99] Simon N Wood. *Generalized Additive Models: An Introduction with R*. chapman and hall/CRC, 2017. ISBN: 1-315-37027-1.

[100] Yukari Yamauchi et al. "Normalizing Flows for Bayesian Posteriors: Reproducibility and Deployment". In: (Oct. 2023). DOI: 10.48550/arXiv.2310.04635. arXiv: 2310.04635 [nucl-th].

[101] Ping Yan and Gerardo Chowell. *Quantitative Methods for Investigating Infectious Disease Outbreaks*. Vol. 70. Texts in Applied Mathematics. Cham: Springer International Publishing, 2019. ISBN: 978-3-030-21922-2 978-3-030-21923-9. DOI: 10.1007/978-3-030-21923-9.

[102] Grace Y. Yi. *Statistical Analysis with Measurement Error or Misclassification*. Springer Series in Statistics. New York, NY: Springer New York, 2017. ISBN: 978-1-4939-6638-7 978-1-4939-6640-0. DOI: 10.1007/978-1-4939-6640-0.

[103] Grace Y. Yi. *Statistical Analysis with Measurement Error or Misclassification*. Springer Series in Statistics. New York, NY: Springer New York, 2017. ISBN: 978-1-4939-6638-7 978-1-4939-6640-0. DOI: 10.1007/978-1-4939-6640-0.

[104] Na Zhu et al. "A Novel Coronavirus from Patients with Pneumonia in China, 2019". In: *New England Journal of Medicine* 382.8 (Feb. 2020), pp. 727–733. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa2001017.

[105] Alexander Ziller et al. "Reconciling Privacy and Accuracy in AI for Medical Imaging". In: *Nature Machine Intelligence* 6.7 (June 2024), pp. 764–774. ISSN: 2522-5839. DOI: 10.1038/s42256-024-00858-y.

Part II

Contributing articles

Chapter 6

Contribution I

Article

Rehms, R., Ellenbach, N., Rehfuess, E., Burns, J., Mansmann, U., & Hoffmann, S. (2024). A Bayesian hierarchical approach to account for evidence and uncertainty in the modeling of infectious diseases: An application to COVID-19. *Biometrical Journal*, 66, 2200341. https://doi.org/10.1002/bimj.202200341

Data and code

The code is available at https://github.com/RaphaelRe/BayesModelCOVID and results are fully reproducible. Results from the generated Markov chains are achieved at https://figshare.com/articles/dataset/MCMC_samples/24183246/1.

Supplementary Material

Supplementary material can be found under https://onlinelibrary.wiley.com/doi/10.1002/bimj.202200341(section: *Supporting Information*).

Copyright information

Author Contributions

S.H., U.M., E.R. and J.B. developed the research idea. **R.R.**, S.H. and N.E. developed the model. The MCMC sampler was implemented and tested by **R.R**., S.H. The code for the simulation study was developed by **R.R**. and N.E. Data collection and preprocessing was done by **R.R**. He also conducted the simulation study and applied the model to the real data. **R.R**. and S.H. wrote the manuscript.
All authors discussed the results and were closely involved in proofreading and revising the manuscript.
Corresponding author is **R.R.**

Received: 5 December 2022 | Revised: 21 August 2023 | Accepted: 24 August 2023

DOI: 10.1002/bimj.202200341

Biometrical Journal

RESEARCH ARTICLE

A Bayesian hierarchical approach to account for evidence and uncertainty in the modeling of infectious diseases: An application to COVID-19

Raphael Rehms[1,2] | **Nicole Ellenbach**[1,2] | **Eva Rehfuess**[1,2] | **Jacob Burns**[1,2] | **Ulrich Mansmann**[1,2,3] | **Sabine Hoffmann**[1,2,3]

[1]Institute of Medical Data Processing, Biometrics and Epidemiology (IBE), Faculty of Medicine, Ludwig-Maximilians-University Munich, Munich, Germany

[2]Pettenkofer School of Public Health, Ludwig-Maximilians-University Munich, Munich, Germany

[3]Department of Statistics, Ludwig-Maximilians-University Munich, Munich, Germany

Correspondence
Raphael Rehms, Institute of Medical Data Processing, Biometrics and Epidemiology (IBE), Faculty of Medicine, Ludwig-Maximilians-University Munich, Marchioninistr. 15, 81377 Munich, Germany.
Email: rrehms@ibe.med.uni-muenchen.de

Funding information
Volkswagen Stiftung, Grant/Award Number: AZ: 99664

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

Infectious disease models can serve as critical tools to predict the development of cases and associated healthcare demand and to determine the set of nonpharmaceutical interventions (NPIs) that is most effective in slowing the spread of an infectious agent. Current approaches to estimate NPI effects typically focus on relatively short time periods and either on the number of reported cases, deaths, intensive care occupancy, or hospital occupancy as a single indicator of disease transmission. In this work, we propose a Bayesian hierarchical model that integrates multiple outcomes and complementary sources of information in the estimation of the true and unknown number of infections while accounting for time-varying underreporting and weekday-specific delays in reported cases and deaths, allowing us to estimate the number of infections on a daily basis rather than having to smooth the data. To address dynamic changes occurring over long periods of time, we account for the spread of new variants, seasonality, and time-varying differences in host susceptibility. We implement a Markov chain Monte Carlo algorithm to conduct Bayesian inference and illustrate the proposed approach with data on COVID-19 from 20 European countries. The approach shows good performance on simulated data and produces posterior predictions that show a good fit to reported cases, deaths, hospital, and intensive care occupancy.

KEYWORDS

Bayesian modeling, COVID-19, hierarchical models, infectious diseases, modeling uncertainty

1 | INTRODUCTION

The experience with severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) causing coronavirus disease 2019 (COVID-19) has underlined both the importance of and the challenges in the modeling of infectious diseases. Infectious disease models can serve as critical tools to predict epidemic development and healthcare demand and to determine when and which nonpharmaceutical interventions (NPIs) should be implemented to slow the spread of an infectious agent. However, the modeling of infectious diseases is complicated by the fact that the main quantity of interest, that is, the number of daily infections, is a latent variable that cannot be observed and therefore has to be estimated by using information on other observable quantities. The number of reported cases, hospital occupancy, and deaths all provide complementary, yet incomplete and sometimes even contradictory, information on the number of infections in a given geographical region. The number of reported cases is prone to underreporting, as it depends on the reliability of tests, the testing capacity, and the employed testing strategy (May, 2020; Pullano et al., 2021). When changes in the testing strategy concur with the introduction or the relaxation of NPIs, they can create severe distortions in the estimation of NPI effects. While modeling disease mortality, which is less prone to underreporting, can avoid biases due to changes in the testing strategy, it is difficult to predict future healthcare demand based on disease mortality. Moreover, the time lag between infection and death may be highly variable, leading to a reduction of statistical power (Sharma et al., 2021) and where (more or less effective) treatment is available, differences in the availability, accessibility, and quality of healthcare may affect case fatality rates. A possible solution for this situation, in which there are several imperfect proxy variables for the number of infections, is to combine information on disease incidence, mortality, and hospital occupancy in a common framework while explicitly accounting for time-varying underreporting and reporting delays.

Bayesian hierarchical approaches can address this challenge by providing a coherent and flexible framework to integrate all available sources of information while accounting for different sources of uncertainty. By combining different submodels through conditional independence assumptions, it is possible to integrate mechanistic assumptions on disease dynamics and submodels describing the relationship between the true (and unknown) number of infections and reported cases, hospital occupancy, and deaths. Additionally, we can borrow information from other geographical regions to stabilize parameter estimates and to improve forecasts on future healthcare demand. Current approaches to assess the effect of NPIs typically either focus on the number of deaths (Flaxman et al., 2020) or the number of cases (Banholzer et al., 2021; Dehning et al., 2020; Islam et al., 2020; Li et al., 2021b). Unwin et al. (2020), Brauner et al. (2021), and Sharma et al. (2021) extend the semimechanistic Bayesian hierarchical model proposed by Flaxman et al. (2020) by including information on reported cases and deaths when inferring the number of new infections. However, these approaches typically only estimate NPI effects for short time periods because they do not explicitly account for differences in host susceptibility over time (due to vaccination or previous infection), seasonality, the prevalence of different virus variants, or time-varying underreporting.

Here, we show how a Bayesian hierarchical approach can be used to integrate the available information on the number of reported cases, the number of deaths, and hospital and intensive care unit (ICU) occupancy in the estimation of the true and unknown number of infections while accounting for underreporting and reporting delays in the number of reported cases. We account for the influence of seasons, previous infections, vaccination coverage, and the prevalence of different virus variants as these factors can have a critical influence on the number of new infections and on disease severity. By doing so, it is possible to use data over long time periods in several countries rather than focusing on short time periods in a single country during which the prevailing variant, vaccination coverage, and the testing strategy remained roughly constant. By allowing for weekday-specific delays in reported cases and deaths (which mainly arise due to reduced reporting during the weekend), we are not required to smooth the analyzed time series and we can estimate the number of infections on a daily basis. We illustrate the proposed approach using data for COVID-19 from 20 European countries and investigate its performance both on simulated data and by assessing how well the model describes reported cases, hospital and ICU occupancy, and deaths through posterior predictive checks.

The rest of the paper is organized as follows. In Section 2, we describe the proposed Bayesian hierarchical model. In Section 3, we present a simulation study to assess the performance of our proposed approach. A case study on the modeling of COVID-19 in 20 European countries is presented in Section 4. In Section 5, we summarize this article with a brief discussion.

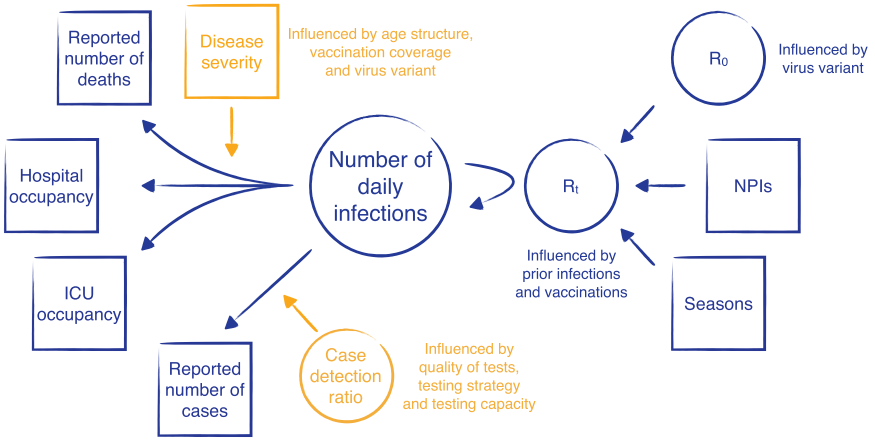


FIGURE 1 Simplified directed acyclic graph (DAG) describing how the true and unknown number of infections is estimated through the four observed time series. Parameters are shown in orange and variables are shown in blue. Unknown quantities that have to be estimated are given in circles and quantities that are either observed or assumed to be known are given in squares.

2 | THE MODEL

In this section, we describe the main elements of the Bayesian hierarchical model. At its core, the model treats the number of true and unknown infections $I_{t,m} \in \mathbb{N}_0$ at time t in geographical region m as a discrete latent variable where t and m are considered as discrete index sets (e.g., days and countries). We describe the model in two parts. First, we describe how we infer the number of true and unknown infections (i.e., the values of the latent variable) from the available information on reported cases, hospital and ICU occupancy, and deaths while accounting for time-varying underreporting, weekday-specific reporting delays, and changes in the severity of the disease due to vaccination coverage and different virus variants. Second, we describe how we estimate the effects of NPIs given the true and unknown number of infections represented through the discrete latent variable while accounting for seasonality, time-varying differences in host susceptibility and changes in the transmissibility of the virus due to different virus variants.

2.1 | Inferring the number of infections

As illustrated in Figure 1, we use information on the reported number of cases, deaths, and hospital and ICU occupancy to estimate the true and unknown number of infections $I_{t,m}$ at time t in geographical region m. Each of these observed time series is linked through a submodel to this discrete latent variable: the reporting model, the death model, and two hospitalization models (normal beds and ICU). Before linking the number of infections to the observable time series, we define a second latent variable, the number of cases $C_{t,m} \in \mathbb{N}_0$ in geographical region m with symptom onset on day t, which is simply a deterministic function of the number of infections $I_{t,m}$ occurring until time t, as described through the following *disease model*:

$$C_{t,m} = \sum_{u \leq t} I_{t,m} \left(F_{\xi^C}(t-u+1) - F_{\xi^C}(t-u) \right),$$

where F_{ξ^C} is the cumulative distribution function of the incubation period. Note that we do not distinguish infections by strength of symptoms. Some infected individuals may even experience so weak symptoms that they are not noticed, and in this case, the incubation period is of merely technical nature. Through the disease model, the number of (symptomatic and asymptomatic) cases becomes a deterministic function of the number of infections, which is simply shifted by the incubation time distribution.

The number of cases $C_{t,m}$ is linked to the number of reported cases $C^R_{t,m}$ through the following *reporting model*:

$$C^R_{t,m} \sim NegBinom\left(\mu^R_{t,m}, \phi^R\right),$$

where

$$\mu^R_{t,m} = \rho_{t,m} \sum_{u<t} C_{u,m}\left(F_{\xi^{R,w}_m}(t-u+1) - F_{\xi^{R,w}_m}(t-u)\right),$$

where $F_{\xi^{R,w}_m}$ is the reporting delay distribution for a specific weekday w in geographical region m. In this model, the number of reported cases follows a negative binomial distribution where the expected number of reported cases on day t is described as the sum of all true cases occurring on some day u before day t weighted by their probability of being reported after $t-u$ days and multiplied by a time-specific underreporting rate $\rho_{t,m}$. We choose a negative binomial distribution rather than a Poisson distribution to allow for overdispersion (controlled by the size parameter ϕ^R). The variance of $C^R_{t,m}$ is then given by $\mu^R_{t,m} + {\mu^R_{t,m}}^2/\phi^R$. Therefore, for high values of ϕ^R relative to $\mu^R_{t,m}$, the distribution resembles a Poisson distribution and low values indicate high overdispersion. The time-specific underreporting rates $\rho_{t,m}$ are modeled through a piece-wise constant function. By accounting for time-varying underreporting and weekday-specific reporting delays, the reporting model allows for discrepancies between the true dynamics of the disease and the number of cases that are reported by health authorities (Höhle & an der Heiden, 2014). In the reporting model, we assume that the delay between symptom onset and day of reporting can be specific to a geographical region m. Since it is very difficult to obtain this information for each region, we use information from a specific geographical region for which we can estimate these weekday-specific reporting delay distributions (in our application to COVID-19, this region is Bavaria in Germany) and adapt them for each location. The whole procedure is as follows. For each weekday of symptom onset, we use a different reporting delay distribution. This means, for instance, that an infected individual with symptom onset on Monday may have another reporting pattern and therefore another reporting delay distribution than an individual with symptom onset on Sunday (since several local authorities do not work on Sundays). Figure S6 in the Supporting Information shows the estimated reference distributions from Bavaria. Furthermore, we individualize these weekday-specific reporting delay distributions for each location with its regional reporting pattern (e.g., some countries do not report any cases on Sundays at all while others do). We achieve this by introducing location and weekday-specific parameters β^w_m to adapt these distributions for weekdays w for each geographical region m. These parameters β^w_m are multiplied with the discretized versions of the Bavarian reporting delay distributions to inflate or deflate the probability mass at the respective time spans that match the weekdays. For example, for a country A that does not report any cases at all on Sundays, the respective β^{Sunday}_A would be estimated as zero. Thus, the reporting delay distribution for symptom onset on Mondays would result in a probability mass of zero on the 6th, 13th, and so on, day, whereby the reporting delay distribution for symptom onset on Tuesday would result in a probability mass of zero on the 5th, 12th, and so on, day. After the multiplication, we renormalize the result to obtain a proper probability distribution. As a consequence, we can account for weekly reporting patterns that are specific to each geographical region m.

Following Flaxman et al. (2020), we describe the number of deaths $D_{t,m}$ occurring on day t in geographical region m as a function of the number of true cases with disease onset prior to t through the following *death model*:

$$D_{t,m} \sim NegBinom\left(\mu^D_{t,m}, \phi^D\right),$$

where

$$\mu^D_{t,m} = \pi^D_{t,m} \sum_{u\le t} C_{u,m}\left(F_{\xi^{D,w}_m}(t-u+1) - F_{\xi^{D,w}_m}(t-u)\right).$$

In this model, $D_{t,m}$ is described by a negative binomial distribution with expected value equal to the sum of the number of true cases with disease onset at time $t-u$, weighted by the probability of dying on the uth day after the onset of symptoms. This latter probability can be obtained by discretizing the probability distribution describing the time until death for patients who died, that is, $F_{\xi^{D,w}_m}$, and multiplying by the infection fatality rate (IFR) $\pi^D_{t,m}$, that is, the probability of dying for an infected individual where this rate can depend on day t and geographical region m. Similarly to $F_{\xi^{R,w}_m}$ in the

reporting model, $F_{\xi_m^{D,w}}$ accounts for weekday effects that can be specific to geographical region m to account for differences in reporting through variables $\beta_m^{D,w}$.

The IFR is a crucial part of the model as we assume that this quantity is fixed and known, but we consider several (potentially time-varying) factors that have an influence on this quantity, namely, the age composition in a country, vaccination rollout, and the prevalence of new variants.

The proposed model operates on an aggregated level with respect to locations m (e.g., countries). However, different locations have different age compositions. As disease severity is strongly correlated with age, we obtain an aggregated IFR by weighting an age-specific IFR with the age structure in each location. We use information on the age strata of each country from O'Driscoll et al. (2021) and the aggregated IFR for four different age groups $s = 1, \dots, 4$ from Staerk et al. (2021), that is, we use an IFR of 0.008% for the age group of 0–34 years, 0.122% for 35–59 years, 0.992% for 60–79 years and 7.274% for an age of 80 or older. The location-specific IFR is then calculated by

$$ifr_m = \sum_{s=1}^{4} w_{s,m} \cdot ifr_s,$$

where $w_{s,m}$ is the proportion of the age category (stratum) of the population of country m and ifr_s defines the age-specific IFR in each stratum s.

As vaccinations substantially reduce the probability of dying, we make the assumption that the IFR at each location m changes as a function of the time-varying vaccination coverage. With growing coverage in the population, the IFR is lowered. It is important to mention that older age groups in the population were vaccinated with a higher priority at the beginning of the vaccination rollout in most countries. We include the effect of vaccinations by reducing the IFR in the different age strata relative to their share of the population. Most countries do not provide enough information about their vaccination progress in the different age groups. We therefore use publicly available data from France (Santé Publique France, 2021) and extrapolate this information to all other countries, because the majority of European countries used a similar vaccination strategy, making it plausible to assume that the evolution of vaccination coverage over time in the different age groups was roughly comparable across different countries. We assume that after the first vaccination, the probability of dying is reduced by 80%, that is, $\beta_{\pi^D}^{vacc} = 0.8$ after a lag of 2 weeks. For example, Haas et al. (2021) found a higher effectiveness against COVID-19-related deaths of the BNT162b2 vaccine. However, since not all countries use the same type of vaccine and one can assume a reduction of the vaccine effectiveness as time goes on, we decide to use this approximation with a lower effectiveness.

Finally, new mutations of the virus can change disease severity. For COVID-19, we include the effect of alpha (B.1.1.7) and delta (B.1.617.2) in the considered time window. Fisman and Tuite (2021) provide information on the severity of these variants of concern. To account for the fact that these variants changed the overall IFR, we combine the time-varying prevalence of these variants of concern with their disease severity. For B.1.1.7, we inflate the IFR by a factor of $\beta_{\pi^D}^{alpha} = 1.51$ and for B.1.617 with $\beta_{\pi^D}^{delta} = 2.08$.

To test the robustness to the assumptions concerning the overall value of the IFR and the influence of variants of concern and of vaccinations on the IFR, we conduct a number of sensitivity analyses (see Section G and Figures S8 and S9 in the Supporting Information for more details and results). Finally, many countries provide information on hospital and ICU occupancy, and it is important to be able to integrate these two additional sources of information in the estimation of the true and unknown number of infections wherever this is possible. We integrate this information through two *hospitalization models* that have a very similar structure as the death model:

$$H_{t,m} \sim NegBinom\left(\mu_{t,m}^H, \phi^H\right),$$

where

$$\mu_{t,m}^H = \pi_{t,m}^H \sum_{u \le t} C_{u,m}\left(F_{\xi^H}(t-u+1) - F_{\xi^H}(t-u)\right),$$

where $\pi_{t,m}^H$ varies over time in the same way as $\pi_{t,m}^D$ to account for vaccination coverage and different virus variants. In contrast to $\pi_{t,m}^D$, however, we can estimate $\pi_{t,m}^H$ for each geographical region and do not have to consider it to be known. By doing so, we can account for differences in medical care and definitions of hospital admissions and ICU admissions that

may vary between geographical regions. To account for vaccination coverage and the influence of virus variants, we define parameters π_m^H that are specific to each geographical region, but that do not change over time, and calculate $\pi_{t,m}^H$ as the product $\pi_m^H \times g_{t,m}$ where $g_{t,m}$ is a fixed quantity representing the effect of vaccinations and virus variants, assuming that these factors modify the severity of the disease in the same time-varying manner as for $\pi_{t,m}^D$. $g_{t,m}$ can therefore constructed from $\pi_{t,m}^D$ as $g_{t,m} = \pi_{t,m}^D / \pi_{1,m}^D$.

We use exactly the same model for hospital (normal beds) and ICU occupancy. For the sake of brevity, we therefore do not present the model for ICU occupancy in detail, but it can be obtained by merely changing the superscripts from H to ICU. The two distributions, F_{ξ^H} and $F_{\xi^{ICU}}$, provide the probability that a person with symptom onset on day t_{onset} occupies a hospital or an ICU unit on day $t_{onset} + t_{delay}$ with $t_{delay} = 1, 2, 3, ...$ For the application to COVID-19, we obtain these two distributions by combining information on the time between symptom onset and hospitalization with information on the time a person occupies a bed or ICU after being hospitalized through Monte Carlo methods. See Sections D.2 of the Supporting Information for a more detailed description of the definition of F_{ξ^H} and $F_{\xi^{ICU}}$.

2.2 | Modeling the effects of NPIs and seasons

As mentioned above, we model the number of true and unknown infections as a discrete latent variable. To describe the dynamics of the infectious disease, we assume that this latent variable follows the following *renewal model*:

$$I_{t,m} \sim NegBinom\big(\mu_{t,m}, \phi^I\big),$$

where

$$\mu_{t,m} = R_{t,m} \sum_{u<t} I_{u,m}\big(F_\gamma(t-u+1) - F_\gamma(t-u)\big).$$

This renewal model describes the number of infected individuals $I_{t,m}$ at each time point t in geographical region m as a function of past infections, the instantaneous reproduction number $R_{t,m}$, and the generation time distribution. To be more specific, the expected number of infections $\mu_{t,m}$ is the sum of the previous infections on the $t-1$ days before t weighted by the corresponding probability mass of the discretized generation time distribution $F_\gamma(t-u+1) - F_\gamma(t-u)$ multiplied by the instantaneous reproduction number $R_{t,m}$ at time t in geographical region m, where the generation time distribution $F_\gamma(t-u+1) - F_\gamma(t-u)$ represents the probability to transmit the infection from one infected individual to another between time $t-u$ and $t-u+1$. Applying the renewal equation to past infections yields the current number of infections $I_{t,m}$ (see, e.g., Fraser et al., 2009), and it can be seen as a more flexible version of the disease dynamics described in classical compartmental models for infectious diseases (Wallinga & Lipsitch, 2007). We assume that the latent variable follows a negative binomial distribution. We set a prior on the size parameter with $\tilde{\phi}^I \sim N^+(0, 0.015)$ where $\tilde{\phi}^I = 1/{\phi^I}^2$. Through this prior assumption, the dispersion is pushed toward smaller values (see Section 2.3 for an explanation). The same prior is also used for the size parameter for the observed time series (i.e., $C_{t,m}^R, D_{t,m}, H_{t,m}, ICU_{t,m}$).

We seed the model for the first day $I_{1,m}$ in each geographical region m through a negative binomial distribution with mean parameter τ_m:

$$I_{1,m} \sim NegBinom\big(\tau_m, \phi^I\big), \tag{1}$$

where we assume a hierarchical model for τ_m, that is, each τ_m follows a truncated Normal distribution around a common parameter τ that follows a Gamma distribution with shape a_τ and scale b_τ.

$$\tau_m \sim N^+(\tau, \sigma_\tau),$$
$$\tau \sim Ga(a_\tau, b_\tau),$$
$$\sigma_\tau \sim N^+(\mu_{\sigma_\tau}, \sigma_{\sigma_\tau}).$$

For a graphical display of the prior on $I_{1,m}$ that this hierarchical structure implies, see Figure S2 in the Supporting Information.

Following Flaxman et al. (2020), Brauner et al. (2021), and Sharma et al. (2021), we describe the effect of K NPIs $\alpha_{k,m}$ through the following model:

$$R_{t,m} = R^0_{t,m} \exp\left(-\sum_{k=1}^{K+3} \alpha_{k,m} \cdot \mathbb{1}_{k,m}(t)\right) \cdot (1 - c^1_{t,m} - c^2_{t,m} \cdot (1 - c^1_{t,m}))$$

with

$$c^1_{t,m} = \frac{\sum_{u<t} I_{u,m}}{N_m} \cdot (1 - \beta^{reinf}),$$

$$c^2_{t,m} = \frac{\sum_{u<t} (Vacc^1_{u,m} \cdot \beta^{vacc1} + Vacc^2_{u,m} \cdot \beta^{vacc2})}{N_m},$$

where $\mathbb{1}_{k,m}(t)$ are indicator variables taking the value of 1 if the kth NPI is active at time t in geographical region m and 0 otherwise. The correction factors $c^1_{t,m}$ and $c^2_{t,m}$ reduce the transmissibility of the virus in the population: $c^1_{t,m}$ corrects for previously infected individuals. Since an infection may not guarantee protection against the infectious agent, we include a parameter β^{reinf} giving the probability of reinfection. The term $c^2_{t,m}$ corrects for vaccination coverage where $\sum_{u<t} Vacc^1_{u,m}$ and $\sum_{u<t} Vacc^2_{u,m}$ are the number of vaccinated individuals in the population at time t and geographical region m and β^{vacc1} and β^{vacc2} represent the probability of infection after a first and second vaccine dose.

Besides NPIs, we also include the effect of seasons (choosing summer as reference category, resulting in $K + 3$ indicator variables in total) where each indicator variable is 1, if the current t corresponds to the according season. Since it is reasonable to assume variations in the effect of NPIs and seasons between different geographical regions, we allow for country-specific effects that are linked through a hierarchical structure. This hierarchical structure makes it possible to share information between regions to infer an overall effect of the NPIs while allowing to estimate individual effects that are specific to each geographical region:

$$\alpha_{k,m} \sim N(\alpha_k, \sigma^2_{\alpha_k}).$$

The basic reproduction number R^0_m may vary over time due to the occurrence of different variants that modify the transmissibility of the virus. We propose a convex combination to construct a time-dependent basic reproduction number R^0_m. For the application to COVID-19, we account for two variants of concern yielding the following formula:

$$\begin{aligned} R^0_{t,m} =& R^0_m \cdot (1 - p^{alpha}_{t,m} - p^{delta}_{t,m}) + \\ & (1 + \beta^{alpha}) \cdot R^0_m \cdot p^{alpha}_{t,m} + \\ & (1 + \beta^{delta}) \cdot R^0_m \cdot p^{delta}_{t,m}, \end{aligned}$$

where $p^{alpha}_{t,m}$ and $p^{delta}_{t,m}$ are the prevalence of the alpha (B.1.1.7) and delta (B.1.617.2) variants, respectively, at each time t in geographical region m. The two unknown parameters β^{alpha} and β^{delta} represent the increased transmissibility of these variants compared to the wild type. We obtain a time variant reproduction number by taking the reproduction number of the original wild type as basis and multiplying it with $(1 + \beta^{alpha})$ and $(1 + \beta^{delta})$ which accounts for the effect of these subsequent variants.

Finally, we allow for variation in the basic reproduction number among geographical regions. We therefore assume reproduction numbers R^0_m that are specific to geographical region m that are again modeled in a hierarchical manner with common mean R_0:

$$R^0_m \sim N(R^0, \sigma^2_R).$$

The proposed model requires a large set of parameters that are either estimable (and possibly with a prior) or have to be specified as a fixed quantity. Table 1 provides an overview of all parameters of the model with their specifications.

TABLE 1 Summary of all parameters used in the model.

Parameters to model disease dynamics			
Parameter	**Description**	**Additional information**	**Prior**
$I_{1,m}$	Initial number of infected individuals	Location-specific (hierarchical)	$NegBinom(\tau_m, \phi^I)$
F_γ	Generation time distribution	$Ga(5, 0.45)$	Fixed
τ_m	Expected mean of $I_{1,m}$	Location-specific (hierarchical)	$N^+(\tau, \sigma_\tau)$
τ	Mean over all τ_m	Shared mean	$Ga(10, 1)$
σ_τ	Variation of τ_m across all m		$N^+(0, 10)$
ϕ^I	Size parameter of infections	Prior on $\tilde{\phi}^I = 1/\phi^{I^2}$	$N^+(0, 0.015)$
$I_{t,m}$	Number of infected individuals at t, m	Location- & time-specific	$NegBinom(\tau_m, \phi)$
R^0_m	Basic reproduction number for each m	Location-specific (hierarchical)	$N(R^0, \sigma_R)$
R^0	Mean R^0 over all m	Shared mean	$N(3.25, 0.05)$
σ_R	Variation of R^0 over all m		$N^+(0, 0.01)$
β^{alpha}	Increased transmissibility of variant	For B1.1.7	$N(0.6, 0.01)$
β^{delta}	Increased transmissibility of variant	For B.1.617.2	$N(1.5, 0.01)$
$p^{alpha}_{t,m}$	Prevalence at time t in location m	For B1.1.7	See supp. info (fixed)
$p^{delta}_{t,m}$	Prevalence at time t in location m	For B.1.617.2	See supp. info (fixed)
$\beta^{vacc1}_{t,m}$	Effectiveness of vaccination	With one dose	0.5 (fixed)
$\beta^{vacc2}_{t,m}$	Effectiveness of vaccination	With two doses	0.35 (fixed)
$\alpha_{k,m}$	Effect of the kth NPI at location m	Location-specific (hierarchical)	$N(\alpha_k, \sigma_{\alpha_k})$
α_k	Mean of the kth NPI over all m	Shared mean	$N(0, 0.3)$
σ_{α_k}	Variation of α_k over all m		$N^+(0, 0.015)$
Parameters to infer the infections			
Parameter	**Description**	**Additional information**	**Prior**
F_{ξ^C}	Incubation time distribution	$Ga(5.68, 0.08)$	Fixed
$F_{\xi^{R,w}_m}$	Reporting delay distribution	See Supp. Info	Fixed
β^w_m	Seasonal reporting at w in m	Weekday-specific	$U(0, 10)$
ϕ^R	Size for reported cases	Prior on $\tilde{\phi}^R = 1/\phi^{R^2}$	$N^+(0, 0.015)$
$\rho_{t,m}$	Reporting ratio at t & m	Piece-wise constant	$U(0, 3)$
$\pi^D_{t,m}$	Infection fatality rate for location m	See Section 2.1	Fixed
ifr_g	IFR for age stratum g	See Section 2.1	Fixed
w_{s_m}	Share of age stratum in location m	See Section 2.1	Fixed
$\beta^{vacc}_{\pi^D}$	Effect of vaccination on IFR		0.8 (fixed)
$\beta^{alpha}_{\pi^D}$	Severity of B.1.1.7		1.51 (fixed)
$\beta^{delta}_{\pi^D}$	Severity of B.1.617.2		2.08 (fixed)
$F_{\xi^{D,w}_m}$	Symptoms-to-death distribution	$Ga(15.93, 0.1)$	Fixed
$\beta^{D,w}_m$	Seasonal reporting at w in m	One for each weekday w	$U(0, 10)$
ϕ^D	Size for reported deaths	Prior on $\tilde{\phi}^D = 1/\phi^{D^2}$	$N^+(0, 0.015)$
π^H	Depends on π^D via $g_{t,m}$		$U(0, 10)$
F_{ξ^H}	Symptoms-to-hospital occupancy	See Supp. Info D.2	Fixed
ϕ^H	Size of hospital occupancy	Prior on $\tilde{\phi}^H = 1/\phi^{H^2}$	$N^+(0, 0.015)$
π^{ICU}	Depends on π^D via $g_{t,m}$		$U(0, 10)$
$F_{\xi^{ICU}}$	Symptoms-to-ICU occupancy	See Supp. Info D.2	Fixed
ϕ^{ICU}	Size of ICU occupancy	Prior on $\tilde{\phi}^{ICU} = 1/\phi^{ICU^2}$	$N^+(0, 0.015)$
$g_{t,m}$	Correction to π^H and π^{ICU}	Derived from $\pi^D_{t,m}$ (Section 2.1)	Fixed

Furthermore, we provide a full directed acyclic graph (Figure 1), a summary of the model, and the full expression of the joint posterior in Section A of the Supporting Information.

