Statistical models for the analysis of response variables with limited data quality due to misclassification and missing information

Dissertation von Felix Günther

München 2021

Statistical models for the analysis of response variables with limited data quality due to misclassification and missing information

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

eingereicht von

Felix Günther

10.04.2021

Erstgutachter:	Prof. Dr. Helmut Küchenhoff
Zweitgutachterin:	Prof. Dr. Iris M. Heid
Drittgutachter:	Prof. Dr. Paul Gustafson
Tag der mündlichen Prüfung:	08.06.2021

Summary

The validity of the results of statistical models depends on the quality of the underlying data. This cumulative thesis addresses problems associated with limited data quality in response variables of statistical models due to missing information or misclassification. Based on two applied research projects it is shown how ignoring limited data quality can bias the results of statistical analyses and how advanced and newly developed statistical models can be used to draw adequate inference from the data.

The first part of the thesis addresses the problem of response misclassification in data on bilateral diseases, i.e., diseases that may affect either one, or two entities of a paired organ. In such data, misclassification in the person-specific disease status may result from ignoring missing information in one of the two entities or due to entity-specific misclassification. Ignoring such misclassification leads to biased parameter estimates in regression models, and it is shown how correct results can be obtained through an adequate maximum likelihood analysis using internal validation data. The research was motivated by work in the field of genetic epidemiology, where information about the occurrence of an eye disease in a large study was available only through an error-prone, automated classification of retinal images. When investigating the association of genetic variants with the disease, it is important to account for the existing misclassification in the disease status, as a varying performance of the classification algorithm may be associated with factors determined by some genetic variants studied, leading to potentially large biases.

The second part of the thesis focuses on infectious disease surveillance and research projects on SARS-CoV-2 surveillance data. To gain situational awareness on the current state of an infectious disease outbreak, it is important to track the epidemic curve, i.e., the number of new disease onsets over time. To assess the current pandemic situation, it is common to interpret the time series of newly reported cases. However, due to reporting delays between disease onset and reporting by the health authorities, the two time series can differ substantially. In this thesis, it is shown, how the epidemic curve can be estimated from available surveillance data in near real-time based on a Bayesian hierarchical model. In addition, work is presented on adjusting case reporting numbers for misclassification in person-specific disease diagnostics and consequences of such misclassification for assessing current dynamics of an infectious disease outbreak.

Zusammenfassung

Eine eingeschränkte Datenqualität kann Ergebnisse statistischer Auswertungen stark beeinflussen und zu fehlerhaften Schlüssen führen. In dieser kumulativen Dissertation werden solche Probleme anhand zweier angewandter Forschungsprojekte aufgezeigt und fortgeschrittene statistische Modelle entwickelt, um trotz eingeschränkter Datenqualität adäquate Schlüsse aus den vorhandenen Daten ziehen zu können.

Der erste Teil der Arbeit handelt von Fehlklassifikation in Daten zu bilateralen Krankheiten, das heißt Krankheiten, die entweder eine oder beide Ausführungen eines paarweise angelegten Organs betreffen können. In solchen Daten kann der personenspezifische Krankheitsstatus sowohl durch das Ignorieren fehlender Informationen in einer der beiden Ausführungen, als auch aufgrund von Fehlklassifikation des Krankheitsstatus der einzelnen Ausführungen, fehlerbehaftet sein. Das Ignorieren der Fehler führt zu verzerrten Parameterschätzungen in Regressionsmodellen. Es wird gezeigt, wie unverzerrte Ergebnisse durch eine adäquate Maximum-Likelihood-Analyse unter Verwendung interner Validierungsdaten erzielt werden können. Die Forschung wurde im Rahmen eines Projektes auf dem Gebiet der genetischen Epidemiologie durchgeführt. In der Studie waren Daten zum Auftreten einer Augenkrankheit nur durch eine fehleranfällige, automatisierte Klassifikation von Fundusbildern verfügbar. Bei der Untersuchung des Zusammenhangs von genetischen Varianten mit der Krankheit sollte die Fehlklassifikation im Krankheitsstatus berücksichtigt werden, da eine variierende Prognosegüte des Klassifikationsalgorithmus mit untersuchten genetischen Varianten in Zusammenhang stehen kann. Dies kann sonst zu substantieller Verzerrung der Schätzung des Zusammenhangs führen.

Der zweite Teil der Arbeit behandelt die Analyse von Meldedaten zur Überwachung von Infektionskrankheiten am Beispiel der COVID-19 Pandemie. Die sogenannte epidemische Kurve, die Zeitreihe der Anzahl an Infizierten im Bezug auf den Tag ihres Krankheitsbeginns, ist ein geeigneter Indikator zur Beschreibung der Ausbreitungsdynamik einer Infektionskrankheit. Aufgrund von Verzögerungen zwischen Krankheitsbeginn und der Meldung eines Falles durch die Gesundheitsbehörden, entspricht die häufig betrachtete Zeitreihe neu gemeldeter Fälle nicht der epidemischen Kurve und beide Zeitreihen können erheblich voneinander abweichen. In dieser Arbeit wird gezeigt, wie die epidemische Kurve aus verfügbaren Meldedaten mittels eines Bayesianischen hierarchischen Modells in Echtzeit geschätzt werden kann. Außerdem wird untersucht, wie groß eine mögliche Verzerrung der Meldedaten durch Fehler in den diagnostischen Tests einzelner Personen sein kann und wie diese Fehler bei der Einschätzung der aktuellen Lage berücksichtigt werden können.

Acknowledgments

First of all, I would like to thank Prof. Helmut Küchenhoff and Prof. Iris M. Heid for giving me the opportunity to write this dissertation. I felt very comfortable working with you throughout and I am very grateful for the great support and supervision I received from you!

Helmut, through my student work at the StaBLab I was able to gain my first experiences in scientific work and research. Without these experiences, this thesis would not exist now.

Iris, our exciting projects, the many helpful and instructive discussions, and the participation in the great conferences and meetings that you have made possible have greatly encouraged me to continue on this path!

I would also like to thank all my colleagues at the Department of Genetic Epidemiology, University of Regensburg, and the StaBLab, LMU Munich, for the great collaboration over the past years!

Many thanks to Prof. Michael Höhle and Andreas Bender for our very stimulating and motivating collaboration and the many fun virtual meetings during special times. I would also like to thank the entire CODAG group at the Department of Statistics at LMU Munich for many inspiring discussions and great collaboration!

Furthermore, I would like to thank Prof. Paul Gustafson for his work as a reviewer of this thesis and Prof. Thomas Augustin as well as Prof. Göran Kauermann for being available as chair for the disputation.

In addition to all the people I have come into contact with through work and this dissertation, I would like to thank a few more people very sincerely:

My friends, especially Pascal, Simon, Christina, Eduard, Joy, Lukas, Miriam, Nico, Valentin, Flo, and Leo. You have made my life in Munich a great one over the last few years!

Anna, thank you so much for the support, great conversations, new perspectives, beautiful hikes, learning to surf, and the best time ever!

My family, Christine, Kristian, Johanna, and Benedikt, for always being there for me and supporting me from the very beginning!

Contents

Chapter 1

Introduction

1.1 Overview

This thesis addresses the analysis of data with limited quality due to constraining factors in the data collection process. These factors may include, e.g., finite time and financial resources, high time pressure in data collection and provision, and general limitations in the methods used to measure and collect data.

The cumulative thesis consists out of four articles that illustrate the topic in two different application areas: first, response misclassification in logistic regression for bilateral disease data and, second, infectious disease modeling and the real-time analysis of surveillance data where important information is partly missing and the available data is potentially error-prone. The contributions of those articles were motivated by applied research projects in the field of genetic epidemiology, more precisely, genome-wide association studies, as well as surveillance of the acute COVID-19 pandemic.

This chapter is a preface to the contributing articles in this thesis. It introduces the scientific context of the articles and presents statistical concepts and models that form the basis for the newly developed methods. Section 1.2 reviews basic concepts from the field of statistical research on measurement error and misclassification, with a focus on response misclassification in Section 1.2.1 and an introduction to misclassification in bilateral disease data in Section 1.2.2. These topics are of direct relevance to the contributing articles in Chapters 2, 3, and 5. Section 1.3 provides an introduction to genome-wide association studies, the application area of the article in Chapter 3. Section 1.4 introduces key concepts from the area of infectious disease surveillance, the topic of the contributing articles in Chapter 4 and 5.

1.2 Measurement error and misclassification

The central task of statistical data analysis is to investigate properties (of the distribution) of a specific random variable Y, or to understand the relationship between a dependent *outcome* or *response* variable Y and one or several explanatory variables (*covariates*) X by means of statistical models. This is done based on multiple observations of these variables (data) collected from different observational units. The statistical research field of measurement error and misclassification is concerned with what happens when the actual variables of interest, Y or X, cannot be measured precisely. Instead, only data on error-prone variables Y^* and/or X^* could be collected. The problem of measurement errors and inaccuracies is manifold and occurs in almost any research area where statistical methods are used.

That measurement error or, more generally, low-quality data can cause problems for statistical analyses and the interpretation of their results is intuitively straightforward and has also been formally demonstrated for a wide variety of analytical methods and model classes. It is established that the consequences of ignoring measurement error can be very different depending on the specific situation, the magnitude and type of the errors and which variables are affected.

In order to characterize a certain measurement error scenario, it is therefore necessary in a first step to distinguish which variable(s) of the respective analysis are affected by measurement error. In a regression context, this can be either the response Y^*, and/or one or more covariates X^*, further covariates Z can be assumed to be measured without error.

In a second step, it is necessary to specify how the existing measurement error can be adequately described based on a *measurement error model.* This measurement error model is usually expressed on the basis of assumptions about the (conditional) distribution of the measured variable and the true (unobserved) variable. For a covariate X measured with error, this can, for example, be done based on an assumption regarding the conditional probability density function of the error-prone variable, given the truth, $d_{X^*|X}(x^*|x)$. A special case is the classical *measurement error model* for continuous random variables, where the observed variable is assumed to equal $X^* = X + U$, where U is a random variable independent of X with expectation zero and fixed variance. If U is assumed to be normally distributed with variance σ_U^2, this is equivalent to assuming that $X^*|X \sim N(X, \sigma_U^2)$. In other situations it might be more reasonable to assume that $E(X^*)$ corresponds to a linear function of X, or that the classical error model holds for a transformation of X^*, e.g., $\log(X^*) = \log(X) + U$, implying a multiplicative error structure on the original scale. In some cases it is also meaningful to characterize the measurement error based on the conditional distribution $d_{X|X^*}(x|x^*)$. This can be adequate when X^* is an observed average or a predicted value and the true X corresponds to an observation-specific realization that exhibits additional variation around X^*. This type of error is called *Berkson error*

(Berkson, 1950). The above considerations can also be applied to categorical variables by utilizing categorical distributions for $d_{X^*|X}(x^*|x)$, where the measurement error model can, for example, be specified by generalized linear models for the respective class probabilities of the conditional distribution (cf. Section 1.2.1). Note that measurement error in binary/categorical variables is often referred to as *misclassification*. The characterization of measurement error or misclassification in outcome variables Y^* can be done in an analogous way, e.g., using the conditional density $d_{Y^*|Y}(y^*|y)$.

When analyzing the association of several variables, for example, in multiple regression, it is important to distinguish between *non-differential* and *differential* measurement error. The mathematical definition of these terms depends on whether the outcome variable or covariates of the model suffer from measurement error. Suppose there are data available on an outcome Y, an error-prone covariate X^*, and additional error-free covariates Z. In such a situation, one speaks of *non-differential error with respect to the outcome* Y when the conditional distribution of the outcome, given all (partly unobserved) variables does not depend on the error-prone variable X^*, that is, $d_{Y|X^*,X,Z}(y|x^*, x, z) = d_{Y|X,Z}(y|x, z)$. If X^* contains additional information about Y, one speaks of differential error with respect to Y. Alternatively, non-differential error can be expressed based on the conditional distribution of X^*: the error is non-differential if $d_{X^*|X,Z,Y}(x^*|x, z, y) = d_{X^*|X,Z}(x^*|x, z)$. That is, if the conditional distribution of X^* is independent of the outcome Y (Keogh et al., 2020). A classical example of differential error is *recall bias* in case-control studies, when, e.g., cases are more aware of certain exposures than controls. For outcomes, one can define differential measurement error with respect to the covariate X as the scenario where $d_{Y^*|Y,X,Z}(y^*|y, x, z) \neq d_{Y^*|Y,Z}(y^*|y, z)$. The distinction between differential and non-differential error is important, as ignoring the existence of differential error yields often a bigger bias for parameter estimates and is more difficult to account for. An example is given in Section 1.2.1 for the case of logistic regression with misclassification in the response.

The general problem of ignoring the existence of measurement error during statistical analyses is that the conditional distributions that are approximated by a statistical model, for example, the conditional distribution of the response given covariates $d_{Y|X}(y|x)$ in a generalized linear model, do not necessarily correspond to the distribution of the observed data $d_{Y|X^*}(y|x^*)$, $d_{Y^*|X}(y^*|x)$ or $d_{Y^*|X^*}(y^*|x^*)$. Ignoring this and estimating the standard regression model based on observed, error-prone data can yield biased estimates that do not adequately describe the association of the true outcome Y and covariates X. The type and extent of bias depends on the actual error-structure in the data (the *true* measurement error model) and can be investigated based on theoretical considerations or simulation studies. It can range from (almost) no consequences, through substantially increased uncertainty in unbiased estimates, to quantitatively and qualitatively strongly biased results.

This has led to the development of a large number of different analytical approaches that can be used to adjust for the existence of measurement error and to eliminate bias.

These methods are based either on prior knowledge and assumptions about the structure of the measurement error or on additional information from data on the performance of the measurement instruments. The latter can stem from *ancillary* or *validation* studies with, for example, different measurement instruments applied to the observational units or repeated measurements per observational unit using the error-prone instrument (Carroll et al. (2006, Ch. 2.3.); Keogh et al. (2020)).

The existing literature on measurement error and ways to account for it in applied analyses is extensive and focuses, partly, on very specific situations. There are, however, also several books and articles covering a broad range of basic and advanced methods and scenarios. Classic references are, for example, Carroll et al. (2006) covering a wide range of topics on measurement error in (non-linear) regression models, also including measurement error in the response of statistical models; Gustafson (2003) on Bayesian methods for measurement error in continuous and categorical covariates; and Grace (2016) covering methods for, i.a., survival data, longitudinal data, multi-state models, case-control studies and response measurement error. Buzas et al. (2005) provide an overview over measurement error in epidemiology, and the two-part article series of Keogh et al. (2020) and Shaw et al. (2020) provides guidance on measurement error and misclassification on a wide range of topics in observational studies in epidemiology.

The following section provides more details on misclassification of categorical and binary response variables, as this topic is of direct relevance to the contributing articles of this thesis.

1.2.1 Misclassification in categorical and binary response variables

Let Y be a categorical random variable with K classes and p_Y be the K-dimensional vector of class probabilities whose entries sum to one. Let Y^* be an observed, error-prone version of Y with L classes, in most cases L equals K. Misclassification in Y^* can be expressed based on the (mis-)classification probabilities $P(Y^* = l|Y = k) = \pi_{l,k},\ l = 1, \dots, L,\ k = 1, \dots, K$. For binary variables Y^* and Y with $l, k \in \{0, 1\}$, the classification probabilities $\pi_{1,1}$ and $\pi_{0,0}$ are often referred to as *sensitivity* and *specificity*. They fully characterize the misclassification model, as $\pi_{0,1} = 1 - \pi_{1,1}$ and $\pi_{1,0} = 1 - \pi_{0,0}$. In such a case, the class probabilities of Y^* are given, based on the law of total probability, as

$$p_{Y^*} = \Pi \times p_Y, \tag{1.1}$$

where Π is the $L \times K$-dimensional matrix with entries $\pi_{l,k}$ in row l and column k.

Estimation of class probabilities

Equation (1.1) shows directly that the class probabilities p_{Y^*} and p_Y are only equal if $L = K$ and $\Pi = I$, that is, if no misclassification is present. Consequently, any estimator of class probabilities based on error-prone data on Y^*, e.g., the maximum likelihood estimator based on relative class frequencies $\hat{p}_{Y^*}$, is biased for the class probabilities p_Y of the true Y. On the other hand, equation (1.1) also indicates that an estimate of $\hat{p}_Y$ can be obtained from $\hat{p}_{Y^*}$ based on assumed or known misclassification probabilities Π as

$$\hat{p}_Y = \Pi^{-1} \times \hat{p}_{Y^*}. \tag{1.2}$$

For the binary case, this corresponds to the results of Rogan and Gladen (1978), who also show that such an estimate of the true classification probabilities $\hat{p}_Y$ is asymptotically unbiased if the classification probabilities in $\hat{\Pi}$ are estimated as binomial fractions from independent samples. Obviously, the misclassification adjustment from equation (1.2) is valid only if Π adequately describes the misclassification process in the current problem. In case of classification probabilities estimated from external data, this is only true if the misclassification process of the analyzed and the independent (*external*) sample are the same. In the contributing article in Chapter 5 of this thesis, we use a similar adjustment for deriving misclassification-adjusted aggregated SARS-CoV-2 case numbers using assumptions regarding the sensitivity and specificity of person-specific SARS-CoV-2 examinations based on PCR tests.

Response misclassification in logistic regression

In logistic regression, a binary response variable Y is modeled based on a set of covariates X by assuming that the observation-specific response, $Y_i,\ i = 1, \dots, n$, is, conditional on the covariates, Bernoulli distributed with success probability π_i, $Y_i|x_i \sim \mathrm{B}(\pi_i)$. Thereby, π_i is modeled based on $\pi_i = H(x_i'\beta)$, where x_i is the observation-specific covariate vector with first entry 1, and β is the vector of covariate effects including an intercept term. $H(\cdot)$ is the logistic response function $H(\cdot) = 1/(1 + exp(-\cdot))$. In the following, it is assumed that the covariates X are observed error-free.

Consider now that the true outcome Y was not observed but instead an error-prone version Y^*. The misclassification model is expressed in terms of the sensitivity, $P(Y_i^* = 1|Y_i = 1, x_i) = \pi_{1,i}$, and specificity, $P(Y_i^* = 0|Y_i = 0, x_i) = \pi_{0,i}$, that might vary between observations i in association with the covariates X. The conditional probability of $P(Y_i^* = 1|x_i) = \pi_i^*$ can then be derived as

$$\begin{aligned}\pi_i^* &= \sum_{y=0,1} P(Y_i^* = 1|x_i, Y_i = y)P(Y_i = y|x_i) \\ &= (1 - \pi_{0i}) + (\pi_{1i} + \pi_{0i} - 1)H(x_i'\beta) = H^*(x_i'\beta). \end{aligned} \tag{1.3}$$

This result is investigated in detail by Neuhaus (1999). In short, it implies that for fixed classification probabilities $\pi_{1,i} = \pi_1 > 0.5$ and $\pi_{0,i} = \pi_0 > 0.5$ (i.e., for non-differential misclassification), the observed Y^* still follows a generalized linear model, but with a different response function compared to standard logistic regression for the error-free Y. Applying standard logistic regression to the observed Y^* corresponds to estimating a misspecified model and leads to biased parameter estimates that do not describe the association between the true Y and covariates X, but are on average attenuated. In case of differential misclassification, i.e., with sensitivity and specificity varying in association with the covariates, the situation becomes more complex. The link function ${H^*}^{-1}(\pi_i^*)$ is not necessarily monotone anymore (Neuhaus, 1999; Grace, 2016, Ch. 8.2.1.) and, in this case, Y^* does not follow a generalized linear model. This can yield bias in any direction for the parameter estimates $\hat{\beta}$.

For fixed and known classification probabilities, equation (1.3) can be used as response function for setting up the Bernoulli likelihood of a generalized linear model. Given the true classification probabilities, this yields unbiased estimates of $\hat{\beta}$. As a consequence of the misclassification, however, the estimator is less efficient compared to logistic regression for the true Y. Such an approach can also be used to perform sensitivity analyses for different assumptions regarding π_1 and π_0. When the sensitivity and specificity are unknown, the parameters of the model (1.3) are only weakly identifiable, even in the case of non-differential misclassification. This is the fundamental problem of response misclassification in (logistic) regression when no further information on the misclassification process is available (see also Carroll et al., 2006, Ch. 15.3.2.).

When external information on the classification probabilities is available, e.g., based on expert knowledge on the performance of the measurement instrument or based on estimates from external validation data, it can be incorporated into the analysis by specifying corresponding priors for $\pi_{1,i}$ and $\pi_{0,i}$ and perform a Bayesian analysis (e.g., Paulino et al., 2003) or by plugging in the (point) estimates in a pseudolikelihood approach (Carroll et al., 2006, Ch. 15.3.2.) ignoring associated uncertainty.

When internal validation data is available (measurements of a *gold-standard* Y and the error-prone Y^*) for a (randomly selected) subset of n^v out of the n observations, it is possible to set up a full likelihood model to simultaneously estimate parameters of the logistic regression model for the true outcome and the misclassification model (Lyles et al. (2011); Carroll et al. (2006, Ch. 15.4.); Grace (2016, Ch. 8.3.)). For observations from the validation data, the joint conditional density of the true and the error-prone response can be factorized into $f_{Y^*,Y|X}(y^*, y|x; \beta, \gamma) = f_{Y^*|Y,X}(y^*|y, x; \gamma) f_{Y|X}(y|x; \beta)$. Here, $f_{Y|X}(y|x; \beta)$ is the (model-based) density of the true outcome given the covariates and (logistic regression) parameters β. The density $f_{Y^*|Y,X}(y^*|y, x; \gamma)$ corresponds to the measurement error (misclassification) model that is governed by the parameters γ. It can be specified to also account for differential misclassification. For observations from the main study data (without information on Y), the conditional density of the error-prone response is given in terms of the true outcome model and the misclassification model by summing over the unobserved true outcome, $f_{Y^*|X}(y^*|x;\beta,\gamma) = \sum_{y=1,2} f_{Y^*|Y,X}(y^*|Y=y,x;\gamma)f_{Y|X}(Y=y|x;\beta)$. This yields the overall likelihood

$$\begin{aligned} L(\beta,\gamma) =& \prod_{i=1}^{n^v} f_{Y^*,Y|X}(y_i^*,y_i|x_i;\beta,\gamma) \times \prod_{j=1}^{n-n^v} f_{Y^*|X}(y_j^*|x_j;\beta,\gamma) \\ =& \prod_{i=1}^{n^v} f_{Y^*|Y,X}(y_i^*|y_i,x_i;\gamma)f_{Y|X}(y_i|x_i;\beta)\times \\ & \prod_{j=1}^{n-n^v} \Big\{ \sum_{y=1,2} f_{Y^*|Y,X}(y_j^*|Y_j=y,x_j;\gamma)f_{Y|X}(Y_j=y|x_i;\beta)\Big\}. \end{aligned} \tag{1.4}$$

Maximization of the likelihood with respect to the parameters (β,γ) gives consistent estimators of the parameters of the true response model and the misclassification model. A correct specification of the misclassification model is, however, crucial for valid results. In the case discussed here, the misclassification model corresponds to modeling the sensitivity and specificity. They might either be assumed to be constant (non-differential misclassification), or can be modeled parametrically based on (a subset of) the covariates X, for example, using the logistic response function.

This approach to adjust for response misclassification in logistic regression falls into the broader class of likelihood-based correction methods (Grace (2016, Ch. 2.5.1., 8.3., 8.4.), Shaw et al. (2020, Sec. 2.1.), Carroll et al. (2006, Ch. 8., 15.4)). This framework can also be used in different measurement error problems, for example, in the case of measurement error of continuous or discrete covariates, continuous outcomes, or measurement error in outcomes as well as covariates. The general approach of factorizing the (conditional) density of the observed data into a model for the (partly unobserved) error-free variables and a measurement error model and integrating over the unobserved true variables remains the same. If, for example, there is data on a true outcome Y, an error-prone covariate X^*, and error-free covariates Z, one can proceed with rewriting the likelihood of the observed data

$$\begin{aligned} f_{Y,X^*|Z}(y,x^*|z;\theta) &= \int f_{Y,X^*|Z,X}(y,x^*|z,X=x;\theta) \times f_{X|Z}(X=x|z;\theta)\,\mathrm{d}x \\ &= \int f_{Y|Z,X,X^*}(y|z,X=x,x^*;\beta) \times f_{X^*|Z,X}(x^*|z,X=x;\gamma)\times \\ &\quad f_{X|Z}(X=x|z;\alpha)\,\mathrm{d}x \\ &= \int f_{Y|Z,X}(y|z,X=x;\beta) \times f_{X^*|Z,X}(x^*|z,X=x;\gamma)\times \\ &\quad f_{X|Z}(X=x|z;\alpha)\,\mathrm{d}x, \end{aligned} \tag{1.5}$$

where the last equality holds in case of non-differential measurement error. The likelihood of an observed data point is therefore factorized in the true data model, a measurement error model, and a model of the true (unobserved) covariate, with corresponding parameters

β, γ, and α, respectively. It is based on the integration over the (conditional) density of the unobserved true covariate. Other measurement error model situations can be tackled in a similar way, correct specification of the different model components is, however, crucial and hard to evaluate. Optimization of the resulting likelihood (including numerical integration) can be challenging and identifiability of the model parameters is questionable without additional information, for example, from validation data.