2.3 | Inference, identifiability, and implementation

The flexibility of the proposed model can come at the cost of nonidentifiability issues. The first obvious problem of identifiability occurs if we try to estimate the IFR $\pi^D_{t,m}$, the probability of being hospitalized π^H_m or being treated in ICU π^{ICU}_m, and the case detection ratios $\rho_{t,m}$ simultaneously. This problem is easily circumvented by considering one of the four parameters as known. As the IFR can be reliably estimated in seroprevalence studies and modified by accounting for factors like the age structure of the population, vaccination coverage, and the prevalence of different variants, we consider this factor known to be able to estimate the three remaining factors. The second identifiability issue arises in the estimation of the number of true and unknown infections. Since we assume that this variable follows a negative binomial distribution where the expected value is a function of the effects of NPIs (that are to be estimated), the model can in theory describe the data through any set of values for these parameters if the dispersion is high (i.e., the size parameter is small). Moreover, the dimension of the latent variable can be rather high depending on the length of the observation window and number of geographical regions (the dimension in the latent variable grows with t and m). We address this issue by assuming an informative prior for the different size parameters in the renewal model, the hospitalization models, and the death model and by splitting $I_{t,m}$ in blocks of 10 for each geographical region m to be able to update each block one at a time. Finally, assuming a hierarchical model structure on $\alpha_{k,m}$, R^0_m, and τ_m has the advantage of stabilizing parameter estimates by using information across countries. This effect is particularly important for the estimation of NPI. Since such interventions are often implemented or relaxed as multicomponent interventions on the same or subsequent days in a country, it is difficult to disentangle their effects if we assume country-specific effects that do not follow a hierarchical structure because the estimated effects would be highly correlated. Using a hierarchical model allows us to account for variation in the effect of these interventions while using the information across countries to reduce the correlation between effect estimates. However, it is difficult to determine the exact amount of shrinkage that should be applied (expressed through the prior distributions on the variance parameters). The choice needs to be transparently reported and tested in sensitivity analyses.

Due to the complexity of the hierarchical model, there is no analytical solution and we use a Metropolis–Hastings algorithm (Hastings, 1970) to sample from the joint posterior distribution. For the simulation study and the application, we fine-tune acceptance rates by using an adaptive phase (Brooks et al., 2011; Roberts & Rosenthal, 2009) and discard a defined number of iterations as burn-in. We apply thinning to reduce the autocorrelation in the generated Markov chains. For more details on the implementation, we refer to Section E of the Supporting Information.

3 | SIMULATION STUDY

3.1 | Data generation and study design

We carry out a simulation study with the aims (1) to assess the correctness of the implemented algorithm, (2) to investigate potential problems concerning the identifiability of model parameters, and (3) to assess the impact of model misspecification concerning age stratification. We simulate date according to the model with prespecified parameters. Afterwards, we apply the proposed model to the generated data sets and compare the results with the known parameters and the latent variable. We generate 100 data sets with 10 geographical regions and an observation period of 600 days for each of them. Thus, each data set contains 6000 rows of data. We specify five artificial interventions with mean effects $\alpha_1 = 0.22, \alpha_2 = 0.25, \alpha_3 = 0.3, \alpha_4 = 0.4, \alpha_5 = 0.45$. This allows the reproduction number to be reduced by roughly 80% when all NPIs are active. To obtain region-specific effects of NPIs, we sample from a Gaussian distribution with the corresponding mean α_k and a standard deviation $\sigma_{\alpha_k} = 0.01$. The basic reproduction number is sampled in the same way using a mean $R^0 = 3.25$ and a standard deviation of $\sigma^2_R = 0.1$. We seed the first day of the pandemic in each region by sampling from a negative binomial distribution with a mean τ_m that is generated from a Gaussian distribution with mean $\tau = 10$ and $\sigma_\tau = 2$. All size parameters of the negative binomial distributions are set to 1000 to obtain stable disease dynamics. To obtain realistic time points at which the NPIs are set to active, we generate data in which the decision on whether an NPI is set to active depends on ICU occupancy: To do so, we generate Bernoulli variables for currently inactive NPIs at

TABLE 2 Average estimated effects of nonpharmaceutical interventions (NPIs) on simulated data. All values (except the coverage) are taken as the mean over all simulated data sets and generated Markov chains for all $\alpha_{k,m}$'s.

Intervention	True value	Estimate	Relative bias (%)	Coverage (%)
NPI1	0.22	0.223	1.312	97.4
NPI2	0.25	0.253	1.140	97.0
NPI3	0.30	0.304	1.189	97.2
NPI4	0.40	0.406	1.462	95.0
NPI5	0.45	0.456	1.534	91.7

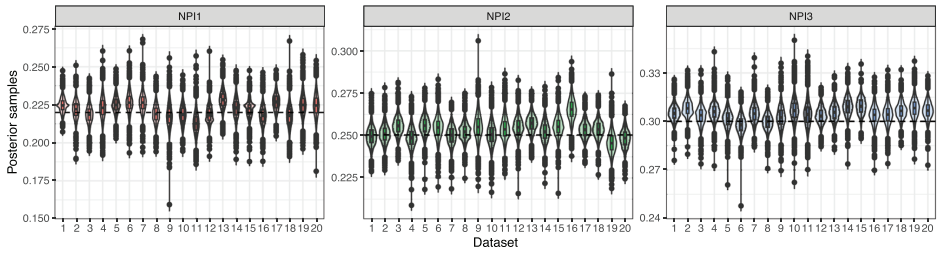


FIGURE 2 Results for the first three NPIs and 20 data sets. The horizontal line is the true mean value.

each t with probability $p_{k,t,m}$ depending on ICU occupancy on $t-1$. In the case an NPI is activated, it remains active for a random time period between 60 and 120 days.

We carry out a second simulation scenario to test the impact of misspecifying the mixing between different age groups (third aim). The proposed model assumes homogeneous mixing by aggregating all time series over the different age groups and only reflects different age compositions via the IFR. We test the impact of this potential misspecification by simulating age-stratified data with diffusion between the age groups and fit the model on the aggregated data. In Section F of the Supporting Information, we provide further details on how the diffusion between age groups is performed. Table S1 and Figure S7 in the Supporting Information shows that aggregating over age strata has only a negligible impact on the estimation of NPI effects.

The data generation is carried out in R version 4.0.4 (R Core Team, 2021). For further details, see the provided R scripts that we used for the data generation.

3.2 | Results on simulated data

We fit the model to each of the 100 data sets where we run two chains with 100,000 iterations and a burn-in of 50,000. We apply thinning by keeping only every 60th iteration. We check convergence by analyzing traceplots and potential scale reduction factors that are always < 1.01 (Gelman & Rubin, 1992). As can be seen in Table 2, the algorithm produces estimates that are very close to the true NPI effects (with a mean relative bias of at most 1.534%) and very high coverage rates. For illustration purposes, we present in Figure 2 the samples from the posterior as violin plot for the first three NPIs and 20 data sets. Figure 3 shows the posterior predictions of the number of (unknown) daily infections (A) and reported cases (B) for one of the 10 regions for one data set. For the simulated data, the model fits very well with a low uncertainty. The results for the missspecified model can be found in Table S1 and Figure S7 of the Supporting Information.

4 | CASE STUDY: MODELING COVID-19 IN 20 EUROPEAN COUNTRIES

4.1 | Data sources

In our case study on COVID-19, we analyze data from 20 European countries (Austria, Belgium, Czechia, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Netherlands, Norway, Poland, Portugal, Slovenia, Spain, Sweden,

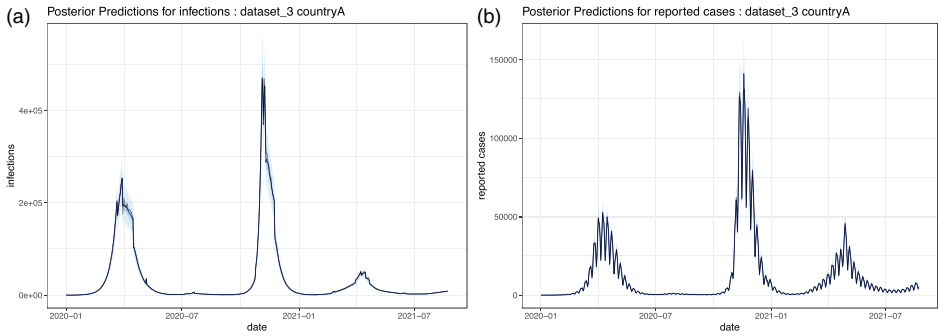


FIGURE 3 Posterior predictions for two time series on simulated data. The (unobserved) infections are shown in A and reported cases in B. Black encodes the true underlying simulated time series. The blue color represents the mean predictions with 95% credible interval.

Switzerland, United Kingdom). Following Flaxman et al. (2020), we define the start of the observation period in each country as 30 days before 10 cumulated deaths were reported. We include data on the entire course of the pandemic until the 31st of October 2021 resulting in a median length of 620 days. We use data on reported cases and deaths from the *Johns Hopkins CSSE COVID-19 Dataset* (Dong et al., 2020). Data on the prevalence of variants of concern and hospital and ICU occupancy are obtained from the *European Centre for Disease Prevention and Control* (European Centre for Disease Prevention and Control, 2022), except for hospital and ICU occupancy for the United Kingdom, which is obtained from the *COVID-19 in the UK dashboard* provided by the UK Health Security Agency (UK Health Security Agency, 2022). As the data on the prevalence of different variants are only available on a weekly basis, we fit a sigmoid function with a squared loss to obtain smooth daily data. More details on this procedure and the resulting time series are presented in Section C.2 with Figures S4 and S5 of the Supporting Information. Data on vaccinations are obtained from *Our World in Data* (Mathieu et al., 2021). Since we use a weighted IFR by age strata, we need information on the number of vaccinations in different age groups. However, very few countries provide information on the age structure of currently vaccinated individuals. We therefore use publicly available data from France and map the relative age-specific vaccination progress to other countries, making the assumption that the prioritization of vaccinations for different age groups evolved roughly in the same manner across different European countries (see Section 2.1). We define the following interventions using information from the *COVID-19 Government Response Tracker* (OxCGRT; Hale et al., 2021) resulting in five NPIs: school closure, gatherings, lockdown, subsequent lockdown, and general behavioral changes. The NPI "school closure" is active when at least some levels of schools and universities (e.g., just high schools) are required to close, and "gatherings" captures the restriction of gatherings to 10 or fewer people. We use two different NPIs depending on whether it was forbidden to leave the house (with possible exceptions such as grocery shopping, and "essential" trips) for the first time ("lockdown") or further times ("subsequent lockdown") because subsequent lockdowns often followed a much more detailed protocol. The last NPI "general behavioral changes" is active from the first time an NPI was implemented in a country and remains active until the end of the observation period. It subsumes many behavioral adaptations that were taken since the beginning of the pandemic and that were respected by a large part of the population in many countries until the end of 2021. These include, for instance, restricting physical contact, working from home wherever possible, higher alertness in case of any respiratory disease symptoms, and the wearing of face masks in some countries. We give a more detailed overview of how we derived these NPIs with the OxCGRT variable coding and the resulting time series (Figure S3) in Section C.1 of the Supporting Information.

4.2 | Challenges in the analysis of the observed time-series

Figure 4 illustrates the challenges in the estimation of the number of daily new infections in a given country by showing the number of reported cases, hospital and ICU occupancy and deaths in the United Kingdom and various factors that have an influence on disease transmission and severity. While the four observed time series show a similar trend during the time between October 2020 and March 2021, they provide rather contradictory information in early-2020 and in late-2021. In particular, the growth rates for the four time series are very different for specific time points (see, e.g., the steep

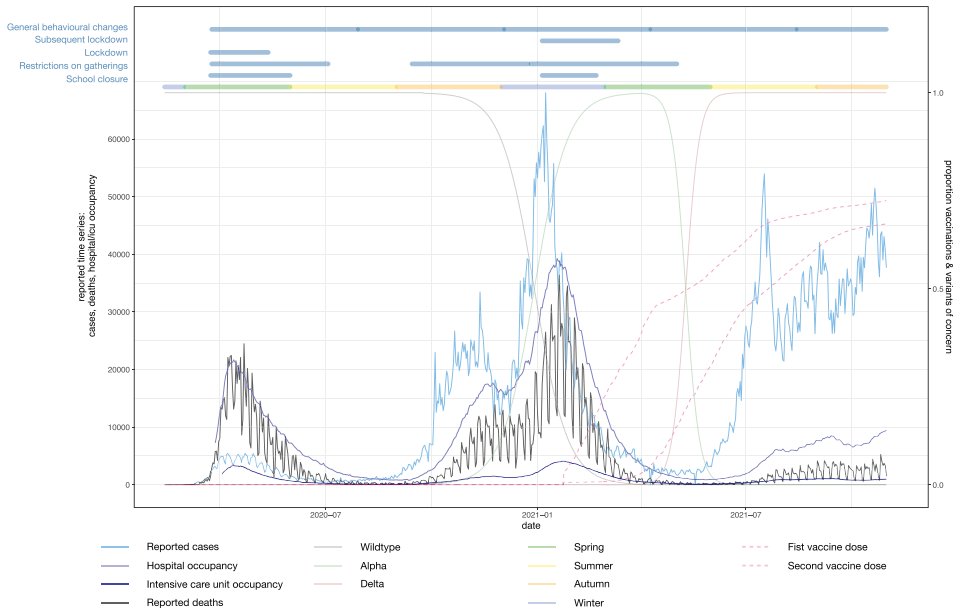


FIGURE 4 Reported cases, hospital occupancy, intensive care unit occupancy, and reported deaths (multiplied by a factor of 20) in the United Kingdom between early-2020 and late-2021. The four observed time series are influenced by the set of nonpharmaceutical interventions that were active at each time point (shown at the top), the season (shown just below), and the number of persons having received a first and second vaccine dose and the prevalence of different virus variants.

increase in reported deaths and hospital occupancy in the early-2020 vs. the more gradual increase in reported cases or the very steep increase in reported cases in July 2021 and the comparably gradual increase in reported deaths). In early-2020, it is obvious that, as in many other countries, only a small proportion of cases were reported because of limited testing capacities. In late-2021, on the other hand, previous infections and vaccinations are likely to have led to fewer severe cases of infections in the population. As a consequence, the numbers of reported deaths and hospital and ICU occupancy are low compared to the number of reported cases. Moreover, the reported time series are not only influenced by the set of NPIs that is active at each time point, but also by the current season with higher infections observed in autumn and winter than in spring and summer and by the prevalence of different virus variants that influence both the transmissibility of the virus and the severity of the disease. Overly simplistic analyses of these time series that only focus on a single indicator of disease transmission and that ignore one or several of the various influencing factors and the weekly patterns in reported cases and deaths may obtain very different answers concerning the same research question, leading to contradictory results that are difficult to communicate to the general public and decision-makers.

4.3 | Results

We run eight chains with a burn-in of 20,000 followed by 50,000 iterations per chain. We apply a thin of 100 resulting in 4000 (i.e., 500×8) samples from the posterior distribution for each parameter. We run a longer adaptive phase with 200 adaptive steps (each with 100 iterations) to get good initial proposal standard deviations. For the final sampling procedure, we again fine-tune these proposals by running 10 adaptive phases (with 50 iterations each). Information about the convergence diagnostic for the parameters of major interest (NPIs and seasonal effects) and further results are presented in Section E of the Supporting Information.

Estimated effects of NPIs Figure 5 provides information on the estimated relative reduction in the reproduction number for NPIs and seasons, respectively. For NPIs, the smallest effect is "school closure" with a credibility interval that includes zero. The most effective NPI is "general behavioral changes," which we defined with the aim to capture several protective measures that were respected by a large portion of the population between the beginning of 2020 and the end

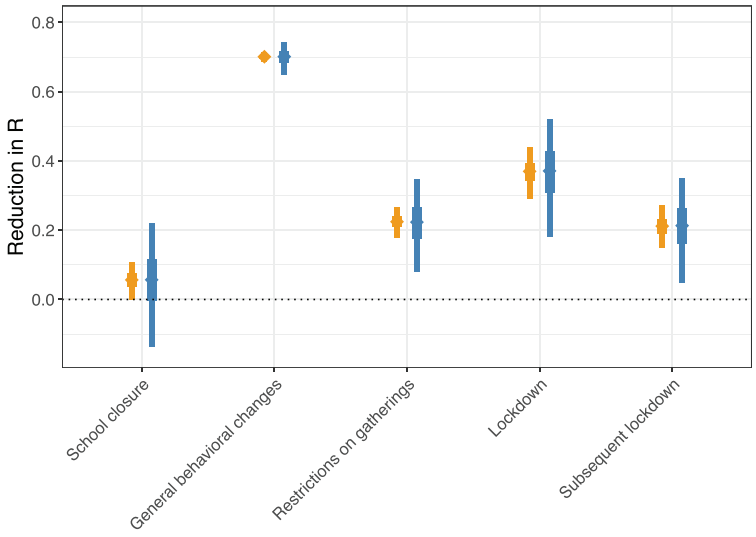


FIGURE 5 Reduction in the reproduction number for NPIs estimated across 20 European countries. Posterior distributions for the mean effects α_k are given in orange. Posterior predictive distribution for $\alpha_{k,m}$ reflecting effect heterogeneity across countries is shown in blue. They are obtained by sampling from a normal distribution with mean α_k and standard deviation σ_{α_k} for each iteration. The 50% and 95% credible intervals are given as bold and normal lines, respectively.

of 2021 including, for instance, working from home wherever possible, higher alertness in case of any respiratory disease symptoms, complying with hygiene recommendations, social distancing, and the wearing of face masks in some countries.

When comparing the effects for the first lockdown with one or several subsequent lockdowns, we can see that the first lockdown is estimated to have a larger effect than subsequent lockdowns, reflecting the fact that the first lockdown was characterized by stronger restrictions and probably better adherence to these than subsequent lockdowns. Figure 6 shows

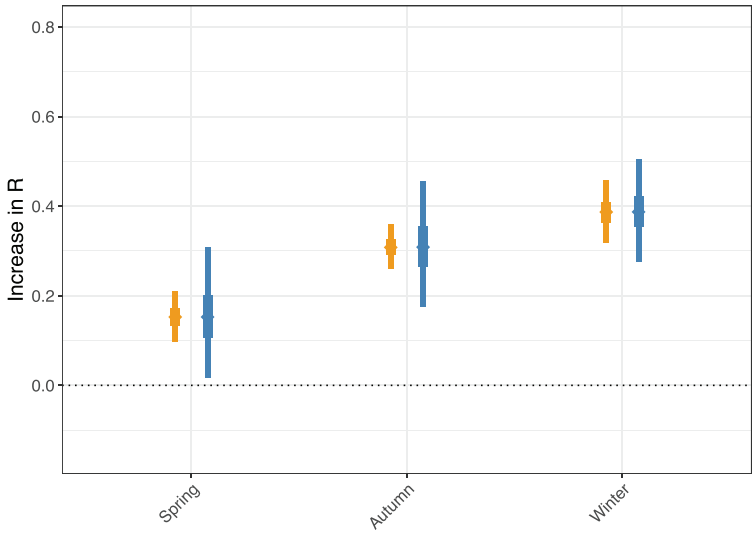


FIGURE 6 Increase in the reproduction number for seasons estimated across 20 European countries. Posterior distributions for the mean effects are given in orange. Posterior predictive distributions are shown in blue. The 50% and 95% credible intervals are given as bold and normal lines, respectively.

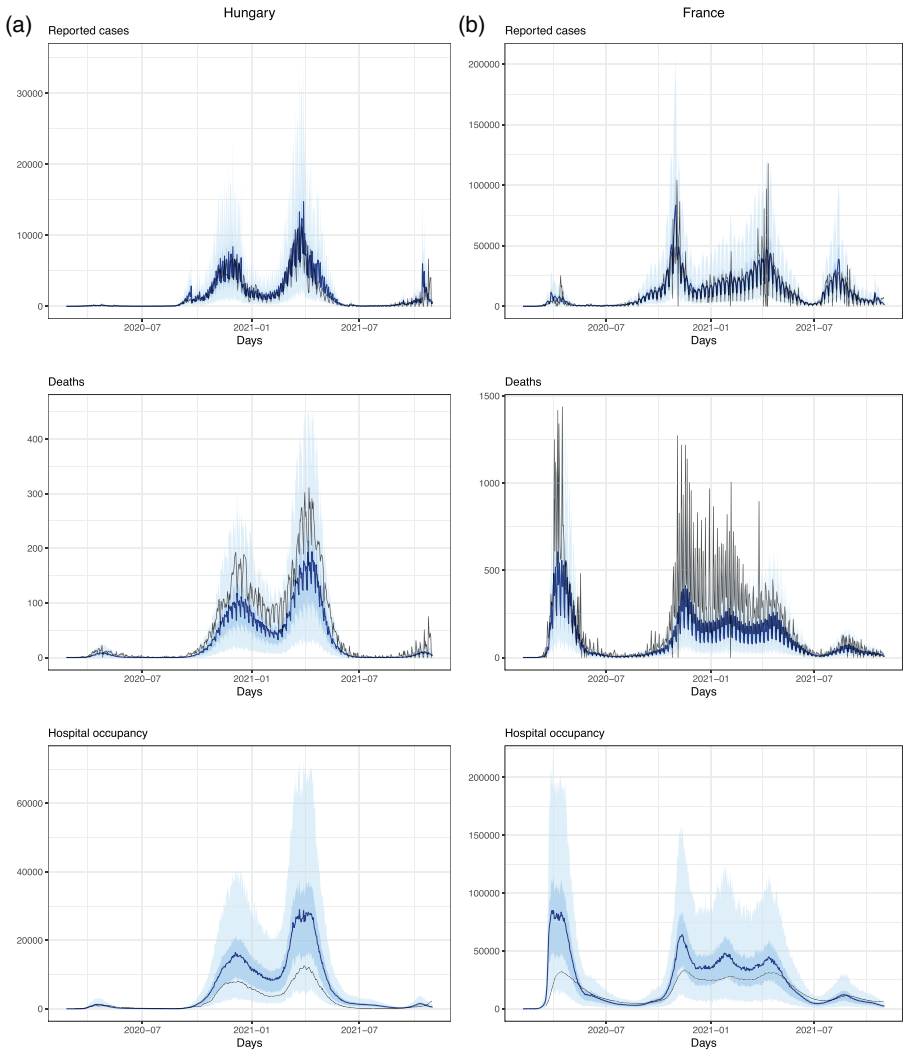


FIGURE 7 Posterior predictions of the reported cases for two countries, Hungary (A) and France (B). The observed time series are given in black and the estimated mean with 50% and 95% credible intervals are shown in blue.

the results for the seasons. As expected, one can observe a strong seasonal influence with an estimated increase in the reproduction number of about 14%, 30%, and 37% for spring, autumn, and winter, respectively.

Model fit and case detection ratios Figure 7 shows posterior predictive checks comparing the observed time series and the posterior predictions for reported cases, hospital occupancy, and deaths for two selected countries, Hungary and France. We chose these two countries, because they represent very different geographical regions in Europe, they differ in their size, they present very different disease dynamics, and data on hospital occupancy were available for both countries. The approach captures the weekly variation in reported cases and deaths that are specific to the two countries. Moreover, it is capable of reproducing the three complementary time series, even though they provide quite contrasting information, in particular, for the first wave. The model shows a tendency to overestimate hospital occupancy during the peaks of the first and second waves of the pandemic for many countries, including Hungary and France. This overestimation might be linked to increases in hospital mortality during the peaks of the first and second waves that are well documented for several countries and have been linked to increasing strain on services that may have led to changes in the case-mix and illness severity of admitted patients (Docherty et al., 2021; Gray et al., 2022; Jassat et al., 2021)

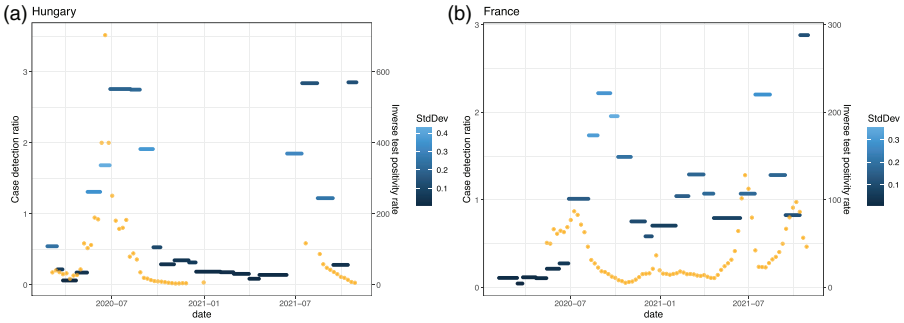


FIGURE 8 Estimated case detection ratios between early-2020 and late-2021 for Hungary (A) and France (B). The shades of blue represent the standard deviation of the posterior with light blue indicating more uncertainty in the estimation. The inverse test positivity rate, which can be interpreted as the number of tests that are performed to detect a case, is given in orange.

Integrating information on reported cases, hospital and ICU occupancy and deaths while accounting for time-varying underreporting in the number of reported cases allow us to estimate variations in case detection ratios that occurred over time. Figure 8 compares these estimated case detection ratios for Hungary and France between early-2020 and late-2021 with the inverse test positivity rate, that is, the number of tests that are performed to obtain a positive test. In general, we observe high underreporting (i.e., very small detection ratios) for the first wave of the pandemic indicating that the true number of infections by far exceeded the reported number of cases. Subsequently, the case detection ratios increased during the summer months and even reached values of over 100%, that is, there were more cases being reported than there were estimated true infections. This might be explained by the fact that the prevalence of the virus was very low during this period and the number of performed tests was very high. In this situation, there may be a nonnegligible proportion of false positive results and we can therefore expect the number of reported cases to exceed the number of true infections due to the imperfect specificity of the tests (Bisoffi et al., 2020; Brownstein & Chen, 2021; Cohen & Kessel, 2020). However, these results must be interpreted with caution as the estimated case detection ratios critically depend on the assumed value of the IFR and on assumptions about how this case fatality rate changes as a function of virus variants and vaccination coverage. In Section H of the Supporting Information, we present in Figure S10 the individual NPI estimations for each country, in Figure S11 the estimated overcontagiousness of the variants of concern, in Figures S12–S14 the posterior predictions for the observed time series, in Figure S15 the estimated infections (latent variable), in Figure S16 the estimated case detection ratios, and in Figure S17 the trace plots for the mean NPIs.

5 | DISCUSSION

We presented a Bayesian hierarchical approach for the modeling of infectious diseases that allows to integrate information on the number of reported cases, hospital and ICU occupancy, and deaths in the estimation of the number of daily new infections. As mentioned in Section 1, previous studies have used various modeling approaches to assess the effect of NPIs on COVID-19 transmission, hospitalizations, and deaths (Banholzer et al., 2021; Brauner et al., 2021; Dehning et al., 2020; Flaxman et al., 2020; Islam et al., 2020; Li et al., 2021a; Sharma et al., 2021; Unwin et al., 2020). Some of these report different findings related to, for example, the magnitude of effect for a specific NPI, as well as the ordering of the relative effectiveness of multiple NPIs. There are numerous other such studies, and systematically identifying and reviewing each of them to compare their findings with those of our study is beyond the scope of this study. Indeed, the principal aim of our study was to show how many of the shortcomings of previous modeling approaches can be overcome by adopting a Bayesian hierarchical approach. Owing to its modular nature, it is possible to model the dynamics of infectious diseases while allowing proper statistical inference and an evaluation of the fit to the observed time series. Moreover, we can integrate the available information while accounting for various sources of uncertainty in this information. By explicitly accounting for time-varying underreporting, seasonality, the spread of different virus variants, vaccination coverage, and previous infections, it is possible to use information over long time periods rather than focusing on short time periods during which these factors remain roughly constant. Using this approach allows for the transparent reporting of model and parameter assumptions and is very flexible: It is thus straightforward to adapt the model to account for additional factors

that might have an influence on disease dynamics. In contrast to most other modeling approaches, our approach explicitly accounts for weekly patterns in the reporting delay distribution for reported cases and deaths, and it is therefore not necessary to smooth these time series. The explicit estimation of new infections on a daily basis is a major advantage if one is interested in the effect of influencing factors that may show variations on a daily scale, for instance, weather conditions (Tosepu et al., 2020), air pollution (Cole et al., 2020), or pollen (Damialis et al., 2021). Due to the flexible combination of different submodels, it would also be straightforward to integrate further information, for example, on the number of tests, on measurements from wastewater, from seroprevalence surveys, or from randomized surveillance testing. Nicholson et al. (2022) combine the latter information with targeted test counts using a stochastic SIR model on weekly aggregated data to obtain fine-scale spatiotemporal prevalence estimates for the United Kingdom.

While the flexible modeling of the observed time series allows to account for different sources of uncertainty, it also comes at the cost of making a number of model and parameter assumptions. Since our modeling approach implicitly gives more weight to reported deaths and ICU and hospital occupancy than to reported cases (by compensating deviations between the true and unknown number of infections and reported cases by differences in case detection ratios for time periods at which the testing strategy changed), our assumptions on how disease severity (and therefore the IFR) is influenced by vaccinations and virus variants play a crucial role in the model. To test the robustness to these assumptions, we conducted extensive sensitivity analyses in which we assessed the degree to which NPI estimates are influenced by variations in the assumptions concerning the overall IFR value and the effect of vaccinations and variants of concern on the IFR. As can be seen in Section G of the Supporting Information, variations in the assumptions concerning these factors has a negligible effect on the estimates of NPI effects.

In our application to COVID-19, we only accounted for the wild type, the alpha, and the delta variant. In principle, it would also be possible to account for the omicron (B.1.1.529) variant using the proposed approach, but there is evidence that this new variant did not only increase the transmissibility of the virus and decrease the severity of the disease, but also entailed changes in the generation time distribution and vaccine efficacy. As a consequence, accounting for omicron would have required a great number of additional assumptions, and it was not in the scope of this work to find reliable information to be able to make all these additional assumptions.

Despite evidence on the importance of asymptomatic infections in the transmission of COVID-19, we did not explicitly distinguish symptomatic and asymptomatic cases in our case study concerning COVID-19. Indeed, it is not clear whether this distinction would necessarily improve the model. This distinction is typically neither made by health authorities in the reporting of cases nor in seroprevalence studies when estimating IFRs. Distinguishing symptomatic and asymptomatic infections would therefore require additional assumptions, in particular, on IFRs that apply only to symptomatic cases, without a clear benefit concerning the insights that we gain from the observable quantities.

Similarly, we assume homogeneous mixing between the different age groups and do not account for age stratification in our model, but it is not clear whether accounting for age stratification would improve our estimates. Indeed, accounting for age stratification would require a great number of additional assumptions, including on the interaction patterns between the age strata in the different countries and age-specific information on the number of reported cases, hospital occupancy, ICU occupancy, and deaths, and this information is only available for a small proportion of the countries that we considered in our application to COVID-19. Even if we had reliable information on mixing patterns between different age groups and age-specific time series, it is not clear whether ignoring these age groups will strongly affect the estimation of NPI effects. In accordance, the results on simulated data presented in Section F of the Supporting Information show that violations of the homogeneous mixing assumption only have a minor influence on NPI estimates. While ignoring age stratification might in general not have a large impact on our NPI estimates, it may lead to an underestimation of the effect of school closures: Since our model relies more on reported deaths and ICU and hospital occupancy than on reported cases, it might not be able to detect an increase or decrease in the number of new infections among children because disease severity in this group is very low. While this reasoning is consistent with some empirical evidence (Fukumoto et al., 2021), others have found differing findings. Studies using various modeling approaches, for example, have reported meaningful decreases in transmission, hospitalization, and deaths due to school closures (Haug et al., 2020; Li et al., 2021a; Liu et al., 2021). Similarly, an overview of systematic reviews, which included and described mainly observational studies, also found that most systematic reviews reported benefits of school closures (Talic et al., 2021). However, each of these studies, as well as many of the underlying studies included in the systematic review, emphasize concerns related to their internal and external validity. Indeed, our model is designed to improve upon multiple assumptions made and approaches taken by such evidence. Nevertheless, our finding that school closures only have a negligible impact on disease dynamics has to be interpreted with caution.

In many countries, the question of whether NPIs should be implemented as a function of reported cases or hospital occupancy was widely debated as both quantities are to some extent unreliable. The proposed Bayesian hierarchical approach provides a framework in which information on both quantities (and on reported deaths and ICU occupancy) can be integrated to predict epidemic development and health care demand in the near future to be able to weigh costs, benefits, and uncertainties in a more robust manner in evidence-informed decision making. While we have developed the approach with reference to COVID-19, the model could easily be adapted to any other known or presently unknown infectious agent.

ACKNOWLEDGMENTS

This work was partially funded by the Volkswagen Stiftung (AZ: 99664). The authors of this work take full responsibilities for its content. The authors thank Anna Jacob for language correction.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest

DATA AVAILABILITY STATEMENT

We provide the code of the Bayesian hierarchical model along with all scripts to generate the data sets for the simulation study and application on 20 European countries. Also, we provide the final processed data sets (simulated and application) directly to run the model and reproduce all results. The code is available at https://github.com/RaphaelRe/BayesModelCOVID or in achieved form at Rehms (2023).

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Raphael Rehms https://orcid.org/0000-0002-3996-0194
Ulrich Mansmann https://orcid.org/0000-0002-9955-8906

REFERENCES

Banholzer, N., Van Weenen, E., Lison, A., Cenedese, A., Seeliger, A., Kratzwald, B., Tschernutter, D., Salles, J. P., Bottrighi, P., Lehtinen, S., Feuerriegel, S., & Vach, W. (2021). Estimating the effects of non-pharmaceutical interventions on the number of new infections with Covid-19 during the first epidemic wave. *PLoS One*, *16*(6), e0252827.

Bisoffi, Z., Pomari, E., Deiana, M., Piubelli, C., Ronzoni, N., Beltrame, A., Bertoli, G., Riccardi, N., Perandin, F., Formenti, F., Gobbi, F., Buonfrate, D., & Silva, R. (2020). Sensitivity, specificity and predictive values of molecular and serological tests for Covid-19: A longitudinal study in emergency room. *Diagnostics*, *10*(9), 669.

Brauner, J. M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., Stephenson, A. B., Leech, G., Altman, G., Mikulik, V., Norman, A. J., Monrad, J. T., Besiroglu, T., Ge, H., Hartwick, M. A., Teh, Y. W., Chindelevitch, L., Gal, Y., & Kulveit, J. (2021). Inferring the effectiveness of government interventions against Covid-19. *Science*, *371*(6531), eabd9338.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. CRC Press.

Brownstein, N. C., & Chen, Y. A. (2021). Predictive values, uncertainty, and interpretation of serology tests for the novel coronavirus. *Scientific Reports*, *11*(1), 1–12.

Cohen, A. N., & Kessel, B. (2020). False positives in reverse transcription pcr testing for sars-cov-2. *medRxiv*.

Cole, M. A., Ozgen, C., & Strobl, E. (2020). Air pollution exposure and Covid-19 in Dutch municipalities. *Environmental and Resource Economics*, *76*(4), 581–610.

Damialis, A., Gilles, S., Sofiev, M., Sofieva, V., Kolek, F., Bayr, D., Plaza, M. P., Leier-Wirtz, V., Kaschuba, S., Ziska, L. H., Bielory, L., Makra, L., Del Mar Trigo, M., & Traidl-Hoffmann, C., COVID-19/POLLEN study group. (2021). Higher airborne pollen concentrations correlated with increased sars-cov-2 infection rates, as evidenced from 31 countries across the globe. *Proceedings of the National Academy of Sciences*, *118*(12), e2019034118.

Dehning, J., Zierenberg, J., Spitzner, F. P., Wibral, M., Neto, J. P., Wilczek, M., & Priesemann, V. (2020). Inferring change points in the spread of Covid-19 reveals the effectiveness of interventions. *Science*, *369*(6500), eabb9789.

Docherty, A. B., Mulholland, R. H., Lone, N. I., Cheyne, C. P., De Angelis, D., Diaz-Ordaz, K., Donegan, C., Drake, T. M., Dunning, J., Funk, S., García-Fiñana, M., Girvan, M., Hardwick, H. E., Harrison, J., Ho, A., Hughes, D. M., Keogh, R. H., Kirwan, P. D., Leeming, G., … ISARIC4C Investigators. (2021). Changes in in-hospital mortality in the first wave of Covid-19: A multicentre prospective observational cohort study using the who clinical characterisation protocol UK. *The Lancet Respiratory Medicine*, *9*(7), 773–785.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track Covid-19 in real time. *The Lancet infectious diseases*, *20*(5), 533–534.

European Centre for Disease Prevention and Control. (2022). Covid-19. https://www.ecdc.europa.eu/en/covid-19

Fisman, D. N., & Tuite, A. R. (2021). Evaluation of the relative virulence of novel sars-cov-2 variants: A retrospective cohort study in Ontario, Canada. *CMAJ: Canadian Medical Association Journal*, *193*(42), E1619.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., Monod, M., Ghani, A. C., Donnelly, C. A., Riley, S., Vollmer, M. A. C., Ferguson, N. M., Okell, L. C., & Bhatt, S., Imperial College COVID-19 Response Team. (2020). Estimating the effects of non-pharmaceutical interventions on Covid-19 in Europe. *Nature*, *584*(7820), 257–261.

Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., Jombart, T., Hinsley, W. R., Grassly, N. C., Balloux, F., Ghani, A. C., Ferguson, N. M., Rambaut, A., Pybus, O. G., Lopez-Gatell, H., … WHO Rapid Pandemic Assessment Collaboration. (2009). Pandemic potential of a strain of influenza a (h1n1): Early findings. *Science*, *324*(5934), 1557–1561.

Fukumoto, K., McClean, C. T., & Nakagawa, K. (2021). No causal effect of school closures in Japan on the spread of Covid-19 in Spring 2020. *Nature Medicine*, *27*(12), 2111–2119.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gray, W. K., Navaratnam, A. V., Day, J., Wendon, J., & Briggs, T. W. (2022). Covid-19 hospital activity and in-hospital mortality during the first and second waves of the pandemic in England: An observational study. *Thorax*, *77*(11), 1113–1120.

Haas, E. J., Angulo, F. J., McLaughlin, J. M., Anis, E., Singer, S. R., Khan, F., Brooks, N., Smaja, M., Mircus, G., Pan, K., Southern, J., Swerdlow, D. L., Jodar, L., Levy, Y., & Alroy-Preis, S. (2021). Impact and effectiveness of mrna bnt162b2 vaccine against sars-cov-2 infections and Covid-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: An observational study using national surveillance data. *The Lancet*, *397*(10287), 1819–1829.

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., & Tatlow, H. (2021). A global panel database of pandemic policies (Oxford Covid-19 government response tracker). *Nature Human Behaviour*, *5*(4), 529–538.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Haug, N., Geyrhofer, L., Londei, A., Dervic, E., Desvars-Larrive, A., Loreto, V., Pinior, B., Thurner, S., & Klimek, P. (2020). Ranking the effectiveness of worldwide Covid-19 government interventions. *Nature Human Behaviour*, *4*(12), 1303–1312.

Höhle, M., & an der Heiden, M. (2014). Bayesian nowcasting during the stec o104: H4 outbreak in Germany, 2011. *Biometrics*, *70*(4), 993–1002.

Islam, N., Sharp, S. J., Chowell, G., Shabnam, S., Kawachi, I., Lacey, B., Massaro, J. M., D'Agostino, R. B., & White, M. (2020). Physical distancing interventions and incidence of coronavirus disease 2019: Natural experiment in 149 countries. *BMJ*, *370*, m2743.

Jassat, W., Mudara, C., Ozougwu, L., Tempia, S., Blumberg, L., Davies, M.-A., Pillay, Y., Carter, T., Morewane, R., Wolmarans, M., von Gottberg, A., Bhiman, J. N., Walaza, S., Cohen, C., & DATCOV author group. (2021). Difference in mortality among individuals admitted to hospital with Covid-19 during the first and second waves in South Africa: A cohort study. *The Lancet Global Health*, *9*(9), e1216–e1225.

Li, Y., Campbell, H., Kulkarni, D., Harpur, A., Nundy, M., Wang, X., & Nair, H. (2021a). The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (r) of sars-cov-2: A modelling study across 131 countries. *The Lancet Infectious Diseases*, *21*(2), 193–202.

Li, Y., Campbell, H., Kulkarni, D., Harpur, A., Nundy, M., Wang, X., Nair, H., & Usher Network for COVID. (2021b). The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (r) of sars-cov-2: A modelling study across 131 countries. *The Lancet Infectious Diseases*, *21*(2), 193–202.

Liu, Y., Morgenstern, C., Kelly, J., Lowe, R., & Jit, M. (2021). The impact of non-pharmaceutical interventions on sars-cov-2 transmission across 130 countries and territories. *BMC Medicine*, *19*(1), 1–12.

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., & Rodés-Guirao, L. (2021). A global database of Covid-19 vaccinations. *Nature Human Behaviour*, *5*, 947.

May, T. (2020). Lockdown-type measures look effective against Covid-19. *BMJ*, *370*, m2809.

Nicholson, G., Lehmann, B., Padellini, T., Pouwels, K. B., Jersakova, R., Lomax, J., King, R. E., Mallon, A.-M., Diggle, P. J., Richardson, S., Blangiardo, M., & Holmes, C. (2022). Improving local prevalence estimates of sars-cov-2 infections using a causal debiasing framework. *Nature Microbiology*, *7*(1), 97–107.

O'Driscoll, M., Dos Santos, G. R., Wang, L., Cummings, D. A., Azman, A. S., Paireau, J., Fontanet, A., Cauchemez, S., & Salje, H. (2021). Age-specific mortality and immunity patterns of sars-cov-2. *Nature*, *590*(7844), 140–145.

Pullano, G., Di Domenico, L., Sabbatini, C. E., Valdano, E., Turbelin, C., Debin, M., Guerrisi, C., Kengne-Kuetche, C., Souty, C., Hanslik, T., Blanchon, T., Boëlle, P. Y., Figoni, J., Vaux, S., Campèse, C., Bernard-Stoecklin, S., & Colizza, V. (2021). Underdetection of cases of Covid-19 in france threatens epidemic control. *Nature*, *590*(7844), 134–139.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rehms, R. (2023). Raphaelre/bayesmodelcovid: Release. https://doi.org/10.5281/zenodo.8047631

Roberts, G. O., & Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, *18*(2), 349–367.

Santé Publique France. (2021). Données relatives aux personnes vaccinées contre la covid-19. https://www.data.gouv.fr/fr/datasets/donnees-relatives-aux-personnes-vaccinees-contre-la-covid-19-1/

Sharma, M., Mindermann, S., Rogers-Smith, C., Leech, G., Snodin, B., Ahuja, J., Sandbrink, J. B., Monrad, J. T., Altman, G., Dhaliwal, G., Finnveden, L., Norman, A. J., Oehm, S. B., Sandkühler, J. F., Aitchison, L., Gavenciak, T., Mellan, T., Kulveit, J., Chindelevitch, L., … Brauner, J. M. (2021). Understanding the effectiveness of government interventions against the resurgence of Covid-19 in Europe. *Nature Communications*, *12*(1), 1–13.

Staerk, C., Wistuba, T., & Mayr, A. (2021). Estimating effective infection fatality rates during the course of the Covid-19 pandemic in Germany. *BMC Public Health*, *21*(1), 1–9.

Talic, S., Shah, S., Wild, H., Gasevic, D., Maharaj, A., Ademi, Z., Li, X., Xu, W., Mesa-Eguiagaray, I., Rostron, J., Theodoratou, E., Zhang, X., Motee, A., Liew, D., & Ilic, D. (2021). Effectiveness of public health measures in reducing the incidence of Covid-19, sars-cov-2 transmission, and Covid-19 mortality: Systematic review and meta-analysis. *BMJ*, *375*, e068302.

Tosepu, R., Gunawan, J., Effendy, D. S., Lestari, H., Bahar, H., Asfian, P., & Ahmad, O. A. I. (2020). Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Science of the Total Environment*, *725*, 138436.

UK Health Security Agency. (2022). Coronavirus (Covid-19) in the UK dashboard. https://coronavirus.data.gov.uk/

Unwin, H. J. T., Mishra, S., Bradley, V. C., Gandy, A., Mellan, T. A., Coupland, H., Ish-Horowicz, J., Vollmer, M. A., Whittaker, C., Filippi, S. L., Xi, X., Monod, M., Ratmann, O., Hutchinson, M., Valka, F., Zhu, H., Hawryluk, I., Milton, P., Ainslie, K. E. C., … Flaxman, S. (2020). State-level tracking of Covid-19 in the United States. *Nature Communications*, *11*(1), 1–9.

Wallinga, J., & Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1609), 599–604.

SUPPORTING INFORMATION

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

How to cite this article: Rehms, R., Ellenbach, N., Rehfuess, E., Burns, J., Mansmann, U., & Hoffmann, S. (2024). A Bayesian hierarchical approach to account for evidence and uncertainty in the modeling of infectious diseases: An application to COVID-19. *Biometrical Journal*, *66*, 2200341. https://doi.org/10.1002/bimj.202200341

Chapter 7

Contribution II

Article

Khazaei, Y., Küchenhoff, H., Hoffmann, Syliqi D. & Rehms R. (2023) Using a Bayesian hierarchical approach to study the association between non-pharmaceutical interventions and the spread of Covid-19 in Germany. *Scientific Reports* 13, 18900. https://doi.org/10.1038/s41598-023-45950-2

Code and Data

The code and preprocessed data is available at https://github.com/RaphaelRe/COVID_NPIs_Germany.

Supplementary Material

Supplementary material can be found at https://doi.org/10.1038/s41598-023-45950-2 (section: *Supplementary Information*).

Copyright information

Author Contributions

Y.K., **R.R.**, H.K. and S.H. conceived and conceptualized the research. **R.R.** and S.H. designed and implemented the original software of the MCMC sampler. **R.R.** adapted and made necessary changes to it. Data collection and preprocessing was done by D.S., Y.K. and **R.R.** The model application and analysis was conducted by **R.R.**
Y.K. and **R.R.** wrote the initial draft that was significantly expanded by S.H. and H.K. All authors discussed the results and were closely involved in proofreading and revision of the manuscript.
All authors discussed the results and were closely involved in proofreading and revision of the manuscript Corresponding author is **Y.K.**

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Using a Bayesian hierarchical approach to study the association between non-pharmaceutical interventions and the spread of Covid-19 in Germany

Yeganeh Khazaei[1]✉, Helmut Küchenhoff[1], Sabine Hoffmann[2,3], Diella Syliqi[1] & Raphael Rehms[2,3]

Non-Pharmaceutical Interventions (NPIs) are community mitigation strategies, aimed at reducing the spread of illnesses like the coronavirus pandemic, without relying on pharmaceutical drug treatments. This study aims to evaluate the effectiveness of different NPIs across sixteen states of Germany, for a time period of 21 months of the pandemic. We used a Bayesian hierarchical approach that combines different sub-models and merges information from complementary sources, to estimate the true and unknown number of infections. In this framework, we used data on reported cases, hospitalizations, intensive care unit occupancy, and deaths to estimate the effect of NPIs. The list of NPIs includes: "contact restriction (up to 5 people)", "strict contact restriction", "curfew", "events permitted up to 100 people", "mask requirement in shopping malls", "restaurant closure", "restaurants permitted only with test", "school closure" and "general behavioral changes". We found a considerable reduction in the instantaneous reproduction number by "general behavioral changes", "strict contact restriction", "restaurants permitted only with test", "contact restriction (up to 5 people)", "restaurant closure" and "curfew". No association with school closures could be found. This study suggests that some public health measures, including general behavioral changes, strict contact restrictions, and restaurants permitted only with tests are associated with containing the Covid-19 pandemic. Future research is needed to better understand the effectiveness of NPIs in the context of Covid-19 vaccination.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has quickly spread globally, with more than 38.4 million cumulative confirmed Covid-19 cases in Germany from the beginning of the pandemic by the end of May 2023, including a total of 174,170 deaths associated with SARS-CoV-2 infection[1]. Starting from March 2020, different bundles of non-pharmaceutical interventions (NPIs), at different times with varying stringency have been implemented to control the transmission of the virus. This was mostly done to protect the most vulnerable individuals from infection and to mitigate the surge of patients requiring hospitalization. By doing so, it aimed to protect the healthcare system from being overwhelmed by a sudden influx of cases. These NPIs included but were not limited to, containment measures such as domestic or international travel bans, individual protection measures like mask-wearing requirements, social distancing measures such as school closing and gathering bans, and health system measures like testing and contact tracing. All the NPIs are considered essential components of public health that people and communities can take to help slow the spread of illnesses[2–4]. However, the effectiveness of these policies remains a subject of debate, requiring further exploration to better understand the relationship between NPI intensity, duration, and their impact. It should be noted that, alongside these mandated policies, volunteer social behavioral changes were also observed during the pandemic. Thus the investigation of the impact of behavioral changes and concurrent NPIs is of immense importance.

[1]Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig-Maximilians-Universität, Munich, Germany. [2]Department of Statistics, Ludwig-Maximilians-Universität, Munich, Germany. [3]Institute of Medical Data Processing, Biometrics and Epidemiology (IBE), Faculty of Medicine, Ludwig-Maximilians-Universität, Munich, Germany. ✉email: yeganeh.khazaei@stat.uni-muenchen.de

In light of the severe social and economic costs[5], affecting individuals' behavior and mental health of these interventions[6], it is crucial to quantify the effects of these measures. During the past two years, there were several attempts to identify the most influential measures across the world including Europe[4,7–14]. However, there is much discussion and controversy around the matter. In 2022, Rehms et al. proposed a Bayesian hierarchical approach, as a common framework, that integrates disease incidence, hospital occupancy, and mortality, as complementary sources of information to get a reliable estimate of the unknown number of infections. The model takes into account that published data suffers from time-varying under-reporting and reporting delays. Moreover, effects on the disease dynamics over time are incorporated: The effect of vaccinations starting on the 8th of December, 2020, and the rise of new variants of concern, which accelerated the spread of the virus and increased its lethality. By explicitly modeling these characteristics, it is possible to look at a larger time horizon and therefore utilize more data, making the results more robust. Hence, one is not forced to look at small periods where constant disease dynamics can be assumed to justify simplified models[15].

Germany has implemented different containment and mitigation strategies starting in March 2020. In this paper, we apply the proposed Bayesian hierarchical approach for all sixteen states of Germany, for a time period of 21 months, and evaluate the effectiveness of different NPIs.

Materials and methods

Data sources and preprocessing

Selection of NPIs

This study includes various NPI time series obtained from the Corona data platform. A team of researchers from different institutes, including the Corona data platform, Infas, Infas 360, and the University of Bonn, funded by the Federal Ministry of Economics and Climate Protection of the Federal Republic of Germany collected regional data on an ongoing basis of all measures and epidemiological-medical as well as socio-economic indicators of all cities and districts[16]. From 23 categories and 1152 subcategories[17], 8 main measures, across 16 different states are selected: "contact restriction (up to 5 people)", "strict contact restriction", "curfew", "events permitted up to 100 people", "mask requirement in shopping malls", "restaurant closure", "restaurants permitted only with test", "school closure" and one additional NPI labeled as "general behavioral changes". Rehms et al. proposed this NPI to account for changes in people's behavior during the pandemic. This NPI is active all the time starting with the activation of the first NPI[15]. It is important to account for such an effect as it makes the other NPIs more comparable (e.g. closing schools or restrictions on gatherings will not have a comparable effect if social distancing in case of respiratory symptoms is practiced). It can therefore be seen as a residual for untracked or latent NPIs which are not directly implemented by the government and are implicitly active through behavioral changes in the population.

The reasons for which we selected this set of NPIs are threefold: they either reflect the characteristics of a specific time span of the pandemic (curfew, events permitted up to 100 people, restaurant closure, restaurants permitted only with test, school closure), they enable us to compare two specific measures with different strictness level (contact restriction (up to 5 people), strict contact restriction), or they evaluate the effectiveness of more long-lasting measures such as mask requirement in shopping malls. Note that not all of the interventions were implemented in all the states. We define the interventions as presented in Table 1.

Reported cases, hospitalizations, and deaths

We use the following official data sets: data on reported cases, hospitalizations, and deaths in Germany are published daily by the RKI on a state level[18]. The RKI is a German federal government agency and scientific institute

Name of NPI	Definition
Contact restriction (up to 5 people)	Max. 5 people, except a household and close family members (private and public settings merged together)
Strict contact restriction	Only persons of a household and close family members (private and public settings merged together)
Curfew	Exit restriction; leaving the apartment only for a valid reason
Events permitted up to 100 people	Indoor public events, up to 100 people
General behavioral changes	This NPI captures many behavioral adaptations people took during the pandemic and is active from the first time an NPI was implemented in a state and remains active until the end of the observation period. The list includes but is not limited to wearing masks, increased engagement in positive/negative health behaviors, working from home, less physical contact, and generally higher vigilance in terms of one's personal health
Mask requirement	In shopping malls and sales outlets
Restaurant closure	Catering establishments of any kind are prohibited. The sale and delivery of takeaway meals are exceptions
Restaurants permitted only with tests	Test-related access restrictions
School closure	Primary/secondary schools and partial/complete school closures merged together (selection of 1 final class and grade, or selection of 2 final classes and grades, or emergency selection of 3 classes and grades for children of certain parent groups, or selection of 4 teaching sessions of only certain subjects). All school holidays for each state were added manually

Table 1. Description of the defined non-pharmaceutical interventions (NPI).

responsible for health reporting, disease control, and prevention. As the national register for Covid-19, it preserves all identified disease cases reported by the local health authorities. In our analysis, we use daily reported cases, the number of new patients admitted to hospitals due to Covid during the past 7 days, and a daily number of deaths due to Covid-19, on a daily basis for each state.

Intensive care unit occupancy
Data on the daily occupancy of Intensive Care Unit (ICU) beds in Germany is made publicly available by the German Interdisciplinary Association for ICU Medicine and Emergency Medicine[19].

The hierarchical model

In this section, we provide a short description of the used model. A more technical description is given in the Supplementary Material Section B. To estimate the effect of the NPIs, we use a Bayesian hierarchical approach proposed by Rehms et al.[15]. Hierarchical models provide a flexible framework to describe complex phenomena through the combination of different submodels. Hereby, each submodel handles another small part of a big problem, making the intractable tractable. To apply the model to German data, we modify the proposed model, making it more flexible and tailored to the data. In the following, we give a short description of the model and its modification. For more insights on the methodological aspects, we refer to the original work of Rehms et al.[15].

The model can be divided in two constituent parts: The first one infers the number of infections for each time point and region from given data. The second one infers, given these infections, the effect of the NPIs. The actual true number of infections can not be observed directly, as official numbers suffer from incomplete and delayed reporting, variations in testing strategies, and more. To infer these actual infections over time, the proposed model uses four different time series as complementary sources of information: reported number of deaths, cases, and the occupancy of hospital beds and ICUs. Each of the series provides individual information on the disease dynamics. For example, one could use the reported deaths to 'calculate back' the number of infected individuals. We do this with all four series linking each of them through individual submodels. As we use a probabilistic Bayesian approach, the uncertainty about each of the linked models is preserved in the inferred number of infections. Given these inferred infections for every day in every region, it is possible to estimate the effect of the NPIs using a renewal equation (see e.g.[20]). The renewal equation formulates the disease dynamics as a function of the reproduction number and the past infections. It can be seen as a flexible version of classical compartment models like the SIR model[21]. The reproduction number in this renewal equation is formulated as a function of a basic reproduction number and the effect of the NPIs. This quantities are then estimated within this framework.

To derive the unobserved infections, the model takes into account that data are reported with delays and suffer from seasonal and structural under-and over-reporting due to weekends and varying testing policies. Moreover, the model does not rely on constant dynamics over time: With the surge of new variants of concern, the contagion process and the infection fatality rate (IFR) change. Both quantities are also affected by the vaccination coverage of the population. By including all these factors, it is possible to use a much larger time horizon, resulting in more usable information to reliably infer the effect of NPIs. It is therefore not necessary to focus on short time periods during which disease dynamics and the degree of underreporting can be assumed to have remained constant. To get estimates of the NPIs, the model is designed in a hierarchical way: The effect of each NPI is estimated for each location (here, the federal states of Germany) separately while sharing a common mean and standard deviation. Therefore, the parameters can borrow information from each other while allowing for an individual effect for each location. The hierarchical formulation gives robust and reliable estimates of the NPIs'. Besides the effect of the NPIs, the effect of the seasons as a proxy for behavioral changes of the population with respect to drifts in weather and temperature are estimated as well. To fit the model of Rehms et al.[15] to the German data, we modify it in two ways. Firstly, as Germany does not provide data about hospital occupancy, we use hospital admissions (for ICU, occupancy data were available). The time-shifting distribution to be used is simplified, as it is not necessary to model the time of occupancy. Secondly, we introduce a new parameter that allows for a change in the fraction of hospitalizations and ICU occupancy on the 1st of July 2020. This gives the model more flexibility to estimate the relationship between hospital data and unobserved infections. As there was only limited experience with COVID-19 in the first wave, the hospitalization pattern may have changed afterwards.

Data preparation

Official data for the first months of the pandemic were not available on the state level. We impute these data points as follows: We considered the sum of the first two weeks of the available data on reported deaths (as it is the most reliable source of information) and calculated the relative proportion of each state compared to all reported deaths in Germany. As aggregated data on the country level were available from the start of the pandemic, we use these estimated proportions to split this aggregated data into the theoretical number of reported deaths and cases for each German state (rounded to integers). This procedure implies that at the beginning of the disease, the infection dynamics are the same in each federal state. Following Rehms et al.[15], we define the start of the observation period in each state as 30 days before the number of reported cumulative deaths reaches or exceeds a count of ten. The first considered dates range from the 18th of February 2020 for Bavaria to the 8th of March 2020 for Mecklenburg-Western Pomerania. The median of the considered days is 615. The observation period ends on 31 October 2021. The merging process of "Berlin & Brandenburg", "Bremen & Niedersachsen" and "Hamburg & Schleswig-Holstein" is described in Supplemental A.1 Section.

Conducting Bayesian inference

The proposed model requires the definition of many different parameters (fixed and variable). To get sensible specifications, we use the same definition as Rehms et al.[15]. Fixed quantities are taken exactly as the specification,

as for Germany, and deployed for all states of Germany. This assumes, for instance, that the infection fatality rate for two persons who are doubly vaccinated and getting infected with the alpha variant of the virus at the same time is the same, regardless of their location. Prior distributions are also set in the same way as in Rehms et al.[15].

Inference of the proposed model is done via Markov Chain Monte Carlo as described in detail in Rehms et al.[15] using a customized Metropolis-Hastings update scheme tailored for the model[22,23]. We run eight chains with 50,000 iterations. Beforehand, each chain used 20,000 samples as burn-in. We only keep each 100th value of each sampled Markov chain to reduce autocorrelation, resulting in a final sample of 4000 data points from the posterior distribution. The presented results are the derived empirical quantities from this model.

Results

We fit the model to our dataset, which included the list of NPIs (as described above), coded as dummy variables with the earliest start at 2020-02-18 until 2021-10-31 for 13 federal states. Each row was complemented by the data on the number of reported cases, hospitalizations, ICU occupancy, deaths, vaccination coverage, seasons, and variants of concern. Using our model, we were able to calculate the effectiveness of each NPI in reducing the instantaneous reproduction number as a percentage value for each measure.

Figure 1 shows the estimated effects of the NPIs as a reduction or an increase in % of the instantaneous reproduction number with a 95% credible interval. The effects are presented as the mean effects over all states. The largest effect is given by general behavioral changes, which reduces the reproduction number by 70% (CI: (68%, 71%)). This effect captures non-observable effects which are not encoded by other NPIs. The effect in the reproduction number was followed by a significant reduction with strict contact restriction by 14% (CI: (3%, 24%)), restaurants permitted only with tests by 13% (CI: (9%, 17%)), contact restriction (up to 5 people) by 12% (CI: (7%, 16%)) and restaurant closure by 8% (CI: (3%, 13%)). Curfew showed a marginal effect by reducing the reproduction number by 7% (CI: (0%, 13%)), while events permitted up to 100 people showed an increase in the reproduction number by 7% (CI: (−14%, −1%)). This increase was observed in school closures by 5% (CI: (−10%, −1%)) and mask requirement in shopping malls and sales outlets by 3% (CI: (−7%, 1%)) as well, albeit not statistically significant. Furthermore, we estimate the effect of the season as a nuisance parameter. One can interpret its result as a relative change to the summer months, which serves as a reference category. We observed a significant negative effect for autumn (an increase in the reproduction number by 28% (CI: (−33%, −23%))

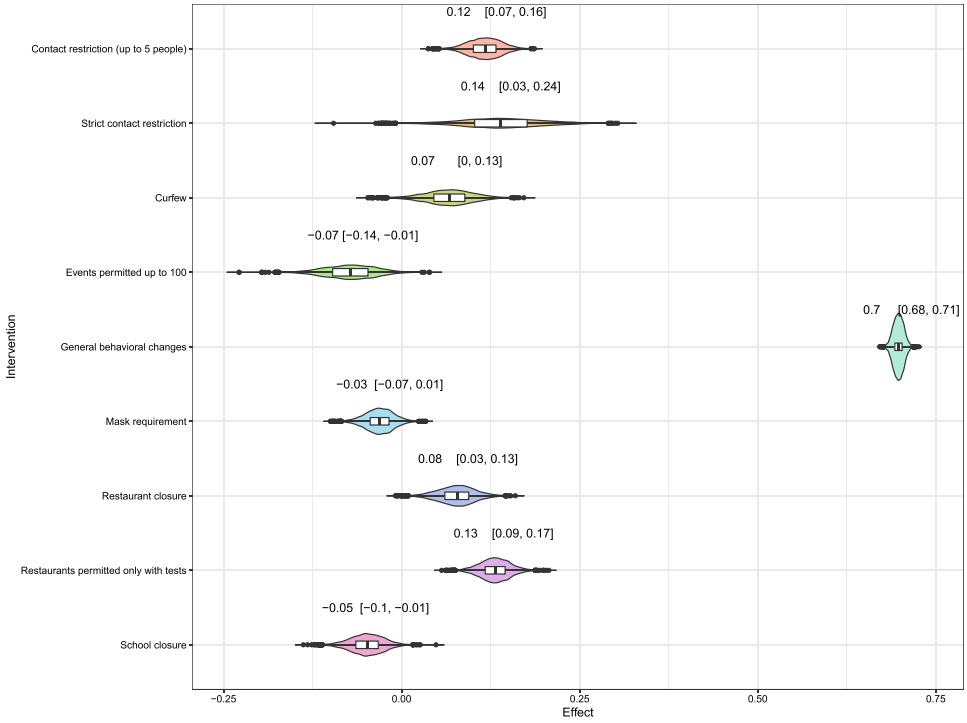


Figure 1. Estimated effects of the defined NPIs. The x-axis gives the relative reduction in % (obtained with the transformation $1 - \exp(-\alpha_k)$ from the original estimated values). The y-axis indicates the defined NPI. The colored area shows the distribution of the estimate. The number above each row shows the estimated mean effect along with a 95%-credible interval. As the numbers indicate a relative reduction, a negative value can be interpreted as a *relative increase*.

and for winter (by 7% (CI: (−13%, −1%)). Spring reduces the reproduction number significantly by 11% (CI: (7%, 14%)).

To test the robustness of the results, we also made three sensitivity analyses by varying crucial parts of the model that could influence the results. However, we found no substantial difference in the estimated effects. Besides the results relating to the NPIs, the model also estimates the unknown number of actual infections in all federal states. We show an aggregated version of these estimations in Fig. 2, i.e. the sum of the estimated infections over all federal states with a 75%- and 95%- credible interval in blue. For comparison, we also provide the reported cases (also aggregated to country level), but shifted by six plus seven days to account for the incubation time and the time until a case is actually reported. The estimation of the infections implies higher under-reporting in phases of large growth rates, in particular in the first and second wave of the pandemic. This effect seems to vanish around July 2021.

Discussion

Since the beginning of the Covid-19 pandemic, many recommendations have been made for citizens and several social measures have been implemented. Besides vaccination, NPIs play a unique role in preventing the pathogen from being transmitted. Previously, the effectiveness of NPIs across different countries was extensively reviewed. Here, we used a data-driven approach to estimate the effects of nine NPIs, from March 2020 to October 2021, in Germany. General behavioral changes were associated with the largest reduction in the effective reproduction number, followed by measures including strict contact restriction, restaurants permitted only with a test, contact restriction (up to 5 people), restaurant closure, and curfew. The current work showed that some NPIs were associated with a clear reduction in the instantaneous reproduction number, which was consistent with the increasing evidence indicating that NPIs are efficient in alleviating and controlling Covid-19 outbreaks. However, some NPIs showed mixed results compared with the existing literature.