The above described situation of response misclassification in logistic regression with observed true covariates X has some central advantages: firstly, integrating over the true response boils down to a summation over the two potential values and, secondly, distributional assumptions for a binary/discrete variable are straight-forward to specify based on the respective class probabilities.

In the contributing articles of Chapter 2 and 3 of this thesis, we present an approach for analyzing error-prone response data on bilateral diseases. It is based on similar considerations as the likelihood-based correction for logistic regression with response misclassification and the following section provides an introduction to the developed model.

1.2.2 Bilateral disease data and response misclassification

The work presented in Chapters 2 and 3 of this thesis was motivated by data on age-related macular degeneration (AMD). AMD is an eye disorder that affects the macular region of the retina, causing progressive loss of central vision and is one of the leading causes of severe irreversible vision loss worldwide (Mitchell et al., 2018). The clinical endpoint of the disease is late AMD, which can appear in two different forms (*neovascular* or *atrophic*). Late AMD is typically preceded by early AMD stages that are clinically asymptomatic and determined by yellowish deposits of extracellular material (drusen) and/or irregularities of the retinal pigment epithelium (hyper-/hypopigmentation). The standard of AMD diagnosis in epidemiological studies is to collect color fundus images of the participants' eyes, which are then manually classified as a disease stage according to standardized protocols. There are various classification systems that differ in their definition of early AMD stages and it is subject of ongoing research which definition of early disease stages predicts progression towards late AMD best (Klein et al., 2014; Brandl et al., 2018; Thee et al., 2020). All stages of AMD can appear in a single or both eyes of a person; it is also possible that both forms of late AMD occur in a single eye. Therefore, AMD is an example of a bilateral disease that can affect neither, one, or both entities of a paired organ.

When interest lies in the investigation of person-specific risk factors for bilateral diseases, a person-specific disease status is often defined as the disease status of the worse entity. In case of binary entity-specific information (each entity is or is not affected), this corresponds to disease occurrence in at least one entity of the organ. Such a *worse-entity status* can then be analyzed with regression models for binary or categorical outcomes.

In Chapter 2 and 3, we analyze binary bilateral disease data on the occurrence of worse-eye *any AMD*. Let $Z_{1,i}, Z_{2,i} \in \{0,1\}$ be the true disease status of each of the two eyes of person $i = 1, \ldots, n$, e.g., $Z_{l,i} = 1$ indicates than an eye is affected by AMD and $Z_{l,i} = 0$ indicates that an eye is not affected by AMD; $l \in \{1,2\}$ for the two eyes of a person. Note that the notation Z was used in Section 1.2 for error-free covariates as it is common in many publications on measurement error. To remain consistent with the notations in the publications from Chapters 2 and 3, we now use Z_l as notation for the entity-specific disease indicators that define the overall outcome. Taken together, $Z_{1,i}$ and $Z_{2,i}$ define four different patterns of disease occurrence $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$. The worse-entity disease status of participant i is defined as $Y_i := max(Z_{1i}, Z_{2i})$, which means that the participant has the disease ($Y_i = 1$) if any or both entities are affected, and no disease ($Y_i = 0$) otherwise.

Using logistic regression to model the worse-entity outcome corresponds to the assumption that $Y_i|x_i \sim \mathrm{B}(\pi_i)$, where the success probability π_i of the Bernoulli distribution is modeled based on a a linear predictor and the logistic response function as $\pi_i = 1/(1 + exp(-x_i'\beta)) = H(\eta_i)$ and consequently $P(Y_i = 1|x_i) = \pi_i$. With respect to the different pattern of disease occurrence, it follows that $P(Z_{1,i} = 0, Z_{2,i} = 0|x_i) = P(Y_i = 0|x_i) = 1 - \pi_i$. Furthermore, one can derive $P(Z_{1,i} = 1, Z_{2,i} = 1|x_i) = P(Z_{1,i} = 1, Z_{2,i} = 1|x_i, Y_i = 1) \times P(Y_i = 1|x_i) = \delta_i \times \pi_i$, where δ_i is the conditional probability of disease in both entities, given disease in at least one entity. Lastly, assuming symmetric probabilities for disease in one but not the other entity gives

$$\begin{array}{c|ll} P(\cdot,\cdot|x_i) & Z_{2,i} = 1 & Z_{2,i} = 0 \\ \hline Z_{1,i} = 1 & \delta_i \times \pi_i & \frac{1-\delta_i}{2} \times \pi_i \\ Z_{1,i} = 0 & \frac{1-\delta_i}{2} \times \pi_i & 1 - \pi_i \end{array} \tag{1.6}$$

as the conditional probability mass function for all potential disease pattern.

Estimation and inference for the regression parameters β can be done based on standard likelihood inference for logistic regression using the observed worse-entity outcomes y_i or by setting up the likelihood based on the entity-specific outcome observations

$$\begin{aligned} L(\delta_i, \beta)_i = \{\delta_i H(\eta_i)\}^{z_{1,i} z_{2,i}} \times \left\{\frac{1-\delta_i}{2} H(\eta_i)\right\}^{z_{1,i}(1-z_{2,i})+(1-z_{1,i})z_{2,i}} \\ \times \left\{1 - H(\eta_i)\right\}^{(1-z_{1,i})(1-z_{2,i})}. \end{aligned} \tag{1.7}$$

Both approaches are equivalent for the estimation of $\hat{\beta}$, and consideration of δ_i is irrelevant when data on Z_1 and Z_2 (and consequently the worse-entity outcome Y) is available for all observational units. The latter formulation of the likelihood becomes, however, relevant when accounting for missing inforation in single entity outcomes (see below).

The true outcome Y in bilateral logistic regression can be affected by misclassification out of two different reasons. First, due to ignoring missing disease information in one of two entities and utilizing the disease classification of a single entity as person-specific response. Second, due to misclassified observations for the single entities. Both problems can be adressed by setting up an an adequate likelihood for the estimation of associations of exposures/risk factors with disease.

In many epidemiological or clinical studies on bilateral diseases, entity-specific disease information is missing for one of two entities in a subset of study participants. On the example of AMD, most studies have a certain fraction of study participants with missing AMD diagnosis in one of two eyes due to fundus images of inferior quality that can not be classified with respect to their AMD status, competing retinal diseases hampering AMD diagnosis, or missing images. This missing data problem is often ignored and the single observed entity, z_{1i} or z_{2i} is used as outcome instead of the actual worse-entity $y_i = max(z_{1i}, z_{2i})$. With this analysis strategy, the binary response Y is inconsistently defined between the two subsets of participants: for individuals with missing information the response corresponds to the occurrence of disease in a single entity, for individuals with full data it corresponds to disease occurrence in at least one of the two entities. This missing data problem can be treated in the context of response misclassification models. Assume that the observed disease status for participants j with missing data in one of two entities is an error-prone observation of the binary outcome, and that the observed disease status is selected randomly from the two entities independent of their disease status, $Y_j^* = z_{r,j}$. The specificity of this observed outcome, $P(Y_j^* = 0|Y_j = 0)$, is one, since, by definition, both eyes (and consequently a randomly selected eye) is unaffected from AMD in case of $Y_j = 0$. The sensitivity is, however, in general smaller than one, $P(Y_j^* = 1|Y_j = 1) = (1+\delta_i)/2 \leq 1$, when the probability of disease in only one entity given disease in at least one entity is bigger than 0 (i.e., as soon as disease occurs occasionally in only one of two entities, that is $\delta_i < 1$). Note that the specificity of $P(Y_j^* = 1|Y_j = 1) = (1 + \delta_i)/2$ implies differential misclassification when the probability of disease in both entities given disease in at least one entity, δ_i, is associated with a risk factor of disease, or when the probability of observing the disease status only in one of the two entities is associated with a risk factor. The latter would imply a specificity that is on average associated with the risk factor due to the changing probability of missing data and the specificity of one for observations without missing data.

The missing data problem can be addressed by optimizing a likelihood that explicitly accounts for the fact of missing information in single entities in a subset of individuals. Based on equation (1.6) one can derive the probability of disease in a single observed entity as $P(Z_{r,j} = 1|x_j) = (1/2 + 1/2 \times \delta_j)H(\eta_j)$ and the likelihood contributions of individuals j with missing information for single entities is

$$L(\delta_j, \beta)_j = \left\{\left(\frac{1}{2} + \frac{1}{2}\delta_j\right)H(\eta_j)\right\}^{z_{r,j}} \times \left\{1 - \left(\frac{1}{2} + \frac{1}{2}\delta_j\right)H(\eta_j)\right\}^{1-z_{r,j}}. \qquad (1.8)$$

The full likelihood is then the product over all contributions of individuals with information on both entities based on equation (1.7) and for the individuals with missing information on single entities based on equation (1.8). The former observations provide information for modeling δ_j, which can, e.g., be done parametrically using the logistic function and accounting for differential misclassification when the (conditional) probability of disease in both entities is associated with risk factors of interest. More details and an application example of this kind of analysis is presented in the contributing article in Chapter 2 of this thesis. There, it is also shown based on simulation studies that the correct specification of the model for δ_j is crucial for obtaining unbiased estimates of the regression parameters β.

The second source of response misclassification in bilateral logistic regression, errors in the entity-specific classification, can be addressed in the framework of likelihood-based correction approaches as well. Assume that instead of the true $(Z_{1,i}, Z_{2,i})$ error-prone observations $(Z^*_{1,i}, Z^*_{2,i})$ were collected and the misclassification process is described based on the classification probabilities

$$\begin{aligned} &P(Z^*_{l,i} = 1 | Z_{l,i} = 1, x_i) = \pi_{1,i} \\ &P(Z^*_{l,i} = 0 | Z_{l,i} = 0, x_i) = \pi_{0,i}, \ l \in \{1,2\}, \end{aligned} \tag{1.9}$$

where the classification probabilities of the observed error-prone outcomes is independent, given the true entity-specific outcomes, $P(Z^*_{1,i} = z^*_{1,i}, Z^*_{2,i} = z^*_{2,i} | Z_{1,i} = z_{1,i}, Z_{2,i} = z_{2,i}, x_i) = P(Z^*_{1,i} = z^*_{1,i} | Z_{1,i} = z_{1,i}, x_i) P(Z^*_{2,i} = z^*_{2,i} | Z_{2,i} = z_{2,i}, x_i)$. The likelihood can then be set up based on the conditional probabilities of the observed data

$$\begin{aligned} P(Z^*_{1,i} = z^*_{1,i}, Z^*_{2,i} = z^*_{2,i} | x_i) = \sum_{z_{1,i}, z_{2,i} \in \{0,1\}} \Big\{ &P(Z^*_{1,i} = z^*_{1,i} | Z_{1,i} = z_{1,i}, x_i) \times \\ &P(Z^*_{2,i} = z^*_{2,i} | Z_{2,i} = z_{2,i}, x_i) \times \\ &P(Z_{1,i} = z_{1,i}, Z_{2,i} = z_{2,i} | x_i) \Big\}, \end{aligned} \tag{1.10}$$

where the first two multiplicative terms correspond to the classification probabilities from equation (1.9) (*misclassification model*) and the last term to the *true data model* as in (1.6). If an error-prone classification is only observed for one (randomly selected) entity, $Z^*_{r,j} = z^*_{r,j}$, the conditional probability of this outcome is given by

$$P(Z^*_{r,j} = z^*_{r,j}) = \sum_{z_{r,j} \in \{0,1\}} P(Z^*_{r,j} = z^*_{r,j} | Z_{r,j} = z_{r,j}, x_j) P(Z_{r,j} = z_{r,j} | x_j), \tag{1.11}$$

where $P(Z_{r,j} = z_{r,j} | x_j)$ is defined as above. As for standard logistic regression with response misclassification, parameters of the misclassification model can not be reliably estimated from observed error-prone data alone, but a likelihood based on equation (1.10) can be used for performing sensitivity analyses under different assumptions on the classification probabilities. Such an analysis is presented and discussed in Chapter 2.

In the contributing article of Chapter 3, we examine data of a study, in which error-prone eye-specific AMD classifications are available based on predictions of an externally pre-trained convolutional neural network for all participants. In addition, the fundus images of a subset of individuals were also manually classified towards their AMD status by an experienced ophthalmologist (*gold standard* classification, considered as true disease status). This corresponds to eye-specific internal validation data, and we show how such information can be utilized to simultaneously estimate the association of risk factors with the true worse-entity disease status (worse-eye AMD) and parameters of the misclassification model, also accounting for differential entity-specific misclassification. Conceptually, this is similar to the likelihood approach for logistic regression with internal validation data as in equation (1.4), but is based on deriving the likelihood contributions for individuals with all possible subsets of available response observations $(z^*_{1,i}, z^*_{2,i}, z_{1,i}, z_{2,i})$. We compare results of misclassification adjusted and unadjusted estimates and show based on the concrete scenario that conclusions drawn from unadjusted analyses can be highly misleading, especially in case of differential misclassification in the entity-specific observations.

1.3 Genome-wide association studies

Genome-wide association studies (*GWAS*) are a widely established and successful experimental design for the identification of the association of genetic variants or regions (*loci*) with human traits and diseases. Detected associations have led to the discovery of novel biological mechanisms and GWAS findings have diverse clinical applications, as they are, for example, increasingly used for identifying individuals at high risk of developing certain diseases (Visscher et al., 2017; Tam et al., 2019; McCarthy et al., 2008). The formation of international consortia and the adoption of relatively standardized evaluation routines for GWAS have made results comparable and relatively easy to combine in meta-analyses. This enables reproducibility of results from single studies and leads to the large sample sizes that are necessary to continue to make progress in the discovery of genetic associations (Evangelou and Ioannidis, 2013).

The general idea of GWAS is to investigate the association of large sets of genetic variants (typically single-nucleotide polymorphisms, *SNPs*) spread over the whole genome for association with a specific trait, e.g., a disease in a case-control setting. SNPs are variations of single nucleotides at specific base positions in the genome that occur for a specific fraction from a population. Therefore, they correspond to locations in the genome where genotypes vary between individuals. At a given SNP position, each of the two nucleotides of the base pair can take one of two alleles, thus one of three allele combinations can occur for an individual (e.g., CC, CT, TT). The number of variants that define variation between genomes is huge, the 1000 Genomes Project Consortium et al. (2015) characterized genomic variation based on 88 million variants and found that a typical genome differs from the reference human genome at 4.1 million to 5.0 million sites (depending on the ancestry). The different variants (SNPs) are, however, not independent of each other but show a correlation structure (*linkage disequilibrium*). To describe the genome of a person, it is therefore not necessary to measure the variants/alleles at each SNP. Instead, it is possible to use SNP-arrays that include, e.g., 200.000 to 2 million SNPs (Visscher et al., 2017) and exploit the linkage disequilibrium for imputing genotypes of unobserved genetic variants.

After imputation, a GWAS is performed by screening all considered genetic variants for association with the phenotype of interest (see below). Associated variants are *identified* by a stringent significance threshold to account for multiple testing. This can lead to the detection of associations (*signals*) at one or several genetic regions. When interpreting the results, it is important to consider that detected variants are not necessarily causally associated with the phenotype of interest and also do not clearly indicate one specific gene or biological mechanism that leads to the detected association. Once a specific locus is identified to be associated with a phenotype, additional steps are required to identify potentially causal variants and their target genes, including statistical, bioinformatic, functional or evolutionary genetic analyses (Gallagher and Chen-Plotkin, 2018; Tam et al., 2019).

1.3.1 Statistical methods in GWAS

In order to estimate and test the association of the individual genetic variants with the phenotype under investigation, different approaches can be considered in principle. For a binary phenotype (as in our application in Chapter 3), it is now most common to estimate the odds ratio per risk allele by logistic regression, assuming that the number of "risk alleles" $\in \{0, 1, 2\}$, or the quantitative *dosage* $\in [0, 2]$ in case of imputed variants, has an additive, linear effect on the log odds of the response. The association can then be tested, for example, by Wald or likelihood ratio tests. The use of logistic regression has the advantage that estimated odds ratios can be compared between different types of studies and further covariates can be taken into account when estimating the genetic association. This provides one solution for the common problem of *population stratification* in the analyzed GWAS data. The background of this problem is that alleles of certain genetic variants occur with different frequencies in different population groups. If the phenotype (e.g., the occurrence of a disease) also occurs at different frequencies in the sub-populations (for other reasons), this leads to confounding of the estimated genetic association. By taking into account the population structure in the data through appropriate covariates (e.g., Price et al., 2006), it is possible to adjust for confounding. It is also possible to include other covariates into the model, e.g., known and strong (clinical) risk factors of disease, for example, the age of the individuals. If those risk factors are independent of the genetic variants, inclusion is not necessary for the estimation of the genetic association but can increase power of the statistical tests for continuous phenotypes and in some situations (e.g., in population based studies) also for binary phenotypes (Pirinen et al., 2012).

Due to the large number of analyzed variants, GWAS suffer from a large *multiple-testing burden* (Tam et al., 2019). In practice, the multiple testing problem is usually addressed by comparing the results of the association tests with the p-value threshold of $p \leq 5 \times 10^{-8}$. Only in this case is one speaking of a *confirmed genome-wide significant* association. This is based on the assumption that approximately 1 million *independent* statistical tests are performed in a (European-descent) GWAS for common variants. Under this assumption, the p-value threshold of $p \leq 5 \times 10^{-8}$ corresponds to controlling the family-wise error rate at ≤ 0.05. The (theoretical) appropriateness of this threshold can be debated, for example, as the use of state-of-the-art sequencing increases genomic coverage and increasingly rare variants are studied. As a result, more *independent* tests are performed in a screening. On the other hand, there are also proposals to increase statistical power based on controlling false discovery rates or incorporating prior information. Overall, however, the fundamental problem exists that very large sample sizes are required to provide convincing evidence for the existence of a true association in genome-wide scans, especially considering the commonly small effect sizes of single variants. Especially for phenotypes that are difficult and complex to collect, this can lead to high costs and be a hurdle for the successful implementation of GWAS.

1.3.2 Measurement error in GWAS

There exists relatively little work on consequences of measurement error and misclassification for GWAS results or applied work on adjusting for measurement errors in such analyses. However, the estimation of association between the phenotypes (response) and genetic variants (covariates) is often based on standard models, e.g., linear regression for quantitative phenotypes and logistic regression for binary phenotypes. Therefore, results on the consequences of measurement error for parameter estimates from these models are in general transferable. Since GWAS are fundamentally about discovering associations between a phenotype and genetic regions, and not about perfectly quantifying the association/effects of individual variants, it can be argued that some bias in the association estimates is in general acceptable and unavoidable. However, especially claiming the existence of false positive associations would be problematic, as this can also lead to a waste of resources in follow-up studies. The reduction of statistical power due to the existence of measurement error is, of course, a fundamental problem and should serve as motivation for high-quality and consistent measurement.

With respect to the genetic variants, data are generally assumed to be rather accurately measured. Genotyping error based on SNP microarrays occurs with rather low frequencies (Hong et al., 2012) and can in most cases assumed to be non-differential with respect to the phenotype. Such an error reduces statistical power for detecting associations, but effects are expected to be rather small and is in not necessary to explicitly account for the error during analysis (Heid et al., 2008). Differential measurement error might, however, occur in case-control studies where genotype data on cases and controls stem from different data sources and can yield increased type-1 error probabilities (Moskvina et al., 2006). In the context of single studies and meta-analyses for GWAS data, standardized protocols for quality control of the genetic data have been proposed (Anderson et al., 2010; Winkler et al., 2014). These are widely used in practice to prevent erroneous conclusions due to limited quality of the collected and imputed genetic data.

Regarding measurement error in the phenotype variable, it is established that a lack of precision in measurement, or misclassification, can affect the association estimates, reducing statistical power in case of non-differential error (Barendse, 2011; Liao et al., 2014; Edwards et al., 2005; Rekaya et al., 2016). One consequence is that between study heterogeneity in the phenotype measurements/definitions can make it difficult to compare their results. This includes the problem that locations of detected associations (*lead variants*) may shift between studies due to inconsistent or erroneous measurement (Barendse, 2011). In Chapter 3 of this thesis, we illustrate the problem of phenotype misclassification in a GWAS for age-related macular degeneration, a phenotype that is expensive to collect because fundus images have to be manually classified by ophthalmologists with respect to the disease. In our example, we instead use a pre-trained neural network to classify the images and perform the GWAS based on the resulting error-prone response data. Such an approach is appealing as it allows cheap phenotype measurement in large studies, but can yield biased results. Using an internal validation data set with manual gold-standard classifications, we investigate the misclassification and adjust the association estimates of detected variants based on the maximum likelihood approach from Section 1.2.2. In doing so, we find a false positive association in the unadjusted analysis, which is due to a differential performance of the classification algorithm with respect to a genetic variant, more specifically with respect to eye color, which is strongly associated with this variant.

1.4 Infectious disease surveillance

The spread of the novel SARS-CoV-2 virus and the resulting COVID-19 pandemic have profound implications for humans and societies worldwide. Since early 2020, policy makers, public health, and scientific institutions have faced the challenge of evaluating existing surveillance data in real-time to assess the current pandemic situation as a basis for appropriate responses. Because the readily available data often do not come from well-designed studies, evaluation and interpretation are challenging, and inadequate analysis can also lead to erroneous conclusions. Key problems include delays in the reporting data, underreporting and lack of representativeness, and also measurement error and misclassification, e.g., in diagnostic tests. The contributing articles in Chapters 4 and 5 discuss ways to account for at least some of those aspects in the real-time analysis of available data. As an introduction, this section presents the available surveillance data in Germany and introduces basic statistical concepts for the real-time analysis of reporting data on the spread of infectious diseases.

1.4.1 Surveillance data

Infectious disease surveillance data, due to their nature, contain some peculiarities that need to be taken into account for an adequate analysis and interpretation of the numbers. In the following, some details of the official reporting data on the COVID-19 pandemic in Germany are presented, which also motivate the contributions of the work from Chapters 4 and 5. Similar issues arise for data on other diseases and from other countries, as the surveillance systems are often similar.

For SARS-CoV-2, each infected person goes through different stages, as illustrated in Figure 1.1 for a symptomatic case. After infection and an incubation period, there is a disease onset, followed by recovery or severe disease progression, potentially leading to death. Testing for SARS-CoV-2 infection, subsequent registration and reporting in case of a positive test result is usually after symptom onset, as testing is in many cases symptom based. However, as part of contact tracing and increasingly available screening tests, some cases may also be identified in the presymptomatic stage and asymptomatic cases might be detected as well.

In Germany, legally defined by the Infection Protection Act (IfSG), different information is recorded for each case, so that each infected case potentially has multiple time stamps in the surveillance data. Each record contains the *registration date* on a specific date when a positive PCR test result is reported to the local health authorities (at the patient's place of residence). Due to the federal structure of the German healthcare system, it is then passed on to the respective state counterpart and finally to the federal health authority. This reporting chain is not generally digitized and, therefore, reporting between each stage can yield delays and the reporting/registration date might vary between different data sources. Those dates are either given in the data or can be derived from comparing data sets of consecutive days. Information on the date of disease onset (symptom onset) is also recorded in the German surveillance data. This is done during registration or retrospectively, based on personal communication between health authorities and affected individuals. This date is, however, only available in about 50% of all cases (as of March, 2021), either because no information was collected or because the corresponding infections are identified in the pre- or asymptomatic stage. In addition to information on reporting dates and onset of disease, publicly available reporting data also contains information on death counts with respect to their time of case registration, but these are not considered further in the work presented here.