We believe that by using the proposed Bayesian hierarchical approach, we can integrate different data sources and this has several benefits such as increasing the amount of data available for estimating NPI effects and reducing biases in the reporting of cases and deaths. Since we are not attempting to deduce the total number of Covid-19 infections, our results are more robust to violations concerning the assumptions about specific infection fatality rates (IFR). Lastly, we allow the effects of all NPIs to vary across states, acknowledging differences in NPI implementation and adherence.

There are some factors that limited our analysis of the estimation of NPI effects. First, defining NPIs proved to be a complex task. According to the Corona data platform, an NPI variable was considered active if the measure was in place at the federal-state level. This information, while exceptionally detailed and organized, was also dependent on the 7-day incidence rate, potentially resulting in a weekly pattern for NPI activation, as outlined in the Supplementary Material Section A. Therefore, we implemented a rigorous decision-making strategy to extract the NPI data. Second, many measures were introduced simultaneously (e.g., the introduction of mask requirements and the prohibition of mass events). Talic et al. noted a similar challenge, where over half of the 72 studies they reviewed couldn't be included in their meta-analysis because they evaluated "packages" of measures in the

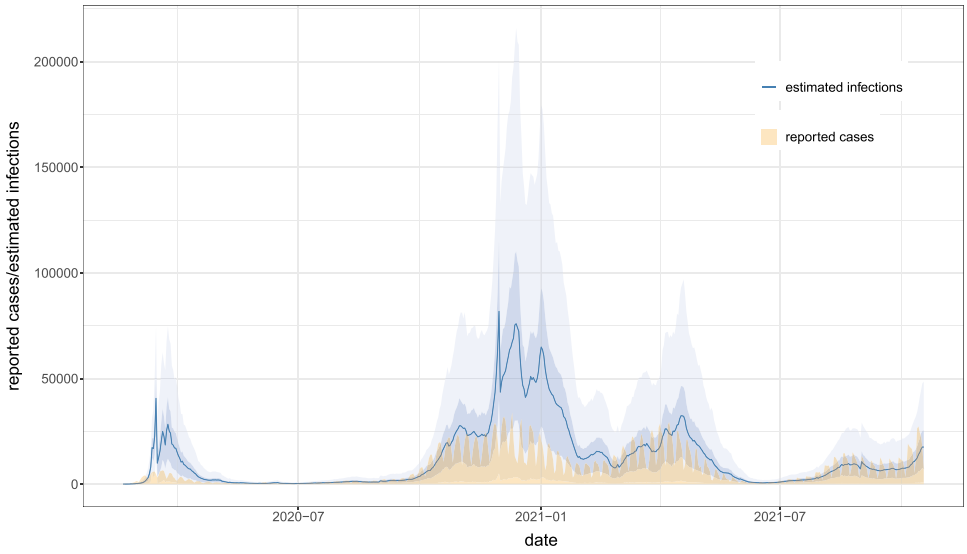


Figure 2. Estimated infections with 75%- and 95%- credible intervals in blue. In yellow we provide the reported cases for a better comparison. We shifted the reported cases by 13 days to the left to get a better comparison between the two curves. This shift reflects the mean time until an infected person is getting symptoms (roughly 6 days) and being reported as an actual case (roughly 7 days) afterward.

form of NPI combinations, making it impossible to assess the effects of each NPI individually[24]. On the other hand, the imposition and relaxation of control measures varied across federal states. For instance, contact restrictions limiting gatherings to a maximum of 5 people in private settings were enforced in Baden-Württemberg in mid-March 2020, while the same restriction was applied in Bavaria in mid-May 2020. Similarly, we anticipated substantial variation in school closure measures since we accounted for school holidays, which are not uniform nationwide across Germany, meaning it is not the same in different states. Consequently, decisions on closing schools were sometimes made in accordance with school holidays; leading to instances of no implementation of school closures provided a holiday in place or sometimes merely extensions of existing holidays. Given the complexities described above, through the hierarchical formulation of the model, it is possible to identify the effect of an NPI, as long as there is some degree of heterogeneity among different locations. A third limitation in our analysis pertains to potential interdependencies among infection dynamics in different states. Unfortunately, our model cannot account for these potential dependencies.

The key strength of this study is twofold. First, we were able to use high-quality and comprehensive daily data on reported cases, hospitalizations, deaths, and occupancy of ICU beds provided by RKI and DIVI through their respective dashboards. Furthermore, a critical asset to our analysis is the Corona data platform, which consistently delivers detailed information on the implementation of NPIs at the state level on a daily basis. This granular dataset empowers us to define NPIs with precision, a fundamental requirement for our investigation. It is worth noting that the data on the platform is methodically based on the Oxford Stringency Index[25], a recognized metric for assessing governmental responses to the pandemic. Since March 2020, the platform has systematically collected official publications pertaining to Covid-19 protective measures and diligently categorized their content into various upper and subcategories. While our study shares a hierarchical data structure with the Oxford Stringency Index, it distinguishes itself through the depth and the content of the respective coding[17]. The second strength is the use of a sophisticated model that comprises a wide variety of aspects, including but not limited to the use of the information from four different daily time series (reported cases, hospitalizations, deaths, and ICU occupancy) to infer reliable disease dynamics. While the model accounts for uncertainties in the information (e.g. under-reporting), it also considers effects like vaccinations and the emergence of new variants of concern making it possible to use information over a relatively long period of time giving more informed estimates. Moreover, we do not need to smooth the observed time series, since we account for variations in daily reporting patterns in reported cases and deaths.

The roles of general behavioral changes or measures and their public adoption during a pandemic have been evaluated before[26–28]. The term general behavioral changes encompasses any actions that contribute to reducing the transmission of Covid-19 and, as a result, aid in containing the pandemic. Hence, this NPI subsumes a large variety of not directly defined NPIs which are more latent and are very difficult to define on an aggregated level. It can include various practices, such as practicing good hygiene by washing or disinfecting hands, following proper cough and sneeze etiquette, and regularly cleaning surfaces. Additionally, it involves engaging in voluntary physical distancing measures, such as staying at home, limiting close contact, and avoiding crowded places. Wearing masks or gloves, staying home when experiencing respiratory symptoms, utilizing testing services, refraining from non-essential travel, and utilizing contact-tracing applications are among other measures included within this term. The definition of general behavioral changes can therefore be quite vague depending on the considered context of a study as it may include some of the mentioned aspects or not. In this work, the NPI general behavioral changes serve as a controlling variable, making other NPIs more comparable to each other by capturing latent NPIs that are not directly defined or implemented. It is therefore important, to interpret the result for this variable with caution. More often than not, general behavioral changes or measures are underevaluated and their consideration in the epidemiological models is limited[26]. In an SEIR model suggested by Khairulbahri, they reported a reduction in infected cases of about 22%, by studying behavioral measures effect, which was consistent with the findings of our study[27]. A reduction in the incidence of Covid-19 associated with physical distancing (75%, CI: (59%, 95%)) was reported in a meta-analysis by Talic et al., which is in line with the results of our study[24]. In addition, Brauner et al. showed limiting gatherings to fewer than 10 people had a large effect size for reducing transmission at the advent of the pandemic (42%, CI: (17%, 60%))[29]. We found similar substantial reductions in the reproduction number for restaurants permitted only with tests and restaurant closure. Ledebur et al.[30] reported restrictions in gastronomy reduced transmissions by about 17%. However we should keep in mind that their analysis focused on less disruptive measures that did not consist of full closures, but rather of restrictions such as mandatory registration of visitors, limits for the opening hours, or the number of people seated at a table. However, on the same note, the effectiveness of fully closing gastronomy has been repeatedly established in the literature[31–33].

We observed marginal effects for cancellation of events beyond 100 people and curfew in our results, which should be interpreted with caution. Previously, in a study on the effectiveness of a nighttime curfew in Hamburg, Germany, the researchers concluded that the curfew was substantially reducing the number of Covid-19 cases[34]. Several other studies found that nighttime curfews reduce mobility, hence they result in fewer Covid-19 infections[35]. It should be considered that we included curfew as an exit restriction; leaving the apartment only for a valid reason, which is considered a harsher intervention than a nighttime curfew. In detail, curfews and cancellation of events beyond 100 people were implemented only over short periods across many states, and mostly, in co-treatments with other NPIs implemented at the same time. This endogeneity of the policies can hinder the estimation of the true effectiveness of NPIs. On the other hand, it has been discussed that strict exit restrictions that limit the lives of citizens might backfire and increase Covid-19 infections[36]. However, there is little and rather mixed evidence on the effectiveness of curfews with varying strictness to contain the Covid-19 pandemic.

Our study showed no evidence for the effectiveness of school closure. Previously, Talic et al. were not able to make a consensus for school closure, due to the high heterogeneity between studies. They qualitatively reported that school closure could be highly effective if implemented early, with low incidence rates of Covid-19[24]. This

is in accordance with Fritz et al who emphasized that the effectiveness of school closure in the case of Covid-19 is inconclusive and high caution should be maintained when interpreting the results, specifically due to many socio-economic and psychological implications to it[37]. Similarly, the same result was reported by Rehms et al. They estimated the smallest effect for school closure with a credibility interval that included zero[15]. Moreover, Isphording et al. showed the number of Covid-19 infections did not increase with school re-openings in the summer of 2020[38]. All in all, with the simultaneous implementation of different public health measures, the results should not be overstated[24,39,40].

Thoroughly examining the effectiveness of interventions presents relevant obstacles in terms of methodology. While simulation studies can investigate different situations, they rely on strong assumptions that may not be easily verifiable and bring a low level of evidence[41,42]. As an alternative approach with potential, we used cross-state modeling that is data-driven and compares the timing of state-wide interventions with the subsequent cases, hospitalization, ICU, or death counts. In the previous works, there was a fairly large variation among the inclusion of different sets of NPIs and methodologies in use. They reported varying results on the effectiveness of public health measures in reducing different outcomes such as incidence, transmission, or mortality. Hence, the comparison between these studies can be impeded by this variation. Talic et al. mainly used observational studies from different countries in their meta-analysis. They further explained the concern hovering around the ability of the mathematical models and their assumptions, to predict the course of virus transmission or the effectiveness of interventions was the main reason they excluded such studies in their meta-analysis[24,43]. Additionally, sophisticated and flexible methods, like Bayesian longitudinal models, were used by some researchers[9,29,44]. For instance, Hunter et al. used Bayesian generalized additive mixed models to adjust for spatial dependency in Covid-19 between nation states, as well as multilevel mixed-effects negative binomial regression model with cases or deaths on a specific day as the outcome variable[45]. Some studies used linear regression, simple correlation coefficients[41], or mixed effects linear regression[42].

Moreover, different outcomes of interest have been reported as well, including the number of confirmed cases, mortality or death rate, or confirmed deaths. Bo et al. evaluated the effectiveness of four types of NPIs on the transmission of Covid-19 by generalised linear mixed model, with city/country-level random intercept in the model to control for clustering effects within the same city/country[46]. At the same time, focusing on short time periods during the pandemic or using either the number of reported cases, intensive care occupancy, new hospital admissions, or deaths, as a single indicator of disease transmission, would give an incomplete picture of the pandemic. All things considered, the settings, methodologies, and results of these studies were inconsistent and the interpretation and application of their findings and methods should be done cautiously. Banholzer et al. pointed out that such huge variation in a plethora of published studies can result in the robustness of the results in different settings, as much as it can hinder conclusive evidence on the effectiveness of NPIs[44].

In conclusion, the current work contributes to the body of evidence on the effectiveness of individual NPI. As previously mentioned, serious deficiencies in the available empirical data are observable. Although our work focused on a data-driven approach to estimate the effects of NPIs, our estimates should not be taken as the final word on NPI effectiveness. Further high-quality original studies with reliable effect estimates are necessary in order to avoid unrealistic expectations or overestimation of the effectiveness of the NPIs.

Data availability

Supplementary materials for this paper are available from the publisher's webpage including additional information on data preprocessing, technical description of the model, and results of sensitivity analyses. The code to run the model is available in the following repository: https://github.com/RaphaelRe/COVID_NPIs_Germany.

Received: 20 June 2023; Accepted: 26 October 2023
Published online: 02 November 2023

References

1. Robert Koch Institute. Daily COVID-19 Cases Data. https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4/page/Bundesländer/ (2023). Accessed: 2023/10/27 08:51:30.
2. World Health Organization. Overview of Public Health and Social Measures in the Context of COVID-19: Interim Guidance. https://www.who.int/publications/i/item/overview-of-public-health-and-social-measures-in-the-context-of-covid-19 (2020). Accessed: 2023/10/27 08:51:30.
3. World Health Organization. Who Coronavirus (COVID-19) Dashboard: Measures, 2021. https://covid19.who.int/measures (2021). Accessed: 2023/10/27 08:51:30.
4. Martini, M., Gazzaniga, V., Bragazzi, N. L. & Barberis, I. The Spanish influenza pandemic: A lesson from history 100 years after 1918. *J. Prev. Med. Hyg.* **60**, E64 (2019).
5. Chakraborty, I. & Maity, P. COVID-19 outbreak: Migration, effects on society, global environment and prevention. *Sci. Total Environ.* **728**, 138882 (2020).
6. Pfefferbaum, B. & North, C. S. Mental health and the COVID-19 pandemic. *N. Engl. J. Med.* **383**, 510–512 (2020).
7. Wibbens, P. D., Koo, W.W.-Y. & McGahan, A. M. Which COVID policies are most effective? A Bayesian analysis of COVID-19 by jurisdiction. *PLoS ONE* **15**, e0244177 (2020).
8. Pozo-Martin, F. *et al.* The impact of non-pharmaceutical interventions on COVID-19 epidemic growth in the 37 OECD member states. *Eur. J. Epidemiol.* **36**, 629–640 (2021).
9. Flaxman, S. *et al.* Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).
10. Hirt, J., Janiaud, P. & Hemkens, L. G. Randomized trials on non-pharmaceutical interventions for COVID-19: A scoping review. *BMJ Evid. Based Med.* **27**, 334–344 (2022).
11. Alo, U. R., Nkwo, F. O., Nweke, H. F., Achi, I. I. & Okemiri, H. A. Non-pharmaceutical interventions against COVID-19 pandemic: Review of contact tracing and social distancing technologies, protocols, apps, security and open research directions. *Sensors* **22**, 280 (2022).
12. Oh, K.-B., Doherty, T. M., Vetter, V. & Bonanni, P. Lifting non-pharmaceutical interventions following the COVID-19 pandemic-the quiet before the storm?. *Expert Rev. Vaccines* **21**, 1541–1553 (2022).

13. Royal Society Expert Working Group. COVID-19: Examining the Effectiveness of Non-pharmaceutical Interventions (2023).
14. Robert Koch Institute. Stopptcovid-Studie: Wirksamkeit und Wirkung von Anti-epidemischen Maßnahmen auf die COVID-19-Pandemie in Deutschland. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/StopptCOVID_studie.html/ (2023). Accessed: 2023/10/27 08:51:30.
15. Rehms, R. *et al.* A Bayesian hierarchical approach to account for reporting uncertainty, variants of concern and vaccination coverage when estimating the effects of non-pharmaceutical interventions on the spread of infectious diseases. *medRxiv*. https://doi.org/10.1101/2022.06.20.22276652 (2022).
16. Healthcare-Datenplattform. Corona Data Platform Project in Germany. https://www.healthcare-datenplattform.de/pages/projekt (2022). Accessed: 2023/10/27 08:51:30.
17. Healthcare-Datenplattform. Development of a Corona Data Platform and (Regional) Analyzes of the SARS-COV-2 Epidemic in Germany. https://www.healthcare-datenplattform.de/dataset/?tags=corona-massnahmen (2022). Accessed: 2023/10/27 08:51:30.
18. Robert Koch Institute. Robert Koch-Institut Github. https://github.com/robert-koch-institut (2022). Accessed: 2023/10/27 08:51:30.
19. DIVI e.V. Daily ICU Occupancy Data for COVID-19 and Non-COVID-19 Patients. https://www.divi.de/register/tagesreport (2021). Accessed: 2023/10/27 08:51:30.
20. Fraser, C. *et al.* Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science* **324**, 1557–1561 (2009).
21. Wallinga, J. & Lipsitch, M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B Biol. Sci.* **274**, 599–604 (2007).
22. Brooks, S., Gelman, A., Jones, G. & Meng, X.-L. *Handbook of Markov Chain Monte Carlo* (CRC Press, 2011).
23. Roberts, G. O. & Rosenthal, J. S. Examples of adaptive MCMC. *J. Comput. Graph. Stat.* **18**, 349–367 (2009).
24. Talic, S. *et al.* Effectiveness of public health measures in reducing the incidence of COVID-19, SARS-COV-2 transmission, and COVID-19 mortality: Systematic review and meta-analysis. *bmj* **375**, e068302 (2021).
25. Oxford Stringency Index. Variation in Government Responses To COVID-19. https://www.bsg.ox.ac.uk/research/publications/variation-government-responses-covid-19 (2022). Accessed: 2023/10/27 08:51:30.
26. Coccia, M. Sources, diffusion and prediction in COVID-19 pandemic: Lessons learned to face next health emergency. *AIMS Public Health* **10**, 145 (2023).
27. Khairulbahri, M. The SEIR model incorporating asymptomatic cases, behavioral measures, and lockdowns: Lesson learned from the COVID-19 flow in Sweden. *Biomed. Signal Process. Control* **81**, 104416 (2023).
28. Safaie, N. *et al.* Investigation of factors affecting COVID-19 and sixth wave management using a system dynamics approach. *J. Healthc. Eng.* **2022**, 4079685 (2022).
29. Brauner, J. M. *et al.* Inferring the effectiveness of government interventions against COVID-19. *Science* **371**, eabd9338 (2021).
30. Ledebur, K. *et al.* Meteorological factors and non-pharmaceutical interventions explain local differences in the spread of SARS-COV-2 in Austria. *PLoS Comput. Biol.* **18**, e1009973 (2022).
31. Sharma, M. *et al.* Understanding the effectiveness of government interventions in Europe's second wave of COVID-19. *MedRxiv*. https://doi.org/10.1101/2021.03.25.21254330 (2021).
32. Fetzer, T. Subsidising the spread of COVID-19: Evidence from the UK's eat-out-to-help-out scheme. *Econ. J.* **132**, 1200–1217 (2022).
33. Glaeser, E. L., Jin, G. Z., Leyden, B. T. & Luca, M. Learning from deregulation: The asymmetric impact of lockdown and reopening on risky behavior during COVID-19. *J. Reg. Sci.* **61**, 696–709 (2021).
34. Apel, J., Rohde, N. & Marcus, J. The effect of a nighttime curfew on the spread of COVID-19. *Health Policy* **129**, 104712 (2023).
35. Ghasemi, A. *et al.* Impact of a nighttime curfew on overnight mobility. *medRxiv*. https://doi.org/10.1101/2021.04.04.21254906 (2021).
36. Sprengholz, P., Siegers, R., Goldhahn, L., Eitze, S. & Betsch, C. Good night: Experimental evidence that nighttime curfews may fuel disease dynamics by increasing contact density. *Soc. Sci. Med.* **286**, 114324 (2021).
37. Fritz, C. *et al.* Statistical modelling of COVID-19 data: Putting generalized additive models to work. *Stat. Model.*https://doi.org/10.1177/1471082X221124628 *(2022).*
38. Isphording, I. E., Lipfert, M. & Pestel, N. Does re-opening schools contribute to the spread of SARS-COV-2? Evidence from staggered summer breaks in Germany. *J. Public Econ.* **198**, 104426 (2021).
39. Vlachos, J., Hertegård, E. & Svaleryd, H. B. The effects of school closures on SARS-COV-2 among parents and teachers. *Proc. Natl. Acad. Sci.* **118**, e2020834118 (2021).
40. Thayer, W. M., Hasan, M. Z., Sankhla, P. & Gupta, S. An interrupted time series analysis of the lockdown policies in India: A national-level analysis of COVID-19 incidence. *Health Policy Plan.* **36**, 620–629 (2021).
41. Mendez-Brito, A., El Bcheraoui, C. & Pozo-Martin, F. Systematic review of empirical studies comparing the effectiveness of non-pharmaceutical interventions against COVID-19. *J. Infect.* **83**, 281–293 (2021).
42. Siedner, M. J. *et al.* Social distancing to slow the US COVID-19 epidemic: Longitudinal pretest-posttest comparison group study. *PLoS Med.* **17**, e1003244 (2020).
43. Holmdahl, I. & Buckee, C. Wrong but useful-what COVID-19 epidemiologic models can and cannot tell us. *N. Engl. J. Med.* **383**, 303–305 (2020).
44. Banholzer, N. *et al.* The estimated impact of non-pharmaceutical interventions on documented cases of COVID-19: A cross-country analysis. *MedRxiv*. https://doi.org/10.1101/2020.04.16.20062141 (2020).
45. Hunter, P. R., Colón-González, F. J., Brainard, J. & Rushton, S. Impact of non-pharmaceutical interventions against COVID-19 in Europe in 2020: A quasi-experimental non-equivalent group and time series design study. *Eurosurveillance* **26**, 2001401 (2021).
46. Bo, Y. *et al.* Effectiveness of non-pharmaceutical interventions on COVID-19 transmission in 190 countries from 23 January to 13 April 2020. *Int. J. Infect. Dis.* **102**, 247–253 (2021).

Acknowledgements
We would also like to thank Theresa Meier for contributing to the data processing.

Author contributions
Conceptualization, H.K., S.H. and R.R.; Methodology, S.H. and R.R.; Software, S.H., R.R. and Y.K.; Validation, S.H. and R.R.; Formal Analysis, R.R. and Y.K.; Resources, H.K. and S.H.; Data Curation, D.S., Y.K. and R.R.; Writing—Original Draft Preparation, Y.K. and R.R.; Writing—Review & Editing, Y.K., R.R., S.H. and H.K.; Visualization, D.S., Y.K. and R.R.; Supervision, S.H. and H.K.; Project Administration, Y.K. and R.R.; Funding Acquisition, H.K. and S.H. All authors have read and agreed to the published version of the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. The work has been partially supported by the Bavarian Health and Food Safety Authority (LGL) and the Volkswagen Stiftung (AZ: 99664). The authors of this work take full responsibility for its content.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-45950-2.

Correspondence and requests for materials should be addressed to Y.K.

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

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

Chapter 8

Contribution III

Article

Rehms, R., Ellenbach, N., Deffner, V., & Hoffmann, S. (2025). Addressing complex structures of measurement error arising in the exposure assessment in occupational epidemiology using a Bayesian hierarchical approach. *arXiv preprint* arXiv:2503.17161.

Code and Data

The code is available at https://github.com/RaphaelRe/Wismut_ME_Bayes. Scripts to generate simulated data are included making the simulation study fully reproducible. The used Wismut data for the application are not provided due to privacy protection. We provide the from the generated Markov chains for the simulation and application at https://doi.org/10.5281/zenodo.15050372.

Supplementary Material

Supplementary material can be found in the same provided document at arXiv:2503.17161 after the main manuscript pages 25-37.

Copyright information

Author Contributions

S.H. and R.R conceived and conceptualized the research. N.E. and V.D. collected and curated the data. N.E. wrote the code for data simulation. **R.R.** and S.H. developed and tested the code for inference. **R.R.** run the simulation study and application on the Wismut cohort data and produced the results. The manuscript is written by **R.R.** and S.H., with contributions from N.E. and V.D.
All authors discussed the results and were closely involved in proofreading and revising the manuscript. Corresponding author is **R.R.**

Addressing complex structures of measurement error arising in the exposure assessment in occupational epidemiology using a Bayesian hierarchical approach

Raphael Rehms[1,*] Nicole Ellenbach[1,2] Veronika Deffner[3]
Sabine Hoffmann[4]

March 24, 2025

Abstract

Exposure assessment in occupational epidemiology may involve multiple unknown quantities that are measured or reconstructed simultaneously for groups of workers and over several years. Additionally, exposures may be collected using different assessment strategies, depending on the period of exposure. As a consequence, researchers who are analyzing occupational cohort studies are commonly faced with challenging structures of exposure measurement error, involving complex dependence structures and multiple measurement error models, depending on the period of exposure. However, previous work has often made many simplifying assumptions concerning these errors. In this work, we propose a Bayesian hierarchical approach to account for a broad range of error structures arising in occupational epidemiology. The considered error structures may involve several unknown quantities that can be subject to mixtures of Berkson and classical measurement error. It is possible to account for different error structures, depending on the exposure period and the location of a worker. Moreover, errors can present complex dependence structures over time and between workers. We illustrate the proposed hierarchical approach on a subgroup of the German cohort of uranium miners to account for potential exposure uncertainties in the association between radon exposure and lung cancer mortality. The performance of the proposed approach and its sensitivity to

[1]Institute for Medical Information Processing, Biometry and Epidemiology, Faculty of Medicine, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML)
[3]Federal Office for Radiation Protection, Germany
[4] Department of Statistics, LMU Munich, Germany
[*]Corresponding author: rrehms@ibe.med.uni-muenchen.de

model misspecification are evaluated in a simulation study. The results show that biases in estimates arising from very complex measurement errors can be corrected through the proposed Bayesian hierarchical approach.

1 Introduction

In occupational epidemiology, researchers are often interested in the association between the cumulative exposure to a specific chemical or physical agent and the time until an event occurs, such as a diagnosis of or death from a certain disease. In this situation, exposure is time-dependent and ongoing, and the exposure history of workers may be collected using different assessment strategies depending on the period of exposure. During the estimation process, measurement errors can arise, leading to complex patterns of exposure uncertainty, where the structure and magnitude of measurement error can vary over time.
In many occupational cohort studies, there are no prospective exposure measurements. As it is often infeasible or too costly to measure or to estimate exposure values for individual workers, many occupational cohort studies rely on job-exposure matrices (JEMs). In general, JEMs provide information about exposure levels for certain job categories or titles [16]. The assigned exposure value may also vary with respect to location and year. Although it is widely acknowledged that exposure measurement error can have deleterious consequences on the validity of statistical inference and may lead to erroneous conclusions, researchers in the field of occupational epidemiology do not often account for these errors [10]. Many of the publications account for measurement error use Bayesian approaches [6, 49, 11, 5, 26, 34], while frequentist methods are also employed [16, 35, 34, 3]. It is often assumed that errors follow a simple structure where deviations of individual exposures for workers from the assigned exposure level in a JEM is described as unshared Berkson error [4, 7, 30, 51]. However, the estimation of the exposure values in a JEM often involves the estimation of multiple uncertain quantities that may be subject to complex structures of measurement error. Exposure values in a JEM are often reconstructed retrospectively by experts [44], leading to measurement errors that may affect several groups of workers and several exposure years at the same time. As a consequence, group-level estimates and individual job conditions may give rise to mixtures of Berkson and classical measurement error with complex dependency structures. While it is difficult to account for complex error structures using available measurement error correction methods, they pose serious threats to the validity of statistical inference in occupational epidemiology. Moreover, it is unlikely that the bias introduced by these complex measurement error structures can be adequately corrected for with methods that assume simple measurement error structures, such as unshared Berkson error. In previous work, we found that uncertainty components shared within workers cause more bias in risk estimation than components of unshared exposure uncertainty and that this can lead to an attenuation of the exposure-response curve for high exposure values [25], a phenomenon that is frequently observed in occupational cohort studies

[24, 50, 51].
The aim of this paper is to demonstrate that even highly complex measurement error structures, which may commonly arise in occupational cohort studies, can be effectively addressed. We illustrate this by proposing a flexible hierarchical Bayesian framework capable of accounting for these complexities, applied to the German cohort of uranium miners. Section 2 describes measurement error structures that typically arise through the prospective and retrospective exposure assessment in occupational epidemiology and the basic methodology to account for them using a Bayesian hierarchical approach. In section 3, we demonstrate the flexibility of the approach as a proof of concept using the German cohort of uranium miners as illustrative example. We account for highly complex structures of measurement errors that arises when studying the association between radon exposure and lung cancer mortality, presenting preliminary results for a selective subgroup of the Wismut cohort [28, 31]. Section 4 presents a simulation study in which we evaluate the performance of the proposed approach and assess its sensitivity to model misspecification. In section 5 we discuss our results and give an outlook for future work.

2 Accounting for measurement error in occupational epidemiology

2.1 Measurement error characteristics in occupational epidemiology

When describing measurement error, one commonly distinguishes classical and Berkson error. A classical measurement error model describes the error prone observed value Z as a function of the true value X and of an error term U that is independent of X. For additive and multiplicative error, classical measurement error can be written as $Z = X + U$ and $Z = X \cdot U$ with $X \perp\!\!\!\perp U$, respectively. Conversely, for Berkson error, the error term U is independent of the observed value Z. Again, we can write $X = Z + U$ and $X = Z \cdot U$ with $Z \perp\!\!\!\perp U$ for additive and multiplicative Berkson error, respectively. A classical measurement error model is often employed to describe the measurement arising from a measurement device or through the estimation of experts whereas a Berkson model describes the situation where the true and unknown value of a quantity of interest deviates from a fixed and observed value. In an occupational cohort study, it is in general plausible to assume that exposure measurement error is non-differential, i.e. that errors are independent of the outcome since it is unlikely that errors arising in the exposure estimation in an occupational cohort depend on the (future) disease status of individual workers.
If a JEM is used, the same exposure level is typically assigned to all workers in a job category (potentially also as a function of year and location). As a consequence, measurement errors that arise from the estimation of this common exposure level will affect all workers in that job category in the same way. In cases where exposure values in a JEM are based on measurements and there

is only one quantity that is to be measured (for instance pesticide levels, airborne contaminants or noise) the error arising through the estimation of the exposure level given in the JEM can be described through a classical measurement error component that is shared among workers in the same job category: $Z(t,j) = X(t,j) + U(t,j)$. Here, every worker who works at year t in job category j will receive the same error $U(t,j)$. In cases where exposure values are reconstructed by experts rather than being based on measurements, the classical error component described above will often be shared for the entire time period for which the estimation was made, leading to a classical measurement error that is shared among workers and years: $Z(p_t,j) = X(p_t,j) + U(p_t,j)$ for all job categories j and years t in the time period p_t.
Deviations of each individual true exposure from the common exposure level can then be described by unshared Berkson error. If the same value is assumed for different locations, the true average value at each location may deviate from this common value, leading to a Berkson error that is shared among all workers in a given location. For instance, an additive shared Berkson error at different locations o (e.g. mining objects) and years t could be written as $X(t,o) = Z(t,o) + U(t,o)$. Finally, even when exposure values are estimated in a prospective fashion, there may be additional uncertain quantities involved in the estimation that are retrospectively estimated by experts. We will describe more complex error structures involving several uncertain quantities in more detail in our application to the German cohort of uranium miners.

2.2 Accounting for measurement error

To account for measurement error using a Bayesian hierarchical approach, one specifies three sub-models and concatenates them to a full joint model using conditional independence assumptions [47]. This is similar to the procedure in a likelihood-based approach [12, 17]. However, Bayesian approaches may be more versatile from a computational perspective when it comes to the correction for complex errors structures. In the following, we will give a short summary of the general Bayesian approach. We start by defining the three required sub-models: The disease model, measurement model and the exposure model.

Disease model: The disease model defines the relation $[Y|X,\theta_1]$ between an outcome Y and one or more covariates of interest X (at least exposure values), where θ_1 is the collection of all parameters of the disease model. We follow Richardson and Gilks [47], to denote (possibly conditional) distributions using squared brackets. In occupational cohort studies, a commonly used outcome is the time until an event occurs, for instance the time until a specific diagnosis or cause-specific death where the covariate of interest would be an exposure to a specific chemical or physical agent.

Measurement model: The measurement model $[Z|X,\theta_2]$ describes the relation between the observed, error prone variable Z and the true, unobserved values X for all uncertain quantities intervening in the calculation of individual exposure values, parameterized by θ_2. With respect to measurement errors that

arise through application of a JEM, a combination of a classical and Berkson error can be used to describe the following situation: In general, the estimation of the exposure for one group or job category is not measured precisely. Therefore, a classical error can be assumed to reflect uncertainty in the estimation of this common exposure level. Moreover, an additional Berkson error can describe deviations of exposure values of individual workers from this common exposure level. This leads to a combination of two different errors. Since multiple uncertain quantities typically intervene in the estimation of an exposure level in a JEM, we may often be faced with combinations of Berkson and classical measurement error in multiple uncertain quantities. We illustrate this situation in more detail in the next section for the Wismut cohort.

Exposure model: The exposure model defines the distribution $[X|\theta_3]$ of the unobserved exposure X. Note, that in the case of a Berkson error, a formulation of an exposure model is not required. θ_3 parameterized the distribution of X. If we assume, for instance, that exposure values follow a normal distribution, the parameters would be $\theta_3 = (\mu_X, \sigma_X^2)$, i.e the assumed mean value and standard deviation of the distribution of X.