The particular temporal structure must be taken into account in any analysis of the

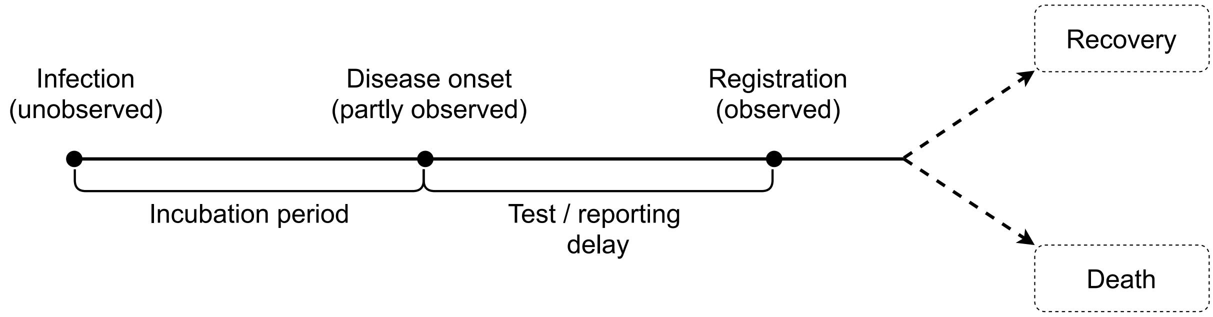


Figure 1.1: Schematic time course of a SARS-CoV-2 infection for an individual with symptomatic course of infection.

surveillance data and can pose a challenge with regard to the temporal allocation and interpretation of the data. Particularly in the context of real-time analysis of the surveillance data, it is necessary to check exactly which numbers are already complete at a suitable reference time and which information is still incomplete and might change in future. If this is not taken into account, the conclusions drawn from the data can be systematically biased. This is relevant, e.g., for the real-time determination of the epidemic curve, as the number of persons must be taken into account for whom disease onset has already occurred but who have not yet been reported (see Section 1.4.2 and the contributing article in Chapter 4). However, similar considerations also apply to seemingly simpler tasks, such as the calculation of reporting incidence, for which a reference date (registration date) must be used on which there is no subsequent change in the reported number of new cases.

There are two other key aspects that need to be considered when interpreting case reporting data. First, such data suffer from underreporting, as not all infected persons are covered by the health and testing system. This is especially true for infectious diseases, for which some of the infected have only mild courses of disease or are even asymptomatic (Gibbons et al., 2014). If underreporting changes over time, for example, due to changes in the testing regime, this can distort conclusions drawn from the data regarding the dynamics of infection spread. Second, the surveillance data consist of the number of individuals who tested positive for the infectious disease. Such diagnostic tests can yield erroneous results and therefore false positives (uninfected individuals) can be included in the data as well as actual infected individuals not being included in the data (false negatives). Erroneous diagnostic tests have consequences for the individuals tested, since in the case of SARS-CoV-2, consequences such as isolation or quarantine are drawn from the test results, or false negative individuals can pass on the infection to personal contacts. The predictive values of the test (probability of correct classification) depend not only on the quality of the test (sensitivity and specificity) but also on the actual prevalence of infection or the *pre-test probability* in the corresponding testing regime (see e.g., Watson et al., 2020, for the concrete example of SARS-CoV-2). Additionally, the erroneous tests, especially with changes in the number of tests performed over time, migh potentially also bias apparent trends in the surveillance data. The plausible magnitude of such bias for the German

SARS-CoV-2 surveillance data is investigated in more detail in the contributing article in Chapter 5.

1.4.2 Real time estimation of the epidemic curve - nowcasting

The epidemic curve is a central indicator for describing the dynamics of the spread of an infectious disease over time. It is defined as the time series of the number of infected with disease onset per time interval (usually per day or week). Compared to the time series of newly reported cases, the epidemic curve better reflects the dynamics of disease spread, since the onset of a person's disease differs from the time of infection only by the incubation period. The (distribution of the) incubation time can plausibly be assumed to be stable over time in many situations. Given a valid estimate of the epidemic curve, it is also possible to estimate the time-series of the number of infections based on *backprojection* methods (e.g., Küchenhoff et al., 2021). For the time series of newly reported cases, there is the additional delay due to testing and reporting that has to be taken into account for interpretation. This delay may also change over time, e.g., if reporting structures change because the burden on health systems varies over time.

For the estimation of the epidemic curve in close to real-time, it is important to account for the reporting delay in the case reporting data. On a specific day T the data on individuals with disease onset on day $t \leq T$ is not complete and underreporting is biggest for days t close to T. Ignoring the reporting delay and focusing on the available data only leads to a systematic underestimation of the number of disease onsets close to the current time T.

Figure 1.2 shows a schematic representation of the available data on a given day T. Let $N_{t,d} = n_{t,d}$ be the number of cases, with disease onset on day t and reported with a delay of d days (case report arrives on day $t + d$). On the "current" day T, information is available on $N(t,T) = \sum_{d=0}^{T-t} n_{t,d}$ cases that had disease onset on day t and are reported until day T. The aim of nowcasting is to predict the unobserved total number of cases with disease onset on day t, $N(t,\infty) = \sum_{d=0}^{\infty} N_{t,d}$, based on information available up until the current day T.

Therefore, the general task of nowcasting is to set up a model for the count data $n_{t,d}$, $t \leq T$, $d \leq D$ that are observed for $t + d \leq T$ and unobserved for $t + d > T$. For identifiability reasons, one defines a maximum relevant delay of $d = D$ and considers each observation with an observed delay $> D$ as having a delay of D days. Based on such a model, it is then possible to draw inference about the quantity of central interest, the number of new disease onsets on day t, $N(t,D) = \sum_{d=0}^{D} n_{t,d}$, $t = 1,\ldots,T$. Different approaches have been proposed in the literature that can be applied to surveillance data with different levels of complexity, e.g., Lawless (1994); Höhle and an der Heiden (2014); Bastos et al.

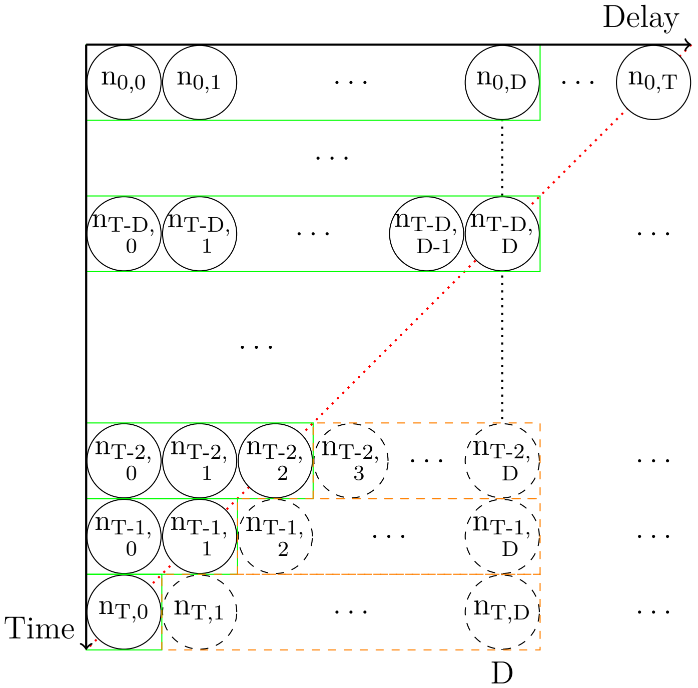


Figure 1.2: Schematic representation of the data situation in a nowcasting task at a given day T. The circles $n_{t,d}$ represent the number of events occurred on day t and reported with a delay of d days. Solid circles correspond to the observed data, dashed circles indicate occurred-but-not-yet-observed events. The green solid boxes indicate the sum of events on day $t \leq T$ that have already been observed, the dashed orange boxes correspond to the sum of events that have not been observed until day T. Within the model, a maximum reporting delay of D days is considered.

(2019); McGough et al. (2020). The need for providing timely and reliable information on the current SARS-CoV-2 pandemic has generated increased interest in nowcasting methods. Nowcasting is used in the official surveillance system in Germany (an der Heiden and Hamouda, 2020), and recent publications on the topic present use cases for nowcasting of infection and death counts in different regions. Partly, they also evaluate the performance of the models and/or suggest methodological developments and adaptations, e.g., Greene et al. (2021); Schneble et al. (2020); Seaman et al. (2020); Hawryluk et al. (2021).

In the contributing article from Chapter 4, we proposed a Bayesian hierarchical nowcasting model for the real-time estimation of the epidemic curve for the Bavarian COVID-19 data that builds up on the work of Höhle and an der Heiden (2014) and McGough et al. (2020). Let the expected number of new disease onsets on day t be $E(N(t, D)) = \lambda_t$. Additionally, let the probability of a reporting delay of d days for an individual with disease onset on day t be $P(\text{delay} = d|\text{disease onset} = t) = p_{t,d}$. We propose to model the reported case counts $N_{t,d}$ as Negative-binomial distributed with expectation

$E(N_{t,d}) = \lambda_t \times p_{t,d}$ and overdispersion parameter ϕ. For modeling λ_t we account for the expected smoothness of the epidemic curve based on a first-order random walk on the log-scale, $\log(\lambda_t)|\lambda_{t-1} \sim \mathrm{N}(\log(\lambda_{t-1}), \sigma^2)$, and model the time-varying reporting delay distribution based on a discrete time hazard model. This flexible nowcasting approach allows to account for changes in the reporting delay distribution over time and allows to directly incorporate specific features of the reporting system, such as reduced reporting of cases on weekends, into the model. Allowing for overdispersion in the reporting numbers enables adequate quantification of uncertainty for the estimation of the epidemic curve. These aspects have been found to be central to good performance in a retrospective evaluation of the model and a comparison with simpler nowcasting approaches.

1.4.3 Estimation of the effective reproduction number R_t

The basic reproduction number R_0 is an established metric for characterizing the transmissibility of an infectious agent in a population. It describes the average number of secondary cases a single infected case produces in a completely susceptible population. It is important to understand that R_0 is not a biological constant of a pathogen, but should be interpreted in relation to the population and, e.g., its contact behavior (Delamater et al., 2019). As an infectious disease outbreak continues over time, the average number of secondary infections from an infected case changes, as the number of susceptibles decreases. Furthermore, the behavior and contact frequency of the population can change over time, e.g., due to implementation or relaxation of containment measures. The (time-varying) effective reproduction number R_t is then used to describe the expected number of infections caused by an infected person at time t and serves as a measure to describe the dynamics of the outbreak over time.

By estimating the effective reproduction number from existing data, the (time series) of $\hat{R}_t$ can be used to retrospectively analyze the course of the spread of an infectious disease or to describe the current infection dynamics in (near) real-time. Estimates of the effective reproduction number can either be derived from fitting mechanistic transmission models/compartmental models to available data, or by directly estimating it from time-series of, e.g., case counts. Two popular approaches for the latter are Wallinga and Teunis (2004) and Cori et al. (2013). These approaches are relatively straight-forward to apply to existing time series of infection numbers and require as a central assumption information on the serial interval (or the generation time interval), that is the distribution of the time between the disease onset (or infection) of two generations of infected individuals. For interpretation, it must be taken into account that both approaches estimate a slightly different version of the effective reproduction number. The method by Wallinga and Teunis (2004) estimates the so-called *case-* or *cohort* reproduction number. It is a direct generalization of the basic reproduction number R_0 and describes the average number of secondary cases infected from a person who had their disease onset at time t. Cori et al.

(2013) estimate the *instantaneous* reproduction number, which uses a backward-looking perspective to describe infectiousness at the current time t, assuming that current infection dynamics behave as they did in the immediate past. This leads in most situations mainly to a *horizontal* time-shift in the estimated effective reproduction number $\hat{R}_t$, but has to be considered for interpretation and comparison of different analyses.

The interpretation of the effective reproduction number is generally straightforward and it is therefore an appealing metric for communicating the current dynamics of disease spread. However, estimating the current reproduction number based on surveillance data can be challenging, in part because of underreporting in the underlying data near the current time (Gostic et al., 2020). In the contributing article of Chapter 4, we therefore propose to estimate $\hat{R}_t$ based on the reporting delay adjusted epidemic curve from the nowcast model. In doing so, we propose to account for the associated uncertainty by repeatedly estimating the effective reproduction number for draws from the posterior of the epidemic curve, and averaging over the results for inference and uncertainty quantification.

1.5 Bibliography

1000 Genomes Project Consortium et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.

an der Heiden, M. and Hamouda, O. (2020). Schätzung der aktuellen Entwicklung der SARS-CoV-2- Epidemie in Deutschland – Nowcasting. *Epidemiologisches Bulletin*, 2020(17):10–15.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564–1573.

Barendse, W. (2011). The effect of measurement error of phenotypes on genome wide association studies. *BMC genomics*, 12(1):1–12.

Bastos, L. S., Economou, T., Gomes, M. F., Villela, D. A., Coelho, F. C., Cruz, O. G., Stoner, O., Bailey, T., and Codeço, C. T. (2019). A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in medicine*, 38(22):4363–4377.

Berkson, J. (1950). Are there two regressions? *Journal of the american statistical association*, 45(250):164–180.

Brandl, C., Zimmermann, M. E., Günther, F., Barth, T., Olden, M., Schelter, S. C., Kronenberg, F., Loss, J., Küchenhoff, H., Helbig, H., et al. (2018). On the impact of different approaches to classify age-related macular degeneration: Results from the German AugUR study. *Scientific reports*, 8(1):1–10.

Buzas, J. S., Stefanski, L. A., and Tosteson, T. D. (2005). Measurement error. In *Handbook of epidemiology*, pages 729–765. Springer.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective.* CRC press.

Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology*, 178(9):1505–1512.

Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T., and Jacobsen, K. H. (2019). Complexity of the basic reproduction number (R0). *Emerging infectious diseases*, 25(1):1.

Edwards, B. J., Haynes, C., Levenstien, M. A., Finch, S. J., and Gordon, D. (2005). Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC genetics*, 6(1):1–12.

Evangelou, E. and Ioannidis, J. P. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6):379–389.

Gallagher, M. D. and Chen-Plotkin, A. S. (2018). The post-GWAS era: from association to function. *The American Journal of Human Genetics*, 102(5):717–730.

Gibbons, C. L., Mangen, M.-J. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M., et al. (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC public health*, 14(1):1–17.

Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., De Salazar, P. M., et al. (2020). Practical considerations for measuring the effective reproductive number, Rt. *PLoS computational biology*, 16(12):e1008409.

Grace, Y. Y. (2016). *Statistical analysis with measurement error or misclassification*. Springer.

Greene, S. K., McGough, S. F., Culp, G. M., Graf, L. E., Lipsitch, M., Menzies, N. A., and Kahn, R. (2021). Nowcasting for Real-Time COVID-19 Tracking in New York City: An Evaluation Using Reportable Disease Data From Early in the Pandemic. *JMIR public health and surveillance*, 7(1):e25538.

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press.

Hawryluk, I., Hoeltgebaum, H., Mishra, S., Miscouridou, X., Schnekenberg, R. P., Whittaker, C., Vollmer, M., Flaxman, S., Bhatt, S., and Mellan, T. A. (2021). Gaussian Process Nowcasting: Application to COVID-19 Mortality Reporting. *arXiv preprint arXiv:2102.11249*.

Heid, I. M., Lamina, C., Küchenhoff, H., Fischer, G., Klopp, N., Kolz, M., Grallert, H., Vollmert, C., Wagner, S., Huth, C., et al. (2008). Estimating the single nucleotide polymorphism genotype misclassification from routine double measurements in a large epidemiologic sample. *American journal of epidemiology*, 168(8):878–889.

Höhle, M. and an der Heiden, M. (2014). Bayesian nowcasting during the STEC O104: H4 outbreak in Germany, 2011. *Biometrics*, 70(4):993–1002.

Hong, H., Xu, L., Liu, J., Jones, W. D., Su, Z., Ning, B., Perkins, R., Ge, W., Miclaus, K., Zhang, L., et al. (2012). Technical reproducibility of genotyping SNP arrays used in genome-wide association studies. *PloS one*, 7(9):e44483.

Keogh, R. H., Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., Küchenhoff, H., Tooze, J. A., Wallace, M. P., Kipnis, V., et al. (2020). STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: part 1—basic theory and simple methods of adjustment. *Statistics in medicine*, 39(16):2197–2231.

Klein, R., Meuer, S. M., Myers, C. E., Buitendijk, G. H., Rochtchina, E., Choudhury, F., de Jong, P. T., McKean-Cowdin, R., Iyengar, S. K., Gao, X., et al. (2014). Harmonizing the classification of age-related macular degeneration in the three-continent AMD consortium. *Ophthalmic epidemiology*, 21(1):14–23.

Küchenhoff, H., Günther, F., Höhle, M., and Bender, A. (2021). Analysis of the early COVID-19 epidemic curve in Germany by regression models with change points. *Epidemiology and Infection*, 149:e68.

Lawless, J. (1994). Adjustments for reporting delays and the prediction of occurred but not reported events. *Canadian Journal of Statistics*, 22(1):15–31.

Liao, J., Li, X., Wong, T.-Y., Wang, J. J., Khor, C. C., Tai, E. S., Aung, T., Teo, Y.-Y., and Cheng, C.-Y. (2014). Impact of measurement error on testing genetic association with quantitative traits. *PloS one*, 9(1):e87044.

Lyles, R. H., Tang, L., Superak, H. M., King, C. C., Celentano, D. D., Lo, Y., and Sobel, J. D. (2011). Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. *Epidemiology (Cambridge, Mass.)*, 22(4):589.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, 9(5):356–369.

McGough, S. F., Johansson, M. A., Lipsitch, M., and Menzies, N. A. (2020). Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking. *PLoS computational biology*, 16(4):e1007735.

Mitchell, P., Liew, G., Gopinath, B., and Wong, T. Y. (2018). Age-related macular degeneration. *The Lancet*, 392(10153):1147–1159.

Moskvina, V., Craddock, N., Holmans, P., Owen, M. J., and O'Donovan, M. C. (2006). Effects of differential genotyping error rate on the type I error probability of case-control studies. *Human heredity*, 61(1):55–64.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, 86(4):843–855.

Paulino, C. D., Soares, P., and Neuhaus, J. (2003). Binomial regression with misclassification. *Biometrics*, 59(3):670–675.

Pirinen, M., Donnelly, P., and Spencer, C. C. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature genetics*, 44(8):848–851.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909.

Rekaya, R., Smith, S., El Hamidi Hay, N. F., and Aggrey, S. E. (2016). Analysis of binary responses with outcome-specific misclassification probability in genome-wide association studies. *The application of clinical genetics*, 9:169.

Rogan, W. J. and Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American journal of epidemiology*, 107(1):71–76.

Schneble, M., De Nicola, G., Kauermann, G., and Berger, U. (2020). Nowcasting fatal COVID-19 infections on a regional level in Germany. *Biometrical Journal*.

Seaman, S., Samartsidis, P., Kall, M., and De Angelis, D. (2020). Nowcasting CoVID-19 Deaths in England by Age and Region. *medRxiv*.

Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., Keogh, R. H., Kipnis, V., Tooze, J. A., Wallace, M. P., Küchenhoff, H., et al. (2020). STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2—More complex methods of adjustment and advanced topics. *Statistics in medicine*, 39(16):2232–2263.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484.

Thee, E. F., Meester-Smoor, M. A., Luttikhuizen, D. T., Colijn, J. M., Enthoven, C. A., Haarman, A. E., Rizopoulos, D., and Klaver, C. C. (2020). Performance of classification systems for age-related macular degeneration in the rotterdam study. *Translational vision science & technology*, 9(2):26–26.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1):5–22.

Wallinga, J. and Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of epidemiology*, 160(6):509–516.

Watson, J., Whiting, P. F., and Brush, J. E. (2020). Interpreting a covid-19 test result. *Bmj*, 369.

Winkler, T. W., Day, F. R., Croteau-Chonka, D. C., Wood, A. R., Locke, A. E., Mägi, R., Ferreira, T., Fall, T., Graff, M., Justice, A. E., et al. (2014). Quality control and conduct of genome-wide association meta-analyses. *Nature protocols*, 9(5):1192–1212.

Chapter 2

Response misclassification in studies on bilateral diseases

Chapter 2 discusses the problem of response misclassification in logistic regression for bilateral disease data. We show how misclassification due to missing disease information in single entities, or due to entity-specific misclassification with known misclassification probabilities can be accounted for in a maximum likelihood analysis.

Contributing article:
Günther, F., Brandl, C., Heid, I. M., & Küchenhoff, H. (2019). Response misclassification in studies on bilateral diseases. *Biometrical Journal, 61(4), 1033-1048.*

Supplementary material:
`https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201900039`

Author contributions:
Küchenhoff, Heid, and Günther devised the research question and work. Günther derived and implemented the maximum likelihood approach, conducted the data analysis, and wrote the first draft of the manuscript. All authors contributed to the interpretation of the results and to the writing and revision of the manuscript.

Received: 14 January 2019 | Revised: 8 April 2019 | Accepted: 19 April 2019

DOI: 10.1002/bimj.201900039

Biometrical Journal

RESEARCH PAPER

Response misclassification in studies on bilateral diseases

Felix Günther[1,2] | Caroline Brandl[2,3] | Iris M. Heid[2] | Helmut Küchenhoff[1]

[1]Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, München, Germany

[2]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany

[3]Department of Ophthalmology, University Hospital Regensburg, Regensburg, Germany

Correspondence
Felix Günther, Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Ludwigstraße 33, 80539 München, Germany; Department of Genetic Epidemiology, University of Regensburg, Franz-Josef-Strauß-Allee 11, 93053 Regensburg, Germany.
Email: felix.guenther@stat.uni-muenchen.de

Funding information
Foundation for the National Institutes of Health, Grant/Award Number: NIH-2017 R01 EY RES511967; Bundesministerium für Bildung und Forschung, Grant/Award Numbers: BMBF 01ER1206, BMBF 01ER1507; Deutsche Forschungsgemeinschaft, Grant/Award Number: DFG HE3690/5-1

Abstract

Misclassification in binary outcomes can severely bias effect estimates of regression models when the models are naively applied to error-prone data. Here, we discuss response misclassification in studies on the special class of *bilateral* diseases. Such diseases can affect neither, one, or both entities of a paired organ, for example, the eyes or ears. If measurements are available on both organ entities, disease occurrence in a person is often defined as disease occurrence in at least one entity. In this setting, there are two reasons for response misclassification: (a) ignorance of missing disease assessment in one of the two entities and (b) error-prone disease assessment in the single entities. We investigate the consequences of ignoring both types of response misclassification and present an approach to adjust the bias from misclassification by optimizing an adequate likelihood function. The inherent modelling assumptions and problems in case of entity-specific misclassification are discussed. This work was motivated by studies on age-related macular degeneration (AMD), a disease that can occur separately in each eye of a person. We illustrate and discuss the proposed analysis approach based on real-world data of a study on AMD and simulated data.

KEYWORDS

age-related macular degeneration, bilateral diseases, maximum likelihood, measurement error, response misclassification

1 | INTRODUCTION

This paper discusses response misclassification in *bilateral diseases*, that is, diseases that can occur in neither, one or both entities of paired organs. Examples are, amongst others, eye diseases or hearing impairment. The modelling of such paired binary outcomes is in general not straightforward, since the disease status of both entities of a person cannot be treated as independent conditional on covariates. Here, we discuss the situation in which interest lies in person-specific risk factors for disease occurrence, like age, sex, or genetic factors. In this case, a person is usually diagnosed as having the disease when at least one entity of the organ is affected. This *worse-entity* outcome can then be analysed with regression models for binary outcomes. Such a definition of the outcome bears, however, problems in case of missing disease information for one of the two entities. A naive analysis, that ignores the missing data problem and uses the disease status of the single observed entity as response for persons with missing information, can lead to misclassification in the outcome compared to the true worse-entity response. A second, related issue is response misclassification resulting from error-prone disease information for the single entities due to imperfect sensitivity or specificity of disease diagnosis. In the remainder of this paper, we will discuss response misclassification in bilateral diseases and its consequences based on the example of age-related macular degeneration (AMD) but our considerations are generally transferable to other diseases.