We can combined the three models using conditional independence assumptions to formulate the unnormalized joint posterior over all unknown quantities by assuming additional prior distributions on the parameters of the models $[\theta_1]$, $[\theta_2]$ and $[\theta_3]$:

$$[\theta_1, \theta_2, \theta_3, X|Y, Z] \propto [\theta_1][\theta_2][\theta_3][Y|X, \theta_1][Z|X, \theta_2][X|\theta_3]. \tag{1}$$

We can use any suitable inference method to obtain the posterior. Most prominent choices are Markov chain Monte Carlo (MCMC) [9], variational inference [8] or integrated nested Laplace approximation [43, 48]. MCMC can be considered as the most versatile approach and it can approximate the posterior with arbitrary accuracy (at least in theory). This usually comes with a substantially higher computational burden. MCMC algorithms generate a Markov chain that has the posterior of interest as stationary distribution. After initializing the chain at an arbitrary state, the chain will converge to the stationary distribution (samples before convergence are typically discarded as burnin). Samples from this Markov chain can then be considered as samples from the posterior. It is common to have a calibration phase at the beginning to sample more efficiently and to thin the chain to obtain less correlated samples. Furthermore, it is good practice to run multiple chains to parallelize computation and to calculate common quality criteria like the $\hat{R}$ statistic [15, 55].

3 Application: Modeling the association between radon exposure and lung cancer mortality in the Wismut cohort

In the following, we will use the German cohort of uranium miners, also referred to as Wismut cohort [28], as an example to illustrate how complex structures of potential exposure uncertainties that may arise in occupational cohort studies can be accounted for through a Bayesian hierarchical approach. The cohort consists of 58,974 workers who were employed between 1946 and 1989 at the Wismut company. We are interested in the association between the exposure of radon gas (or rather the decay products) and lung cancer mortality. It is generally acknowledged that the exposure to radon progeny is a relevant cause of lung cancer [53]. For all exposure years in the cohort, individual exposure estimates were based on a JEM [33, 32], which provides estimated annual exposure values to radon progeny for a reference activity with 2000 working hours. These values were then multiplied by a so-called activity weighting factor that can be summarized as a correction factor for the different radiation exposures associated with different activities. Further, the estimated annual exposure is multiplied by a working time factor to adjust for deviations in the number of standard working hours from a reference [31]. The individual annual exposure is afterwards calculated by combining these quantities (see section 3.2 that explains this in more detail as part of the measurement model).

3.1 Disease model

In the Wismut cohort, we are interested in the association between radon exposure and lung cancer mortality. We use a survival outcome, where lung cancer mortality is a right-censored variable (Y_i, δ_i) for each miner $i \in \{1, \dots n\}$, where Y_i denotes the attained age in years and δ_i is the censoring indicator. Attained age is left truncated (at the time of entry into the cohort) and radon exposure, denoted as $X_i^{\text{cum}}(t)$, is a time-varying covariate that accumulates over years t. Two popular model choices in radiation epidemiology are the proportional hazards (PH) and the excess hazard ratio (EHR) model. The instantaneous hazard for miner i in a PH model is defined as

$$h_i(t; \boldsymbol{\lambda}, \beta) = h_0(t, \boldsymbol{\lambda}) \exp(\beta \cdot X_i^{\text{cum}}(t)),$$

where h_0 is the baseline hazard and β the effect of the exposure on the instantaneous hazard. The baseline hazard does not depend on the covariates, but may depend on time and a set of parameters $\boldsymbol{\lambda}$. In the case of the EHR model, th e instantaneous hazard $h_i(t; \boldsymbol{\lambda}, \beta)$ is modeled as

$$h_i(t; \boldsymbol{\lambda}, \beta) = h_0(t, \boldsymbol{\lambda})(1 + \beta \cdot X_i^{\text{cum}}(t)).$$

The EHR model implies a constraint on β as the hazard must be positive. In both cases, we assume a simplified linear model without effect-modifying variables and we model the baseline hazard assuming an *explicitly modeled functional form* as the correction for measurement error through a hierarchical model requires the formulation of the full likelihood. We choose a flexible piecewise-constant function as model for the hazard baseline, parameterized through $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_4\}$, i.e. $h_0(t, \boldsymbol{\lambda}) = \lambda_k \quad \forall t \in I_k = (s_{k-1}, s_k]$, where I_k is the time interval corresponding to the baseline hazard of λ_j that is defined through the partitions $0 = s_0 < s_1 < s_2 < s_3 < s_4$ [27, 14, 39, 37]. Following Hoffmann et al. [26], we use $s_1 = 40, s_2 = 55, s_3 = 75, s_4 = 104$ as break points and define the priors on the parameters as

$$\beta \sim N(0, \sigma_\beta^2)$$

and

$$\lambda_k \sim Ga(\alpha_k^\lambda, \beta_k^\lambda) \qquad \forall k \in \{1, \ldots 4\},$$

where each λ_k has an individual prior specification through shape and scale parameters α_k^λ and β_k^λ to reflect a stepwise increase in the baseline hazard. The prior on β is chosen to be uninformative while the parameters of the baseline hazard are chosen to be informative (see first part of section C in the supplementary for a list of all chosen prior parameters of the disease model).

3.2 Measurement model for the Wismut cohort

In the Wismut cohort, the exposure values in the JEM were estimated through different methods depending on the time period and the type of workplace (underground, open pit, milling or surface). A mining location is also referred to as an object in the Wismut cohort. Based on the preliminary work of Küchenhoff et al. [31], Ellenbach et al. [13] characterize, quantify and develop measurement models to describe the characteristics of exposure uncertainties arising in the exposure assessment for all time periods and types of workplaces (see Figure 4 in section A for an overview). We consider five of these different measurement models for the Bayesian approach: M1, M2, M2_Expert and M3 for underground miners (depending on the time period and the availability of measurements for radon gas and radon progeny), and M4 for miners employed at surface areas affiliated to mining locations. Besides, we assume no measurement error for miners working in pure surface objects without exposure to radon. We do not consider the error structures arising in processing companies and in open pit mining objects and exclude miners who ever worked in either of these two types of mining locations, leading to a sub sample of 48,534 miners. Furthermore, we exclude all miners whose working histories include measurement model MX_Expert_WLM, which was defined by Ellenbach et al.[13] for cases in which information to reconstruct the exposure values according to the other measurement models was lacking, finally leading to a selective sub sample of 34,809 miners. The cumulated exposure of a miner is often derived considering different measurement models, since most miners worked in more than just one time period or changed

the location over time. For the sake of readability, in this section, we describe the measurement model M2 as it represents a typical error structure that may arise through the use of a JEM. We refer to the supplementary material A for the full specification of all measurement models.
Measurement model M2 was employed for workers in underground mining objects located in the federal states Saxony and Thuringia, Germany and development objects in Saxony in the exposure assessment period 1955/56 to 1965 in Saxony and 1955/56-1974 in Thuringia. In this exposure assessment period, exposure values were calculated using the following formula:

$$E(t,o,j) = 12 \cdot C_{Rn}(p_{t,o}) \cdot \tau(t,o) \cdot f(p_{o,j}) \cdot w\left(p_t\right) \cdot g\left(p_{t,o}\right), \tag{2}$$

where $E(t,o,j)$ denotes the estimated annual exposure to radon for a worker who conducted activity j in location o and year t. $12 \cdot C_{Rn}(p_{t,o}) \cdot \tau(t,o)$ is the estimated annual radon gas concentration for the reference activity (being a hewer) and 2000 annual working hours. To obtain the annual exposure to radon progeny, the exposure is multiplied by an activity weighting factor $f(p_{o,j})$, which corrects for the fact that most activities had lower exposure to radon than a hewer, as well as by a working time factor $w(p_t)$, which modifies the reference working time of 2000 hours to obtain a smaller or higher amount of working hours. By multiplying with an equilibrium factor $g\left(p_{t,o}\right)$, the measured radon gas exposure is converted to radon progeny exposure in working level months (WLM), which is the historical unit of radon exposure in cohorts of uranium miners and related to the potential alpha energy concentration [38]. We call these different quantities *uncertain factors* as they are considered to be potentially error-prone. With a slight abuse of notation, we define variables p_t, $p_{t,o}$ and $p_{o,j}$ to express the dependence structures arising from the fact that many of the uncertain factors were not estimated for individual years t and locations o, but instead a common value was used for several years, locations and activities (e.g. $p_{t,o}$ uses a common value for several years t and locations o). Although there were ambient radon gas concentration measurements $C_{Rn}(p_{t,o})$ available for most years and locations in measurement model M2, there are some years and locations for which there were no measurements. For these cases, measurements from different years or locations were extrapolated and sometimes adjusted with a transfer factor $\tau(t,o)$. In general, the use of a common value or extrapolated measurements for several years, locations, and activities leads to a shared classical measurement error for these years, locations, and activities.
For the radon gas concentrations, we assume a classical error component that describes potential uncertainty in the measurement process. Since the average radon gas concentration is the result of a large number of measurements, we assume that the average of the measurements is distributed normally around its true value, following the central limit theorem. Additionally, we assume a Berkson error component only for those years and location without ambient radon gas measurements but used the transferred values from other years or locations. For the activity weighting, the working time and the equilibrium factors, we assume both a classical and a Berkson measurement error component to describe the potential uncertainty in the estimation of a common value and

the variability around this common value for several years, objects and activities, respectively:

$$C_{Rn}(p_{t,o}) = \mathcal{C}_{Rn}(p_{t,o}) + U_{\mathcal{C},c}(p_{t,o})$$
$$C'_{Rn}(t,o) = \mathcal{C}_{Rn}(p_{t,o}) \cdot U_{\mathcal{C}'_{Rn},B}(t,o) \cdot \tau(t,o) \quad \text{(only if values were transferred)}$$

$$f(p_{o,j}) = \varphi(p_{o,j}) \cdot U_{\varphi,c}(p_{o,j})$$
$$\varphi'(t,o,j) = \varphi(p_{o,j}) \cdot U_{\varphi',B}(t,o,j)$$

$$w(p_t) = \omega(p_t) \cdot U_{\omega,c}(p_t)$$
$$\omega'(t,o) = \omega(p_t) \cdot U_{\omega',B}(t,o)$$

$$g(p_{t,o}) = \gamma(p_{t,o}) \cdot U_{\gamma,c}(p_{t,o})$$
$$\gamma'(t,o) = \gamma(p_{t,o}) \cdot U_{\gamma',B}(t,o)$$

where $C'_{Rn}(t,o)$, $\varphi'(t,o,j)$, $\omega'(t,o)$, and $\gamma'(t,o)$ are the true values of the radon gas concentration, the activity weighting factor, the working time factor, and the equilibrium factor respectively. $C_{Rn}(p_{t,o})$, $\varphi(p_{o,j})$, $\omega(p_t)$ and $\gamma(p_{t,o})$ are the true average (level) values for each of them and $C_{Rn}(p_{t,o})$, $f(p_{o,j})$, $w(p_t)$ and $g(p_{t,o})$ are the level values that were estimated by experts. For additive errors, we assume an error term that follows a normal distribution, i.e. $U_{\mathcal{C},c}(p_{t,o}) \sim N(0, \sigma^2_{\mathcal{C},c}(p_{t,o}))$, for the radon gas concentration while we assume a log-normal distributed error for multiplicative errors. That is $\log(U_{\mathcal{C}'_{Rn},B}(t,o)) \sim N\left(-\frac{1}{2}\sigma^2_{\mathcal{C}'_{Rn},B}(t,o), \sigma^2_{\mathcal{C}'_{Rn},B}(t,o)\right)$ for the Berkson error of the radon gas concentration. We assume that the error distributions for the classical and the Berkson error components of the other uncertain factors also follow a log-normal distribution analogous to the Berkson error of the radon gas concentration.
The *true* exposure of a miner i who is employed in activity j in object o and year t is then given by:

$$X_i(t,o,j) = 12 \cdot C'_{Rn}(t,o) \cdot \varphi'(t,o,p_j) \cdot \omega'(t,o) \cdot \gamma'(t,o) \cdot l_i(t,o,j) \qquad (3)$$

where the factor $l_i(t,o,j)$ accounts for the individual time that a miner worked in object o and activity j in year t. Equation (3) resembles equation (2). However, it is important to note that now the different factors are *latent variables* that are estimated in addition to the parameters of interest.
Given the individual exposure X_i, a cumulated individual exposure X_i^{cum} must be calculated as exposure accumulates over time. For details and an example, see section B.2 of the supplementary material. We show the full measurement error model M2 in Figure 1 as a directed acyclic graph (DAG). See Ellenbach et al. [13] for the DAGs of the other measurement models.

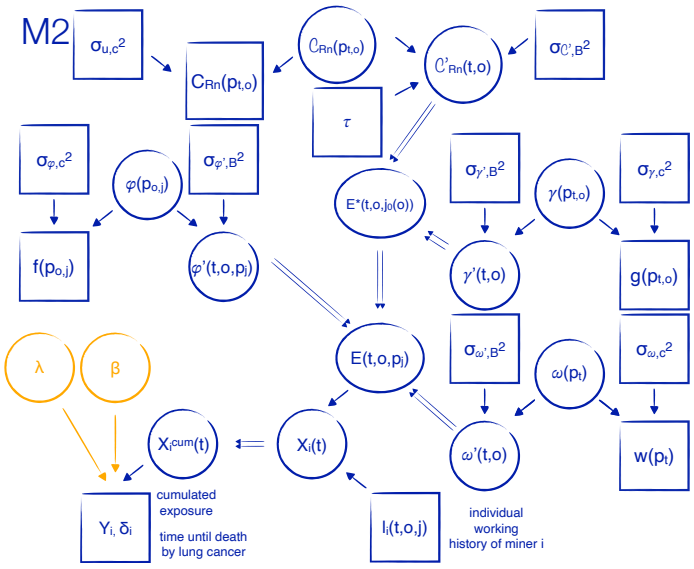


Figure 1: DAG of measurement model M2. With circles we represent unknown quantities and with boxes observed/fixed quantities. Single arrows indicate probabilistic dependencies between the different quantities and double arrows deterministic dependencies. The DAG does not include the prior parameters and parameters of the exposure model.

3.3 Exposure model for the Wismut cohort

In the Wismut cohort, the uncertain factors vary depending on the measurement model. We describe the exposure models for measurement model M2 and refer to section C of the supplementary material for the exposure models of the other measurement models. For M2, we do not only need to define exposure models for the true and unknown radon gas concentration $C_{Rn}(p_{t,o})$, but also for the other uncertain factors, namely the activity weighting factor $\varphi'(t,o,j)$, the working time factor $\omega'(t,o)$ and the equilibrium factor $\gamma'(t,o)$. The radon gas concentration is assumed to follow a truncated Gaussian distribution with varying priors over time on the mean and standard deviation to reflect varying concentration levels over the years, i.e.

$$
\begin{aligned}
C_{Rn}(p_{t,o}) &\sim N^{+}(\mu_{C_{Rn}}(t), \sigma^2_{C_{Rn}}(t)) \\
\mu_{C_{Rn}}(t) &\sim N(\mu_{C_{Rn}}, \sigma^2_{C_{Rn}}) \\
\sigma_{C_{Rn}}(t) &\sim N^{+}(\mu_{C_{Rn}}, \sigma^2_{C_{Rn}}).
\end{aligned}
$$

The other uncertain factors $\varphi(p_{o,j}), \omega\,(p_t)$, and $\gamma\,(p_{t,o})$ are assumed to follow a generalized Beta distribution with modified support to an appropriate range that represents realistic values:

$$\begin{aligned}\varphi(p_{o,j}) &\sim B_{[0,1.3]}(a_\varphi, b_\varphi),\\ \omega\,(p_t) &\sim B_{[0.6,1.5]}(a_\omega, b_\omega),\\ \gamma\,(p_{t,o}) &\sim B_{[0.05,0.8]}(a_\gamma, b_\gamma),\end{aligned}$$

where we use $B_{[lo,up]}(a,b)$ to denote a Beta distribution with parameters a and b and write the modified support in subscript with $[lo, up]$ and we allow for additional flexibility by setting the following priors on the shape parameters of the Beta distributions (a and b):

$$a_m, b_m \sim N(\mu_{B_m}, \sigma^2_{B_m}), \quad \text{for} \quad m \in \{\varphi, \omega, \gamma\}.$$

3.4 Inference using an efficient MCMC algorithm

After defining the disease, measurement and exposure model, we can use them to write the unnormalized joint posterior for all workers and years where the exposure assessment is based on measurement model M2. To get a compact form, we use again squared brackets to denote the probability density function (PDF) of a random variable.

$$\begin{aligned}&[\boldsymbol{\theta}, X^{cum} \mid \cdot] \propto\\ &[\beta][\boldsymbol{\lambda}]\,[a_\omega]\,[b_\omega]\,[a_\gamma]\,[b_\gamma]\,[a_\varphi]\,[b_\varphi] \prod_t [\mu_{\mathcal{C}_{Rn}}(t)] \prod_t [\sigma_{\mathcal{C}_{Rn}}(t)] \times\\ &\prod_{i,t} [Y_i \mid \boldsymbol{\lambda}, \beta, X_i^{\text{cum}}\,(t)] \times\\ &\prod_{i,t} [X_i^{\text{cum}}\,(t) \mid X_i(t)] \times\\ &\prod_{i,t} \big[X_i(t) \mid \mathcal{C}'_{R_n}(t,o), \varphi'(t,o,j), \gamma'(t,o), \omega'(t,o), l_i(t,o,j)\big] \times\\ &\prod_{t,o} \big[\omega'(t,o) \mid \sigma^2_{\omega',B}, \omega\,(p_t)\big] \prod_{p_t} \big[w\,(p_t) \mid \sigma^2_{\omega,c}, \omega\,(p_t)\big] \prod_{p_t} [\omega\,(p_t) \mid a_\omega, a_\omega] \times\\ &\prod_{t,o} \big[\gamma'(t,o) \mid \sigma^2_{\gamma',B}, \gamma\,(p_{t,o})\big] \prod_{p_{t,o}} \big[g\,(p_{t,o}) \mid \sigma^2_{\gamma,c}, \gamma\,(p_{t,o})\big] \prod_{p_{t,o}} [\gamma\,(p_{t,o}) \mid a_\gamma, a_\gamma] \times\\ &\prod_{t,o,j} \big[\varphi'(t,o,j) \mid \sigma^2_{\varphi',B}, \varphi(p_{o,j})\big] \prod_{p_{o,j}} \big[f(p_{o,j}) \mid \sigma^2_{\varphi,c}, \varphi(p_{o,j})\big] \prod_{p_{o,j}} [\varphi(p_{o,j}) \mid a_\varphi, a_\varphi] \times\\ &\prod_{t,o} \big[\mathcal{C}'_{R_n}(t,o) \mid \sigma^2_{\mathcal{C},c}, \mathcal{C}_{R_n}(p_{t,o})\big] \prod_{p_{t,o}} \big[\mathcal{C}_{R_n}(p_{t,o}) \mid \mu_{\mathcal{C}_{Rn}}(t), \sigma^2_{\mathcal{C}_{Rn}}(t)\big]\end{aligned}$$

where $\boldsymbol{\theta}$ denotes the collection of all latent quantities, i.e. $\boldsymbol{\theta} = (\beta,\ \boldsymbol{\lambda},\ a_\omega,\ b_\omega,\ a_\gamma,\ b_\gamma,\ a_\varphi,\ b_\varphi,\ \mu_{\mathcal{C}}(t),\ \sigma_{\mathcal{C}}(t),\ \mathcal{C}'_{R_n}(t,o),\ \varphi'(t,o,j),\ \gamma'(t,o),\ \omega'(t,o),\ \mathcal{C}_{R_n}(p_{t,o}),$

$\varphi(p_{o,j})$, $\gamma\,(p_{t,o})$, $\omega\,(p_t)$). Note that the fourth and fifth line do not represent a probabilistic, but a *deterministic* relationship that expresses the connection between the individual cumulated exposure of a worker and the value of the uncertain quantities intervening in the calculation of the yearly exposure values.
We implement a custom Metropolis-Hastings (MH) algorithm with component-wise updates [40, 9] to sample from the posterior. We briefly describe the key-points of the MCMC algorithm. More details are presented in section B in the supplementary material. All uncertain quantities are treated as latent variables that are updated at each sampling step of the MCMC algorithm one at a time, starting with the parameters of the disease model. For the parameters of the disease model, it is possible to condition on the cumulative latent exposure X_i^{cum} leading to a simplified MH-ratio where only the disease model has to be evaluated. The update of the latent exposure poses more challenges: After proposing a new value for one of the uncertain factors, formula (3) can be used to calculate the latent exposure and to evaluate the disease model. However, this requires two extra steps:
1) Because the quantities are affected by shared classical and Berkson measurement error within and between workers, caution is necessary as each error is shared across different domains defined through the dependency structures specified in the measurement model. For instance, the classical error part of the working time factor ω, varies only *over periods* p_t (i.e. it is shared for multiple years), while the Berkson error is defined over *every year* t *and object* o. As a consequence, the uncertain factor has to be *mapped* to its corresponding domain before it can be used for the calculation of the true and unknown exposure values.
2) Given a new proposed state, the calculation in (3) returns only annual exposure values. However, as the exposure of a worker accumulates over time, it is necessary to calculate the cumulative exposure vector for each individual worker over all the working years every time a yearly exposure value is proposed.
We used sparse matrix multiplication to solve both challenges. This is computationally efficient because only non-zero values are stored and used for the mapping and cumulation.
We implement the MCMC update scheme in `python3` [54] in an object-oriented fashion using mainly the standard numerical library `numpy` [18]. Statistical distributions and sparse matrix functionalities rely on `scipy` [56].

3.5 Results

In order to obtain 4000 independent samples from the posterior distribution, we generate samples from eight independent chains with 100,000 iterations each and and thin them by keeping only every 200th sample). Beforehand, we tune the chains using 100 adaptive phases with 50 samples each to obtain better sampling quality and run further 50,000 iterations as burnin. Figure 2 and Table 1 show the results for a proportional hazards and an EHR model. We present as point estimates the empirical mean and median and the 95%-highest density interval (credible interval) as measure of uncertainty. The HDI represents the 95% of

the most credible values [29].

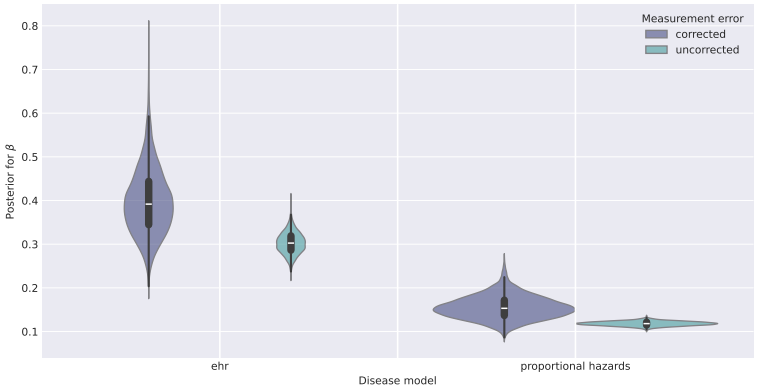


Figure 2: Violin plots of the samples from the posterior distribution for an EHR (left) and proportional hazards model (right). Results that account for measurement error are plotted in blue and the results from the naive models without measurement error correction are plotted in green.

Disease model	Correction	Mean	Median	HDI (95%)
Proportional hazards	Corrected	0.1553	0.1532	[0.1066, 0.2032]
	Uncorrected	0.1181	0.1181	[0.1088, 0.1275]
EHR	Corrected	0.3977	0.3917	[0.2614, 0.5352]
	Uncorrected	0.3028	0.3025	[0.2567, 0.3471]

Table 1: Summary statistics of the parameter β (association between radon exposure and lung cancer mortality) using a proportional hazards or an excess hazards (EHR) model for the application on data of a selective subgroup of the German uranium miners cohort. We present the empirical mean, median and HDI calculated from the sampled posterior.

It is observable that without accounting for the assumed measurement error structure, the estimated association between radon exposure and lung cancer mortality in the selective subgroup of the cohort is underestimated by about 23.95% for the proportional hazards model and 23.86% for the EHR model (when considering the empirical mean of the posterior distribution as point estimate). This indicates that in this example, accounting for measurement error results in an increase in the point estimate of the risk. However, it also considerably increases the uncertainty of the risk estimate. We provide the

convergence analysis in section D of the supplementary material.

4 Simulation study

We conduct a simulation study to assess whether the proposed Bayesian hierarchical approach can produce reliable results when accounting for complex structures of measurement error that may typically arise in occupational cohorts. In the following, we will follow the structure proposed in Morris et al. [42].

Aims

1) We want to ensure that an unbiased estimate for the parameter of interest can be inferred; 2) We aim to test the frequentist properties of the proposed approach by verifying if the 95% credible interval from the posterior samples adequately covers the parameter of interest under the correct model assumptions; 3) We want to assess the sensitivity to model misspecification by estimating the extent to which an incorrect specification of the distributions of the measurement model can influence the results.

Data-generating mechanisms

We generate 100 data sets per scenario, where each data set simulates the working history and survival times of 5,000 miners. The simulation study uses *not* only measurement model M2, but generates data over *all* measurement models that were used for the application to the real data of the Wismut cohort, only with the small simplification of no reference objects and thus no Berkson error component for radon measurements. To generate data sets that follow the assumed probabilistic models of the different measurement models (see section A of the supplementary materials), we first randomly draw 5000 miners from a simplified cohort data set (due to data protection reasons) using only the information on whether a miner worked at the Wismut company and whether he was exposed to radon in the respective year. All miners are randomly sampled into different objects and different activities. Secondly, we sample the true average values of all uncertain factors, as well as all their classical and Berkson errors from the respective distributions. We take the dependency structures into account and generate shared errors accordingly. For example, for the working time factor, we sample as many true mean values $\omega(p_t)$ from a Beta distribution and as many multiplicative classical errors $U_{\omega,c}(p_t)$ from a log-normal distribution, as there are different values for p_t. For the multiplicative Berkson errors $U_{\omega',B}(t,o)$, we sample a separate value for each year t and each object o from a log-normal distribution. We then obtain the true values of the uncertain factors by multiplying the sampled true average values with the sampled Berkson errors (e.g., $\omega'(t,o) = \omega(p_t) \cdot U_{\omega',B}(t,o)$). By multiplying (or adding in the case of an additive measurement error) the sampled classical errors with the

sampled true average values, we obtain the observed values of the uncertain factors (e.g., $w(p_t) = \omega(p_t) \cdot U_{\omega,c}(p_t)$). Then, the miners' true exposures and their error-prone observed exposures are calculated using, respectively, the true or observed values of the uncertain factors according to the formula for the respective measurement model (see section A of the supplementary materials). The Bayesian hierarchical approach uses the observed exposures for measurement error correction and for the uncorrected naive estimate. The true exposures, on the other hand, are used to generate the survival times. In particular, we generate the censored time until death by lung cancer according to a PH model as a function of a miner's radon exposure in WLM as time-varying covariate using a method that relies on the generation of truncated piecewise exponential random variables, initially proposed in Zhou [57] and further extended by Hendry [23] and Montez-Rath et al. [41]. For the exact implementation of the simulation code in `R` [45], we refer to the accompanying git repository.

Estimand

Our estimand is the parameter of interest β representing the association between radon exposure and lung cancer mortality. In particular, we consider the samples drawn for β as estimates of the posterior distribution.

Methods

We define three different scenarios for the simulation study. The first scenario (S1) simulates data assuming $\beta = 0.3$, whereas the second scenario (S2) uses $\beta = 0.6$. For both scenarios we apply the proposed model with measurement error (ME) correction, as well as a naive one without ME correction to the respective simulated data. Both scenarios should test the correctness of the approach covering aims 1) and 2). The third scenario (S3) is designed for aim 3): We test the robustness of the model against wrong distributional assumptions by investigating the impact of assuming a log-normal distribution for the radon concentration measurements for those measurement models where the data is simulated using a (truncated) normal distribution and vice versa assuming a (truncated) normal distribution for models where the data is simulated using a log-normal distribution. Furthermore, we want to investigate whether a misspecification of the distributional assumptions for the exposure models specified for the uncertain factors other than radon concentration (e.g. working time) impacts model performance in a significant way: Instead of flexible Beta distributions (with additional priors on a and b), we force the model to use a fixed uniform distribution implying $a = b = 1$ for the latent factors while using the standard data generating process. All scenarios are fitted using solely a proportional hazards model for the exposure-disease relationship and no EHR to keep the computational cost feasible. For every scenario, we generated 100 data sets. For scenarios S1 and S2 we used the Bayesian approach exactly as described in section 3, and for scenario S3 we only modified the exposure models to account for the wrong distributional assumptions. For scenario S1 and S2, we also run

a model on the true, unknown values of the uncertain factors, as they were measured without any error. For this we solely use formula (3) to calculate the exposure and use only the disease model. It can therefore be seen as a reference where one would expect very accurate estimates. Due to convergence problems for some data set in the second scenario (S2), only 99 or 97 data sets are used for S2 (see section D in the supplementary materials).

Performance measures

The main performance measure is the bias between the point estimate (empirical mean) of the posterior distribution and the true value of β. We quantify it by calculating the absolute and the relative bias. Furthermore, we estimate the mean squared error (MSE). We calculate these quantities using the empirical mean of the posterior distribution as point estimate $\hat{\beta}$. Our secondary performance measure is the proportion of the coverage of the true β value (that was used to generate the data) in the interval estimate of β, estimated by the empirical 95%-HDI.

Results

Table 2 shows the results for the different scenarios with their Monte Carlo standard error in parentheses, that quantifies the simulation uncertainty due to using a finite number of simulations [42]. The results are in line with the results of the application on the data of the Wismut cohort: ignoring the measurement error may lead to some bias. Through measurement error correction, this bias can be eliminated for both scenarios S1 and S2 (aim 1) achieving a bias level that is nearly as good as fitting a model directly to the true values without measurement error. However, when accounting for measurement errors, the 95%-HDI for the risk estimate becomes wider, even leading to overcoverage in scenarios with $beta = 0.3$. Moreover, the results for S3 imply that a potential misspecification with respect to the exposure distribution on the radon gas measurements or on other uncertain factors has only a negligible impact on the estimates (aim 3). Further, we show the results of the posterior estimates for the first 20 simulated data sets graphically over all considered scenarios in Figure 3 (naive estimates and measurement error correction). Looking at the mean and 95%-HDI, one can see that the measurement error correction provides good results while having slightly wider intervals caused by the higher uncertainty induced through the error (aim 3). In section E of the supplementary material, we analyze and discuss the convergence of the presented results.

5 Discussion

In this work, we proposed a Bayesian hierarchical approach to account for complex structures of measurement error in occupational cohort studies. These error structures can involve both multiple potentially uncertain quantities that may be subject to complex mixtures of Berkson and classical measurement error

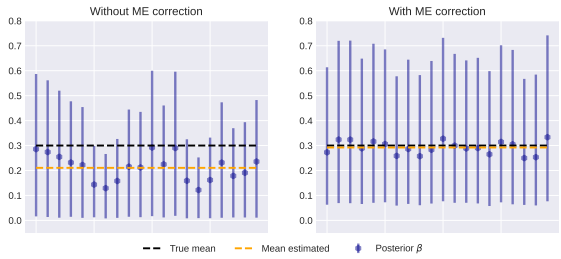
	S 1 ($\beta = 0.3$)			S 2 ($\beta = 0.6$)			S 3 ($\beta = 0.3$)	
	naive	ME cor.	true	naive	ME cor.	true	misspec1	misspec2
Absolute bias	-0.08	0.004	-0.002	-0.207	0.022	-0.003	-0.002	-0.004
	(0.005)	(0.003)	(0.001)	(0.01)	(0.006)	(0.001)	(0.003)	(0.003)
Relative bias	-0.266	0.014	-0.007	-0.344	0.037	-0.005	-0.007	-0.013
	(0.018)	(0.011)	(0.002)	(0.017)	(0.009)	(0.001)	(0.01)	(0.01)
MSE	0.0092	0.001	0.0001	0.0529	0.0036	0.0001	0.0009	0.0009
	(0.001)	(0.0001)	(0.0)	(0.0043)	(0.0006)	(0.0)	(0.0001)	(0.0001)
Coverage	0.12	0.99	0.95	0.031	0.959	0.959	0.98	1.0
	(0.325)	(0.099)	(0.218)	(0.172)	(0.198)	(0.199)	(0.14)	(0.0)

Table 2: Absolute/relative bias, mean squared error (MSE) and the coverage using 95%-HDIs over 100 data sets (99 or 97 for S2, see section E in the supplementary materials) for the different scenarios, i.e. S1 with $\beta = 0.3$, S2 with $\beta = 0.6$ and S3 with two misspecified models using also $\beta = 0.3$. 'misspec1' assumes a log-normal distribution for the radon concentration measurements while data is simulated with a (truncated) normal or vice versa depending on the measurement model. 'misspec2' assumes a uniform distribution in the exposure model on other multiplicative factors. The column with 'true' calculates the model without measurement error correction on the true unobserved values and is therefore a reference model.