AMD is a degenerative disorder of the central retina and a leading cause of severe vision impairment in the older population (Lim, Mitchell, Seddon, Holz, & Wong, 2012). The clinical endpoint of the disease is late AMD, which can appear as a neovascular complication characterised by choroidal/sub-retinal ingrowth of diseased blood vessels or an atrophic form known as geographic atrophy of the retinal pigment epithelium. Late AMD is typically preceded by early AMD stages that are clinically asymptomatic and determined by differently sized yellowish deposits of extracellular material (drusen) and/or irregularities of the retinal pigment epithelium (hyper/hypopigmentation). The standard in epidemiological studies is to collect colour fundus images of the participants' eyes, which are then manually classified as a disease stage according to standardised protocols. There are various classification systems that differ in their definition of early AMD stages and it is still subject of research which definition of early disease stages predicts progression towards late AMD best (Klein et al., 2014; Brandl et al., 2018). All stages of AMD can appear in neither, one, or both eyes of a study participant; it is also possible that both forms of late AMD occur in the same single eye. The complexity of diagnosis gives rise to methodological questions regarding the definition of disease status in statistical models. Researchers often use logistic regression to estimate effects of risk factors for a binary outcome of early or late or any AMD. Most studies have a subset of participants where disease information is missing in one of the two eyes due to various reasons: colour fundus images can be ungradable with respect to AMD because of low image quality with regard to brightness, contrast, and focus or competing retinal diseases blurring the image (e.g., glaucoma). A common approach is to utilise the worse-eye disease status to define the participants' disease status when disease information is available for both eyes and the single eye disease status of the participants when disease information is available only for one eye (naive analysis). This approach corresponds to a definition of the binary response as *AMD in at least one eye* for participants without missing data. Compared to this response definition, an outcome derived from a single eye is misclassified towards no AMD, if the observed eye is unaffected, but the second unobserved eye of the study participant is affected by AMD. The response observations of study participants with missing disease diagnosis in one of two eyes are thus potentially misclassified compared to the "true" worse-eye disease status if the observed eye is unaffected.

From the literature on measurement error, it is known that response misclassification in regression models leads to biased effect estimates if it is not accounted for (see, e.g., Carroll, Ruppert, Stefanski, & Crainiceanu, 2006, Chapter 15). If the misclassification is non-differential, that is the misclassification probabilities do not depend on covariates, this leads, on average, to attenuated effect estimates. If the misclassification is differential, it is not possible to make any general statements regarding the direction of bias and spurious associations can occur (Neuhaus, 1999). In case of AMD and response misclassification from utilizing the single eye disease status for persons with disease information only available in one eye, differential misclassification can occur because of two reasons: (a) if the probability of missing single eye information varies between study participants, the probability of observing only an unaffected single eye when the worse eye is affected can vary as well; (b) if the (conditional) probability of being affected by AMD in both eyes (given AMD in at least one eye) varies between individual study participants, the probability of observing only an unaffected single eye when the worse eye is affected varies as well, even if the probability of missing single eye diagnosis is constant. Such settings appear quite plausible. This gives rise to serious concerns regarding the estimates of the naive modelling strategy and motivates the following derivation of a maximum likelihood approach that explicitly considers missing single eye diagnosis within analysis. Our approach is based on a consistent definition of the binary response as AMD occurrence in at least one eye (worse-eye diagnosis) and considers potential response misclassification for study participants with missing disease information in single eyes. In Section 2.1, we provide the likelihood of modelling worse-eye AMD occurrence by logistic regression, expressed in terms of observed single eye disease statuses per participant. Since we assume that similar data situations can also occur for data on different diseases, we formulate the derivations with respect to the more general notion of bilateral diseases. Section 2.2 focuses on missing disease diagnosis in single entities (e.g., in one of two eyes in the case of AMD) and the consequences of ignoring them in the naive modelling strategy. We derive the model-based conditional probability of disease in randomly selected single entities and propose to optimise a likelihood based on the derived conditional probabilities to estimate regression parameters. In Section 2.3, we discuss additional misclassification in the single entity disease diagnosis and, in Sections 3 and 4, we compare the different estimation approaches in a real data example of the AugUR study on AMD (Age-related diseases: Understanding genetic and nongenetic influences—a study at the University of Regensburg, Stark et al., 2015) and based on simulated data. In Section 5, we discuss the proposed modelling approach and its assumptions and give recommendations for application.

2 | METHODS AND MODELS

2.1 | General approach of modelling binary bilateral disease data

We assume a disease that we examine for each of two entities of paired organs, from which we deduce the disease status of the person. Let $Z_{1i}, Z_{2i} \in \{0, 1\}$ be the true and here assumed as observable disease status of each of the two entities for

TABLE 1 Probability mass function of disease patterns conditional on covariate vector $\mathbf{x}_i$, with $\pi_i = \text{H}(\mathbf{x}_i^t\boldsymbol{\beta})$ and $\delta_i = P(Z_{1i} = 1, Z_{2i} = 1|Y_i = 1, \mathbf{x}_i)$

$\mathbf{P}(\cdot,\cdot|\mathbf{x}_i)$	$\mathbf{Z_{2i} = 1}$	$\mathbf{Z_{2i} = 0}$
$Z_{1i} = 1$	$\delta_i\pi_i$	$\frac{1-\delta_i}{2}\pi_i$
$Z_{1i} = 0$	$\frac{1-\delta_i}{2}\pi_i$	$1 - \pi_i$

study participant i, where $Z_{li} = 1$ indicates disease at entity l of participant i and $Z_{li} = 0$ represents a healthy entity ($l = 1, 2$; $i = 1, \ldots, n$). Taken together, Z_{1i} and Z_{2i} define four different patterns of disease occurrence (0,0), (0,1), (1,0), and (1,1). In case of AMD, Z_{li} corresponds to the observed AMD disease status in eye l of participant i. The true worse-entity disease status of participant i is defined as $Y_i := \max(Z_{1i}, Z_{2i})$, which means that the participant has the disease ($Y_i = 1$) if any of the two entities or both are affected, and no disease ($Y_i = 0$) otherwise. In our AMD example, this corresponds to the AMD status in the worse eye, which is commonly used as the AMD status of a person.

When we are interested in factors that are associated with the disease status, we model Y_i as response based on a generalised linear model for binary outcomes, for example, a logistic regression model. The probability of disease occurrence in participant i is modelled depending on a k-dimensional person-specific covariate vector $\mathbf{x}_i$, $\pi_i := \text{P}(Y_i = 1|\mathbf{x}_i) = \text{H}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}) = \text{H}(\mathbf{x}_i^t\boldsymbol{\beta}) = \text{H}(\eta_i)$. $\text{H}(\cdot)$ is an appropriate response function, for example, the logistic function $\text{H}(\cdot) = 1/\{1 + \exp(-\cdot)\}$. As usual in the context of generalised additive models, this allows a quite flexible modelling of covariate effects including, e.g., dummy effects for binary or categorical covariates, linear and nonlinear effects of continuous covariates, and the inclusion of potential interaction terms. In the AMD example, one might be interested in the estimation of the association of age, sex, life-style factors like smoking, or genetic factors with AMD.

The disease patterns (Z_{1i}, Z_{2i}) follow a four-categorical distribution, where the conditional probability of no disease at both entities, given the covariate vector $\mathbf{x}_i$ is given by

$$\text{P}(Z_{1i} = 0, Z_{2i} = 0|\mathbf{x}_i) = \text{P}(Y_i = 0|\mathbf{x}_i) = 1 - \pi_i.$$

The conditional probability of disease at both entities can be written as

$$\begin{aligned}\text{P}(Z_{1i} = 1, Z_{2i} = 1|\mathbf{x}_i) &= \text{P}(Z_{1i} = 1, Z_{2i} = 1|Y_i = 1, \mathbf{x}_i)\text{P}(Y_i = 1|\mathbf{x}_i)\\ &= \text{P}(Z_{1i} = 1, Z_{2i} = 1|Y_i = 1, \mathbf{x}_i)\pi_i.\end{aligned}$$

We introduce the notation $\delta_i := \text{P}(Z_{1i} = 1, Z_{2i} = 1|Y_i = 1, \mathbf{x}_i)$ for the conditional probability of disease at both entities given covariates $\mathbf{x}_i$ and disease in participant i, so the conditional probability of disease at both entities is

$$\text{P}(Z_{1i} = 1, Z_{2i} = 1|\mathbf{x}_i) = \delta_i\pi_i.$$

Under the assumption that the conditional probability of disease in one, but not the other entity is symmetric, $\text{P}(Z_{1i} = 1, Z_{2i} = 0|\mathbf{x}_i) = \text{P}(Z_{1i} = 0, Z_{2i} = 1|\mathbf{x}_i)$, the probability mass function of the conditional two entity disease status distribution can be written concisely as given in Table 1.

This probability mass function implies a specific correlation structure between the disease status in each of the two entities of participant i depending on $\mathbf{x}_i$ and δ_i: the bigger δ_i, the bigger the (Pearson) correlation ϕ_i between the entities' disease statuses given a fixed person-specific disease probability. If $\delta_i = 1$ for all participants, all individuals affected by the disease are affected at both entities and the correlation coefficient ϕ of Table 1 is 1. In this case, the knowledge of the disease status in one entity perfectly predicts the disease status of the second and no information is lost from only observing the disease status in a single entity. Figure 1 illustrates the correlation between the disease status of each of the two entities depending on different person-specific disease probabilities π_i and different values of δ_i.

Based on the probability mass function of Table 1, the conditional probabilities of disease in only one specific entity, given Y_i is without loss of generality $\text{P}(Z_{1i} = 1, Z_{2i} = 0|Y_i = 1, \mathbf{x}_i) = (1 - \delta_i)/2$ and it is possible to set up the likelihood function of the parameters $(\boldsymbol{\delta}_i, \boldsymbol{\beta})$. The contribution of a single study participant i to this likelihood is given by

$$L(\delta_i, \boldsymbol{\beta})_i = \{\delta_i\text{H}(\eta_i)\}^{z_{1i}z_{2i}} \times \left\{\frac{1-\delta_i}{2}\text{H}(\eta_i)\right\}^{z_{1i}(1-z_{2i})+(1-z_{1i})z_{2i}} \times \{1 - \text{H}(\eta_i)\}^{(1-z_{1i})(1-z_{2i})}. \quad (1)$$

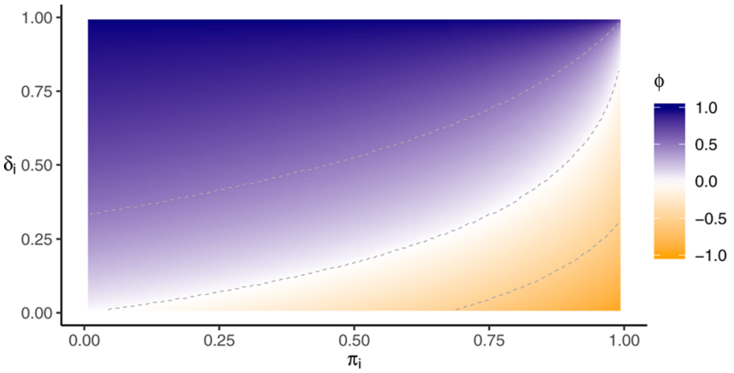


FIGURE 1 Illustration of the correlation coefficients ϕ_i between two binary single entity disease statuses for different probabilities of disease occurrence in at least one entity, π_i, and probabilities of disease at both entities given disease in at least one entity δ_i

The full likelihood function is the product over the contribution of all n participants and the regression parameters can be estimated by minimizing the negative log-likelihood with respect to $\boldsymbol{\beta}$. The score functions (partial derivatives of the log-likelihood) with respect to the regression coefficients $\boldsymbol{\beta}$ are independent of δ_i and identical to the score functions of a logistic regression model with response $Y_i = \max(Z_{1i}, Z_{2i})$ (see (A4) in the appendix). Consequently, the two different approaches to estimate $\boldsymbol{\beta}$ are equivalent. If disease diagnosis is available for each of the two entities of all participants, δ_i are nuisance parameters for the estimation of $\boldsymbol{\beta}$ based on the likelihood of (1). A more detailed derivation of the likelihood and the following formulas can be found in the appendix.

2.2 | Partly missing response observations

In epidemiological or clinical studies, entity-specific disease information is often missing for a subset of study participants due to various reasons. On the example of AMD, most studies have a proportion of study participants with missing AMD diagnosis in one of two eyes due to fundus images of inferior quality disenabling diagnosis (e.g., too bright/dark, lack of contrast, not capturing the whole macular region), competing retinal diseases hampering AMD diagnosis, or missing images. Here, we discuss how to deal with this missing data problem in cross-sectional data. The naive analysis strategy is to ignore the missing data: To utilise the disease status of the single observed entity, Z_{1j} or Z_{2j}, as response for study participants j with missing diagnosis for one entity and the worse-entity diagnosis, $\max(Z_{1i}, Z_{2i})$, for study participants i with diagnosis in both entities. When applying this strategy, the binary response Y is inconsistently defined between the two subsets of participants: In the former, subset the response corresponds to the occurrence of disease in a single entity, in the latter to disease occurrence in at least one of two entities.

We propose to discuss this missing data problem in the context of response misclassification. As before, we define the true response Y to indicate the person-specific disease status as the worse-entity disease status, $Y = \max(Z_1, Z_2)$. For participants for which both diagnoses, Z_1 and Z_2, are available, the true response Y is here considered as known. For participants with missing disease information in one entity, we cannot observe this response, but only a potentially misclassified response Y^*, the disease status in a single entity, Z_1 or Z_2. The specificity of Y^*, $P(Y^* = 0|Y = 0)$, is 1, since a single entity is necessarily healthy if both entities are unaffected. The sensitivity of Y^* is, however, $P(Y^* = 1|Y = 1) \leq 1$, since it can happen that an unaffected entity is observed while the diagnosis of the second, affected entity is missing.

This resembles the situation of misclassified data with an internal validation sample that is discussed in the measurement error literature: The subsample with missing single disease diagnosis can be referred to as “main study sample” for which only a potentially misclassified response Y^* is observed. The subsample with information on both entities represents the “validation sample” for which the true responses $Y = \max(Z_1, Z_2)$, as well as some kind of misclassified response Y^*, the disease diagnosis Z_1 and Z_2 of both single entities separately, are observed. Those terms are somewhat confusing in the context of bilateral diseases since observations of study participants with disease diagnosis for both entities should be the standard and participants with missing information on a single entity the exception instead of the “main study sample”. We will therefore call the participants from the “main study sample” the participants with missing data and the participants from the “validation sample” the participants with full data when referring to the missing data problem in datasets on bilateral diseases.

It is known that a naive application of regression models to partly misclassified binary responses yields biased effect estimates. In case of misclassified data with an internal validation sample, different modelling strategies exist to obtain meaningful results. A simple strategy is to focus only on the validation data subsample and estimate the model based on the observed true responses. This yields valid results if *selection into validation data* is independent of the true disease status and risk factors, for example, in case of AMD if the probability of obtaining disease diagnosis for both eyes is constant for all study participants. It also yields unbiased effect estimates if selection into validation data depends on risk factors that are considered within the regression model estimated on the validation data (e.g., if the probability of obtaining disease diagnosis for both eyes depends on the well known risk factor age but age is considered as covariate in the regression model) (Carroll et al., 2006, Chapter 15.5). This approach is straightforward to implement, since it corresponds simply to applying the model of choice to a subset of the data and can be a good solution for analysis. It ignores, however, a potentially big part of the collected data and might therefore be suboptimal.

Here, we introduce an approach to obtain unbiased estimates for the regression parameters based on all collected data. It is based on a likelihood with different contributions of participants with full data and participants with missing data. The participants with full data provide information regarding the potential misclassification from observing only the disease status of a single entity, which can be used for an appropriate consideration of the observations with missing data.

The general idea of such an approach is that the likelihood contribution for observations with information on Y and Y^* (*validation data* or participants with full data) can be constructed based on the decomposition of the joint distribution of $\mathrm{P}(Y, Y^*|X)$ into a model for the misclassification process and a model of the true response given covariates, $\mathrm{P}(Y, Y^*|X) = \mathrm{P}(Y^*|Y, X)\mathrm{P}(Y|X)$. For observations with information only on Y^* (*main study data* or participants with missing data), it is possible to set up their likelihood contribution correspondingly by applying the law of total probability, $\mathrm{P}(Y^*|X) = \sum_y \mathrm{P}(Y^*|Y = y, X)\mathrm{P}(Y = y|X)$ (see, for example, Lyles et al., 2011; Carroll et al., 2006, Chapter 15.4).

To apply such an approach to data of paired organs with a subset of participants with missing diagnosis in one of two entities, we make the assumption that the entity with missing diagnosis is missing independently of the entity's true disease status. That is, we assume that the observed disease status for participants with missing data is assumed to be the disease status of a single entity Z_r, selected randomly from the two entities with disease statuses Z_1 and Z_2. For participants with full data, we can replace the conditional probabilities $\mathrm{P}(Y^*|Y, X)$ from the previous paragraph by $\mathrm{P}(Z_1, Z_2|Y, X)$, and for participants with missing diagnosis, we replace $\mathrm{P}(Y^*|Y = y, X)$ by $\mathrm{P}(Z_r|Y = y, X)$. The corresponding derivations are given in the appendix, where we show in (A5), that for study participant j with missing data

$$\mathrm{P}(Y_j^* = 1|\mathbf{x}_j) = \mathrm{P}(Z_{rj} = 1|\mathbf{x}_j) = \left(\frac{1}{2} + \frac{1}{2}\delta_j\right)\mathrm{H}(\eta_j).$$

Therefore, the conditional probability of observed disease for persons with missing data depends on the conditional probability of disease at both entities, given disease in at least one entity, δ_j. Participants with full data provide information on this probability. If we had $\delta_j = 1$, we would yield $\mathrm{P}(Z_{rj} = 1|\mathbf{x}_j) = \mathrm{H}(\eta_j)$. This is, in fact, the assumption of the naive modelling strategy, but δ_j is usually smaller than one. For AMD, it is known that persons can be affected only in one eye and be healthy in the other, which violates the assumption of the naive model.

Based on the derived conditional probabilities, the likelihood contributions are: For each observation $i = 1, \dots, n^{\text{full}}$ with full data, the likelihood contribution is, as already given in (1),

$$L(\delta_i, \boldsymbol{\beta})_i^{\text{full}} = \left\{\delta_i \mathrm{H}(\eta_i)\right\}^{z_{1i}z_{2i}} \times \left\{\left(\frac{1}{2} - \frac{1}{2}\delta_i\right)\mathrm{H}(\eta_i)\right\}^{z_{1i}(1-z_{2i})+(1-z_{1i})z_{2i}} \times \left\{1 - \mathrm{H}(\eta_i)\right\}^{(1-z_{1i})(1-z_{2i})}; \tag{2}$$

each observation $j = 1, \dots, n^{\text{miss}}$ from the subsample with missing data contributes with

$$L(\delta_j, \boldsymbol{\beta})_j^{\text{miss}} = \left\{\left(\frac{1}{2} + \frac{1}{2}\delta_j\right)\mathrm{H}(\eta_j)\right\}^{z_{rj}} \times \left\{1 - \left(\frac{1}{2} + \frac{1}{2}\delta_j\right)\mathrm{H}(\eta_j)\right\}^{(1-z_{rj})}. \tag{3}$$

The complete likelihood function is then given by

$$L(\boldsymbol{\delta}, \boldsymbol{\beta}) = \prod_{i=1}^{n^{\text{full}}} L_i^{\text{full}} \times \prod_{j=1}^{n^{\text{miss}}} L_j^{\text{miss}}.$$

If the conditional probability of disease at both entities given disease, δ_i, is assumed to be constant for all study participants, it is straightforward to optimise the log-likelihood with respect to the parameters $(\delta, \boldsymbol{\beta})$ (the subscript i is here used for all $n^{\text{full}} + n^{\text{miss}}$ observations). If the probability is assumed to potentially change with specific characteristics of the study participants, this can be considered by specifying a parametric model for it. It can, for example, be modeled by again using the logistic function of a linear predictor, $\delta_i = \mathrm{H}(\gamma_0 + \gamma_1 u_{1i} + \dots + \gamma_m u_{mi}) = \mathrm{H}(\mathbf{u}_i^t \boldsymbol{\gamma})$, where $\mathbf{u}_i$ is an m-dimensional vector of observed participant characteristics that might or might not be similar to the covariates $\mathbf{x}_i$ of the main model for disease. In that case, the log-likelihood has to be optimised with respect to the parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$. If the vector of covariates $\mathbf{u}_i$ for modelling δ_i contains only an intercept term, this corresponds to assuming a constant δ for all analysed subjects.

2.3 | Additional misclassification in entity-specific disease diagnosis

So far, we considered misclassification in the person-specific worse-entity disease status Y from missing disease information in one of the two entities, while we assumed the entity-specific disease statuses, Z_1, Z_2, and Z_r, to be observed without error. A second source of response misclassification can result from a misclassification of the entity-specific disease status: Instead of the true disease status Z_1, Z_2, or Z_r, only a potentially misclassified disease status Z_1^*, Z_2^*, or Z_r^* is observed. The misclassification process can be described by the sensitivity and specificity of diagnosis

$$\mathrm{P}(Z_{li}^* = 1 | Z_{li} = 1) = \pi_{1i} \text{ and}$$
$$\mathrm{P}(Z_{li}^* = 0 | Z_{li} = 0) = \pi_{0i},$$

respectively, with $l \in \{1, 2, r\}$ and i referring to all $n^{\text{miss}} + n^{\text{full}}$ observations. By applying the law of total probability iteratively, the conditional probabilities $\mathrm{P}(Z_{1i}^* = z_{1i}^*, Z_{2i}^* = z_{2i}^* | \mathbf{x}_i)$ and $\mathrm{P}(Z_{ri}^* = z_{ri}^* | \mathbf{x}_i)$ can be derived and used to set up a likelihood with the following contribution for full and missing data participants:

$$\begin{aligned}
L(\pi_{1i}, \pi_{0i}, \delta_i, \boldsymbol{\beta})_i^{\text{full}} &= \left[(1-\pi_{0i})^2 + \left\{(1-\pi_{0i})\pi_{1i}(1-\delta_i) + \pi_{1i}^2\delta_i - (1-\pi_{0i})^2\right\}\mathrm{H}(\eta_i)\right]^{z_{1i}^* z_{2i}^*} \\
&\times \Bigg[\pi_{0i}(1-\pi_{0i}) + \Bigg\{\left(\pi_{1i}\pi_{0i} + (1-\pi_{1i})(1-\pi_{0i})\right)\left(\frac{1}{2} - \frac{1}{2}\delta_i\right) \\
&+ \pi_{1i}(1-\pi_{1i})\delta_i - \pi_{0i}(1-\pi_{0i})\Bigg\}\mathrm{H}(\eta_i)\Bigg]^{(1-z_{1i}^*)z_{2i}^* + z_{1i}^*(1-z_{2i}^*)} \\
&\times \left[\pi_{0i}^2 + \left\{(1-\pi_{1i})\pi_{0i}(1-\delta_i) + (1-\pi_{1i})^2\delta_i - \pi_{0i}^2\right\}\mathrm{H}(\eta_i)\right]^{(1-z_{1i}^*)(1-z_{2i}^*)}, \\
L(\pi_{1j}, \pi_{0j}, \delta_j, \boldsymbol{\beta})_j^{\text{miss}} &= \left[(1-\pi_{0j}) + (\pi_{1j} + \pi_{0j} - 1)\left(\frac{1}{2} + \frac{1}{2}\delta_j\right)\mathrm{H}(\eta_j)\right]^{z_{rj}^*} \\
&\times \left[\pi_{0j} + (1-\pi_{1j} - \pi_{0j})\left(\frac{1}{2} + \frac{1}{2}\delta_j\right)\mathrm{H}(\eta_j)\right]^{1-z_{rj}^*}.
\end{aligned} \tag{4}$$

The sensitivity and specificity, π_{1i} and π_{0i}, are in general unknown and can vary between study participants, for example, if specific participant characteristics make diagnosis more difficult or if subsets of participants are examined by different investigators. Often it is, however, not possible to estimate those classification probabilities based on collected data, since the true entity-specific disease status of the participants is unknown. The likelihood can then be used for estimation making assumptions regarding the classification probabilities (in sensitivity analyses), with knowledge from external data, or in simulation studies with known parameters.