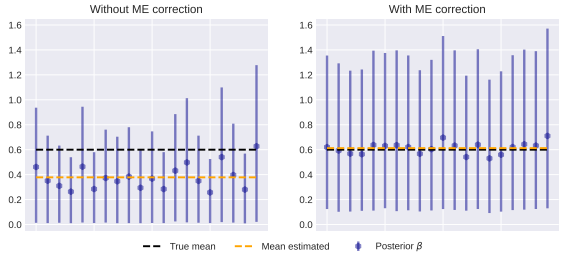
and multiple measurement error models to tailor the assumed measurement error structures to the exposure assessment strategies that were used for different workers and at different exposure periods. We illustrated the approach on data of the Wismut cohort where and showed on simulated data that the proposed approach is able to produce reliable results under the assumed data generating processes.

However, a number of limitations have to be considered in the interpretation of our results. Like any statistical method, the proposed approach to account for measurement error may stand and fall with its implicitly and explicitly stated assumptions. In the simulation study, we investigated how assuming an additive error when the error is actually multiplicative and vice versa would affect our results. We tested the robustness to this misspecification of the measurement model, as a broad body of literature suggests a multiplicative error [36, 52, 19, 20, 22, 21, 1, 2, 3] while we chose an additive error for measurements of radon gas concentration and radon progeny whenever multiple measurements were averaged.

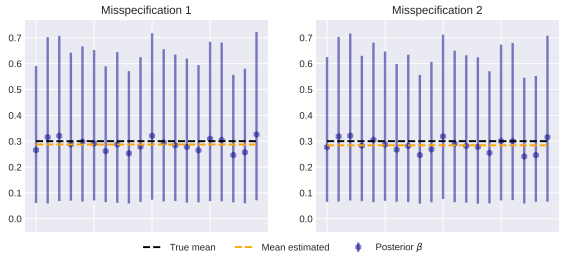
We only considered a simplified model for the association between one (time-varying) exposure and an outcome, but ignored effect modifying variables that are known to be important in the association between radon exposure and lung cancer mortality. This requires more future work and was not the focus of this paper. Hence, the presented results should be interpreted as a proof of concept and illustration and rather not as an answer to the question what the actual effect of the variable of interest is. However, the used Bayesian hierarchical model would provide enough flexibility, to account for potential confounding and effect modifying variables in future work.



(a) $\beta = 0.3$, left: naive, right: ME corrected



(b) $\beta = 0.6$, left: naive, right: ME corrected



(c) $\beta = 0.3$, left: misspecification first setting, right: misspecification second setting, both with ME correction

Figure 3: Empirical mean and 95%-HDI (blue) derived from the posterior on the 20 first data sets of the simulation study. Horizontal line: mean value, yellow is the estimated empirical mean over all posterior means of β and black denotes the true value of β.

Conflicts of Interest

The authors declare no conflicts of interest.

Data and code availability

We provide the full code of the implemented MCMC sampler. We also provide all required files to re-run the simulation study. Moreover, we provide the code that was used to run the application on the Wismut data. The repository can be found at `https://github.com/RaphaelRe/Wismut_ME_Bayes`. The actual data of the Wismut cohort cannot be shared due to privacy protection. However, we share the generated Markov chains for both, application and simulation [46] that can be used to produce all presented results.

References

[1] R. S. Allodji, K. Leuraud, S. Bernhard, S. Henry, J. Bénichou, and D. Laurier. Assessment of uncertainty associated with measuring exposure to radon and decay products in the French uranium miners cohort. *Journal of Radiological Protection*, 32(1):85–100, 2012.

[2] R. S. Allodji, K. Leuraud, A. C. Thiébaut, S. Henry, D. Laurier, and J. Bénichou. Impact of measurement error in radon exposure on the estimated excess relative risk of lung cancer death in a simulated study based on the French Uranium Miners' Cohort. *Radiation and Environmental Biophysics*, 51(2):151–163, 2012.

[3] R. S. Allodji, A. Thiébaut, K. Leuraud, E. Rage, S. Henry, D. Laurier, and J. Bénichou. The performance of functional methods for correcting non-Gaussian measurement error within Poisson regression: corrected excess risk of lung cancer mortality in relation to radon exposure among French uranium miners. *Statistics in Medicine*, 31(30):4428–4443, 2012.

[4] B. G. Armstrong. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occupational and Environmental Medicine*, 55(10):651–656, 1998.

[5] S. M. Bartell, G. B. Hamra, and K. Steenland. Bayesian analysis of silica exposure and lung cancer using human and animal studies. *Epidemiology*, 28(2):281–287, 2017.

[6] M. Belloni, C. Guihenneuc, E. Rage, and S. Ancelet. A Bayesian hierarchical approach to account for left-censored and missing radiation doses prone to classical measurement error when analyzing lung cancer mortality due to γ-ray exposure in the French cohort of uranium miners. *Radiation and Environmental Biophysics*, 59:423–437, 2020.

[7] R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713 – 1723, 2005.

[8] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[9] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

[10] I. Burstyn. Occupational epidemiologist's quest to tame measurement error in exposure. *Global Epidemiology*, 2:100038, 2020.

[11] I. Burstyn, P. Gustafson, J. Pintos, J. Lavoué, and J. Siemiatycki. Correction of odds ratios in case-control studies for exposure misclassification with partial knowledge of the degree of agreement among experts who assessed exposures. *Occupational and environmental medicine*, 75(2):155–159, 2018.

[12] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement error in nonlinear models: A modern perspective*. Chapman Hall, Boca Raton, 2006.

[13] N. Ellenbach, R. Rehms, S. Hoffmann, et al. Ermittlung der Unsicherheiten in der Strahlenexpositionsabschätzung in der Wismut-Kohorte-Teil 2-Vorhaben 3618S12223. 2023.

[14] L. Fahrmeir and A. Hennerfeind. Nonparametric Bayesian hazard rate models based on penalized splines. Technical report, Discussion paper Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München, 2003.

[15] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472, 1992.

[16] S. Greenland, H. J. Fischer, and L. Kheifets. Methods to explore uncertainty and bias introduced by job exposure matrices. *Risk Analysis*, 36(1):74–82, 2016.

[17] P. Gustafson. *Measurement error and misclassification in statistics and epidemiology - Impacts and Bayesian adjustments*. Chapman & Hall/CRC, 2004.

[18] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020.

[19] I. Heid, H. Küchenhoff, J. Wellmann, M. Gerken, L. Kreienbrock, and H.-E. Wichmann. On the potential of measurement error to induce differential bias on odds ratio estimates: an example from radon epidemiology. *Statistics in Medicine*, 21:3261–3278, 2002.

[20] I. M. Heid, H. Küchenhoff, J. Miles, L. Kreienbrock, and H. E. Wichmann. Two dimensions of measurement error: Classical and Berkson error in residential radon exposure assessment. *Journal of Exposure Analysis and Environmental Epidemiology*, 14:365–377, 2004.

[21] W. Heidenreich, L. Tomasek, B. Grosche, K. Leuraud, and D. Laurier. Lung cancer mortality in the European uranium miners cohorts analyzed with a biologically based model taking into account radon measurement error. *Radiation and Environmental Biophysics*, 51(3):263–275, 2012.

[22] W. F. Heidenreich, E. G. Luebeck, and S. H. Moolgavkar. Effects of exposure uncertainties in the TSCE model and application to the Colorado miners data. *Radiation Research*, 161(1):72–81, 2004.

[23] D. J. Hendry. Data generation for the Cox proportional hazards model with time-dependent covariates: a method for medical researchers. *Statistics in Medicine*, 33:436–454, 2014.

[24] I. Hertz-Picciotto and A. H. Smith. Observations on the dose-response curve for arsenic exposure and lung cancer. *Scandinavian Journal of Work, Environment & Health*, 19:217–226, 1993.

[25] S. Hoffmann, D. Laurier, E. Rage, C. Guihenneuc, and S. Ancelet. Shared and unshared exposure measurement error in occupational cohort studies and their effects on statistical inference in proportional hazards models. *PloS one*, 13(2):e0190792, 2018.

[26] S. Hoffmann, E. Rage, D. Laurier, P. Laroche, C. Guihenneuc, and S. Ancelet. Accounting for Berkson and classical measurement error in radon exposure using a Bayesian structural approach in the analysis of lung cancer mortality in the French cohort of uranium miners. *Radiation Research*, 187(2):196–209, 2017.

[27] J. G. Ibrahim, M.-H. Chen, and D. Sinha. *Bayesian survival analysis*. Springer, New York, 2001.

[28] M. Kreuzer, M. Schnelzer, A. Tschense, L. Walsh, and B. Grosche. Cohort profile: the German uranium miners cohort study (WISMUT cohort), 1946-2003. *International Journal of Epidemiology*, 39(4):980–7, 2010.

[29] J. Kruschke. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. 2014.

[30] H. Küchenhoff, R. Bender, and I. Langner. Effect of Berkson measurement error on parameter estimates in Cox regression models. *Lifetime Data Analysis*, 13(2):261–272, 2007.

[31] H. Küchenhoff, V. Deffner, M. Aßenmacher, H. Neppl, C. Kaiser, D. Güthlin, et al. Ermittlung der Unsicherheiten der Strahlenexpositionsabschätzung in der Wismut-Kohorte - Teil I - Vorhaben 3616S12223. Ressortforschungsberichte zum Strahlenschutz, 2018. Bundesamt für Strahlenschutz (BfS).

[32] F. Lehmann. Job-Exposure-Matrix "Ionisierende Strahlung im Uranerzbergbau der ehemaligen DDR.", Version 06/2004. Technical report, 2004.

[33] F. Lehmann, L. Hambeck, K.-H. Linkert, H. Lutze, H. Meyer, H. Reiber, A. Reinisch, H.-J. Renner, T. Seifert, and F. Wolf. *Belastung durch ionisierende Strahlung im Uranerzbergbau der ehemaligen DDR: Abschlußbericht zu einem Forschungsvorhaben.* Hauptverband der Gewerblichen Berufsgenossenschaften:Sankt Augustin, 1998.

[34] M. P. Little, D. Kwon, K. Doi, S. L. Simon, D. L. Preston, M. M. Doody, T. Lee, J. S. Miller, D. M. Kampa, P. Bhatti, et al. Association of chromosome translocation rate with low dose occupational radiation exposures in us radiologic technologists. *Radiation research*, 182(1):1–17, 2014.

[35] M. P. Little, A. Patel, N. Hamada, and P. Albert. Analysis of cataract in relationship to occupational radiation dose accounting for dosimetric uncertainties in a cohort of us radiologic technologists. *Radiation research*, 194(2):153–161, 2020.

[36] J. H. Lubin, Z. Y. Wang, L. D. Wang, J. D. Boice, Jr, H. X. Cui, S. R. Zhang, S. Conrath, Y. Xia, B. Shang, J. S. Cao, and R. A. Kleinerman. Adjusting lung cancer risks for temporal and spatial variations in radon concentration in dwellings in Gansu Province, China. *Radiation Research*, 163(5):571–9, 2005.

[37] A. Majumdar. *Maximum likelihood estimation of measurement error models based on the Monte Carlo EM algorithm.* PhD thesis, State University of New York at Buffalo, 2007.

[38] J. W. Marsh, E. Blanchardon, D. Gregoratto, W. Hofmann, K. Karcher, D. Nosske, and L. Tomášek. Dosimetric calculations for uranium miners for epidemiological studies. *Radiation Protection Dosimetry*, 149(4):371–383, 2012.

[39] S. Martino, R. Akerkar, and H. Rue. Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics*, 38(3), 2011.

[40] N. Metropolis. The beginning of the Monte Carlo method. *Los Alamos Science Special Issue*, pages 125–130, 1987.

[41] M. E. Montez-Rath, K. Kapphahn, M. B. Mathur, A. A. Mitani, D. J. Hendry, and M. Desai. Guidelines for generating right-censored outcomes

from a Cox model extended to accommodate time-varying covariates. *Journal of Modern Applied Statistical Methods*, 16(1):6, 2017.

[42] T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102, 2019.

[43] S. Muff, A. Riebler, L. Held, H. Rue, and P. Saner. Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(2):231–252, 2015.

[44] S. Peters. Although a valuable method in occupational epidemiology, job-exposure matrices are no magic fix. *Scandinavian Journal of Work, Environment & Health*, 46(3):231–234, 2020.

[45] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024.

[46] R. Rehms. MCMC sampling results, `https://doi.org/10.5281/zenodo.15050372`. Mar. 2025.

[47] S. Richardson and W. R. Gilks. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, 138(6):430–442, 1993.

[48] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):319–392, 2009.

[49] A. B. Singer, M. Daniele Fallin, and I. Burstyn. Bayesian correction for exposure misclassification and evolution of evidence in two studies of the association between maternal occupational exposure to asthmagens and risk of autism spectrum disorder. *Current environmental health reports*, 5:338–350, 2018.

[50] L. Stayner, K. Steenland, M. Dosemeci, and I. Hertz-Picciotto. Attenuation of exposure-response curves in occupational cohort studies at high exposure levels. *Scandinavian Journal of Work, Environment & Health*, 29:317–324, 2003.

[51] K. Steenland, C. Karnes, L. Darrow, and V. Barry. Attenuation of exposure-response rate ratios at higher exposures: A simulation study focusing on frailty and measurement error. *Epidemiology*, 26(3):395–401, 2015.

[52] D. Stram, B. Langholz, M. Huberman, and D. Thomas. Correcting for exposure measurement error in a reanalysis of lung cancer mortality for the Colorado Plateau Uranium Miners cohort. *Health Physics*, 77(3), 1999.

[53] UNSCEAR. 2019 Report, Annex B – Lung cancer from exposure to radon. *United Nations*, New York, 2020.

[54] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[55] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian analysis*, 16(2):667–718, 2021.

[56] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[57] M. Zhou. Understanding the Cox regression model with time-change covariates. *The American Statistician*, 55(2):153–155, 2001.

Chapter 9

Contribution IV

Article

Schalk, D.[†], Rehms, R.[†], Hoffmann, V. S., Bischl, B., & Mansmann, U. (2024). Distributed non-disclosive validation of predictive models by a modified ROC-GLM. BMC Medical Research Methodology, 24(1), 190. https://doi.org/10.1186/s12874-024-02312-4

Data and code

The code for the simulation study can be found at https://github.com/difuture-lmu/simulations-distr-auc. The use case can be found at https://github.com/difuture-lmu/datashield-roc-glm-demo. Code and data is fully accessible and results are fully reproducible.

Supplementary Material

Supplementary material can be found at https://doi.org/10.1186/s12874-024-02312-4 (section: *Supplementary Information*).

Copyright information

Author Contributions

U.M. conceived the research. D.S. and **R.R.** implemented the methods, the simulation study, and prepared the use case. D.S. and **R.R.** wrote the manuscript with valuable contributions of all coauthors. All authors discussed the results and were closely involved in proofreading and revising the manuscript. Corresponding author is U.M.

[†]Shared first author

Schalk *et al.*
BMC Medical Research Methodology (2024) 24:190
https://doi.org/10.1186/s12874-024-02312-4

BMC Medical Research Methodology

RESEARCH Open Access

Distributed non-disclosive validation of predictive models by a modified ROC-GLM

Daniel Schalk[1,3,4†], Raphael Rehms[2†], Verena S. Hoffmann[2], Bernd Bischl[1,4] and Ulrich Mansmann[1,2,3*]

Abstract

Background Distributed statistical analyses provide a promising approach for privacy protection when analyzing data distributed over several databases. Instead of directly operating on data, the analyst receives anonymous summary statistics, which are combined into an aggregated result. Further, in discrimination model (prognosis, diagnosis, etc.) development, it is key to evaluate a trained model w.r.t. to its prognostic or predictive performance on new independent data. For binary classification, quantifying discrimination uses the receiver operating characteristics (ROC) and its area under the curve (AUC) as aggregation measure. We are interested to calculate both as well as basic indicators of calibration-in-the-large for a binary classification task using a distributed and privacy-preserving approach.

Methods We employ DataSHIELD as the technology to carry out distributed analyses, and we use a newly developed algorithm to validate the prediction score by conducting distributed and privacy-preserving ROC analysis. Calibration curves are constructed from mean values over sites. The determination of ROC and its AUC is based on a generalized linear model (GLM) approximation of the true ROC curve, the ROC-GLM, as well as on ideas of differential privacy (DP). DP adds noise (quantified by the ℓ_2 sensitivity $\Delta_2(\hat{f})$) to the data and enables a global handling of placement numbers. The impact of DP parameters was studied by simulations.

Results In our simulation scenario, the true and distributed AUC measures differ by $\Delta\text{AUC} < 0.01$ depending heavily on the choice of the differential privacy parameters. It is recommended to check the accuracy of the distributed AUC estimator in specific simulation scenarios along with a reasonable choice of DP parameters. Here, the accuracy of the distributed AUC estimator may be impaired by too much artificial noise added from DP.

Conclusions The applicability of our algorithms depends on the ℓ_2 sensitivity $\Delta_2(\hat{f})$ of the underlying statistical/predictive model. The simulations carried out have shown that the approximation error is acceptable for the majority of simulated cases. For models with high $\Delta_2(\hat{f})$, the privacy parameters must be set accordingly higher to ensure sufficient privacy protection, which affects the approximation error. This work shows that complex measures, as the AUC, are applicable for validation in distributed setups while preserving an individual's privacy.

Keywords Area under the ROC curve, Distributed computing, Medical tests, ROC-GLM

†Daniel Schalk and Raphael Rehms contributed equally to this work.

*Correspondence:
Ulrich Mansmann
mansmann@ibe.med.uni-muenchen.de
Full list of author information is available at the end of the article

Introduction

Medical research needs trust that the use of confidential patient data follows principles of privacy protection. However, depending on the released data, breaches of the patient's privacy may occur [16]. Even when a patient gives informed consent that the researcher can have access to his/her pseudonymized patient data, it is necessary to keep data in a protected environment and to process it accordingly. Privacy-preserving modeling protects sensitive patient data [1].

Typically, multi-center studies in medicine or epidemiology collect the data in a central study database and perform the analyses in a specifically protected environment following the informed consent of the study subjects. However, this requires an administratively challenging and time-consuming trustworthy data-sharing process.

Using only anonymous and aggregated data for analysis can alleviate the administrative load for data sharing. Distributed data networks in clinical studies allow to leverage routinely collected electronic health data and thus streamline data collection. Non-disclosing distributed analysis is an important part of this concept. It enables statistical analyses without sharing individual patient data (IPD) between the various sites of a clinical study or sharing IPD with a central analysis unit. Non-disclosing distributed analyses protect patient data privacy and enhance data security, making this a potentially advantageous approach for medical research involving sensitive patient data. However, algorithms are needed to support robust multivariable-adjusted statistical analysis without the need to centralize IPD.

As a part of the German Medical Informatics Initiative[1] (MII) the Data Integration for Future Medicine (DIFUTURE) consortium [21] undertakes distributed data network studies and provides tools as well as algorithms for non-disclosing distributed analyses. DIFUTURE's specific objective is to provide digital tools for individual treatment decisions and prognosis and to develop distributed algorithms for the discovery and validation of prognostic and predictive rules. In the following paper, we investigate how the area under the curve (AUC) and its confidence intervals (CIs) proposed by DeLong et al. [6] behave if the computed AUC uses a generalized linear model (GLM) approach of Pepe [19] in a distributed differential privacy framework. We can also determine and view the ROC using distributed analyses.

The concept of differential privacy was operationalized by Dwork [7]. An algorithm is considered to be differential private if an observer cannot determine based solely on the output whether a particular individual's information was used in the computation. Differential privacy ensures protection of patient data privacy, as differential private algorithms are more likely to resist identification and re-identification attacks [8] than alternative approaches.

The ROC curve and its AUC in pooled IPD testing data as well as assessing the quality of calibration [27] is the state-of-the-art of prognostic/predictive validation techniques in a binary classification setting. In general, IPD transfer requires specific patient consent, and data protection laws apply. Here, we present a non-disclosing distributed ROC-GLM, which we use to calculate the ROC curve, its AUC, and the respective CIs. These methods and their implementation in DataSHIELD framework [10] allow analyses in which IPD does not leave its secured environment. This way, only noisy IPD under differential privacy or anonymous and aggregated statistics are shared, thereby preventing the identification of individuals. We also demonstrate that assessing the calibration of binary classification rules based on distributed calculation is a straightforward task.

We motivate our approach by looking at the binormal classification case, where individuals with negative or positive outcome have $\mathcal{N}(\mu_0, \sigma_0^2)$ or $\mathcal{N}(\mu_1, \sigma_1^2)$ distributed scores with $\mu_0 < \mu_1$. With $a = (\mu_1 - \mu_0)/\sigma_1$ and $b = \sigma_0/\sigma_1$ it holds that $\mathrm{ROC}(t) = \Phi(a + b \cdot \Phi^{-1}(t))$ and $AUC = \Phi(a/(1 + b^2)^{0.5})$. In the case of non-normal score distribution, the ROC-GLM allows to approximate the respective ROC and AUC by using the same expressions where a and b are estimated from a probit regression. Furthermore, the ROC-GLM approach allows a simultaneous estimation of ROC curves and AUCs over a set of subgroups defined by covariates [19].[2]

Contribution The work herein proposes new privacy-preserving algorithms adapted to the distributed data setting for the ROC-GLM [18], the AUC derived therefrom, and its CIs for that AUC. To validate the algorithms, we provide a simulation study to assess estimation accuracy. We compare the results with those from the standard procedure. Furthermore, we apply the proposed algorithms to validate a given prognostic rule on data of breast cancer patients.

We describe how the concept of the distributed ROC analysis can be incorporated into the ROC-GLM by using differential privacy. We generate privacy-preserving survivor function that can be communicated without

[1] www.medizininformatik-initiative.de

[2] Note, that the estimation of a ROC-GLM is not unbiased in the non-normal case. We provide an illustrative counterexample in Appendix A.6. which uses gamma-distributed score values in the outcome groups.

the threat of privacy breaches. Furthermore, we outline a distributed Fisher scoring algorithm [14] that estimates parameters for the ROC-GLM. In addition, we describe a privacy protecting distributed calibration approach and demonstrate that distributed GLM model building does not impose specific algorithmic challenges. Furthermore, we introduce a distributed version of the Brier score [5] and the calibration curve [28]. The bycatch of the principles described is a privacy protected version of the binormal ROC and its AUC.

Related literature

Boyd et al. [3] calculate the AUC under differential privacy using a symmetric binormal ROC function. However, our approach is more general and allows extension to non-parametric data with multiple covariates. While they derive the AUC from the ROC parameters, we also use integration techniques. In addition, we provide CIs for the AUC. Ünan et al. [25] use homomorphic encryption to calculate the ROC curve. Their approach does not provide CIs or an extension to multiple covariates. To the best of our knowledge, a modified ROC-GLM algorithm for non-disclosing distributed analyses has so far not been developed.

Background

Throughout this paper, we consider binary classification, with 1 for a case with the trait(s) of interest (i.e., "diseased", "success", "favorable") and 0 for the remaining cases (i.e., lacking trait(s) of interest, "healthy", "no success", "unfavorable"). Furthermore, $f(\boldsymbol{x}) \in \mathbb{R}$ is the true score based on a true but unknown function f for a patient with a feature vector $\boldsymbol{x}$ (the individual realization of an underlying random vector $\boldsymbol{X}$). In this paper, the score can also express a posterior probability with $f(\boldsymbol{x}) \in [0,1]$. The function f is estimated by a statistical (classification) model $\hat{f} : \mathbb{R}^p \to \mathbb{R}$. The estimated individual score for a subject with feature or covariate vector $\boldsymbol{x} \in \mathbb{R}^p$ is $\hat{f}(\boldsymbol{x})$. The training or validation data set used to fit or validate $\hat{f}$ is denoted as $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ with $y_i \in \{1, 0\}$. The score $\hat{f}(\boldsymbol{x})$ and a threshold value $c \in \mathbb{R}$ are used to define a binary classifier: $\mathbb{1}_{[c,\infty)}(\hat{f}(\boldsymbol{x}))$. On an observational level, $\boldsymbol{x}_{1,i}$ and $\boldsymbol{x}_{0,i}$ indicate the i^{th} observation that corresponds to a positive or negative output y. The number of observations in $\mathcal{D}$ with output 1 and 0 are denoted by n_1 and n_0. The set of scores that corresponds to the positive or negative output is denoted by $\mathcal{F}_1 = \{\hat{f}(\boldsymbol{x}_{1,i}) \mid i = 1, \ldots, n_1\}$ and $\mathcal{F}_0 = \{\hat{f}(\boldsymbol{x}_{0,i}) \mid i = 1, \ldots, n_0\}$, with $\mathcal{F}_{1,i} = \hat{f}(\boldsymbol{x}_{1,i})$ and $\mathcal{F}_{0,i} = \hat{f}(\boldsymbol{x}_{0,i})$.

ROC curve and AUC

To quantify the quality of a binary classifier, we use the true positive rate (TPR) and false positive rate (FPR) with values between 0 and 1: $\mathsf{TPR}(c) = P(f(\boldsymbol{X}) \geq c \mid Y = 1)$ and $\mathsf{FPR}(c) = P(f(\boldsymbol{X}) \geq c \mid Y = 0)$ for threshold $c \in \mathbb{R}$ [18]. These probability functions are also known as positive or negative *survivor functions* $S_1(c) = \mathsf{TPR}(c)$ and $S_0(c) = \mathsf{FPR}(c)$. The ROC curve is defined as $\text{ROC}(t) = S_1(S_0^{-1}(t))$. The AUC as a measure of discrimination between the two distributions of the positive and negative class is given as $AUC = \int_0^1 \text{ROC}(t)\, dt$ [30].

Empirical calculation of the ROC curve and AUC

The calculation of the empirical ROC curve uses the *empirical survivor functions* $\hat{S}_1$ and $\hat{S}_0$. These functions are based on the empirical cumulative distribution functions (ECDF) $\hat{F}_1$ of $\mathcal{F}_1$ and $\hat{F}_0$ of $\mathcal{F}_0$: $\hat{S}_1 = 1 - \hat{F}_1$ and $\hat{S}_0 = 1 - \hat{F}_0$. The set of possible values of the empirical TPR and FPR are given by $\mathcal{S}_1 = \{\hat{S}_1(\hat{f}(\boldsymbol{x}_{0,i})) \mid i = 1, \ldots, n_0\}$ and $\mathcal{S}_0 = \{\hat{S}_0(\hat{f}(\boldsymbol{x}_{1,i})) \mid i = 1, \ldots, n_1\}$ and are also called *placement values*. These values standardize a given score relative to the class distribution [19]. The set $\mathcal{S}_1$ represents the positive placement values and $\mathcal{S}_0$ the negative placement values.

The empirical version of the $\text{ROC}(t)$ is a discrete function derived from the placement values $\mathcal{S}_1 \subseteq \{0, 1/n_1, \ldots, (n_1 - 1)/n_1, 1\}$ and $\mathcal{S}_0 \subseteq \{0, 1/n_0, \ldots, (n_0 - 1)/n_0, 1\}$. The empirical AUC is a sum over rectangles of width $1/n_0$ and height $\hat{S}_1(\hat{f}(\boldsymbol{x}_{0,i}))$ ([19], p.106):

$$\widehat{AUC} = {n_0}^{-1} \sum_{i=1}^{n_0} \hat{S}_1(\hat{f}(\boldsymbol{x}_{0,i})). \tag{1}$$

Equation (1) is the empirical analogue of the expectations of the placement values, i.e. $AUC = E(S_1(f(\boldsymbol{x})))$. The term $\hat{f}(\boldsymbol{x}_{0,i})$ is the score of the estimated statistical model for the negative output $\boldsymbol{x}_{0,i}$. The empirical AUC is a function of the empirical survivor function $\hat{S}_1$ evaluated at the score values for all negative outputs $\boldsymbol{x}_{0,i}$.

The empirical AUC is equivalent to the Mann-Whitney U-statistic and inherits the respective distributional properties. For a sufficient large sample of n_0 and n_1, it converges to the normal distribution.

CI for the empirical AUC

CIs are calculated following [6]. The variance of the empirical AUC is determined by:

$$\widehat{\mathsf{var}}(AUC) = \frac{\widehat{\mathsf{var}}(\mathcal{S}_1)}{n_0} + \frac{\widehat{\mathsf{var}}(\mathcal{S}_0)}{n_1}. \tag{2}$$

An asymmetric confidence interval which guarantees values within the interval (0,1) is derived

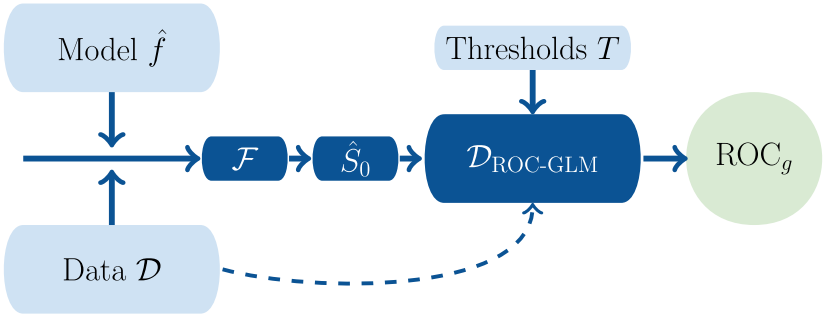


Fig. 1 The ROC-GLM($\mathcal{D}$) procedure starts with the data ($\mathcal{D}$) and a model f for predicting scores Y. It calculates the survivor function $\hat{S}_{\tilde{D}}$ and determines the intermediate data $\mathcal{D}_{\text{ROC-GLM}}$. The probit regression estimates the parameters

from a symmetric confidence interval for logit AUC $\text{ci}_\alpha(\text{logit}(AUC))$ using the logit^{-1} transformation (p 107, [19]):

$$\text{ci}_\alpha(\text{logit}(AUC)) = \text{logit}(\widehat{AUC}) \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sqrt{\widehat{\text{var}}(AUC)}}{\widehat{AUC}\left(1-\widehat{AUC}\right)}. \tag{3}$$

The term Φ^{-1} denotes the quantile function of the standard normal distribution. The term of the standard error is a direct consequence of the application of the delta rule to logit(AUC).

Statistical testing can be conducted based on that CI. For example, the hypothesis $H_0 : AUC \leq a_0$ vs. $H_1 : AUC > a_0$ with a significance level of α can be tested by checking whether $\text{logit}(a_0) < a, \forall a \in \text{ci}_\alpha$ to reject H_0.

The ROC-GLM

The ROC-GLM interprets the ROC curve as a GLM ([19], Section 5.5.2): $\text{ROC}_g(t|\boldsymbol{\gamma}) = g(\boldsymbol{\gamma} h(t))$, with link function $g : \mathbb{R} \to [0,1], \eta \mapsto g(\eta)$, coefficient vector $\boldsymbol{\gamma} \in \mathbb{R}^l$, and covariate vector $h : \mathbb{R} \to \mathbb{R}^l, t \mapsto \boldsymbol{h}(t) = (h_1(t), \ldots, h_l(t))^\mathsf{T}$. In general this estimator is not unbiased (see for example Appendix A.6).

Estimating the ROC-GLM uses an intermediate data set $\mathcal{D}_{\text{ROC-GLM}} = \{(u_{ij}, \boldsymbol{h}(t_j)) \mid i = 1, \ldots, n_1, j = 1, \ldots, n_T\}$ with covariates $\boldsymbol{h}(t_j)$, a set of thresholds $T = \{t_1, \ldots, t_{n_T}\}$, and binary response $u_{ij} \in \{0,1\}$, $u_{ij} = \mathbb{1}_{(\hat{S}_0(\mathcal{F}_{1,i}),\infty)}(t_j) = \mathbb{1}_{(-\infty,\mathcal{F}_{1,i}]}(\hat{S}_0^{-1}(t_j))$. The simplest ROC-GLM uses the two-dimensional vector $\boldsymbol{h}(t)$ with $h_1(t) = 1$ and $h_2(t) = \Phi^{-1}(t)$. Setting the link function to $g = \Phi$ results in the binormal form $\text{ROC}_g(t|\boldsymbol{\gamma}) = \Phi(\gamma_1 + \gamma_2 \Phi^{-1}(t))$. It is equivalent to a probit regression with response variable u_{ij} and covariate $\Phi^{-1}(t_j)$. A common strategy for choosing the set of thresholds T is to use an equidistant grid.