3 | EXAMPLE: MAXIMUM LIKELIHOOD APPROACH APPLIED TO CROSS-SECTIONAL DATA ON AMD AND RISK FACTORS IN THE AugUR STUDY

We illustrate the proposed maximum likelihood approach based on data of a cross-sectional survey within the AugUR study on AMD. Detailed information on this data can be found in Stark et al. (2015). Briefly, the AugUR study is a prospective study in the elderly population in the city and county of Regensburg in Eastern Bavaria, Germany. It explicitly aims at collecting data of persons with an age of ≥ 70 years to build a database that enables the investigation of genetic and nongenetic risk factors for late-onset diseases like type-2 diabetes, cardiovascular complications, and eye diseases like AMD. Here, we analyse the baseline survey, which was conducted from 2013 to 2015, that includes an interview based questionnaire, bio-banking, and physical examinations for each participant. Data on at least one gradable fundus image and genetic and nongenetic covariates is available for $1,034$ participants, which constitute the analysed data here. AMD diagnosis was performed based on the Three Continent AMD Consortium Severity Scale (Klein et al., 2014). This grading system was developed in 2014 and differentiates no AMD from mild early, moderate early, and severe early AMD stages (according to respective drusen sizes and area and/or pigmentary abnormalities) as well as from late AMD (defined as neovascularization and/or geographic atrophy). To illustrate our approach, we focus on modelling the probability of any AMD (early or late AMD) and define any AMD in a person as any AMD in at least one eye (worse-eye definition), as described above.

TABLE 2 Estimated coefficients of logistic regression models for modelling the occurrence of AMD in at least one eye on AugUR data by (I) naive logistic regression, (II) logistic regression based on 880 participants with available diagnosis for both eyes, (III) maximum likelihood with constant $\hat{\delta} = \mathrm{H}(\hat{\gamma}_0) = 0.59$, (IV) ML with varying $\hat{\delta}_i = \mathrm{H}(\mathbf{x}_i^t\hat{\boldsymbol{\gamma}})$, (V) ML with varying $\hat{\delta}_i = \mathrm{H}(\mathbf{x}_i^t\hat{\boldsymbol{\gamma}})$ and assumed sensitivity of 0.95 and specificity of 0.97

		(I)		(II)		(III)		(IV)		(V)	
		Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
$\hat{\beta}$	(Int.)	−2.28	0.17	−2.14	0.18	−2.23	0.17	−2.21	0.17	−3.08	0.28
	Age	0.47	0.10	0.52	0.11	0.50	0.11	0.47	0.10	0.59	0.13
	Sex:f	0.00	0.16	−0.07	0.17	−0.03	0.16	−0.02	0.16	0.05	0.22
	rs10490924	0.86	0.13	0.87	0.14	0.88	0.13	0.86	0.13	1.21	0.17
	rs1061170	0.69	0.11	0.67	0.12	0.70	0.11	0.68	0.11	0.94	0.15
	Age×sex:f	−0.09	0.15	−0.16	0.16	−0.11	0.15	−0.09	0.15	−0.11	0.20
$\hat{\gamma}$	(Int.)	–	–	–	–	0.35	0.14	−0.51	0.31	2.33	1.69
	Age	–	–	–	–	–	–	0.64	0.20	1.56	0.92
	Sex:f	–	–	–	–	–	–	0.01	0.30	−0.53	0.92
	rs10490924	–	–	–	–	–	–	0.42	0.23	−0.61	0.72
	rs1061170	–	–	–	–	–	–	0.49	0.22	0.06	0.53
	Age×sex:f	–	–	–	–	–	–	−0.31	0.29	−1.01	0.98

Of the 1,034 AugUR participants 154 participants (14.9%) have missing AMD information in one of two eyes. Ignoring missing single eyes, 237 participants (22.9%) were affected by AMD in at least one or the single observed eye. Focusing on the participants with missing eyes, 23 (14.9%) have AMD in the single observed eye. Of the 880 participants with disease diagnosis for both eyes, 214 (24.3%) are affected by AMD in at least one eye. Of those, 87 (40.7%) are affected in only one eye. Of all 1,760 eyes of the participants with diagnosis for both eyes, 341 (19.4%) are affected by AMD.

We aim to quantify the effects of the covariates age, sex, and lead variants of two known genetic loci on the occurrence of AMD in at least one eye by estimating a logistic regression model with linear effects of the standardised age and the binary sex, their interaction, and linear (additive) effects of two genetic lead variants (genotyped SNPs, number of effect alleles $\in \{0, 1, 2\}$).

All computations where performed in R, version 3.5.1 (R Core Team, 2018). Code to perform and reproduce the analysis is available in the Supporting Information (https://onlinelibrary.wiley.com/doi/10.1002/bimj.201900039). The proposed analysis approach for logistic regression on bilateral disease data is implemented in an R package `bilateralogistic` that can be found there, as well. The data available on the web page is, however, a simulated dataset mimicking the original data since the original data of the AugUR study cannot be published online due to data protection reasons. Results based on the original code and the simulated data deviate, therefore, slightly from the results shown here. The code used for simulating the artificial data is part of the Supporting Information as well. Interested researchers can get access to the original data after getting into contact with the corresponding author and signing a data privacy statement.

Table 2 shows the estimated coefficients of five different models. Model (I) corresponds to the naive logistic regression analysis where the disease status of the single observed eye is used as response for participants with missing diagnosis in one eye and the worse-eye disease status for study participants with diagnosis for both eyes. Model (II) is a logistic regression model using the worse-eye disease status as response. It is estimated based on the subset of participants for which diagnosis in both eyes is available and corresponds therefore to the validation data (full data) only strategy. Models (III) and (IV) apply the maximum likelihood estimation of Section 2 to all data assuming a constant and varying conditional probability of AMD in both eyes, δ_i, respectively. In Model (IV), the δ_i's are modelled based on the same linear predictor as used in the model for AMD occurrence and using the logistic response function. If there exists external information on misclassification in the observed (eye-specific) diagnosis, such information should be considered within the analysis. Here, we do not have such information available, the performed manual disease diagnosis of the fundus images based on standardised protocols can be seen as a gold standard diagnosis. Therefore, the available data would normally be analysed without assuming misclassification in the observed data. Nevertheless, we decided to present for illustrative purposes the estimates of Model (V), in which we assume a constant sensitivity and specificity of the single eye diagnosis of 0.95 and 0.97, respectively.

The estimated parameters in Table 2 are additive effects on the log odds for the occurrence of AMD in at least one eye ($\hat{\boldsymbol{\beta}}$) and on the log odds for the occurrence of AMD in two eyes, given AMD in at least one eye ($\hat{\boldsymbol{\gamma}}$ in Models (IV) and (V)), for

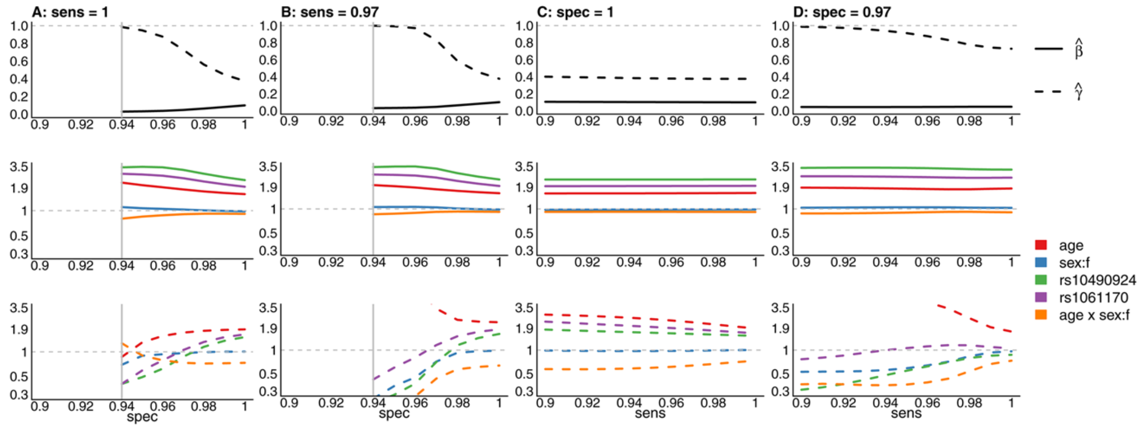


FIGURE 2 Results of Model (IV) on AugUR data under various assumptions of misclassification in single eye specific diagnosis with constant sensitivity and specificity. In panel (A) and (B) estimates are shown for fixed sensitivity of 1 and 0.97 and varying specificity between 1 and 0.94. In panel (C) and (D) for fixed specificity of 1 and 0.97 and varying sensitivity between 1 and 0.9. Shown are the estimated probabilities of AMD in at least one eye H($\hat{\boldsymbol{\beta}}_0$) and AMD in two eyes, given AMD in at least one eye H($\hat{\gamma}_0$) for males of mean age and without effect alleles (black solid/dotted line) in the first row. The second row shows the estimated odds ratios $exp(\hat{\boldsymbol{\beta}}_k)$ in the main logistic regression model (with outcome AMD on at least one eye) for changes of one unit in sd(age) (red), sex (blue), one positive effect allele of SNPs rs10490924 (green), rs1061170 (purple), and the age-sex interaction (orange) under varying assumptions for sensitivity and specificity of diagnosis in single eyes; y-axis is on log scale. The third row shows the odds ratios $exp(\hat{\gamma}_k)$ of the misclassification model with outcome parameter δ_i. Solid lines refer to estimates of the main logistic regression model, dotted lines to the misclassification model.

a one unit increase in the respective covariates. Looking at the results of Model (IV), we see evidence for a higher risk of AMD occurrence with increasing age. The odds ratio of AMD occurrence for a one standard deviation (≈ 5 years) increase in age is estimated as $exp(0.47) = 1.60$ for men and $exp(0.38) = 1.46$ for women. Furthermore, we estimated increased probabilities for being affected by AMD on both eyes, given any AMD occurrence with increasing age. The odds ratio for men is $exp(0.64) = 1.90$ and for women $exp(0.33) = 1.39$ with an age increase of 5 years. We also find strong effects of the genetic variants rs10490924 and rs1061170. Both increase the probability of AMD occurrence and for being affected by AMD in both eyes.

To test for differences in, for example, sex, it is possible to conduct a likelihood ratio test against the nested model without the main and age-interaction effect of sex (on π_i and δ_i). The corresponding test statistic is ≈ 1.66 and is under the null hypotheses $\chi^2_{df=4}$ distributed, which corresponds to a p-value of 0.80. Therefore, we do not have evidence for substantial differences in AMD occurrence between men and women.

Comparing the estimated regression parameters $\hat{\boldsymbol{\beta}}$ of Models (I)–(IV), we find that they are quite similar. The estimated $\hat{\boldsymbol{\gamma}}$ parameters of Model (IV) indicate that the assumptions for modelling the observations with missing diagnosis in one of two eyes based on Models (I) and (III) are questionable. The overall resulting estimates with respect to the occurrence of AMD in at least one eye, $\hat{\boldsymbol{\beta}}$, are, however, qualitatively similar. Comparing the results of Model (V) to Model (IV), we find that assuming a constant sensitivity of 0.95 and specificity of 0.97 for diagnosis in single eyes leads to effect estimates $\hat{\boldsymbol{\beta}}$ that are bigger in absolute value compared to Model (IV) with sensitivity and specificity of 1. The estimated parameters $\hat{\boldsymbol{\gamma}}$ vary as well, some change also their direction. The associated standard errors of $\hat{\boldsymbol{\gamma}}$ are, however, relatively big and the uncertainty in the effect estimates is therefore is rather high.

Figure 2 continues such kind of sensitivity analyses and illustrates the results of Model (IV) under various assumptions for single eye specific misclassification with constant misclassification probabilities for all study participants. Results are shown depending on the assumed combination of sensitivity and specificity in the single eye diagnosis. In each panel (A)–(D) of the figure, the upper plots show the estimated predicted probabilities for the occurrence of AMD (solid line) and for AMD in both eyes, given at AMD in at least one eye (dotted line) for a reference participant (male, mean age, no positive genetic effect alleles) for fixed sensitivity and varying specificity ((A) and (B)), and fixed specificity and varying sensitivity ((C) and (D)). The coloured lines in the second and third row represent the estimated odds ratios for a one unit increase of the respective covariates on the occurrence of AMD (second row) and the occurrence of AMD in both eyes given any occurrence of AMD (third row) on a log scale.

Assuming a decreasing specificity of diagnosis, the effect estimates on AMD occurrence get bigger in absolute value (and therefore odds ratios are more different from 1; second row, panels (A) and (B)). The intercept $\hat{\beta}_0$, and therefore the estimated probability of AMD occurrence decreases with decreasing specificity (solid line, upper plot). A specificity < 1 implies that some of the single eyes that were observed as being affected by AMD are in fact not affected. The smaller the specificity, the bigger the probability that an eye observed to be affected is actually unaffected. While a smaller specificity leads on average to a decreased estimated probability of being affected in at least one eye, it results, however, in a bigger probability of being affected in both eyes, given AMD occurrence in at least one eye (dotted line, upper plot). The smaller the specificity, the lower the overall estimated probability of AMD occurrence. The *remaining* probability mass accumulates in the eyes with relatively biggest (observed) probability for AMD occurrence, what corresponds to the increased odds ratios. If we assume a specificity of 0.94 or 0.95 (with assumed sensitivity of 1 or 0.97, respectively), the estimated probability of AMD in both eyes, given AMD in at least one eye, $\hat{\delta}_i$, is near 1 for most study participants and the probability of observing AMD in at least one eye is quite low. The effect estimates $\hat{\boldsymbol{\gamma}}$ are in general quite unstable with varying assumptions regarding the specificity of the single eye diagnosis. At first sight, the assumption of a constant single eye specificity less than 0.95 might appear plausible. However, such an assumption is, combined with a high sensitivity, problematic with respect to the observed data. The fraction of affected single eyes in the dataset is not huge and a low specificity implies that a relevant fraction of the observed affected eyes are in fact healthy since $1 - \pi_0$ of all unaffected eyes are assumed to be falsely classified as diseased. The resulting estimates are strongly influenced by the model specification and the assumption of constant single eye misclassification probabilities.

Parts (C) and (D) of Figure 2 show the results of reducing the assumed sensitivity in the single eye observation process from 1 to 0.9. With respect to the estimated $\hat{\boldsymbol{\beta}}$'s (row 2), we observe hardly any differences. The estimated $\hat{\gamma}$'s change and the average probability of AMD occurrence in two eyes increases. A single eye specific sensitivity smaller than one implies that some of the truly affected single eyes are graded as non-affected. The (model based) probability of such an observation depends on the probability of AMD occurrence in a single eye. Therefore, an assumed sensitivity smaller than one places a higher probability of unobserved AMD in eyes that have in general a relatively bigger probability of being affected by AMD. This assumption gets incorporated into estimation, and yields an increased probability of AMD in both eyes, given AMD in at least one eye. If the assumed sensitivity is further reduced, the estimated $\hat{\boldsymbol{\beta}}$ coefficients change as well, and the estimated probability of AMD in at least one eye increases. With, for example, a constant sensitivity of only $\pi_1 = 0.75$, the probability of AMD in at least one eye is estimated as 0.364 for males of mean age with one effect allele at each genetic locus instead of approximately 0.338 with $\pi_1 = 1$.

4 | SIMULATION STUDY

4.1 | Design of simulation study

To further analyse the behaviour of the different modelling strategies, we performed a simulation study. We sampled data mimicking studies on bilateral diseases the following way: Binary response data of worse-entity disease occurrence was simulated for 1,000 "participants" based on a logistic regression model with two independent standard normal distributed covariates x. If a (true) response $Y_i = 0$ was sampled for participant i, the disease status of both single entities Z_{1i}, Z_{2i} are set to zero. If $Y_i = 1$, a Bernoulli random variable was sampled with probability δ_i to define whether both entities of the participant were affected ($Z_{1i} = Z_{2i} = 1$). If only one entity was affected for participant i, Z_{1i} or Z_{2i} was chosen randomly and set to 1. For a differing proportion of randomly selected participants, the disease diagnosis of a randomly selected entity was set to missing. Additional entity-specific misclassification was incorporated with constant sensitivity and specificity π_1, π_0.

4.1.1 | Simulation scenarios

We sampled data based on two different scenarios: The outcome occurrence of worse-entity disease Y was sampled based on a logistic regression model with parameters $\boldsymbol{\beta} = (0, 1, -.75)'$ in both scenarios. δ_i is specified based on the logistic function as well with a linear predictor $\mathbf{x}_i'\boldsymbol{\gamma}$. In Scenario 1, $\boldsymbol{\gamma}$ was set to $\boldsymbol{\gamma} = (-.5, 0, 0)'$, which corresponds to a constant $\delta = 0.38$. In Scenario 2, we set $\boldsymbol{\gamma} = (-.5, 0.75, 0)'$, which corresponds to δ_i's with an empirical mean of 0.39 and standard deviation of 0.16, where δ_i varies with x_1.

Entity-specific misclassification is simulated in four different ways, (A) no misclassification, (B) (constant) sensitivity $\pi_1 = 1$, specificity $\pi_0 = 0.95$, (C) $\pi_1 = 0.95$, $\pi_2 = 1$, (D) $\pi_1 = 0.95$, $\pi_2 = 0.95$.

For each of the eight different combinations (1A–2D) data was sampled with an expected fraction of participants with missing disease status in one of two entities of 20%, 50%, and 80%. This was done by sampling a Bernoulli random variable with probability of 0.2, 0.5, and 0.8 for each "participant," where 1 indicates full available data and 0 lead to removal of the disease status in a randomly selected entity.

For each combination of scenario, misclassification in entity-specific diagnosis, and probability of missing single entity disease status, we sampled 5,000 datasets.

4.1.2 | Estimated models in simulation study

For each dataset we estimated the following seven models: (1) Naive logistic regression model using worse-entity disease status of participants with information on both entities and single entity disease status for participants with missing information in a single entity, (2) logistic regression with worse-entity disease status as response on subset of participants with disease status available for both entities (focus on *full data only*), (3) maximum likelihood estimation of Section 2 with constant $\hat{\delta}$ estimated from data, (4) maximum likelihood estimation of Section 2 with potentially varying $\hat{\delta}_i$ modelled via the logistic function $\mathrm{H}(\mathbf{x}_i'\hat{\boldsymbol{\gamma}})$, (5)–(7) as Model (4) but with additionally assumed constant entity-specific misclassification with sensitivity $\pi_1 = 1$, specificity $\pi_0 = 0.95$; $\pi_1 = 0.95$, $\pi_0 = 1$; and $\pi_1 = 0.95$, $\pi_0 = 0.95$, respectively.

Source code to reproduce the results or change simulation settings is available as Supporting Information (https://onlinelibrary.wiley.com/doi/10.1002/bimj.201900039).

4.2 | Results of simulation study

The results of the simulation study can be found in Tables S1–S4 in the online Supporting Information. For each effect estimate in $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, we report the average point estimate, standard deviation of point estimates, their average bias, mean squared error, and coverage frequencies of 95%-confidence intervals (CIs). Figures S1–S8 illustrate the empirical distributions of estimated regression parameters. In the following, we describe some central findings.

4.2.1 | Performance of naive model

The naive modelling strategy (Model 1) yields biased estimates. In simulation Scenario 1, the estimated regression parameters are on average attenuated, which is expected given the constant probability of missing single entity diagnosis and constant δ. This corresponds to a non-differential response misclassification with constant sensitivity < 1 and a specificity of 1 for the true response. The bigger the fraction of participants with missing diagnosis in single entities, the lower the sensitivity of the observation process and the bigger the bias. For a fraction of only 0.2 participants with available diagnosis in both entities, the average bias of β_1 is in 1A, for example, −0.28, which is quite substantial given the true slope effect of 1. If there exists additional non-differential single entity misclassification (Scenarios 1B–1D), the slope estimates are even slightly more attenuated.

In Scenario 2, the probability of being diseased in both entities, δ_i, varies with the covariate X_1. The bigger x_{1i} the bigger δ_i. Compared to Scenario 1, the bias in $\hat{\beta}_1$ of the naive Model 1 is reduced, since the effect of not observing the disease in affected entities is somewhat reduced by the fact that with increasing x_1, the probability of being affected in both entities is increased as well (β_1 and γ_1 are effects of x_1 in the same direction). This is not explicitly considered within the model, but reduces the bias of $\hat{\beta}_1$ in Model 1. The $\hat{\beta}_2$'s of Model 1 are on average still quite strongly attenuated.

Coverage frequencies of the 95%-CIs are in both scenarios below the desired level.

4.2.2 | Models considering missing diagnosis in entities

The models accounting for response MC due to missing diagnosis in one of two entities (i.e., focus on validation data only, and MC adjusted maximum likelihood estimation with constant/potentially varying parameter $\delta_{i/j}$, Models 2–4) yield unbiased parameter estimates in Scenario 1A. Coverage frequencies of the 95%-CIs vary between 94% and 96% indicating a performance as expected.

The variance of the estimates (and consequently the MSE) is biggest in Model 2 and smallest in Model 3. Model 2 utilises less data than Models 3 and 4, while Model 3 estimates only one parameter for modelling δ instead of the three of Model 4. This is sufficient in Scenario 1, since we used a constant δ in the data generating process; the differences between Models 3 and 4 are very small.

In Scenario 2, Models 2 and 4 yield in 2A unbiased estimates, as well. Model 3 assumes a constant parameter δ and yields on average inflated estimates of $\hat{\beta}_1$, since it does not account for the increase of δ_i with increasing X_1. The 95%-CIs for $\hat{\beta}_1$ cover the true effect estimate only in 71–94%, depending on the fraction of participants with a missing entity. $\hat{\beta}_2$ is not biased from the model misspecification since X_2 is unrelated to δ_i in the data generating process and X_1 and X_2 are independent.

Misclassification in the diagnosis for single entities (Scenarios 1B–1D, 2B–2D) biases effect estimates of Models 2–4, that assume a correct diagnosis in the single entities.

4.2.3 | Misclassification in diagnosis of single entities

If the assumptions regarding entity-specific sensitivity and specificity are correct, the optimization of the corresponding likelihood yields unbiased estimates with correct coverage frequencies of the CIs (Models 5–7 in Scenarios 1B–1D and 2B–2D, respectively). Falsely assuming entity-specific misclassification yields, however, biased parameter estimates. The slope parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are in 1A and 2A (without entity-specific misclassification) on average inflated compared to the truth. If misclassification exists, but the sensitivity and/or specificity are falsely specified, this leads to biased estimates.

5 | SUMMARY, DISCUSSION, AND GUIDE FOR EPIDEMIOLOGISTS

Statistical models need clear definitions of the response variable and covariates to yield meaningful and interpretable results. Without missing values, the general approach of modelling binary bilateral disease data based on the worse disease status per participant corresponds to the definition of a binary response that indicates the occurrence of disease in at least one entity. If diagnosis is missing in one of the two entities for some study participants, the naive modelling strategy of ignoring those missing values and utilizing the observed disease status of the single entity as response yields a misspecified model in which the response is not consistently defined between the two different subsets. This leads to biased estimates compared to a model with the worse disease status as response. The bigger the fraction of study participants with missing diagnosis in one entity, the bigger the resulting bias.

Here, we derived an approach to avoid this bias by performing a maximum likelihood estimation in which we explicitly consider the conditional probability mass function of the binary disease status of a randomly selected entity for study participants with missing information in one of the two entities. We model concurrently the probability of disease in at least one entity and the probability of disease in both entities given disease in at least one entity based on observed participant characteristics. The latter part of the model specifies the potential response misclassification for participants with missing single entity disease information. It is a crucial part of the modelling approach and has to contain the relevant covariates to yield unbiased estimates. In the context of modelling the occurrence of AMD, it seems plausible that the conditional probability of AMD in both eyes, given AMD in at least one eye, δ_i, depends on characteristics like age, which is established as main risk factor of disease onset and progression.