The estimated ROC curve $\text{ROC}_g(t|\hat{\boldsymbol{\gamma}})$ results from the estimated model parameters $\hat{\boldsymbol{\gamma}}$. The AUC from the ROC-GLM $\widehat{AUC}_{\text{ROC-GLM}}$ is the integral $\widehat{AUC}_{\text{ROC-GLM}} = \int_0^1 \text{ROC}_g(t|\hat{\boldsymbol{\gamma}})\, dt$. Here, we use the `R`-function `integrate` [20] or the explicit formula $AUC = \Phi(a/(1+b^2)^{0.5})$. Figure 1 visualizes the single steps of the ROC-GLM algorithm.

Differential privacy

Differential privacy (DP) is a theoretical framework which provides formal guarantees to restrict privacy leakage of individual information when statistical analysis is performed on the data [9, 26]. One of the most prominent DP approaches adds noise $\boldsymbol{r}$ to a deterministic algorithm to obtain a randomized version $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$ with domain $\mathcal{X}$ (e.g., $\mathcal{X} = \mathbb{R}^p$) and target domain $\mathcal{Y}$ (e.g., $\mathcal{Y} = \mathbb{R}$ in regression). Formally speaking a mechanism $\mathcal{M}$ is (ε, δ) -differential private, if for any subset of outputs $R \subseteq \mathcal{Y}$, the property $P(\mathcal{M}(\boldsymbol{x}) \in R) \leq \exp(\varepsilon) P(\mathcal{M}(\boldsymbol{x}') \in R) + \delta$ holds for two adjacent inputs.[3] The value of ε controls how much privacy is guaranteed. Intuitively, this means that for a small ε, applying the randomized algorithm $\mathcal{M}$ on two datasets that only differ in one data point, the typical output (i.e. a high probability) of $\mathcal{M}$ for both datasets has to be nearly the same while a larger value of ε would allow that the typical output could differ more. The value of δ can be interpreted as the probability that ε-differential privacy is broken (see [8]). Hence, δ has to be set to a small value that should be at least less than the inverse number of data points. We provide an interpretation of the privacy parameter ε in Appendix A.3.

[3] In theory, multiple definitions of adjacent inputs exist. Throughout this article, adjacent inputs are based on a histogram representation $\tilde{\boldsymbol{x}} \in \mathbb{N}^p$ and $\tilde{\boldsymbol{x}}' \in \mathbb{N}^p$ of two input vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$. Two inputs are adjacent if the ℓ_1 norm of $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}'$ is equal to one: adjacent $\boldsymbol{x}, \boldsymbol{x}' \Leftrightarrow \|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}'\|_1 = 1$ (cf., [9]).

We add normally distributed noise $\boldsymbol{r}$ to $\hat{f}$ to obtain a private version of the estimated scores $\hat{f}(\boldsymbol{x})$ (i.e. *Gaussian mechanism*): $\mathcal{M}(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}) + r$. Hence, the obfuscated values of the survivor function $\tilde{\mathcal{F}}_1 = \{\mathcal{M}(\boldsymbol{x}_{1,i}) \mid i = 1, \ldots, n_1\}$ and not the original score values $\mathcal{F}_1$ are used for further calculations. The noise $\boldsymbol{r}$ follows a zero-mean Gaussian $\mathcal{N}(0, \tau^2)$, where its variance is set to the minimal value that guarantees a certain level of privacy. Balle and Wang [2] propose the *analytic Gaussian mechanism* which searches numerically for a minimal value of τ such that a defined level of privacy (ε, δ) for a given ℓ_2-sensitivity is achieved. The sensitivity of an algorithm is defined as $\Delta_2(\hat{f}) = \sup_{\text{adjacent } \boldsymbol{x}, \boldsymbol{x}'} \|\hat{f}(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}')\|_2$. Within this work, we first calculate the ℓ_2-sensitivity of the prediction model $\hat{f}$ to determine possible values of the privacy parameters (see Correctness of the AUC inferred from ROC-GLM and distributed ROC-GLM section). Given these parameters, we subsequently determine the minimal required amount of noise τ for the analytic Gaussian mechanism. We provide further details and a visualization of the Gaussian mechanism in Appendix A.2.

Distributed ROC-GLM

General principles

A total of K data sets are distributed over a network of K sites: $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}$. Each data set $\mathcal{D}^{(k)}$ consists of $n^{(k)}$ observations $(\boldsymbol{x}_i^{(k)}, y_i^{(k)})$. The j^{th} component of the i^{th} feature vector of the k^{th} site is denoted by $x_{j,i}^{(k)}$. The i^{th} outcome on site k is $y_i^{(k)}$. We assume (1) the single data have empty intersections and (2) the union of the distributed data is a subset of the full but inaccessible data set:

$$\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}^{(k)}, \quad n = n^{(1)} + \cdots + n^{(K)} \tag{4}$$

Instead of calculating the ROC-GLM for one local data set, we want to calculate the ROC-GLM on K distributed data sets $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}$. All shared information must comply with the following non-disclosing principles:

- **A1** Given the value q, the *privacy level*, an aggregation $a : \mathbb{R}^d \mapsto \mathbb{R}$, $\boldsymbol{v} \to a(\boldsymbol{v})$ is admissible for sharing the value $a(\boldsymbol{v})$ if $d \geq q \in \mathbb{N}$. The *privacy level* requests a minimum number of values on which $a(\boldsymbol{v})$ is derived. In the distributed setup, the aggregation $a(\boldsymbol{v}^{(k)})$ with $n^{(k)}$ unique values in $\boldsymbol{v}^{(k)}$ shared from each of the K sites can then be further processed. Values $a(\boldsymbol{v}^{(k)})$ can be shared if $n^{(k)} \geq q$.
- **A2** Differential privacy [7] is used to ensure non-disclosive IPD via a noisy representation.

Distributed Brier score and calibration curve Calibration of a probabilistic (or scoring) classifier is often addressed by the Brier score [5] or a calibration curve [28]. Both can be calculated by considering criterion **A1**.

Brier score: The Brier score (BS) is the mean squared error of the true 0-1-labels and the predicted probabilities of belonging to class 1. For the Brier score, the score $\hat{f}(\boldsymbol{x}) \in [0, 1]$ is given as posterior probability. The Brier score is calculated by:

$$\text{BS} = n^{-1} \sum_{i=1}^{n} \left(y_i - \hat{f}(\boldsymbol{x}_i)\right)^2 \tag{5}$$

Hence, having a prediction model $\hat{f}$ at each of the K sites, we can calculate the Brier score by:

1. Calculating the residuals $e_i^{(k)}$ based on the true label $y_i^{(k)}$ at site k and the predicted probabilities $\hat{f}(\boldsymbol{x}_i^{(k)})$: $e_i^{(k)} = y_i^{(k)} - \hat{f}(\boldsymbol{x}_i^{(k)}), \forall i = 1, \ldots, n^{(k)}$.
2. Calculating $a_{\text{sum}}(\boldsymbol{e}^{(k)} \circ \boldsymbol{e}^{(k)})$, with $\boldsymbol{e}^{(k)} = (e_1^{(k)}, \ldots, e_{n^{(k)}}^{(k)})^{\intercal} \in \mathbb{R}^{n_k}$, the element-wise product $\circ$, and aggregation $a_{\text{sum}}(\boldsymbol{v}^{(k)}) = \sum_{i=1}^{n^{(k)}} v_i^{(k)}$.
3. Sending $a_{\text{sum}}(\boldsymbol{e}^{(k)} \circ \boldsymbol{e}^{(k)})$ and $n^{(k)}$ (if $n_k \geq q$) to the host, who finally calculates $\text{BS} = n^{-1} \sum_{k=1}^{K} a_{\text{sum}}(\boldsymbol{e}^{(k)} \circ \boldsymbol{e}^{(k)})$.

Calibration curve: To calculate a calibration curve, we discretize the domain of the probabilistic classifier $\hat{f}$ in $[0, 1]$ into n_{bin} bins (for example, $n_{\text{bin}} + 1$ equidistant points p_i from 0 to 1 to construct the n_{bin} bins $b_l = [p_l, p_{l+1})$ for $l = 1, \ldots, n_{\text{bin}} - 1$ and $b_{n_{\text{bin}}} = [p_{n_{\text{bin}}}, p_{n_{\text{bin}}+1}]$ for $l = n_{\text{bin}}$). The calibration curve is the set of 2-dimensional points $p_{\text{cal},l} = (\text{pf}_l, \text{tf}_l)$, with $\text{tf}_l = |\mathcal{I}_l|^{-1} \sum_{i \in \mathcal{I}_l} y_i$ as the true fraction of $y_i = 1$ in bin b_l and $\text{pf}_l = |\mathcal{I}_l|^{-1} \sum_{i \mathcal{I}_l} \hat{f}(\boldsymbol{x}_j)$ as the predicted fraction for outcome 1 in b_l. The set $\mathcal{I}_l$ describes the observations for which the prediction $\hat{f}(\boldsymbol{x}_i)$ falls into bin b_l: $\mathcal{I}_l = \{i \in \{1, \ldots, n\} \mid \hat{f}(\boldsymbol{x}_i) \in b_l\}$. A probabilistic classifier $\hat{f}$ is well-calibrated if the points $p_{\text{cal},l}$ are close to the bisector.

In the distributed setup, the points $p_{\text{cal},l}$ are constructed by applying the distributed mean to both points for each bin at each site:

1. Set all $b_1, \ldots, b_{n_{\text{bin}}}$, and communicate them to the sites.
2. Calculate the values $c_{l,\text{pf}}^{(k)} = a_{\text{sum}}(\{\hat{f}(\boldsymbol{x}_i^{(k)}) \mid i \in \mathcal{I}_l^{(k)}\})$ and $c_{l,\text{tf}}^{(k)} = a_{\text{sum}}(\{y_i^{(k)} \mid i \in \mathcal{I}_l^{(k)}\})$ for all $l = 1, \ldots, n_{\text{bin}}$.
3. Send $\{(c_{l,\text{tf}}^{(k)}, c_{l,\text{pf}}^{(k)}, |\mathcal{I}_l^{(k)}|) \mid k = 1, \ldots, K, l = 1, \ldots, n_{\text{bin}}\}$ to the host if $|\mathcal{I}_l^{(k)}| \geq q$.

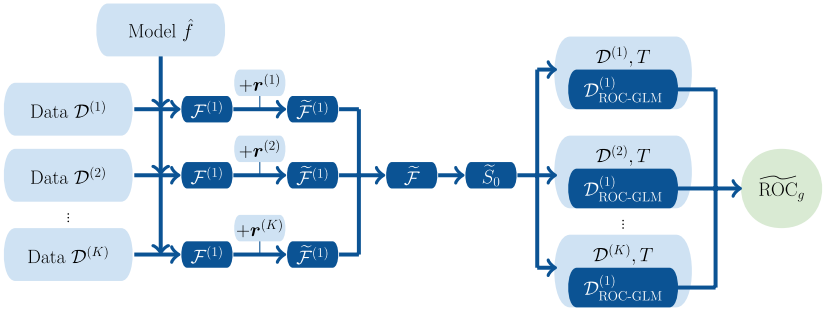


Fig. 2 The distributed ROC-GLM procedure (distrROCGLM) calculates the distributed approximation $\widetilde{\text{ROC}}_g$ of ROC_g. The sites (here $K = 3$) communicate scores with added noise. Centrally, the global negative survivor function $\tilde{S}_{\tilde{\mathcal{D}}}$ is determined and returned to the sites. Finally, the distributed probit regression operates on local intermediate data $\mathcal{D}^{(k)}_{\text{ROC-GLM}}$

4 The host calculates the calibration curve $p_{\text{cal},l}$ by aggregating the elements $\text{tf}_l = (\sum_{k=1}^K |\mathcal{I}_l^{(k)}|)^{-1} \sum_{k=1}^K c_{l,\text{tf}}^{(k)}$ and $\text{pf}_l = (\sum_{k=1}^K |\mathcal{I}_l^{(k)}|)^{-1} \sum_{k=1}^K c_{l,\text{pf}}^{(k)}$ for $l = 1, \ldots, n_{\text{bin}}$.

The distributed ROC-GLM Two aspects are of relevance when building the distributed version of the ROC-GLM (distrROCGLM): (1) The distributed version of the empirical survivor function and (2) a distributed version of the probit regression. Figure 2 shows details of the general procedure. The starting point of the distributed ROC-GLM is the private data $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}$ on the K sites.

The global survivor function $\hat{S}_0$ is approximated by $\tilde{S}_0$ (Approximating the global survivor functions section) using principle **A2**. The computation of $\tilde{S}_0$ depends on the level of privacy induced by the (ε, δ) DP parameters (Differential privacy section). The accuracy of the AUC as well as its CI depends on the choice of ε and δ. The global survivor function $\tilde{S}_0$ is transmitted to each of the K sites and allows calculation of a local version of the intermediate data set $\mathcal{D}^{(k)}_{\text{ROC-GLM}}$ (see The ROC-GLM section). The distributed probit regression complies with principle **A1** and produces the distributed ROC-GLM parameter estimates (see Distributed GLM section). Using the ROC-GLM of these parameters, denoted by $\widetilde{\text{ROC}}_g$, allows calculation of the approximated AUC, denoted by $\widetilde{AUC}_{\text{ROC-GLM}} = \int_0^1 \widetilde{\text{ROC}}_g(t|\hat{\boldsymbol{y}})\, dt$. Finally, the CIs can be calculated based on a variance estimation, which also complies with principle **A2** (see Distributed CIs for the AUC based on the Score function section).

The distributed GLM model building Distributed GLM section describes the federation of the Fisher Scoring algorithm and explains how it can be applied under principle **A2**. Therefore, distributed privacy protected GLM model building does not pose specific challenges.

Approximating the global survivor functions

The privacy-preserving calculation of the global negative survivor function $\hat{S}_0$ needs special attention. It is prohibited to directly communicate score values $\mathcal{F}_0^{(k)}$ from the local sites to the central analyst. Instead, we propose to calculate an approximation $\tilde{S}_0$: First, we determine the ℓ_2-sensitivity of the prediction model $\hat{f}$ and set the value of ε and τ. Then, we generate a noisy representation $\tilde{\mathcal{F}}_0^{(k)} = \mathcal{F}_0^{(k)} + \boldsymbol{r}^{(k)}$ of the original score values $\mathcal{F}_0^{(k)}$ at each site. Second, the noisy scores are communicated to the host and pooled to $\tilde{\mathcal{F}}_0 = \bigcup_{k=1}^K \tilde{\mathcal{F}}_0^{(k)}$ to calculate an approximation $\tilde{S}_0$ of the global survivor function. Third, (ε, δ) DP allows sharing $\tilde{S}_0$ with all sites. Forth, the local sites calculate the global placement values and create the intermediate data set used by the distributed probit regression.

Distributed GLM

For distributed calculation of the GLM, we use an approach described by [14] and adjust the optimization algorithm of GLMs – the Fisher scoring – at its base to estimate parameters without performance loss. This approach complies with **A1**.

The basis of the ROC-GLM is a probit regression (and therefore a GLM) with $\mathbb{E}(Y \mid X = x) = g(x^\top \theta)$ and link function g, response variable Y, and covariates X. The Fisher scoring is an iterative descending technique $\hat{\theta}_{m+1} = \hat{\theta}_m + \mathcal{I}^{-1}(\hat{\theta}_m)\mathcal{V}(\hat{\theta}_m)$ that uses second order gradient information. The components are the score vector $\mathcal{V}(\hat{\theta}_m) = [\partial \ell_\theta(y, x)/\partial\theta]_{\theta=\hat{\theta}_m} \in \mathbb{R}^p$ and the observed Fisher

information $\mathcal{I}(\hat{\theta}_m) = [\partial\mathcal{V}(\theta)/\partial\theta]_{\theta=\hat{\theta}_m} \in \mathbb{R}^{p\times p}$ based on the log likelihood $\ell_\theta(\mathcal{D}) = \sum_{i=1}^n \log(f_Y(y_i, x_i))$. A common stop criterion (as used in `R` function `glm` [22]) to determine whether the Fisher scoring has converged is when the relative improvement $|dev_m - dev_{m-1}|/(|dev_m| + 0.1)$ of the deviance $dev_m = -2\ln(\ell_{\hat{\theta}_m}(\mathcal{D}))$ is smaller than a value a. The default value used in the `glm` function of `R` is $a = 10^{-8}$.

Sufficiently large non-overlapping data at the K sites (each subject contributes information only at a unique site) implies the additive structure of the global score vector $\mathcal{V}(\theta_m)$ and Fisher information $\mathcal{I}(\theta_m)$. With the site-specific score vector $\mathcal{V}_k(\theta_m)$ and Fisher information $\mathcal{I}_k(\theta_m)$, it holds:

$$\mathcal{V}(\hat{\theta}_m) = \sum_{k=1}^K \mathcal{V}_k(\hat{\theta}_m) \tag{6}$$

$$\mathcal{I}(\hat{\theta}_m) = \sum_{k=1}^K \mathcal{I}_k(\hat{\theta}_m) \tag{7}$$

Distributed CIs for the AUC based on the Score function

The distributed calculation of the global sample mean $(\text{distrAVG}(\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(K)}))$ complies with **A1** as well as the distributed version of the sample variance $\widehat{\text{var}}(\boldsymbol{v}) = (n-1)^{-1}\sum_{i=1}^n (v_i - \bar{v})^2$. In the first step, the sample mean is calculated using $\bar{v} = \text{distrAVG}(\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(K)})$ and shared with all K sites. In the second step, each site calculates the aggregation $a_{\text{var}}(\boldsymbol{v}^{(k)}) = \sum_{i=1}^{n^{(k)}} (v_i^{(k)} - \bar{v})^2$, which is further aggregated to the sample variance $\widehat{\text{var}}(\boldsymbol{v}) = (n-1)^{-1}\sum_{k=1}^K a_{\text{var}}(\boldsymbol{v}^{(k)})$: $\text{distrVAR}(\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(K)})$. The operations distrAVG and distrVAR fulfill **A1** if $n^{(k)} \geq q, \forall k \in \{1, \ldots, K\}$.

The operation distrVAR provides a non-disclosing distributed CIs for the global AUC. As described in Empirical calculation of the ROC curve and AUC and CI for the empirical AUC sections, the calculation of the approximated CI requires both approximated survivor functions $\tilde{S}_0$ and $\tilde{S}_1$ (see Approximating the global survivor functions section). A distributed CI $\widetilde{\text{ci}}_\alpha$ to approximate ci_α follows from Formula (3).

Simulation study

General considerations

The aim of the simulation study is to understand the effect of the noise (introduced by DP) on the AUC estimate of the distributed ROC-GLM and its DeLong confidence intervals. We take the global empirical AUC [11, 17] as a proxy for the true AUC of the underlying data generating process. Our goal is not to construct better estimates for the true AUC, but to study the difference between our distributed approach to the empirical AUC on the the pooled data.

In this context, we assess the bias of the distributed approach and measure the difference $\Delta AUC = AUC - \widetilde{AUC}_{\text{ROC-GLM}}$ between the empirical AUC on pooled data (Empirical calculation of the ROC curve and AUC section) and the distributed ROC-GLM $\widetilde{AUC}_{\text{ROC-GLM}}$ (General principles section).

To evaluate CI related bias, we calculate the error Δci_α based on the symmetric difference between ci_α proposed by DeLong et al. ([6], see Sect. 3.3) and our non-disclosing distributed approach $\widetilde{\text{ci}}_\alpha$ (Distributed CIs for the AUC based on the Scor function section). We study $\Delta\text{ci}_\alpha = |\widetilde{\text{ci}}_{\alpha,l} - \text{ci}_{\alpha,l}| + |\widetilde{\text{ci}}_{\alpha,r} - \text{ci}_{\alpha,r}|$, with indices l and r denoting the left and right side of the CI, respectively.

We explore the following research questions:

Question 1– Correctness of the AUC inferred from ROC-GLM and distributed ROC-GLM.(Correctness of the AUC inferred from ROC-GLM and distributed ROC-GLM section): Which privacy parameters ε and δ result in $|\Delta AUC|$ below 0.01?

Question 2–Correctness of the AUC CIs inferred from ROC-GLM and distributed ROC-GLM. (Correctness of the AUC CIs inferred from ROC-GLM and distributed ROC-GLM section): Which privacy parameters ε and δ result in Δci_α below 0.01?

Data generation

In order to avoid the specification of the score distributions in both outcome groups, we simulate data as follows. We generate uniformly distributed AUC values between 0.5 and 1. (1) The population size n is randomly chosen from $\{100, 200, \ldots, 2500\}$. (2) For each $i \in \{1, \ldots, n\}$, the *true* prediction scores are generated from the uniform distribution $\mathcal{F}_i \sim U[0, 1]$. Next, (3) the class membership $y_i \in \{0, 1\}$ is determined by $y_i = \mathbb{1}(\mathcal{F}_i \geq 0.5)$. This results in a perfect discrimination by scores between positives and negatives (AUC=1). (4) The perfect ordering of the class values with respect to individual scores is broken by flipping labels randomly. A set of indexes $\mathcal{I}$ of size $\lfloor \gamma n \rfloor$ is selected for which the corresponding labels are replaced by $y_i \sim \text{Ber}(0.5), \forall i \in \mathcal{I}$. The fraction γ is sampled from a $U[0.5; 1]$ distribution. (5) For comparison, the empirical AUC is calculated from the vector of scores $\mathcal{F}$ and flipped labels y. (6) The non-disclosing distributed

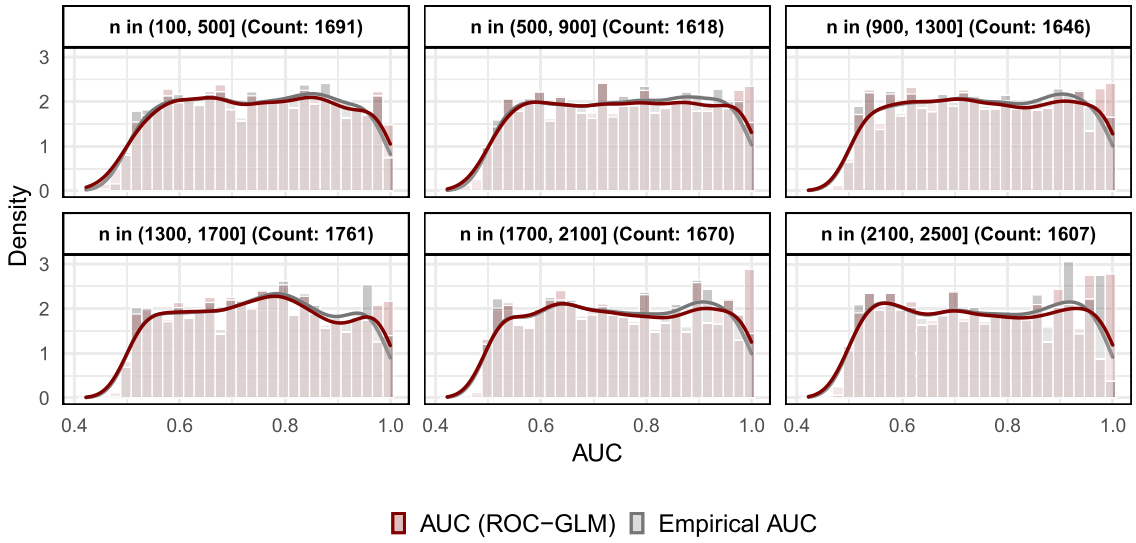


Fig. 3 Densities of 10 000 simulated values of the empirical and non-distributed ROC-GLM AUC. The Densities are grouped according data sizes *n*

process described in General principles section is based on 5 centers and produces the $\widetilde{AUC}_{\text{ROC-GLM}}$ and $\widetilde{ci}_{0.05}$. The examined values for the distributed ROC-GLM are described in Correctness of the AUC inferred from ROC-GLM and distributed ROC-GLM section. The simulation is repeated $N^{sim} = 10000$ times.

Figure 3 shows the empirical distribution of the empirical as well as ROC-GLM-based AUC values depending on the sizes of n. The distribution of the empirical AUC values is close to the uniform distribution over the range of 0.5 to 1. The behaviour of AUC estimates at the borders can be explained as follows: To obtain an AUC value of one, it is necessary to keep all original class labels y. However, this happens rarely, due to the randomized assignment of the observations chosen in $\mathcal{I}$. The same applies to AUC values close to 0.5. An AUC value of 0.5 appears if the class labels are completely randomized. This is also a rare event.

Results

Correctness of the AUC inferred from ROC-GLM and distributed ROC-GLM

ROC-GLM Figure 3 shows a nearly perfect overlap of the means of the simulated empirical as well as the non-distributed ROC-GLM AUC values in the range of values between 0.6 and 0.8. Nevertheless, the behaviour at the right border results from numerical problems of the probit regression on data containing only very few values of zero and mostly values of 1.

Table 1 shows summary statistics of $(AUC - AUC_{\text{ROC-GLM}})$ organized by bins of the empirical AUC of width 0.025. In **Question 1**, an absolute difference below 0.01 is requested, which is fulfilled over the whole AUC range. Mean and median differences ranging from 0.5 to 0.95 fulfill this requirement, whereas for empirical AUC values between 0.95 and 0.975 slightly larger differences are observed. Moreover, for the lower bins, the difference is always positive while it is negative for the higher bins. This is in line with the example from Appendix A.6 for biased ROC-GLM estimation.

The results suggest that there are systematic deviations. Thus, we use as an alternative measure the ℓ_1 -norm that quantifies the discrepancy between the estimated empirical and the estimated GLM formulation of the ROC curve: the (absolute) area between both curves over the whole range t, that is, $discr_s = \int_0^1 |\text{ROC}_g(t|\hat{\boldsymbol{y}}) - \hat{S}_1(\hat{S}_0^{-1}(t))|dt$ for a data set $s \in \{1, \ldots, N^{sim}\}$. Figure 4 shows the empirical distribution of the defined measure over all simulations.

It can be seen that the difference are in general less than 5% (mean: 0.032, 25%- quantile: 0.021, 75% -quantile: 0.043). The small discrepancy with respect to the AUC is explained by the fact that there are areas where the empirical AUC is above the ROC-GLM and vice versa which compensate each other (see for example the left panel of Fig. 9 and Figure S4 in the appendix where this regions can be seen).

Distributed ROC-GLM In the following, we investigate the accuracy of the AUC estimated by the distributed

Table 1 Minimum, 0.25-quantile/1st quantile, median, mean, 0.75-quantile/3rd quantile, maximum, standard deviation, and the differences $AUC - AUC_{\text{ROC-GLM}}$ of the bins containing the respective subset of the 10000 empirical AUC values

Emp. AUC (Bin)	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Sd.	Count
(0.5, 0.525]	−0.0044	−0.0001	0.0005	0.0040	0.0014	0.0506	0.0100	431
(0.525, 0.55]	−0.0052	0.0001	0.0006	0.0027	0.0011	0.0986	0.0123	505
(0.55, 0.575]	−0.0031	0.0003	0.0009	0.0014	0.0015	0.1298	0.0080	465
(0.575, 0.6]	−0.0018	0.0006	0.0012	0.0015	0.0017	0.1567	0.0072	482
(0.6, 0.625]	−0.0044	0.0009	0.0015	0.0014	0.0020	0.0064	0.0010	485
(0.625, 0.65]	−0.0039	0.0012	0.0017	0.0017	0.0022	0.0069	0.0010	501
(0.65, 0.675]	−0.0031	0.0013	0.0018	0.0018	0.0023	0.0068	0.0011	503
(0.675, 0.7]	−0.0022	0.0012	0.0018	0.0018	0.0023	0.0064	0.0010	465
(0.7, 0.725]	−0.0082	0.0010	0.0016	0.0016	0.0023	0.0070	0.0012	523
(0.725, 0.75]	−0.0031	0.0008	0.0015	0.0014	0.0021	0.0087	0.0012	485
(0.75, 0.775]	−0.0058	0.0004	0.0011	0.0010	0.0018	0.0053	0.0013	501
(0.775, 0.8]	−0.0053	−0.0003	0.0004	0.0005	0.0012	0.0088	0.0015	523
(0.8, 0.825]	−0.0061	−0.0013	−0.0002	−0.0004	0.0005	0.0045	0.0016	476
(0.825, 0.85]	−0.0125	−0.0023	−0.0013	−0.0014	−0.0003	0.0059	0.0019	484
(0.85, 0.875]	−0.0111	−0.0037	−0.0026	−0.0025	−0.0014	0.0074	0.0020	520
(0.875, 0.9]	−0.0136	−0.0056	−0.0044	−0.0043	−0.0030	0.0076	0.0023	534
(0.9, 0.925]	−0.0195	−0.0080	−0.0065	−0.0065	−0.0052	0.0066	0.0026	515
(0.925, 0.95]	−0.0193	−0.0105	−0.0091	−0.0089	−0.0076	0.0056	0.0030	481
(0.95, 0.975]	−0.0227	−0.0138	**−0.0113**	**−0.0113**	−0.0093	0.0067	0.0037	503
(0.975, 1]	−0.0180	−0.0093	−0.0062	−0.0064	−0.0034	0.0013	0.0039	529

Bold values indicate that these AUC bins show absolute differences larger 0.01 and provide a negative answer to **Question 1**. The count column indicates the number of simulated AUC values per bin

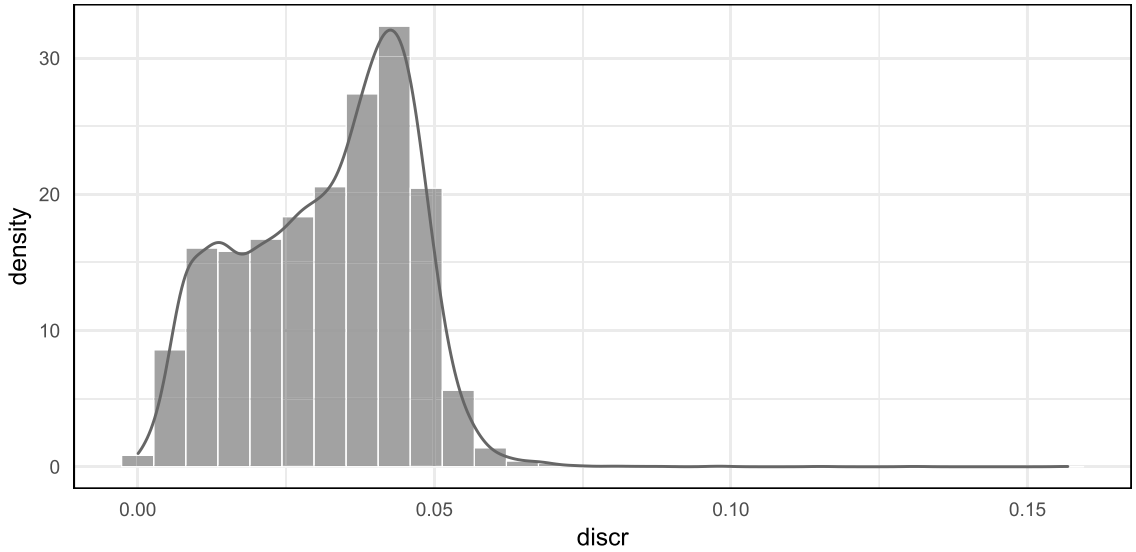


Fig. 4 Distribution of area between the empirical ROC curve and the ROC-GLM curve. The distribution of the alternative discrepancy measure $discr_s$ is estimated from all simulated datasets $s \in \{1, \ldots, N^{sim}\}$

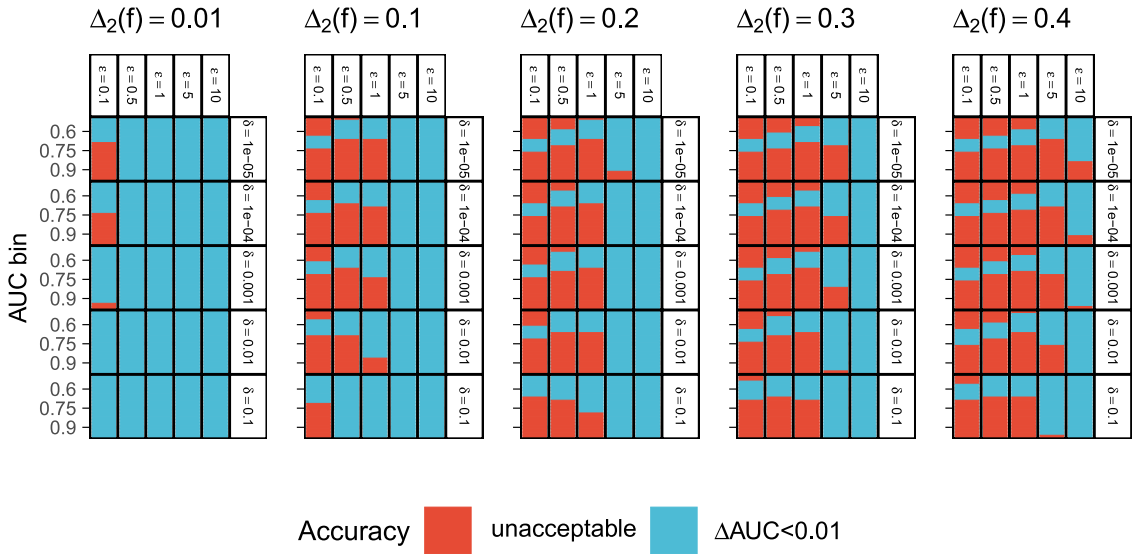


Fig. 5 Absolute difference $|\Delta AUC|$ (mean absolute error, MAE): Combinations of privacy parameters (ε, δ): Each rectangle contains empirical AUC bins of size 0.025 (cf. Table 1) and visualizes the mean of the absolute difference $|\Delta AUC|$ (mean absolute error, MAE) of the distributed ROC-GLM AUC compared to the empirical AUC per bin. Each rectangle corresponds to one simulation setting $(\Delta_2(\hat{f}), \varepsilon, \delta)$. The MAE per bin is categorized according to the required precision, with blue visualizing an MAE ≤ 0.01 (**Question 1**) while red shows an unacceptable accuracy measured as MAE larger than 0.01

ROC-GLM. The respective DP parameters (ε and δ) must be determined in such a way that the answer to **Question 1** is positive. The data are distributed over five sites: The simulated prediction scores $\mathcal{F}$ and true classes y are randomly split into $K = 5$ parts $\mathcal{F}^{(1)}, \ldots, \mathcal{F}^{(5)}$ and $y^{(1)}, \ldots, y^{(5)}$. Our simulation setting uses $\varepsilon \in A_\varepsilon = \{0.1, 0.5, 1, 5, 10\}$ and $\delta \in A_\delta = \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. Due to the Gaussian mechanism, we must also take the ℓ_2-sensitivity into account as the added noise depends on it. Since we do not have an analytical description of the score function $\hat{f}$, we can not determine $\Delta_2(\hat{f})$ explicitly in this simulation. We assume $\Delta_2(\hat{f}) \in A_{\Delta_2(\hat{f})} = \{0.01, 0.1, 0.2, 0.3, 0.4\}$. For the simulation, each setting of the grid $A_\varepsilon \times A_\delta \times A_{\Delta_2(\hat{f})}$ is evaluated by simulating 10000 data sets (cf. Data generation section) and hence obtaining 10000 $\widetilde{AUC}_{\text{ROC-GLM}}$ values that are compared to the respective empirical AUC.

Figure 5 shows the simulation results for different (ε, δ) combinations. The absolute difference of $(AUC - AUC_{\text{distributed ROC-GLM}})$ is checked for having a value below 0.01. The results are based on 10000 simulation runs for 25 (ε, δ) combinations and for each $\Delta_2(\hat{f}) \in \{0.01, 0.1, 0.2, 0.3, 0.4\}$. The variance of the added noise to the scores is determined by the analytic Gaussian mechanism from [2]. The figure reveals that the bias between empirical and distributed ROC-GLM AUC depends heavily on the ℓ_2-sensitivity. The smaller the sensitivity, less noise is required to ensure a certain level of privacy. Correspondingly, smaller choices of privacy parameters can and should be used to ensure privacy. Very small values of ε lead often to unreliable results (except for a very small $\Delta_2(f)$ in combination with higher values of δ). For larger values of ε the results depend (besides the sensitivity) on δ. For instance, the evaluation of the AUC on an algorithm with sensitivity $\Delta_2(f) = 0.1$ and $\varepsilon = 0.5$ would only be reliable with a very high value of $\delta = 0.1$ while a value of $\delta = 10^{-5}$ would be possible for $\Delta_2(f) = 0.01$ with $\varepsilon = 0.5$. For higher values of $\Delta_2(f)$, one has to fall back to higher values of ε. For example, consider a hypothetical dataset with 5000 records and an algorithm with $\Delta_2(f) = 0.3$. In this case one has to accept $\varepsilon = 10$ to guarantee a reliable estimate of the AUC while δ should be set to a small value.

Correctness of the AUC CIs inferred from ROC-GLM and distributed ROC-GLM

The respective results in terms of acceptable (ε, δ) combinations are shown in Fig. 6. In general, acceptable (ε, δ) combinations under **Question 1** are also acceptable under **Question 2**. Therefore, we recommend using the more restrictive settings described in the previous

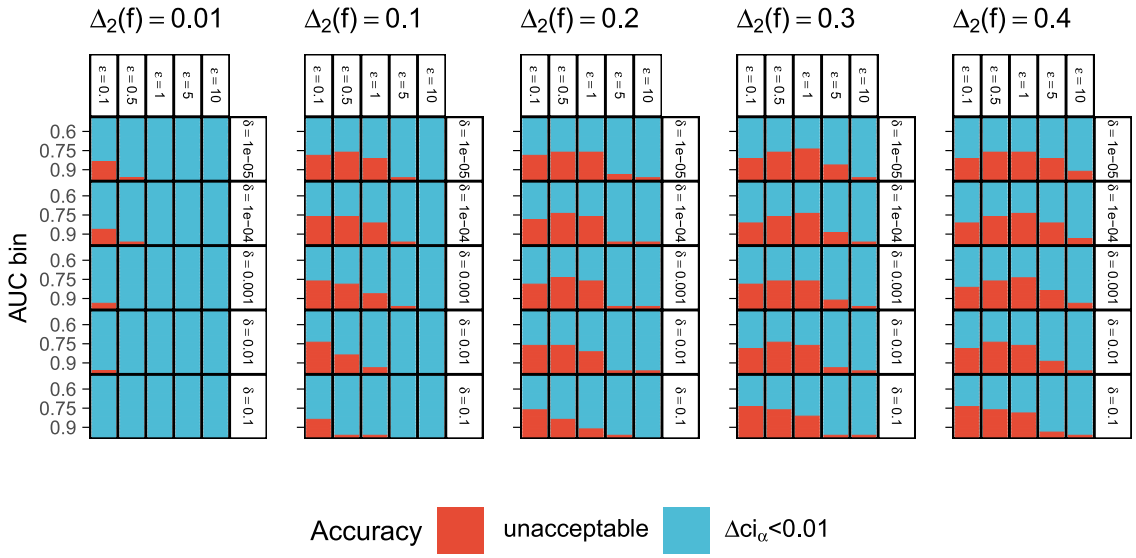


Fig. 6 Mean relative error $\Delta\text{ci}_{0.05}$: Combinations of the privacy parameters ε and δ and their applicability depending on $\Delta_2(\hat{f})$. Each rectangle contains empirical AUC bins of size 0.025 (cf. Table 1) and visualizes the mean of the relative error $\Delta\text{ci}_{0.05}$ of the distributed CI $\tilde{ci}_{0.05}$ compared to $\text{ci}_{0.05}$. Blue shows accuracy values with $\Delta\text{ci}_{0.05} \leq 0.01$ (**Question 2** applies), while red visualizes inaccuracies of $\Delta\text{ci} > 0.01$

Correctness of the AUC inferred from ROC-GLM and distributed ROC-GLM section for the AUC estimation of the distributed ROC-GLM.

Data analysis

We develop a prognostic model on a pooled data and validate its predictive performance on a distributed test data set. We also compare the distributed validation results to results derived from the pooled analysis (see Comparison with pooled data section). As a privacy level, we choose a value of $q = 5$ (see General principles section, **A1**).

About the data The public data set from the German Breast Cancer Study Group [24] can be found in the `TH.data` package [12]. The dataset consists of records from 686 breast cancer patients to assess the effect of hormonal therapy on survival. Besides the binary variable hormonal treatment (`horTH`), the data contains information on age (`age`), menopausal status (`menostat`), tumor size (in mm, `tsize`), tumor grade (`tgrade`), number of positive nodes (`pnodes`), progesterone receptor (in fmol, `progrec`), estrogen receptor (in fmol, `estrec`), recurrence-free survival time (in days, `time`), and censoring indicator (0- censored, 1- event, `cens`).

We split the data into a training data (60 % , 412 observations) and split the remaining (40 %, 250 observations) into 5 parts $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(5)}$ with $n^{(1)} = 51$, $n^{(2)} = 45$, $n^{(3)} = 55$, $n^{(4)} = 46$, and $n^{(5)} = 53$ that are used for the distributed validation. The f of interest is $p(t|\boldsymbol{x}) = P(T > t|X = \boldsymbol{x})$: Probability of surviving time point t without recurrence based on covariates $\boldsymbol{x}$. We choose $t = 730$ (two years). Since we evaluate the binary predictor *patient survives at least t days without recurrence*, we omit 24 patients censored before 730 days from the validation sets. As censoring is assumed to be independent and does not introduce selection bias. For both sets, train and test, roughly 25% of the observations encountered an event before 730 days. We provide the Kaplan-Meier curves of the used training and test data in Appendix A.4. The predicted scores are the survival probabilities $\hat{y}_i = \hat{f}(\boldsymbol{x}_i) = \hat{p}(730|\boldsymbol{x}_i)$ with $\boldsymbol{x}_i \in \cup_{k=1}^{K} \mathcal{D}^{(k)}$. The corresponding binary variable y_i equals 0 if the patient dies in [0, 730] or a recurrence was observed, and y_i equals 1 if otherwise. Therefore, a high value for the survival probability $\hat{y}_i$ ideally corresponds to a binary outcome of 1.

About the model We choose a random survival forest [4, 13] using the `R` package `ranger` [29] as a prognostic model $\hat{f}$ for the survival probability $p(t|\boldsymbol{x})$. With the exception of the number of trees (which is set to 20), the random forest was trained with the default hyperparameter settings of the `ranger` implementation. The model formula is given by

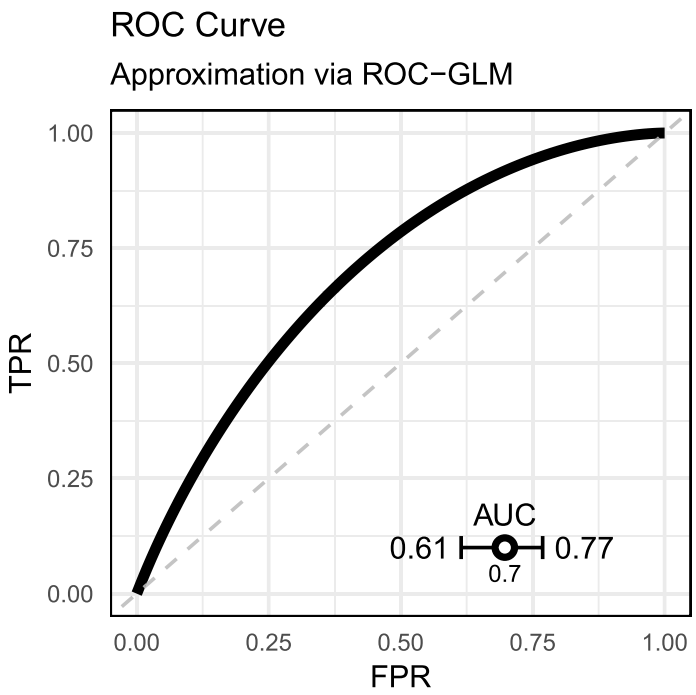


Fig. 7 ROC curve estimated by the distributed ROC-GLM

$$\text{Surv (time, cens)} \sim \text{horTh + age + tsize} + \text{tgrade + pnodes + progrec + estrec.}$$

About the implementation The implementation is based on the DataSHIELD [10] framework and is provided by an `R` package called `dsBinVal` (github.com/difuture-lmu/dsBinVal). Further details about these methods and privacy considerations can be found in the respective GitHub README.

Aim of the analysis The main goal of the analysis is to test the hypothesis that the true AUC is significantly larger than 0.6 as the minimal prognostic performance of the model $\hat{f}$. The significance level is set to $\alpha = 0.05$:

$$H_0 : AUC \leq 0.6 \text{ vs. } H_1 : AUC > 0.6 \tag{8}$$

To test the hypothesis, we estimate the AUC with $\widetilde{AUC}_{\text{ROC-GLM}}$ using the distributed ROC-GLM as well as the approximated CI $\widetilde{\text{ci}}_{0.05}$. We reject H_0 if $AUC > 0.6$, $\forall AUC \in \widetilde{\text{ci}}_{0.05}$.

Analysis plan In the following, (1) we start with the calculation of the ℓ_2-sensitivity (Choice of the privacy parameters section). Depending on the result and the size of the data, we set the privacy parameters ε and δ using the algorithm from [2]. Next, (2) we continue with fitting the distributed ROC-GLM and calculating the approximation of the AUC's confidence interval (Calculation of the distributed ROC-GLM section). At this point, we are able to make a decision about the hypothesis in Eq. (8). In a final step, (3) we demonstrate how to check the calibration of the model using the distributed Brier score and calibration curve (Checking the calibration section).

Choice of the privacy parameters

Given the model and the data set, the ℓ_2-sensitivity is $\Delta_2(\hat{f}) = 0.178$. The results of Correctness of the AUC inferred from ROC-GLM and distributed ROC-GLM section, imply $\varepsilon = 5$ and $\delta = 0.01$ to obtain a reliable estimation.

Calculation of the distributed ROC-GLM

The fit of the ROC-GLM results in parameter estimates of $\gamma_1 = 0.79$ and $\gamma_2 = 1.16$. The AUC obtained from the ROC curve using these parameters is $AUC_{\text{ROC-GLM}} = 0.697$ with $\widetilde{\text{ci}}_{0.05} = [0.615, 0.769]$. The results are visualized in Fig. 7.

Based on the given CI, we significantly reject H_0 for H_1 and hence assume the true AUC to be greater than 0.6.

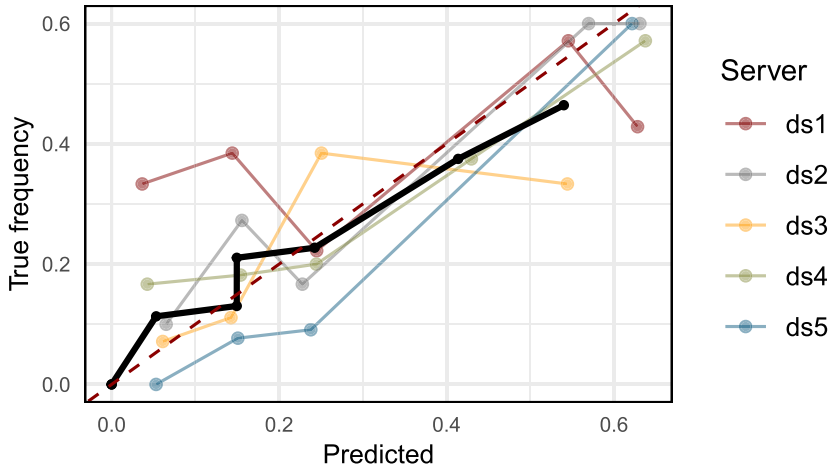


Fig. 8 Distributed calibration curve (bold line) and calibration curves of the individual sites using 10 bins. Note that aggregated values from the site are only shared if one bin contains more than 5 values. See Appendix A.5 for tables containing the numbers of values per bin

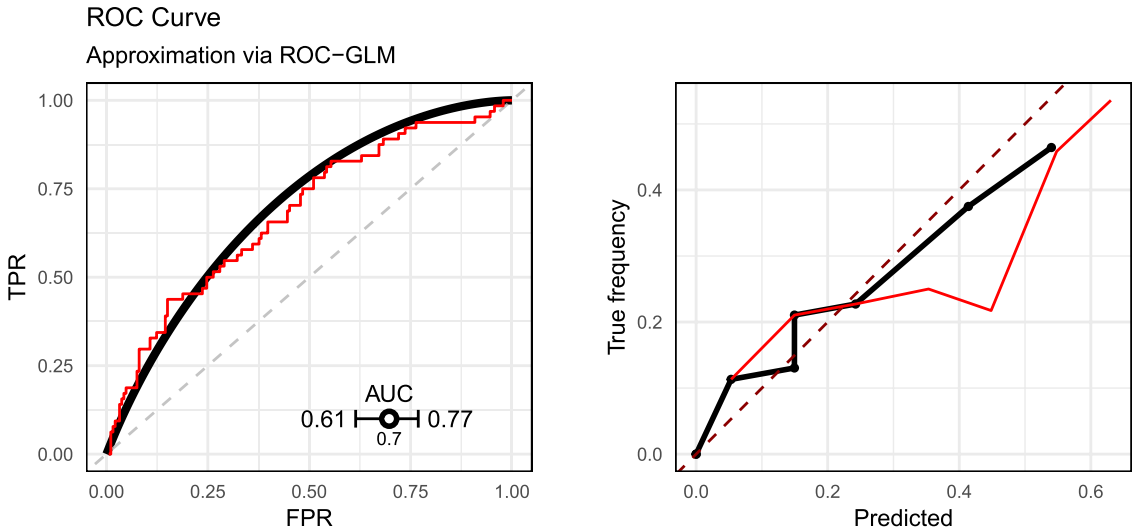


Fig. 9 Comparison of the empirical ROC curve with ROC curve obtained by the distributed ROC-GLM (left). Comparison of the calibration curve when calculated on the pooled scores compared with the distributed calibration curve (right). The thin (red) curves are the lines on the pooled data

Checking the calibration

The Brier score of $\hat{f}$ calculates to BS $= 0.184$ and indicates a good but not perfect calibration. We further assume our model to be not calibrated perfectly. Figure 8 shows the distributed calibration curve as well as the individual calibration curves per site. Furthermore, we observe that the range of the calibration curve does not cover the whole range of the scores $\hat{f}(x) \in [0, 1]$. This indicates that our model does not predict scores close to 1. We want to highlight that, due to privacy reasons, not all score values were included in the calculation; aggregated values are only shared if they consist of at least 5 elements. The table in Appendix A.5 shows the number of elements per bin and site.

Comparison with pooled data

Comparison of both ROC curves (empirical ROC on the pooled sample and the distributed ROC-GLM) (Fig. 9, left) shows an acceptable fit of the ROC-GLM. However, by scrutinizing the plot more closely, one can see that there is a discrepancy between the empirical ROC curve and the estimated ROC-GLM: For a small FPR, the curve from the ROC-GLM is below the empirical one. On the other hand, a similar trend is observed for high values of the FPR in the opposite direction. This refers to differences also observed in the example in Appendix A.6. The resulting AUC values are $\widetilde{AUC}_{\text{ROC-GLM}} = 0.697$ and $AUC = 0.679$ with $|\Delta AUC| = 0.018$. The CIs of the approximated CI $\widetilde{\text{ci}}_{0.05} = [0.615, 0.769]$ and the CI on the pooled scores $\text{ci}_{0.05} = [0.598, 0.751]$ reveals a slightly more optimistic CI estimation in the distributed setup. The error of the CI calculates to $\Delta\text{ci}_{0.05} = 0.034$.

The distributed calibration curve shows an overlap with the calibration curve in areas where all data are allowed to be shared. For bins where this is not the case, the distributed calibration curve differ. Still, the tendency of over- or underestimation of the distributed calibration curve corresponds to one of the pooled curves. The bins for which the full information was received are [0, 0.1], (0.1, 0.2], and (0.2, 0.3] (cf. Appendix A.5 Table S1). For all other bins, at least one site was not allowed to share the aggregated values. The pooled calibration curve shows potential overprediction which is not is reflected by the distributed curve.

The Brier score of the pooled and distributed approach is equal.

Reproducibility considerations

All experiments were conducted using `R` version 4.1.2 on a Linux machine with an Intel(R) Core(TM) i7-8665U CPU @ 1.90GHz processor. The package used to run the simulation was `batchtools` [15]. The code to reproduce all results as well as all simulation results is available in a GitHub repository[4]. The repository contains a README file with further details and a script to install all packages with the respective version used when the benchmark was conducted.

The code to conduct the data analysis is given in a separate GitHub repository[5]. The repository contains the data, an installation of all necessary packages, as well as code to set up the publicly available DataSHIELD server[6] to run the analysis[7].

[4] github.com/difuture-lmu/simulations-distr-auc

[5] github.com/difuture-lmu/datashield-roc-glm-demo

[6] Available at opal-demo.obiba.org. The reference, username, and password are available at the OPAL documentation opaldoc.obiba.org/en/latest/resources.html in the "Types" section.

[7] We cannot guarantee the functionality of the DataSHIELD server or if it will be publicly available forever. However, we keep the repository up-to-date by using continuous integration, which is triggered automatically every week. This system also reports errors that occur if the analysis cannot be conducted on the test server anymore. Further information can be found in the README file of the repository.

Discussion

Distributed non-disclosing (i.e., privacy-preserving) strategies for data analysis are highly relevant for data-driven biomedical research. Since the analyses can be considered anonymous, current legal data protection frameworks allow their use without requesting specific consent. Protecting privacy by appropriate means is fundamental when using personal data for research. Distributed approaches also enable taking part in broader network structures without additional administrative work concerning data protection issues. Privacy-preserving distributed computation allows researchers to digitally cooperate and leverage the value of their data while respecting data sovereignty and without compromising privacy. Besides the privacy preservation in algorithms that are backed up with security mechanisms, it is worth noting that software is also a key player in privacy-preserving analysis. For example, most models fitted with the statistical software `R` attach data directly to the model object. Sharing these objects without caution gives analysts direct access to the training data (cf., e.g., [23]).

International activity has been dedicated to setting up distributed non-disclosing analysis frameworks, which implement machine learning approaches into a distributed analysis scheme. However, our impression is that algorithms for distributed *validation* of these learning algorithms are lacking.

In this paper, we specifically focused on the assessment of discrimination and calibration of learning algorithms with a binary outcome. The discrimination is estimated by a ROC curve and its AUC. We also provide CIs to the distributed AUC estimate. The distributed estimation process is based on *placement values* and *survivor functions*. They represent qualities of the global distribution of score values (aggregated over all centers). To do this in a non-disclosing way, we applied differential privacy techniques. With the creation of the placement values and the transmission of this information to the local server, we applied a distributed version of the ROC-GLM approach to estimate the ROC curve and its AUC in a distributed way. We used a straightforward approach for the distributed GLM estimation. However, we acknowledge that there may be more efficient approaches.

The proposed method implements a combination of aggregation and differential privacy (DP) with privacy parameters (ε, δ). DP offers a solution to exchange critical information privately to other sites, but a part of the information is lost through the induced noise of the privacy mechanism. The balance between utility (i.e. accurate estimates) and privacy must be carefully weighted. The results suggest, that for algorithms with a small sensitivity, the estimates stay reliable. However, for a higher sensitivity this is not the case. In general, a higher value

of δ may lead to more flexibility (and therefore to a higher privacy level) with respect to ε, e.g. setting $\delta = 0.1$. This suggests, that ε-DP is broken in 10% of the cases. It is questionable, whether this is an acceptable value.

We discuss broadly the potential bias in the approximation of the ROC curve by the distributed GLM approach and show results in Table 1 and Fig. 9. We focus on a binary measure of bias ($|\Delta AUC| < 0.01$) and did not address bias issues in detail. We did not explore how bias may be assessed by choosing different metrics (like relative measures). We did not explore aspects of unbalanced datasets and there effect on metrics like negative/positive predictive value. Hence, a more comprehensive analysis of the proposed method is necessary: even though the presented simulation studies provides valuable insights into the proposed method, it lacks of a in-depth detailed analysis. It is missing a comparison of the empirical ROC curve and its distributed ROC-GLM counterpart in terms of a ℓ_1-metric.

Besides the potential bias of the ROC-GLM, the simulation study of the DP parameters considers only a selected range of configurations and does not further investigate their impact beyond the binary threshold. Additionally, the application limits itself only to one exemplary scenario with one dataset and one defined algorithm. Therefore, it can rather be seen as a didactic example. An in-depth examination of various classification tasks with different characteristics of data and classifiers under real-world conditions are necessary. Hence, future work is required to address the mentioned points in a comprising simulation study and a range of application settings.

Furthermore, a reviewer pointed to the potential anti-conservative effect of the proposed procedure. Figure 9 (left panel) suggests to reject the Null-hypothesis that the AUC is below 0.6 on a 5%-level while the result given in Comparison with pooled data section for the 95%-CI on the pooled data contains 0.6.

In view of these critical points, we therefore recommend applying the proposed method with caution at the moment. It is a straight-forward and pragmatic way to validate data in a federated manner while preserving privacy. Moreover we provide `R` code that directly implements the proposed method in the DataSHIELD framework. However, the previously mentioned problems imply, that the software should not be used as a black-box tool. It can serve as a low-level entry to investigate these issues for a specific setting.

We also want to highlight, that the proposed strategy cannot be used to develop a full machine learning model on distributed data. We focus exclusively on validating an already trained model, using data from other sites only once for this specific context. In general, applying a DP algorithm many times on the same data implies a higher privacy loss. See for example Section 3.5 in [9] about composition theorems in DP.

The procedure proposed can be summarized as follows: (1) The validation of an algorithm requires that it is known and can be shared. (2) The calculation of $\Delta_2(f)$ provides essential input to determine the DP setting. It can be derived from the data at hand and the algorithm under validation. The selection of the DP parameters (ε, δ) depends on the setting and use-case specific features. (3) The user has also to specify the level of privacy for the aggregation (i.e. the minimal number of unique values q to be shared aggregated) under project specific requirements. It is recommended to apply the proposed procedure on settings with large datasets at the different sites.

We mainly concentrate on the validation of a prediction model while the property of the ROC-GLM is not fully explored. We do not address specific features of the ROC-GLM estimates and ignore aspects of unbiasdness and consistency. We demonstrate that the approximation of the AUC by the distributed ROC-GLM estimates introduces bias which needs to be controlled and assessed. The approach creates a bias and needs a pragmatic assessment of whether it is acceptable or not. If the proposed approach is used in an analysis, this aspect must be clearly described in the corresponding analysis plan and its impact on the analysis must be discussed. Our example shows that the proposed approach produces an overly liberal result.

But, it can be seen as an advantage of the proposed strategy that the privacy protecting aspects are also helpful for subgroup analyses. Moreover, the proposed approach makes it straightforward to develop distributed privacy protected GLM based classification models since the log-likelihoods consist of site specific independent additive parts. The procedure described in Distributed GLM section can also be applied to federated privacy protecting model building activities in the family of generalized linear models.

Abbreviations

AUC	Area under the curve
CI	Confidence interval
DP	Differential privacy
FPR	False positive rate
GLM	Generalized linear model
IPD	Individual patient data
MII	Medical Informatics Initiative
ROC	Receiver operating characteristics
TPR	True positive rate

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02312-4.

Supplementary Material 1.

Acknowledgements
This work was supported by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A and Grant Nos. 01ZZ1804C (DIFUTURE, MII) as well as 01ZZ2316E (PrivateAIM, MII). The authors of this work take full responsibility for its content. We thank both reviewers for their helpfull and constructive feedback.

Authors' contributions
DS wrote the manuscript, implemented the methods and the simulation study, prepared the use case, developed the statistical analysis scripts and also created all graphics, interpreted the results from the simulation study and from the use case. In addition, he created the GitHub repositories. RR wrote the manuscript, implemented the methods and the simulation study, prepared the use case, developed the statistical analysis scripts and also created all graphics, interpreted the results from the simulation study and from the use case. In addition, he made significant contributions to the GitHub repositories. VH provided substantial assistance in writing the manuscript and interpreting the results. BB provided substantial assistance in writing the manuscript and interpreting the results. UM had the initial idea of the distributed AUC calculation. He wrote the manuscript, implemented the methods and the simulation study, prepared the use case, and interpreted the results from the simulation study and from the use case.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was supported by the German Federal Ministry of Education and Re- search (BMBF) under Grant No. 01IS18036A and Grant Nos. 01ZZ1804C (DIFUTURE, MII) as well as 01ZZ2316E (PrivateAIM, MII).

Availability of data and materials
The simulated datasets generated during the current study are available on GitHub, https://github.com/difuture-lmu/simulations-distr-auc.

Data availability
The example data and simulated datasets generated during the current study are available on GitHub.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Statistics, LMU Munich, Munich, Germany. [2]Institute for Medical Information Processing, Biometry and Epidemiology, LMU Munich, Munich, Germany. [3]DIFUTURE (DataIntegration for Future Medicine, www.difuture.de), LMU Munich, Munich, Germany. [4]Munich Center for Machine Learning (MCML), LMU Munich, Munich, Germany.

Received: 1 April 2024 Accepted: 20 August 2024
Published online: 29 August 2024

References

1. Arellano AM, Dai W, Wang S, Jiang X, Ohno-Machado L. Privacy policy and technology in biomedical data science. Ann Rev Biomed Data Sci. 2018;1:115–29.
2. Balle B, Wang YX. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In: International Conference on Machine Learning. Red Hook: PMLR; 2018. pp. 394–403.
3. Boyd K, Lantz E, Page D. Differential privacy for classifier evaluation. In: Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security. ACM Digital Library (Association for computing machinery); 2015. pp. 15–23. https://doi.org/10.1145/2808769.2808775.
4. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
5. Brier GW, et al. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1–3.
6. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;837–45.
7. Dwork C. Differential privacy. In: International Colloquium on Automata, Languages, and Programming. Venice: Springer; 2006. pp. 1–12.
8. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography conference. Berlin: Springer; 2006. p. 265–84. https://doi.org/10.1007/11681878_14.
9. Dwork C, Roth A, et al. The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci. 2014;9(3–4):211–407.
10. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014;43(6):1929–44.
11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
12. Hothorn T. TH.data: TH's Data Archive. 2021. R package version 1.1-0.
13. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008;2(3):841–60. https://doi.org/10.1214/08-AOAS169.
14. Jones EM, Sheehan NA, Gaye A, Laflamme P, Burton P. Combined analysis of correlated data when data cannot be pooled. Stat. 2013;2(1):72–85.
15. Lang M, Bischl B, Surmann D. batchtools: Tools for R to work on batch systems. J Open Source Softw. 2017;2(10). https://doi.org/10.21105/joss.00135.
16. Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. J Am Med Inform Assoc. 2010;17(3):322–7.
17. Mason SJ, Graham NE. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Q J R Meteorol Soc J Atmos Sci Appl Meteorol Phys Oceanogr. 2002;128(584):2145–66.
18. Pepe MS. An interpretation for the ROC curve and inference using GLM procedures. Biometrics. 2000;56(2):352–9.
19. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford University Press; 2003. https://global.oup.com/academic/.
20. Piessens R, Doncker-Kapenga E, Überhuber CW, Kahaner DK. Quadpack: a subroutine package for automatic integration. Springer Series in Computational Mathematics. Springer Berlin, Heidelberg; 2012. https://doi.org/10.1007/978-3-642-61786-7.
21. Prasser F, Kohlbacher O, Mansmann U, Bauer B, Kuhn KA. Data Integration for Future Medicine (DIFUTURE). Methods Inf Med. 2018;57(S01):e57–65. https://doi.org/10.3414/ME17-02-0022.
22. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2021.
23. Schalk D, Irmak On B, Hapfelmeier A, Mansmann U, Hoffmann VS. Model Transportability and Privacy Protection. 2022. 31st International Biometric Conference. https://github.com/schalkdaniel/talk-ibc-2022/blob/main/model-transportability-and-privacy-protection.pdf. Accessed 25 Aug 2024.

24. Schumacher M, Bastert G, Bojar H, Hübner K, Olschewski M, Sauerbrei W, et al. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. J Clin Oncol. 1994;12(10):2086–2093.
25. Ünal AB, Pfeifer N, Akgün M. ppAURORA: Privacy Preserving Area Under Receiver Operating Characteristic and Precision-Recall Curves with Secure 3-Party Computation. Springer Nature Switzerland: Network and System Security. ArXiv, 2102. 2021.
26. Vadhan S. The Complexity of Differential Privacy. In: Lindell, Y. (eds) Tutorials on the Foundations of Cryptography. Information Security and Cryptography. Springer, Cham. 2017;347–450. https://doi.org/10.1007/978-3-319-57048-8_7.
27. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):1–7.
28. Vuk M, Curk T. ROC curve, lift chart and calibration plot. Metodoloski Zvezki. 2006;3(1):89.
29. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017;77(1):1–17. https://doi.org/10.18637/jss.v077.i01.
30. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993;39(4):561–77.

Publisher's Note

Nutzung von großen Sprachmodellen

Zur Anfertigung dieser Dissertation wurden große Sprachmodelle (Large Language Models) genutzt. Diese wurden ausschließlich herangezogen, um Vorschläge für sprachliche Korrekturen auf Basis bereits verfasster Inhalte zu generieren. Es kamen folgende Modelle zum Einsatz: o3-mini-high (OpenAI), GPT-4o (OpenAI).

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 11.05.2025 Raphael Rehms