The proposed analysis is based on several assumptions: (I) a logistic regression model for disease occurrence in at least one entity, and (II) a logistic regression model for disease occurrence in both entities, given disease in at least one; (III) for observations with missing diagnosis in one of two entities randomly missing disease information with respect to the true disease status of the single entities, and (IV) the transferability of δ_i from the study participants with diagnosis for both entities (validation/full data) to the participants with a missing diagnosis in a single entity (main study/missing data subset). Assumption (I) appears to be a reasonable approach to model binary bilateral disease data with respect to person-specific risk factors and is standard for the analysis of binary responses. In the context of progressing diseases like AMD it is, however, necessary to thoroughly think about the definition of the (binary) response. If different factors are associated with the onset of early AMD stages and the progression to late stages, a response definition of "any AMD" as in Section 3 might be problematic and a separate analysis of the different disease stages or a multinomial modelling approach might be preferable. Assumption (II) is also related to the definition of response/disease within the statistical model and can be seen as an approach to deal with the two separate disease diagnoses for each study participant. The explicit modelling of the conditional probability of being affected in both entities, given disease occurrence in at least one entity can yield interesting and practically relevant insights into disease occurrence/progression and allows us to consider potentially differential response misclassification from missing diagnosis in single entities. As described in Section 2, the modelling strategy implies a specific correlation structure of the two single entities within each participant. This correlation structure is different to other modelling approaches like the estimation of a generalised linear mixed model where each binary single entity disease status would represent a separate observation and the correlation structure of the observations could be considered by specifying an appropriate random effects structure. A comparison of such approaches could be an interesting future research project. Our proposed modelling approach implies identical disease probabilities for both entities of each study participant. If entity-specific information should be incorporated into analysis, alternative modelling approaches have to be considered. Assumption (III) is central to the proposed modelling approach: If the probability of missing disease information per entity is structurally related to the true disease status of the entities, the likelihood of the proposed model is misspecified for the data subset with partly missing response data and the analysis can yield erroneous results. If diagnosis is, however, mainly missing independently of the true disease status, the proposed approach should produce meaningful results. In case of AMD, this is the case if diagnosis in single eyes is missing (mainly) because of randomly occurring low quality fundus images or other reasons that are unrelated to the actual AMD disease status of the eyes. The modelling approach can, in general, successfully deal with differing probabilities for missing disease information between the study participants that might be related to risk factors of disease occurrence as long as the entity with missing disease information within the study participant is missing independently of the true disease status. As in each model that adjusts for measurement error or misclassification based on validation data, the model controlling the misclassification process has to be transferable from the validation to the main study data. Here, this is strongly connected to Assumption (III) and a correct specification of the model for the conditional probability of disease in both entities given disease in at least one entity, δ_i. It might be helpful to have this transferability assumption explicitly in mind when thinking about potential covariates that are related to disease occurrence and the misclassification process (i.e., the conditional probability of disease in both entities δ_i), which should be considered in the respective linear predictors, and/or covariates that are related to selection into validation data, that is, the probability to obtain disease information for both entities.

We discussed consequences of additional misclassification in single entity diagnosis. If such misclassification occurs randomly with constant sensitivity and specificity and is ignored, existing associations in the data are blurred and effect estimates are on average attenuated. Differential misclassification can also lead to inflated estimates. If there exists indication for a big amount of response misclassification, this raises serious concerns regarding the conclusions from statistical models applied to such data. We derived the likelihood of the proposed model also based on potentially error-prone single entity disease diagnosis and showed in the simulation study that we can obtain unbiased estimates with known (constant) sensitivity and specificity of the observation process. It is, however, not possible to estimate the relevant misclassification probabilities without additional data. The analysis of the example data of the AugUR study showed that it can be challenging to make reasonable assumptions for the misclassification probabilities. This gets even more difficult if the entity-specific diagnosis suffers from differential misclassification. The practical value of such (sensitivity) analyses is therefore rather limited, a naive assumption of error-free diagnosis is, on the other hand, questionable as well. If information on entity-specific misclassification is available, the derived likelihood can be used to obtain reasonable effect estimates. One application scenario in the context of AMD could be the modelling of response data that results from an automated and error-prone disease classification of retinal images based on neural networks (e.g., Grassmann et al., 2018) if reliable information on the performance of the automated diagnosis is available. Further research could be spent on the incorporation of information from repeated (independent) classification of the same fundus images to consider the variation and potential misclassification of the diagnosis within the models.

The analysis of the example data of the AugUR study revealed only slight differences between the naive modelling approach and the more complex maximum likelihood analysis considering misclassification. A quite natural question for an analyst dealing with data of bilateral diseases is to ask in which situation the additional effort of the more complex analysis, including an implementation of the maximum likelihood optimization, is really necessary. In the simulation study, we showed that the bias of coefficient estimates increases with an increasing fraction of study participants with missing diagnosis in single entities. The fraction of participants with only one graded eye in the AugUR baseline study is 15%, which might be quite low compared to other studies. If response misclassification occurs only because of (ignoring) missing single entity diagnosis, the specificity of the response observation process is 1, and only participants with an observed single entity disease status $Y^* = 0$ might be misclassified. This further decreases the number of potentially misclassified response observations compared to the fraction of participants with missing single entity diagnosis. In studies with only a small fraction of missing diagnoses and/or a high fraction of observed cases in the fraction of participants with missing diagnosis, the complex maximum likelihood analysis considering MC might not be indispensable. It is in general rather easily possible to compare the results of the naive modelling approach ignoring response misclassification with the results of applying the same model to validation/full data with available diagnosis for both entities only. If the results of both models differ substantially, this would be rather suspicious and indicate problems in the results of the naive analysis resulting from the missing data.

In general, we strongly encourage researchers to think about measurement quality when analyzing data on bilateral diseases. Based on the created R package `bilateralogistic` that is provided as part of this paper, the implementation of the maximum likelihood optimization considering response misclassification is not too complicated for a sufficiently experienced analyst. If the results of a naive modelling approach and the maximum likelihood approach match, this can increase trust in the substantial conclusions of the analysis. If not, the results of the latter should in general be more trustworthy and additional insights can be gained from identifying factors that drive the differences and are related to the occurrence of response misclassification in the data.

ACKNOWLEDGEMENTS

The AugUR study and ophthalmologist's work was funded by the German Federal Ministry of Education and Research (BMBF 01ER1206, BMBF 01ER1507 to I.M.H.); the computing infrastructure was financed by NIH (NIH-2017 R01 EY RES511967 to I.M.H.) and institutional budget of the University of Regensburg and the LMU Munich; methodological work was funded by the DFG (DFG HE3690/5-1 to I.M.H.).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Felix Günther https://orcid.org/0000-0001-6582-1174

REFERENCES

Agresti, A. (2002). *Categorical data analysis* (Vol. 2). New York: Wiley.

Brandl, C., Zimmermann, M. E., Günther, F., Barth, T., Schelter, S. C., Kronenberg, F., … Heid, I. M. (2018). On the impact of different approaches to classify age-related macular degeneration : Results from the German AugUR study. *Scientific Reports*, *8*(1), 8675.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Grassmann, F., Mengelkamp, J., Brandl, C., Harsch, S., Zimmermann, M. E., Linkohr, B., … Weber, B. H. (2018). A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology*, *125*(9), 1410–1420.

Klein, R., Meuer, S. M., Myers, C. E., Gabrielle, H. S., Rochtchina, E. B., Choudhury, F., … Klein, B. E. K. (2014). Harmonizing the classification of age-related macular degeneration in the three-continent AMD consortium. *Ophthalmic Epidemiology*, *21*(1), 14–23.

Lim, L. S., Mitchell, P., Seddon, J. M., Holz, F. G., & Wong, T. Y. (2012). Age-related macular degeneration. *Lancet*, *379*, 1728–1738.

Lyles, R. H., Tang, L., Superak, H. M., King, C. C., Celentano, D. D., Lo, Y., & Sobel, J. D. (2011). Misclassification in logistic regression: An illustration. *Epidemiology*, *22*(4), 589–597.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, *86*(4), 843–855.

R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Stark, K., Olden, M., Brandl, C., Dietl, A., Zimmermann, M. E., Schelter, S. C., … Heid, I. M. (2015). The German AugUR study: Study protocol of a prospective study to investigate chronic diseases in the elderly. *BMC Geriatrics*, *15*(1), 130.

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

How to cite this article: Günther F, Brandl C, Heid IM, Küchenhoff H. Response misclassification in studies on bilateral diseases. *Biometrical Journal*. 2019;61:1033–1048. https://doi.org/10.1002/bimj.201900039

APPENDIX

A.1 Derivation of the likelihood functions

A.1.1 General approach

The contribution of participant i to the likelihood is given by the conditional probability mass function of $\mathrm{P}_{\gamma,\beta}(Z_{1i} = z_{1i}, Z_{2i} = z_{2i}|\mathbf{x}_i)$ evaluated at the observed data as a function of the regression parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Those parameters relate a vector of covariates $\mathbf{x}_i$ to the conditional probabilities of $\mathrm{P}(Z_{1i} = 1, Z_{2i} = 1|Y_i = 1, \mathbf{x}_i) = \mathrm{H}(\mathbf{x}_i'\boldsymbol{\gamma})$ and $\mathrm{P}(Y_i = 1|\mathbf{x}_i) = \mathrm{H}(\mathbf{x}_i'\boldsymbol{\beta})$, where $Y_i := \max(Z_{1i}, Z_{2i})$ and $\mathrm{H}(\cdot)$ are adequate response functions, for example, the logistic function. $\mathrm{P}_{\gamma,\beta}(Z_{1i} = z_{1i}, Z_{2i} = z_{2i}|\mathbf{x}_i)$

can be expressed as

$$\begin{aligned}
\mathrm{P}_{\gamma,\beta}(Z_{1i}=z_{1i},Z_{2i}=z_{2i}|\mathbf{x}_i) &= \mathrm{P}_{\gamma}(Z_{1i}=z_{1i},Z_{2i}=z_{2i}|Y_i=1,\mathbf{x}_i)\mathrm{P}_{\beta}(Y_i=1|\mathbf{x}_i)\\
&\quad + \mathrm{P}_{\gamma}(Z_{1i}=z_{1i},Z_{2i}=z_{2i}|Y_i=0,\mathbf{x}_i)\mathrm{P}_{\beta}(Y_i=0|\mathbf{x}_i). \qquad \text{(A1)}
\end{aligned}$$

The joint probabilities of Z_{1i} and Z_{2i} given Y_i and $\mathbf{x}_i$ can be derived from Table 1 to be

$$\begin{aligned}
&\mathrm{P}_{\gamma}(Z_{1i}=1,Z_{2i}=1|Y_i=1,\mathbf{x}_i)=\delta_i\\
&\mathrm{P}_{\gamma}(Z_{1i}=0,Z_{2i}=1|Y_i=1,\mathbf{x}_i)=\mathrm{P}_{\gamma}(Z_{1i}=1,Z_{2i}=0|Y_i=1,\mathbf{x}_j)=\frac{1}{2}-\frac{1}{2}\delta_i\\
&\mathrm{P}_{\gamma}(Z_{1i}=0,Z_{2i}=0|Y_i=1,\mathbf{x}_i)=0\\
&\mathrm{P}_{\gamma}(Z_{1i}=0,Z_{2i}=0|Y_i=0,\mathbf{x}_i)=1. \qquad \text{(A2)}
\end{aligned}$$

Assuming a logistic regression model for modelling $\mathrm{P}_{\beta}(Y_i=1|\mathbf{x}_i)=\mathrm{H}(\mathbf{x}_i'\boldsymbol{\beta})=\mathrm{H}(\eta_i)$, with $\mathrm{H}(\cdot)$ the logistic function, we can plug the conditional probabilities of (A2) into (A1) and yield

$$\begin{aligned}
&\mathrm{P}_{\gamma,\beta}(Z_{1i}=1,Z_{2i}=1|\mathbf{x}_i)=\delta_i\mathrm{H}(\eta_i)\\
&\mathrm{P}_{\gamma,\beta}(Z_{1i}=1,Z_{2i}=0|\mathbf{x}_i)=\mathrm{P}_{\gamma,\beta}(Z_{1i}=0,Z_{2i}=1|\mathbf{x}_i)=\left(\frac{1}{2}-\frac{1}{2}\delta_i\right)\mathrm{H}(\eta_i)\\
&\mathrm{P}_{\gamma,\beta}(Z_{1i}=0,Z_{2i}=0|\mathbf{x}_i)=1-\mathrm{H}(\eta_i). \qquad \text{(A3)}
\end{aligned}$$

The full likelihood is the product over all single contributions and therefore given by

$$\begin{aligned}
L(\boldsymbol{\gamma},\boldsymbol{\beta})=\prod_{i=1}^{n}&\left\{\delta_i\mathrm{H}(\eta_i)\right\}^{z_{1i}z_{2i}}\times\left\{\left(\frac{1}{2}-\frac{1}{2}\delta_i\right)\mathrm{H}(\eta_i)\right\}^{z_{1i}(1-z_{2i})+(1-z_{1i})z_{2i}}\\
&\times\left\{1-\mathrm{H}(\eta_i)\right\}^{(1-z_{1i})(1-z_{2i})}.
\end{aligned}$$

The corresponding log-likelihood is

$$\begin{aligned}
l(\boldsymbol{\gamma},\boldsymbol{\beta})=\sum_{i=1}^{n}&z_{1i}z_{2i}\log\left\{\delta_i\mathrm{H}(\eta_i)\right\}+\{z_{1i}(1-z_{2i})+(1-z_{1i})z_{2i}\}\log\left\{\left(\frac{1}{2}-\frac{1}{2}\delta_i\right)\mathrm{H}(\eta_i)\right\}\\
&+(1-z_{1i})(1-z_{2i})\log\left\{1-\mathrm{H}(\eta_i)\right\}.
\end{aligned}$$

The score functions with respect to $\boldsymbol{\beta}_k$ are

$$s(\boldsymbol{\beta}_k)=\sum_{i=1}^{n}\{z_{1i}z_{2i}+z_{1i}(1-z_{2i})+(1-z_{1i})z_{2i}\}x_{ik}-\mathrm{H}(\eta_i)x_{ik}, \qquad \text{(A4)}$$

which are identical to the score functions of a logistic regression model with response $Y=\max(Z_1,Z_2)$ (see, for example, Agresti, 2002, eq. 5.17).

A.1.2 Missing diagnosis in single entities

If diagnosis for a single entity is missing for study participant j, we assume that the disease status of a randomly selected entity $Z_{rj}=z_{rj}$ was observed. We can derive the conditional probability as

$$\mathrm{P}(Z_{rj}=z_{rj}|\mathbf{x}_j)=\sum_{z_{1j},z_{2j}=0,1}\mathrm{P}(Z_{rj}=z_{rj}|Z_{1j}=z_{1j},Z_{2j}=z_{2j})\mathrm{P}(Z_{1j}=z_{1j},Z_{2j}=z_{2j}|\mathbf{x}_j).$$

Plugging the conditional probabilities of (A3) in yields, under the assumption of randomly missing entities $\mathrm{P}(Z_{rj} = 1|Z_{1j} = 1, Z_{2j} = 0) = \mathrm{P}(Z_{rj} = 1|Z_{1j} = 0, Z_{2j} = 1) = 0.5$:

$$\mathrm{P}(Z_{rj} = 1|\mathbf{x}_j) = \left(\frac{1}{2} + \frac{1}{2}\delta_j\right)\mathrm{H}(\eta_j). \tag{A5}$$

It is then possible to set up a likelihood based on n^{full} observations with available diagnosis for both entities and n^{miss} observations with diagnosis missing in one of two entities

$$L(\boldsymbol{\gamma},\boldsymbol{\beta}) = \prod_{i=1}^{n^{\text{full}}} \left\{\delta_i \mathrm{H}(\eta_i)\right\}^{z_{1i}z_{2i}} \times \left\{\left(\frac{1}{2} - \frac{1}{2}\delta_i\right)\mathrm{H}(\eta_i)\right\}^{z_{1i}(1-z_{2i})+(1-z_{1i})z_{2i}} \times \left\{1 - \mathrm{H}(\eta_i)\right\}^{(1-z_{1i})(1-z_{2i})}$$
$$\times \prod_{j=1}^{n^{\text{miss}}} \left\{\left(\frac{1}{2} + \frac{1}{2}\delta_j\right)\mathrm{H}(\eta_j)\right\}^{z_{rj}} \times \left\{1 - \left(\frac{1}{2} + \frac{1}{2}\delta_j\right)\mathrm{H}(\eta_j)\right\}^{1-z_{rj}}.$$

A.1.3 Additional misclassification in single entity diagnosis

In this scenario, a potentially misclassified disease diagnosis Z_1^*, Z_2^*, or Z_r^* is observed instead of the true disease status Z_1, Z_2, or Z_r of the entities. The misclassification process can be described by the sensitivity and specificity of the single entity diagnosis

$$\mathrm{P}(Z_{li}^* = 1|Z_{li} = 1) = \pi_{1i}$$
$$\mathrm{P}(Z_{li}^* = 0|Z_{li} = 0) = \pi_{0i};\ l \in \{1, 2, r\}; i = 1, \dots, n^{\text{full}} + n^{\text{miss}}.$$

We further assume that the observation process of the two entities of a single subject i is independent, that is

$$\mathrm{P}(Z_{1i}^* = z_{1i}^*, Z_{2i}^* = z_{2i}^*|Z_{1i} = z_{1i}, Z_{2i} = z_{2i}) = \mathrm{P}(Z_{1i}^* = z_{1i}^*|Z_{1i} = z_{1i})\mathrm{P}(Z_{2i}^* = z_{2i}^*|Z_{2i} = z_{2i}).$$

The conditional probabilities of $\mathrm{P}(Z_{1i}^* = z_{1i}^*, Z_{2i}^* = z_{2i}^*|\mathbf{x}_i)$ can be derived as

$$\begin{aligned}
\mathrm{P}(Z_{1i}^* = z_{1i}^*, Z_{2i}^* = z_{2i}^*|\mathbf{x}_i) &= \sum_{z_{1i},z_{2i}=0,1} \mathrm{P}(z_{1i}^*, z_{2i}^*|Z_{1i} = z_{1i}, Z_{2i} = z_{2i}, \mathbf{x}_i)\mathrm{P}(Z_{1i} = z_{1i}, Z_{2i} = z_{2i}|\mathbf{x}_i)\\
&= \mathrm{P}(z_{1i}^*|Z_{1i} = 0, \mathbf{x}_i)\mathrm{P}(z_{2i}^*|Z_{2i} = 0, \mathbf{x}_i)\\
&\quad + \Big[\Big\{\mathrm{P}(z_{1i}^*|Z_{1i} = 1, \mathbf{x}_i)\mathrm{P}(z_{2i}^*|Z_{2i} = 0, \mathbf{x}_i)\\
&\quad + \mathrm{P}(z_{1i}^*|Z_{1i} = 0, \mathbf{x}_i)\mathrm{P}(z_{2i}^*|Z_{2i} = 1, \mathbf{x}_i)\Big\}\left(\frac{1}{2} - \frac{1}{2}\delta_i\right)\\
&\quad + \mathrm{P}(z_{1i}^*|Z_{1i} = 1, \mathbf{x}_i)\mathrm{P}(z_{2i}^*|Z_{2i} = 1, \mathbf{x}_i)\delta_i\\
&\quad - \mathrm{P}(z_{1i}^*|Z_{1i} = 0, \mathbf{x}_i)\mathrm{P}(z_{2i}^*|Z_{2i} = 0, \mathbf{x}_i)\Big]\mathrm{H}(\eta_i).
\end{aligned} \tag{A6}$$

The conditional probability of a randomly selected single entity of participant j is given by

$$\begin{aligned}
\mathrm{P}(Z_{rj}^* = z_{rj}^*|\mathbf{x}_j) &= \sum_{z_{rj}=0,1} \mathrm{P}(z_{rj}^*|Z_{rj} = z_{rj}, \mathbf{x}_j)\mathrm{P}(Z_{rj} = z_{rj}|\mathbf{x}_j)\\
&= \sum_{z_{rj}=0,1} \Big\{\mathrm{P}(z_{rj}^*|Z_{rj} = z_{rj}, \mathbf{x}_j)\\
&\quad \times \sum_{z_{1j},z_{2j}=0,1} \mathrm{P}(Z_{rj} = z_{rj}|Z_{1j} = z_{1j}, Z_{2j} = z_{2j}) \times \mathrm{P}(Z_{1j} = z_{1j}, Z_{2j} = z_{2j}|\mathbf{x}_j)\Big\}\\
&= \mathrm{P}(z_{rj}^*|Z_{rj} = 0, \mathbf{x}_j) + \Big\{\mathrm{P}(z_{rj}^*|Z_{rj} = 1, \mathbf{x}_j) - \mathrm{P}(z_{rj}^*|Z_{rj} = 0, \mathbf{x}_j)\Big\}\left(\frac{1}{2} + \frac{1}{2}\delta_j\right)\mathrm{H}(\eta_j).
\end{aligned} \tag{A7}$$

The likelihood can be set up by multiplying (A6) and (A7) evaluated at the observed data of all study participants

$$
\begin{aligned}
L(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{i=1}^{n^{\text{full}}} & \left[(1-\pi_{0i})^2 + \left\{(1-\pi_{0i})\pi_{1i}(1-\delta_i) + \pi_{1i}^2\delta_i - (1-\pi_{0i})^2\right\} \mathrm{H}(\eta_i)\right]^{z_{1i}^* z_{2i}^*} \\
& \times \Big[\pi_{0i}(1-\pi_{0i}) + \Big\{ \left(\pi_{1i}\pi_{0i} + (1-\pi_{1i})(1-\pi_{0i})\right) \left(\frac{1}{2} - \frac{1}{2}\delta_i\right) \\
& + \pi_{1i}(1-\pi_{1i})\delta_i - \pi_{0i}(1-\pi_{0i}) \Big\} \mathrm{H}(\eta_i)\Big]^{(1-z_{1i}^*)z_{2i}^* + z_{1i}^*(1-z_{2i}^*)} \\
& \times \left[\pi_{0i}^2 + \left\{(1-\pi_{1i})\pi_{0i}(1-\delta_i) + (1-\pi_{1i})^2\delta_i - \pi_{0i}^2\right\} \mathrm{H}(\eta_i)\right]^{(1-z_{1i}^*)(1-z_{2i}^*)} \\
& \times \prod_{j=1}^{n^{\text{miss}}} \left[(1-\pi_{0j}) + (\pi_{1j} + \pi_{0j} - 1)\left(\frac{1}{2} + \frac{1}{2}\delta_j\right) \mathrm{H}(\eta_j)\right]^{z_{rj}^*} \\
& \times \left[\pi_{0j} + (1-\pi_{1j} - \pi_{0j})\left(\frac{1}{2} + \frac{1}{2}\delta_j\right) \mathrm{H}(\eta_j)\right]^{1-z_{rj}^*}.
\end{aligned}
$$

Chapter 3

Chances and challenges of machine learning-based disease classification in genetic association studies

Chapter 3 illustrates the problem of response misclassification in genome-wide association studies for case-control phenotypes. We perform a GWAS screening using an error-prone AMD phenotype obtained from predictions of a pre-trained neural network ensemble. Afterwards, we perform follow-up analyses accounting for response misclassification based on a maximum likelihood approach utilizing internal validation data. In doing so, we find evidence for a false positive association signal from the standard analysis, i.e., when ignoring misclassification. This is due to the existence of differential response misclassification with respect to a specific genetic variant.

Contributing article:
Guenther, F., Brandl, C., Winkler, T. W., Wanner, V., Stark, K., Kuechenhoff, H., & Heid, I. M. (2020). Chances and challenges of machine learning-based disease classification in genetic association studies illustrated on age-related macular degeneration. *Genetic Epidemiology, 44(7), 759-777.*

Supplementary material:
`https://onlinelibrary.wiley.com/doi/10.1002/gepi.22336`

Author contributions:
Heid, Küchenhoff, and Günther devised the research question and work. Günther derived

the maximum likelihood approach with internal validation data and performed the data analysis with support by Winkler for the genome-wide association screen. Brandl performed manual disease classification in internal validation data. Stark performed follow-up investigation on the *HERC2* variant. Wanner performed a literature search on machine learning based phenotypes in GWAS. Günther and Heid created the first draft of the manuscript. All authors contributed to the interpretation of the results and to writing and revising the manuscript.

Received: 11 December 2019 | Revised: 18 June 2020 | Accepted: 21 June 2020

DOI: 10.1002/gepi.22336

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

Chances and challenges of machine learning-based disease classification in genetic association studies illustrated on age-related macular degeneration

Felix Guenther[1,2] | Caroline Brandl[1,3] | Thomas W. Winkler[1] | Veronika Wanner[1] | Klaus Stark[1] | Helmut Kuechenhoff[2] | Iris M. Heid[1]

[1]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany

[2]Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig Maximilian University of Munich, Munich, Germany

[3]Department of Ophthalmology, University Hospital Regensburg, Regensburg, Germany

Correspondence

Helmut Kuechenhoff, Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig Maximilian University of Munich, 80539 Munich, Germany.
Email: kuechenhoff@stat.uni-muenchen.de

Iris M. Heid, Department of Genetic Epidemiology, University of Regensburg, 93053 Regensburg, Germany.
Email: iris.heid@klinik.uni-regensburg.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: DFG HE 3690/5-1; National Institutes of Health, Grant/Award Number: NIH R01 EY RES 511967

Abstract

Imaging technology and machine learning algorithms for disease classification set the stage for high-throughput phenotyping and promising new avenues for genome-wide association studies (GWAS). Despite emerging algorithms, there has been no successful application in GWAS so far. We establish machine learning-based phenotyping in genetic association analysis as misclassification problem. To evaluate chances and challenges, we performed a GWAS based on automatically classified age-related macular degeneration (AMD) in UK Biobank (images from 135,500 eyes; 68,400 persons). We quantified misclassification of automatically derived AMD in internal validation data (4,001 eyes; 2,013 persons) and developed a maximum likelihood approach (MLA) to account for it when estimating genetic association. We demonstrate that our MLA guards against bias and artifacts in simulation studies. By combining a GWAS on automatically derived AMD and our MLA in UK Biobank data, we were able to dissect true association (*ARMS2*/*HTRA1*, *CFH*) from artifacts (near *HERC2*) and identified eye color as associated with the misclassification. On this example, we provide a proof-of-concept that a GWAS using machine learning-derived disease classification yields relevant results and that misclassification needs to be considered in analysis. These findings generalize to other phenotypes and emphasize the utility of genetic data for understanding misclassification structure of machine learning algorithms.

KEYWORDS

age-related macular degeneration (AMD), genome-wide association study, machine learning-based disease classification, response misclassification, UK Biobank

Helmut Kuechenhoff and Iris M. Heid jointly supervised this work.

1 | INTRODUCTION

Imaging technology allows for noninvasive access to detailed disease features in large studies and genome-wide association studies (GWAS) on such disease phenotypes can be expected to accelerate knowledge gain. However, image-based disease classification can be challenging for large sample sizes due to time-intensive, tiresome manual inspection. This limitation can be overcome by automated disease classification via machine learning and particularly deep learning algorithms. Such emerging approaches (Litjens et al., 2017) can classify diseases effortlessly also for huge sample sizes as needed for GWAS or other Omics approaches.

Deep learning algorithms require enormous input data with available gold standard classification, to "learn" classification reliably. Once trained and tested, the algorithms can be applied to external image data, but they cannot critically reflect unusual findings or incorporate unforeseen aspects, for which the human eye and brain have unmet capability. At the current time, input data to train algorithms are limited and often specific to a certain setting (e.g., patients from a clinic). Some characteristics that appear useful for disease classification in one setting might be misinterpreted in another, which can hamper transferability of trained models; a topic discussed as dataset shift or domain shift (Csurka, 2017; Heinze-Deml & Meinshausen, 2017; Moreno-Torres, Raeder, Alaiz-Rodríguez, Chawla, & Herrera, 2012). Most predictions of deep learning algorithms for image-based disease classification will be error-prone and the structure of misclassification will generally be unknown. When using automated disease classification as outcome for association analyses and GWAS, the underlying response misclassification is usually unaccounted for, giving rise to biased effect estimates and potentially false-positive associations (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Hausman, Abrevaya, & Scott-Morton, 1998; Neuhaus, 1999). Extent and structure of the misclassification process can be assessed by *internal validation data*, that is, a subset of participants with both automated and gold standard classification, which can also be utilized to account for response misclassification in statistical models (Carroll et al., 2006; Lyles et al., 2011).

At present, it is unclear whether machine learning-based disease classification is of any utility for association analyses, particularly for detecting disease signals in GWAS. We thus set out to evaluate machine learning-derived disease classification in GWAS on the example of age-related macular degeneration (AMD) and we developed a statistical approach accounting for the implied response misclassification. AMD is an ideal role model, as a common disease ascertained via imaging of the central retina (Klein et al., 2014) and with particularly strong known genetic effects (Fritsche et al., 2016). The manual grading of images for AMD requires a substantial effort by trained staff and is currently an obstacle for homogeneous disease classification within and across large studies. For example, in UK Biobank (Bycroft et al., 2018), >135,000 color fundus images are available for >68,000 study participants, but there is no manually classified AMD available so far. Several machine learning algorithms have been emerging to classify AMD: they show promising performance, but still yield misclassified predictions, have acknowledged issues due to domain shift or insufficient sample size for training, or lack validation in external studies (Burlina et al., 2017; Grassmann et al., 2018; Peng et al., 2019; Ting et al., 2017). So far, there is no GWAS on fundus image ascertained AMD available in UK Biobank, manually classified or machine learning based.

2 | MATERIALS AND METHODS

2.1 | Machine learning-based disease classification in GWAS as misclassification problem

We consider a binary disease Y, for which each individual has a true status of disease (disease yes/no). A *gold standard* classification often involves manual grading of medical images via trained medical staff, which is considered here to correspond to the true disease classification. When applying a trained machine learning algorithm on medical images, we yield an automated disease classification Y^* for each individual. For an individual i with true disease status $Y_i = y_i$, the classification can either be correct or wrong ($y^*_i = y_i$, or $y^*_i \neq y_i$). If a gold standard classification is available (for at least a subset of study participants, internal validation data), the performance of the algorithm can be quantified by cross-tabulation of the observed error-prone y^* and the gold-standard classification y across all participants in the validation substudy (confusion matrix); the (mis-) classification process can be characterized by classification probabilities $P(Y^*{=}k|Y = l)$, for $l,k \in \{0, 1\}$. For $l = k = 1$ and $l = k = 0$, these probabilities correspond to the sensitivity and specificity of the algorithm, respectively.

In the following, we focus on *bilateral diseases* due to our motivating example of an eye disease (AMD): for each individual i, two entity-specific binary disease variables $Z_{1i}, Z_{2i} \in \{0, 1\}$ (here: AMD per eye) are used to define the binary person-specific disease status as the "worse entity disease status" $Y_i := \max(Z_{1i}, Z_{2i})$, corresponding to "AMD in at least one eye" versus "AMD in none of the two eyes" in our example. The error-prone machine learning-based classification of entity-specific disease $Z^*{}_{1i}, Z^*{}_{2i}$, will propagate to error-prone person-specific disease status, $Y_i^* = \max(Z_{1i}^*, Z_{2i}^*)$, when compared to the manually graded "true" Y_i.

We were interested in evaluating the potential and consequences of such automatically classified disease in GWAS. The standard approach in GWAS is logistic regression for modeling the association of a genetic variant (observed as genotypes ∈{0,1,2} or imputed allelic dosages ∈[0,2]) with a binary disease status, usually adjusted for other covariates like age, sex, and genetic principal components; Wald tests are used to test for genetic association, accounting for multiple testing by judging at a Bonferroni-corrected significance level of $p < 5 \times 10^{-8}$. When the association of the genetic variant with the true disease status Y (here: manually classified person-specific AMD) follows a logistic regression model, a naïve usage of the error-prone disease status Y* (here: automatically derived person-specific AMD) in standard logistic regression corresponds to the utilization of a misspecified model for the observed data (*naïve association analysis*). This has known consequences of decreased power, biased (genetic) association estimates, and potentially false-positive associations (Carroll et al., 2006; Hausman et al., 1998; Neuhaus, 1999). With additional information on the misclassification process, it is possible to correct for the bias and inflated type-I error. However, it is in general not possible to recover power lost due to misclassification.

2.2 | MLA to adjust for response misclassification in bilateral disease

In contrast to classical diseases and logistic regression (Carroll et al., 2006; Hausman et al., 1998; Neuhaus, 1999), no method is currently available to adjust for response misclassification in bilateral diseases. As described previously (Günther, Brandl, Heid, & Küchenhoff, 2019), the conceptual challenge is to account for two types of misclassification: (a) entity-specific misclassification that propagates to an error-prone person-specific disease status; and (b) person-specific misclassification from a missing disease status in one of the two entities. We thus developed an MLA to account for the fact that we are using an error-prone response $Y^*{}_i := \max(Z^*{}_{1i}, Z^*{}_{2i})$, $Z^*{}_{1i}, Z^*{}_{2i} \in \{0,1\}$, in the association analysis, while the true disease $Y_i := \max(Z_{1i}, Z_{2i})$, $Z_{1i}, Z_{2i} \in \{0,1\}$ is assumed to follow a logistic regression model.

Details are provided in Appendix A. The general idea of the MLA is to factorize the likelihood of the observed, error-prone response data into two parts, the model for the association between risk factor and true (but in general unobserved) response (*true association model*) and a model for the misclassification process (*misclassification model*). We adapted this well-established methodology for analyzing misclassified binary response data (Carroll et al., 2006; Lyles et al., 2011) to the scenario of bilateral disease with a "worse-entity" disease definition (i.e., the person-specific disease status is defined as the status of the worse entity). We assume conditional independence of the classification in the two entities $z^*{}_{li}, z^*{}_{2i}$ of an individual i, given the true disease status. This assumption can be checked by validation data. Then, we have

$$P(z_{1i}^*, z_{2i}^* | x_i) = \sum_{z_{1i}, z_{2i} \in \{0,1\}} \underbrace{P(z_{1i}^* \mid z_{1i}, x_i) \times P(z_{2i}^* \mid z_{2i}, x_i)}_{\text{misclassification model}} \\ \times \underbrace{P(z_{1i},\ z_{2i}, \mid x_i)}_{\text{true association model}}.$$

The *misclassification model* is characterized by the sensitivity and specificity of the entity-specific classification process; the *true association model* is the assumed logistic regression model for the person-specific disease status. When internal validation data are available, the parameters of both models can be estimated jointly by optimizing a likelihood with different contributions of participants with only the error-prone response and participants in the validation data with true and error-prone response available.

Our developed approach allows us to adjust for both the entity-specific misclassification from an automated classification and the misclassification of the person-specific status when one entity is ungradable. Altogether, we model four parameters in the MLA: (a) the conditional probability of worse entity disease given the covariate of interest; (b) the probability of disease in both entities conditional on the disease in at least one entity (to adjust for missing information of one of two entities); as well as (c) the sensitivity and (d) the specificity of the entity-specific misclassification process. For each parameter, the conditional probabilities are modeled using the logistic function (as in standard logistic regression) allowing for a dependency on a parameter-specific set of person-specific covariates. An open source R (R Core Team, 2019) implementation is available.

2.3 | Simulation study to investigate the performance of the MLA

We repeatedly simulated association data for a standard normal covariate X and a (true and error-prone) binary outcome of a *bilateral* disease. To do this, we (a) sampled the true, person-specific worse entity status associated with X for 5,000 individuals, (b) derived the true entity-specific disease status (e.g., manual eye-specific AMD classification) given assumptions, (c) sampled the entity-specific error-prone disease status (e.g., automated AMD classification), and (d) derived an error-prone, person-specific disease status. Afterward, we removed the true disease status for 4,000 individuals, yielding a subset of 1,000 with both true and error-prone disease status available (validation data). In different simulation scenarios, we varied sensitivity and specificity of the entity-specific classification. Classification probabilities were either constant for all individuals (nondifferential misclassification) or varying with X (differential misclassification). We also varied the fraction of individuals with missing classification in one of two entities (25–75%). Data were sampled with or without an effect of X on the true person-specific response Y ($\beta_Y \in \{0,\ 1\}$, log odds ratio [OR]) and on the probability δ of having disease in both entities given disease in at least one entity ($\beta_\delta \in \{0,\ 1\}$, log OR). We estimated the covariate effect using the naive analysis (logistic regression, which ignores misclassification) and the developed MLA1 and MLA2 accounting for response misclassification without (MLA1) and with allowing (MLA2) for differential misclassification, respectively. To compare the performance of the naïve analysis and the derived MLA, we investigated the distribution of effect estimates $\hat{\beta}_Y$ across 1,000 simulation runs in each scenario, computed the mean squared error of estimates relative to true effects, frequencies of rejected tests for no association, and coverage frequencies of 95% confidence intervals (CI). A detailed description of the simulation study, data sampling, and estimated models is given in Appendix B.

2.4 | UK Biobank study information and data

UK Biobank recruited ~500,000 individuals aged 40–69 years from across the United Kingdom. Genetic data are available from the Affymetrix UK Biobank Axiom Array imputed to the Haplotype Reference Consortium (McCarthy et al., 2016) and the UK10K haplotype resource (Walter et al., 2015); details described elsewhere (Bycroft et al., 2018). The UK Biobank baseline data contains 135,500 fundus images of 68,400 individuals. The images are taken with the Topcon 3D OCT-1000 Mark II system with a field angle of 45° without application of mydriasis (Keane et al., 2016). The images can be utilized for automated or manual AMD classification; however, there is no image-based AMD classification publicly available so far.

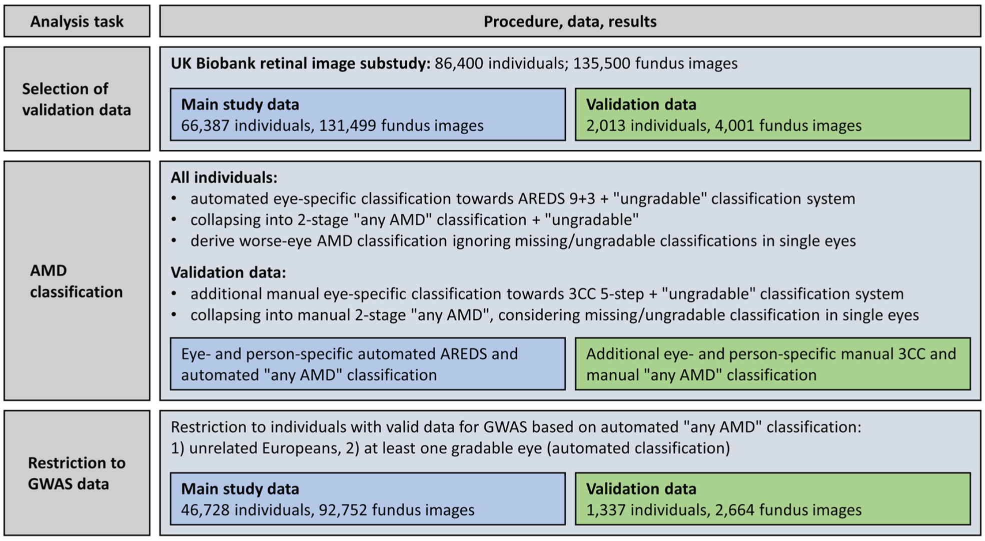
2.5 | AMD classification in UK Biobank derived from a machine learning algorithm and manually

We performed an automated AMD classification for 68,400 individuals with available fundus images in UK Biobank with additional manual classification in a subset of 2,013 participants, as described in Figure 1.

In epidemiological studies, AMD is usually classified per eye via manual grading of color fundus images by trained graders using established classification systems. One such system is the nine-step Age-Related Eye Disease Study (AREDS) severity scale (Davis et al., 2005), which defines early AMD combining a six-step drusen area scale with a five-step pigmentary abnormality scale and is therefore particularly detailed and time-consuming when applied manually. Another more recent system is the Three Continent AMD Consortium severity scale (3CC; Klein et al., 2014), which defines early AMD based on drusen size, drusen area, and the presence of pigmentary abnormalities and is thus more practical to apply manually. While the definition of "advanced AMD" is fairly robust across systems, each system defines "early" or "intermediate" AMD differently, but provides a clear assignment strategy to "no," "early/intermediate," or "advanced AMD" (or "no" and "any AMD").

To obtain an eye-specific AMD status for the 135,500 images of the UK Biobank (≤1 image per eye; 67,100 individuals with images for both eyes, 1,300 with image for only one eye), we applied a published convolutional neural network ensemble (Grassmann et al., 2018) to the fundus images following recommendations of the authors. The ensemble was trained to classify each image into the AREDS nine-step severity scale or three additional categories for advanced AMD (GA, NV, mixed GA + NV, "AREDS9 + 3 steps") or "ungradable." From this, we derived the person-specific automated AMD status as the AMD status of the worse eye (i.e., the higher score of the AREDS9 + 3) or as the status of the only eye, if applicable. We collapsed AREDS AMD severity steps 2–9 or any of the three advanced AMD categories to "any AMD."

To generate internal validation data, we selected a subset of UK Biobank individuals for additional manual grading. When randomly sampling participants, one would expect to catch only few AMD individuals; we thus selected (a) persons with high genetic risk score for AMD based on the known 52 variants for advanced AMD (Fritsche et al., 2016; >99th percentile, $n = 829$); (b) persons with low genetic risk score (<1st percentile, $n = 828$); and (c) persons with self-reported AMD not



Analysis task	Procedure, data, results	
Selection of validation data	**UK Biobank retinal image substudy:** 86,400 individuals; 135,500 fundus images	
	Main study data 66,387 individuals, 131,499 fundus images	**Validation data** 2,013 individuals, 4,001 fundus images
AMD classification	**All individuals:** • automated eye-specific classification towards AREDS 9+3 + "ungradable" classification system • collapsing into 2-stage "any AMD" classification + "ungradable" • derive worse-eye AMD classification ignoring missing/ungradable classifications in single eyes **Validation data:** • additional manual eye-specific classification towards 3CC 5-step + "ungradable" classification system • collapsing into manual 2-stage "any AMD", considering missing/ungradable classification in single eyes	
	Eye- and person-specific automated AREDS and automated "any AMD" classification	Additional eye- and person-specific manual 3CC and manual "any AMD" classification
Restriction to GWAS data	Restriction to individuals with valid data for GWAS based on automated "any AMD" classification: 1) unrelated Europeans, 2) at least one gradable eye (automated classification)	
	Main study data 46,728 individuals, 92,752 fundus images	**Validation data** 1,337 individuals, 2,664 fundus images

FIGURE 1 Schematic diagram of AMD classification and analyzed data. 3CC, Three Continent AMD Consortium severity scale; AMD, age-related macular degeneration; AREDS, Age-Related Eye Disease Study severity scale; GWAS, genome-wide association studies

already selected ($n = 356$). Results of the machine learning-based AMD classification were not used to select individuals into the validation subset and we can therefore validly estimate the algorithm's classification performance (sensitivity/specificity given the manual classification).

Each of the two eyes of the selected 2,013 individuals was manually classified for AMD according to the 3CC system (Klein et al., 2014) by a trained ophthalmologist (five AMD categories, 1 for no AMD, 3 for early, 1 for advanced AMD, and 1 "ungradable"). We collapsed the five AMD categories to "any AMD," "no AMD," or "ungradable" and derived a person-specific AMD status as the AMD status of the worse eye. Assuming neglectable misclassification in the eye-specific manually classified AMD status, this corresponds to the true person-specific AMD status if both eyes are manually gradable or one eye is manually ungradable and the second eye is manually graded as having AMD. If one eye is ungradable and the second, gradable eye is manually classified as "no AMD," the true person-specific disease status is unknown.

We derived eye-specific as well as person-specific confusion matrices based on the detailed (AREDS9 + 3 and five-category 3CC) and collapsed classifications. To conduct the GWAS with automatically derived "any AMD," we restricted the data with available automated AMD classification to unrelated individuals of European ancestry with valid GWAS data (see below), and derived the confusion matrices also for the restricted validation data.

2.6 | Genetic association analyses for AMD without and with accounting for misclassification

We performed a GWAS on the automatically derived "any AMD" versus "no AMD" in unrelated UK Biobank participants (relatedness status >3rd degree) of European ancestry (self-report "White," "British," "Irish," or "Any other White background") as recommended (Loh, Kichaev, Gazal, Schoech, & Price, 2018). For each variant, we applied standard logistic regression (i.e., the naïve analysis ignoring misclassification in the automatically derived AMD status) under the additive genotype model and applied a Wald-test as implemented in QUICKTEST (Kutalik et al., 2011). We included age and the first two genetic principal components as covariates. We excluded variants with low minor allele count (MAC < 400, calculated as MAC = 2 × N × MAF, sample size N, minor allele frequency MAF) or with low imputation quality (rsq < 0.4) yielding 11,567,158 analyzed variants. To correct for potential population stratification, we applied a Genomic Control correction ($\lambda = 1.01$ based on the analyzed variants excluding the 34 known AMD loci; Devlin, Roeder, & Devlin, 2013).

We selected genome-wide significant variants ($p_{GC} < 5.0 \times 10^{-8}$), clumped them into independent regions (≥500 kB between independent regions) and selected the variant with lowest p value in each region ("lead variant"). We also selected 21 of the 34 reported lead variants from the established advanced AMD loci, for which we had ≥80% power to detect them in a UK Biobank sample size of 3,544 cases and 44,521 controls with Bonferroni-adjusted significance—under the assumption that the reported effect sizes for advanced AMD were the true effect sizes and ignoring any misclassification in the AMD classification (Appendix C). Information on linkage disequilibrium in Europeans was obtained from LDLink (Machiela & Chanock, 2015). Enrichment of directionally consistent or enrichment of nominally significant association for the 21 reported lead variants (when compared to the reported direction in literature) was tested based on the Exact Binomial test for H_0:Prob = .5 or H_0:Prob = .05, respectively.

To evaluate the robustness of the genetic association upon accounting for the misclassification, we applied the derived MLAs for the selected variants. For this, we modeled the conditional probability of AMD depending on age, genetic variant, and two genetic principal components (as in the naïve analysis). The MLAs accounted for the misclassification of the eye-specific automated classification and for the person-specific misclassification from missing AMD status in one of two eyes. For the misclassification process of the eye-specific automated classification (quantified by sensitivity and specificity), we allowed for a linear association with age and modeled two scenarios for the association with the genetic variant: (a) no association (nondifferential, MLA1) or (b) linear association (differential misclassification, MLA2). We compared association estimates of the naive analysis with MLA1- and MLA2-analysis and judged significance at Bonferroni-corrected significance levels for a family-wise error rate of 0.05. To allow for comparisons across different models, we did not apply genomic control correction for these comparative analyses. In addition, we evaluated the robustness of findings from the naïve analysis for the selected lead variants upon adjusting for 20 instead of 2 genetic principal components.

To follow-up on the *HERC2* lead variant finding (see Section 3), we quantified lightness of fundus images by calculating gray levels for the "RGB" fundus images (weighted sum of R, G, and B values, $0.30 \times R + 0.59 \times G + 0.11 \times B$, as implemented in IrfanView).

3 | RESULTS

3.1 | Linking misclassification theory to machine learning disease classification

We here establish the usage of machine learning-derived disease classification in genetic association analyses as a response misclassification problem in logistic regression (see Section 2). We present a newly developed maximum likelihood approach (MLA) for *bilateral diseases* like AMD (see Section 2). This includes two versions: (a) assuming *nondifferential misclassification* (MLA1, i.e., no dependency of misclassification probabilities on the covariate of interest, here the genetic variant) and (b) allowing for *differential misclassification* (MLA2, i.e., dependency on the covariate of interest). There are existing MLAs for considering response misclassification in logistic regression using internal validation data (Carroll et al., 2006; Lyles et al., 2011): these MLAs refer to *classic diseases* where the misclassification is on the person-specific disease status. Our developed approach provides a general framework for bilateral diseases with entity-specific misclassification that propagates to person-specific disease misclassification. Our approach also allows for missing classification in one of two entities, which is a second source of bias in association analyses for bilateral diseases as reported previously (Günther et al., 2019). We exemplify our approach on machine learning-derived AMD compared to manually graded AMD. Since machine learning algorithms for AMD are trained on images with human manual AMD grading as benchmark, we assume the manual classification to be gold standard.

We evaluated the performance of the naïve analysis and our developed MLA1 and MLA2 in a simulation study with different misclassification scenarios. By this, we documented substantial bias when the naïve analysis was applied to misclassified data, which was comparable to the theory for classic (nonbilateral) diseases (Carroll et al., 2006; Neuhaus, 1999). Naïve association estimates were biased toward zero in case of nondifferential misclassification and in any direction in case of differential misclassification. In the latter scenario, we observed a lack of type I error control for the naïve analysis. Furthermore, we showed our MLA1 and MLA2 to effectively remove bias and keep type I error when specified correctly (Tables 1 and S1 and Appendix D). In case of differential misclassification, MLA1 (assuming nondifferential misclassification) yields biased estimates and a lack of type I error control as well, comparable to the naïve analysis.

3.2 | AMD in UK Biobank based on automated classification and validation data

We applied a published convolutional neural network ensemble (Grassmann et al., 2018) to automatically derive eye- and person-specific AMD classifications for 68,400 UK Biobank participants with fundus images at

TABLE 1 Simulation results on effect estimates and empirical type I error in naïve and MLA-analysis

Simulation scenario						$\hat{\beta}_Y$						Percent with $p < .05$			Cov. Freq.		
Sens.	Spec.	Percent miss.	β_Y	β_{sens}	β_{spec}	Naïve		MLA1		MLA2		Naive	MLA1	MLA2	Naive	MLA1	MLA2
						Mean	RMSE	Mean	RMSE	Mean	RMSE						
Nondifferential misclassification																	
0.9	0.9	0.25	0	0	0	0.00	0.03	0.00	0.04	0.00	0.04	5.3%	4.6%	4.6%	94.7%	95.4%	95.4%
0.9	0.9	0.25	1	0	0	0.73	0.27	1.00	0.05	1.00	0.05	100%	100%	100%	0.0%	96.5%	96.3%
0.9	0.9	0.75	1	0	0	0.69	0.31	1.00	0.06	1.00	0.07	100%	100%	100%	0.0%	94.4%	93.5%
0.8	0.8	0.25	1	0	0	0.56	0.44	1.00	0.06	1.00	0.07	100%	100%	100%	0.0%	95.0%	95.0%
0.8	0.9	0.25	1	0	0	0.68	0.32	1.00	0.05	1.00	0.06	100%	100%	100%	0.0%	97.0%	95.9%
0.9	0.8	0.25	1	0	0	0.61	0.39	1.00	0.06	1.00	0.06	100%	100%	100%	0.0%	95.3%	94.8%
Differential misclassification																	
0.9	0.9	0.25	0	−1	1	−0.38	0.38	−0.46	0.46	0.00	0.05	100%	100%	4.7%	0.0%	0.0%	95.3%
0.9	0.9	0.25	1	1	−1	1.14	0.14	1.39	0.40	1.00	0.06	100%	100%	100%	4.8%	0.0%	95.1%

Note: We evaluated the performance of naïve and MLA analysis of a quantitative covariate X and a binary bilateral disease Y, for example, person-specific AMD, simulating various scenarios. For each scenario, we sampled 1,000 datasets à 5,000 individuals, 4,000 with only error-prone eye-specific AMD classification, and 1,000 with additional true AMD classification. Shown are performance measures from three models, naïve analysis, MLA1, or MLA2 assuming nondifferential/differential misclassification regarding X, respectively, in various simulation scenarios. For the eight scenarios shown here, we assumed no association of X with δ, the probability of AMD in both eyes given ≥1 affected eye; results were similar when modeling an association of *X* with δ, see Table S1. For each model and scenario, we report mean effect estimates $\hat{\beta}_Y$, log OR per unit increase in standard-normal *X*, over all simulation runs, and the associated root mean squared error (RMSE), fraction of nominally significant effect estimates (% with $p < .05$), and coverage frequencies of 95% CI. %miss., fraction of randomly selected individuals with missing AMD classification in one of two eyes; sens/spec, average sensitivity and specificity of error-prone, eye-specific AMD classification; β_Y, log OR of *X* on true AMD; β_{sens}, log OR of *X* on the sensitivity; β_{spec}, log OR of *X* on the specificity of the eye-specific misclassification process.

Abbreviations: AMD, age-related macular degeneration; MLA, maximum likelihood approach; OR, odds ratio; RMSE, root mean square error.

TABLE 2 Confusion matrices comparing manual and automated AMD classification per eye and per person

(a) Per eye (4,001 eyes, 2,013 individuals)				
	Automated classification			
Manual	Ungradable	No AMD	Any AMD	Sum
Ungradable	813 (74%)	185 (17%)	103 (9%)	1101 (100%)
No AMD	107 (4%)	2207 (90%)	138 (6%)	2452 (100%)
Any AMD	20 (4%)	103 (23%)	325 (73%)	448 (100%)
(b) Per person (1,337 individuals)				
	Automated classification			
Manual classification		No AMD	Any AMD	Sum
Ungradable/NA[a] (NA)		210 (80%)	53 (20%)	263 (100%)
No AMD		750 (91%)	72 (9%)	822 (100%)
Any AMD		58 (23%)	194 (77%)	252 (100%)

Note: Shown are absolute numbers and conditional classification probabilities, that is, in row i and column j,P(automated = j | manual = i) as %, with i, j = "Ungradable," "No AMD," "Any AMD": (a) for all eyes in the validation data; 4,001 eyes of 2,013 individuals. (b) For all individuals in the overlap between validation data and GWAS; 1,337 individuals, all gradable with automated classification.

Abbreviations: AMD, age-related macular degeneration; GWAS, genome-wide association studies.

[a]NA, true AMD status based on worse eye not available, since one eye was manually ungradable and the second AMD-free.

baseline (135,000 eyes; Table S2a). From this, we derived eye-specific "any AMD" status (i.e., any early AMD stage or advanced AMD versus AMD-free) and person-specific "any AMD" status based on the worse eye (see Section 2). Among the 68,400 participants, 10,128 were ungradable for AMD in both eyes by the automated classification (i.e., missing person-specific AMD status by the automated classification, 14.8%), 4,870 were classified as "any AMD" and 53,402 as AMD-free (Table S2b). Among the 58,272 automatically gradable participants (of these: 20.2% automatically gradable only in one eye), 8.4% had AMD and 91.6% were AMD-free. This included 48,065 unrelated individuals of European ancestry with GWAS data (3,544 "any AMD" cases, 44,521 AMD-free controls; 19.8% with only one eye gradable; Table S2b).

To quantify the performance of automated AMD classification, we manually classified AMD in a subset as internal validation data (4,001 images, ≤1 image per eye, 2,013 individuals). When comparing automated to manual (true) "any AMD" status, we found an eye-specific sensitivity of 73% and specificity of 90% in the full validation data and a person-specific sensitivity of 77% and specificity of 91% among the participants in the GWAS (Table 2). We found no structural differences between the full validation data and when restricting to the GWAS data (1,337 individuals, Table S3a,b). Both, the manual and automated classification included the category "ungradable." Among the 4,001 eyes, 1,101 were manually ungradable, of which the automatic classification yielded 74% as ungradable as well, but classified 9% as AMD and 17% as AMD-free, which raises concerns about these classifications. In summary, we found the automated classification to yield reasonable, but error-prone results.

3.3 | GWAS on automated AMD classification in naïve analysis identifies two loci

While we have some idea about the extent of the misclassification from validation data and about its impact on genetic association estimates from simulations, it is unclear whether the automated any AMD classification is "good enough" for GWAS. We conducted a GWAS for person-specific automatically derived "any AMD" in UK Biobank (3,544 "any AMD" cases; 44,521 controls) applying logistic regression as usual, which is without accounting for misclassification (naïve analysis). We found 53 variants with genome-wide significance ($p_{GC} < 5.0 \times 10^{-8}$) spread across two distinct loci (defined as lead variant and proxies +/− 500 kB, Figure 2a,b; Table S4a): the known *ARMS2/HTRA1* locus (lead variant here rs370974631, $p_{GC} = 3.1 \times 10^{-20}$, effect allele frequency [EAF] = 0.23) and an unknown locus for AMD near *HERC2* (lead variant rs12913832, $p_{GC} = 4.7 \times 10^{-16}$, EAF = 0.23). This *ARMS2/HTRA1* lead variant was highly correlated to the reported lead variant for advanced AMD, rs3750846, and effect estimates were

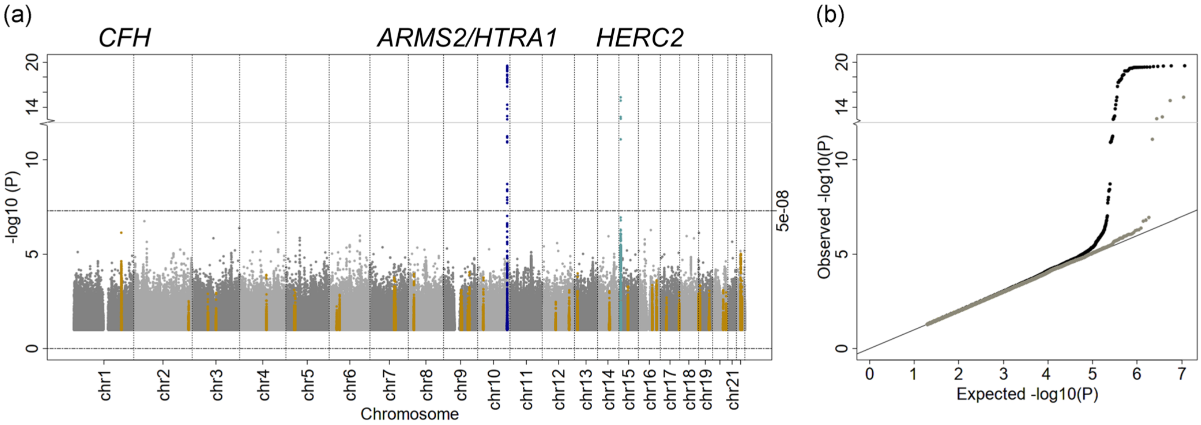


FIGURE 2 GWAS results in UK Biobank based on automatically derived "any AMD" from naïve analysis. Association analyses were conducted using the error-prone, machine learning-derived AMD classification in UK Biobank participants with 3,544 "any AMD" cases and 44,521 controls via logistic regression adjusted for age and two genetic principal components, the *naïve analysis* ignoring misclassification. Shown are (a) Manhattan plot of 11,567,158 analyzed variants; dark blue: genome-wide significant and previously established (Fritsche et al., 2016) locus, light blue: unknown genome-wide significant locus, orange: other 33 previously established loci for advanced AMD), and (b) expected versus observed −log10 *p* values; black: all variants, gray: all variants outside the 34 previously reported loci. 3CC, Three Continent AMD Consortium; AMD, age-related macular degeneration; GWAS, genome-wide association studie

directionally consistent ($r^2 = .93$; Table S4b). The next best known locus is the *CFH* locus, which showed close to genome-wide significance here (smallest *p* value $p_{GC} = 7.0 \times 10^{-7}$, rs6695321, EAF = 0.62): rs6695321 is in linkage disequilibrium with two reported *CFH* variants (rs61818925, rs570618: $r^2 = .63$ or $r^2 = .40$, D′ = 0.81 or D′ = 1.00, EAF = 0.58 or 0.36, respectively; Table S4b) suggesting that rs6695321 captures the signals of these two reported variants.

Among the reported lead variants of the 34 advanced AMD loci (Fritsche et al., 2016), we had ≥80% power to detect 21 of these with Bonferroni-adjusted significance (Table S5). When comparing effect sizes of these 21 variants from this analysis on "any AMD" in UK Biobank with reported effect sizes for advanced AMD, we found 15 with directional consistency ($p_{\text{Bin}} = 0.078$) and 7 with directionally consistent nominal significance ($p_{\text{Bin}} = 4.9 \times 10^{-5}$; Figure 4a and Table S4c). The overall smaller effect sizes for automated "any AMD" compared to reported effect sizes for advanced AMD can be explained by a bias from misclassified automated AMD and by smaller effect sizes for early AMD merged into the definition of "any AMD." For the other 13 of the 34 variants, we refrained from interpreting results due to lack of power in this analysis (Table S4c). Results were similar when adjusting for 20 instead of 2 genetic principal components (data not shown). While the yield of only few known AMD signals in this UK Biobank GWAS may be disappointing, this is not fully unexpected given an effective sample size (Ma, Blackwell, Boehnke, Scott, & GoT2D investigators, 2013) of 13,130 and a power estimate of ~80% (assuming no misclassification and reported effect sizes) to detect associations with genome-wide significance for only 6 of the 34 established variants (*CFH*, *C2/CFB/SKIV2L, ARMS2/HTRA1*, *C3, APOE, SYN3/TIMP3*; Table S5).

In summary, our GWAS on automated AMD in UK Biobank detected the established *ARMS2/HTRA1* locus, an unknown locus around *HERC2* with genome-wide significance, and the established *CFH* locus to some extent.

3.4 | Applying the developed MLA to account for misclassification for selected variants

Due to our simulation results and theory (Carroll et al., 2006; Neuhaus, 1999), we expected our GWAS on automated (error-prone) AMD to yield biased estimates and, when the misclassification was differential toward the genetic variant, even potentially false signals. We applied our developed MLAs for 26 selected variants: (a) the three lead variants detected here with (near) genome-wide significance (*CFH*: rs6695321, *ARMS2/HTRA1:* rs370974631, *HERC2:* rs12913832), (b) the three reported independent variants in the *CFH* locus with MAF ≥ 5% (rs61818925, rs570618, rs10922109; two of these

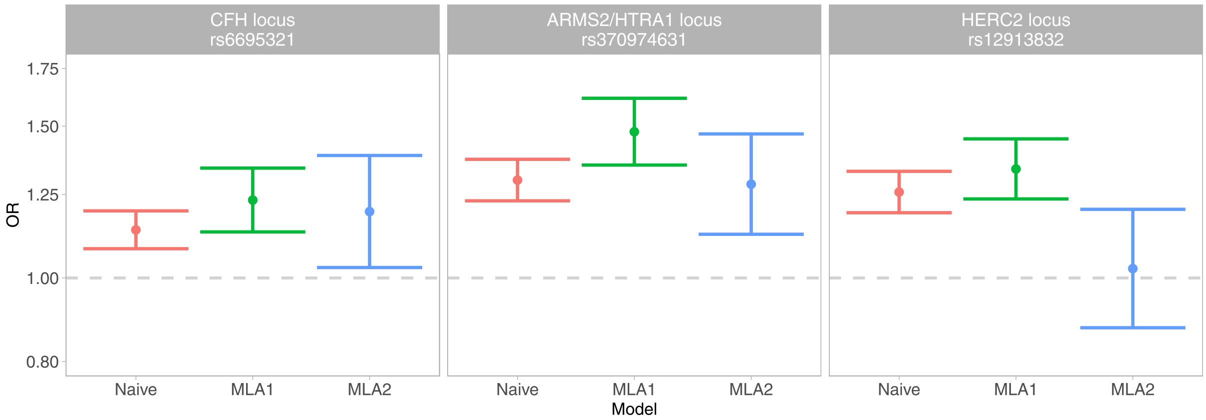


FIGURE 3 Genetic effect estimates for the three lead variants in UK Biobank without and with accounting for misclassification. Shown are genetic effect estimates (odds ratios [OR]) and 95% confidence intervals for three lead variants from the GWAS on automated AMD classification with 3,544 “any AMD” cases and 44,521 controls from three models: without accounting for the misclassification; *naïve analysis*, red. With accounting for nondifferential misclassification, that is, no dependency on the genetic variant; MLA1, green. And accounting for a differential misclassification, that is, dependency on the genetic variant; MLA2, blue. Both MLAs accounted for missing AMD information in one of two eyes and a misclassification associated with age. *Y*-axis is on log-scale. AMD, age-related macular degeneration; GWAS, genome-wide association studies; MLA, maximum likelihood approach

correlated to the here identified *CFH* lead variant), and (c) the other 20 of the 34 reported lead variants (Fritsche et al., 2016), for which we had reasonable power in this analysis (including 1 reported *ARMS2/HTRA1* variant correlated to here identified variant). This yielded a total of ~23 independent variants.

Our MLAs estimated simultaneously (a) sensitivity and specificity of the eye-specific misclassification process and (b) genetic association accounting for the misclassification. With regard to sensitivity and specificity, we found (a) an overall sensitivity of 64.5% (95% CI: 60.1%, 68.7%) and a specificity of 98.6% (98.4%, 98.8%), that is, a false-negative “any AMD” proportion of 35.5% and a false-positive of 1.4%; (b) few evidence for an association of the sensitivity with any selected variant ($p > .05/(23 \times 2) = 1.09 \times 10^{-3}$) and no association with the specificity, except for two variants: *HERC2* lead variant, rs12913832, and the reported *CFH* lead variant rs10922109 ($\mathrm{OR}_{\mathrm{spec}} = 0.64$, $p_{\mathrm{spec}} = 7.38 \times 10^{-9}$ and $\mathrm{OR}_{\mathrm{spec}} = 1.36$, $p_{\mathrm{spec}} = 2.29 \times 10^{-4}$, respectively; Table S6 and

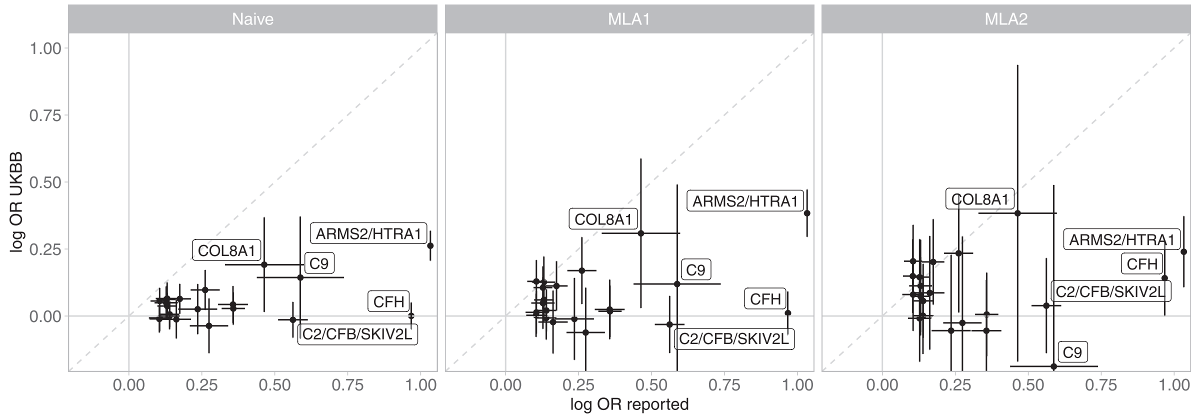


FIGURE 4 Comparison of 21 reported genetic effect estimates for advanced AMD with estimates for automatically derived “any AMD” from UK Biobank without and with accounting for misclassification. We selected the 21 reported AMD lead variants, for which we had ≥80% power to detect them in this UK Biobank sample size with Bonferroni-adjusted significance. Shown are log OR effect estimates and 95% confidence intervals reported for advanced AMD on *x*-axis versus UK Biobank estimates for automatically derived “any AMD” on *y*-axis from the naïve analysis (logistic regression ignoring misclassification), MLA1, and MLA2. AMD, age-related macular degeneration; MLA, maximum likelihood approach; OR, odds ratio

Appendix E). Therefore, we found a misclassification that was associated with some genetic variants (differential), which could induce bias into either direction as well as a severe lack of type I error control.

When comparing genetic association estimates from our MLA1 and MLA2 with the naïve analysis for our three detected lead variants, we found interesting patterns (Figure 3 and Table S7a). (a) For *CFH* and *ARMS2/HTRA1*, we found consistent effect estimates across the three analyses, with larger confidence intervals when using the more complex models MLA1 or MLA2. (b) For *HERC2*, MLA1 yielded comparable results to the naïve analysis, but when accounting for differential misclassification (MLA2), the effect vanished (MLA2: OR = 1.03, $p = .76$; MLA1: OR = 1.34, $p = 1.11 \times 10^{-12}$; naïve: OR = 1.26, $p = 4.16 \times 10^{-16}$). The results of MLA1 for this variant were as expected, since a model considering nondifferential misclassification leads in general, by assumption, to larger estimates and widened confidence intervals if any misclassification is present.

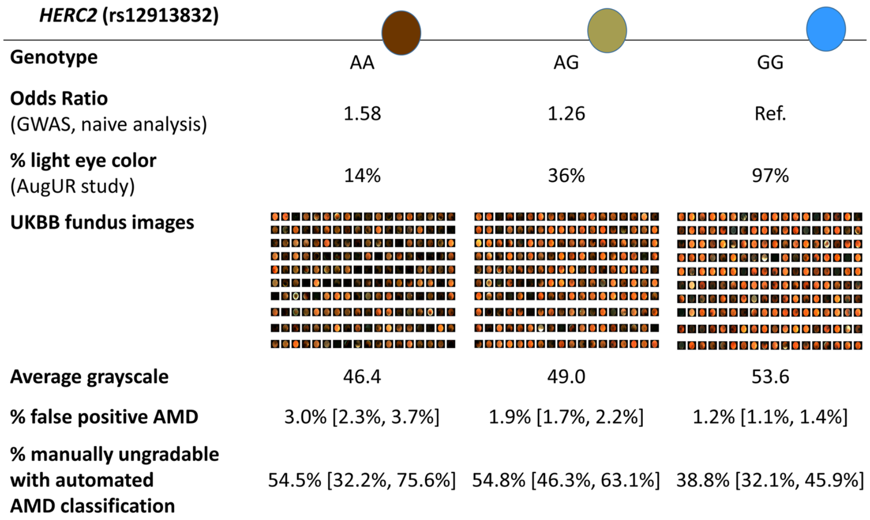
When applying MLA1 and MLA2 to the three reported *CFH* locus variants and the further 20 of the 34 reported lead variants, we found the following (Table S7b,c): (a) Effect estimates for all three *CFH* variants increased when applying MLA2 compared to the naïve analysis. This was particularly interesting for the reported *CFH* lead variant rs10922109, where we now observed a nominally significant association into the reported direction (MLA2: OR = 1.15, $p = .047$; naïve: OR = 1.00, $p = .98$; Table S7c). This is in line with the observed association of the specificity and this *CFH* variant. (b) For the other 20 reported lead variants, many variants showed increased effect estimates by MLA2 compared to the naïve analysis (effect estimates mostly more comparable to reported effect sizes; Fritsche et al., 2016; Figure 4c). Altogether, MLA results confirmed the *CFH* and *ARMS2/HTRA1* loci and unmasked the *HERC2* finding as false positive.

3.5 | Misclassification depended on eye and fundus image color

Interestingly, our *HERC2* lead variant, rs12913832, is precisely the variant for which the G allele was considered causal for blue eyes (Sturm et al., 2008). We were able to support this in our AugUR (Brandl et al., 2018; Stark et al., 2015) study ($n = 1026$; reported "light eye color" for 14%, 36%, or 97% of participants with A/A, G/A, or G/G, respectively). Eye color is discussed as AMD risk factor, but the debate is on blue eyes to increase risk due to increased susceptibility to UV-radiation (Chakravarthy et al., 2010), which is in contrast to our observation of brown eyes to increase AMD risk and a challenge for interpreting this finding. It was interesting to see the *HERC2* rs12913832 association vanish when accounting for rs12913832-associated misclassification. This was in line with the observed strong association of the specificity with this variant ($\mathrm{OR}_{\mathrm{spec}} = 0.64$ per A allele; Table S6a) resulting in 3.0%, 1.9%, or 1.2% of false-positive AMD classifications among persons with A/A, A/G, or G/G, respectively. This notion of a larger misclassification among A/A versus G/G individuals was further supported by the larger fraction of manually ungradable images that were deemed gradable by the automatic classification among A/A versus G/G (54.5% vs. 38.8%, respectively; Figure 5). When visually inspecting fundus images per genotype group, the images for A/A had a darker appearance than those for A/G or G/G (Figure 5), which we were able to quantify by means of average gray level per image of 46.4, 49.0, or 53.6, respectively. Therefore, the *HERC2* signal appeared to be an artifact due to a larger misclassification for brown eyes linked to darker fundus images. One may hypothesize that the darker eye color had reduced light exposure during fundus photography, which gave rise to darker images and more misclassified AMD-free eyes. The notion of a differential misclassification due to eye color was further supported by the fact that the full *HERC2* signal disappeared by modeling a misclassification dependency on the causal variant for eye color (rs12913832; Figure S1a,b), while some signal remained when modeling a misclassification dependency on the respective *HERC2* variant in the model (Figure S1c). In summary, we found the MLA2 not only to effectively remove the artifact signal of the naïve GWAS, but also to help understand the dependencies of the misclassification.

4 | DISCUSSION

GWAS on machine learning-derived classification of imaging-based diseases, like AMD, can be expected to accelerate knowledge gain and drug target development (Nelson et al., 2015), since it will enable substantially increased sample sizes and refined, homogeneous phenotyping. To this date, there was no GWAS reported using a machine learning-derived classification for AMD or any other imaging-based disease—to the best of our knowledge. We here present a GWAS on machine learning-derived AMD in UK Biobank highlighting chances and challenges. By this GWAS on AMD combined with an evaluation of emerging genetic signals via our newly developed MLA, we were able to detect known AMD loci and to distinguish true loci from artifacts.



***HERC2* (rs12913832)**			
Genotype	AA	AG	GG
Odds Ratio (GWAS, naive analysis)	1.58	1.26	Ref.
% light eye color (AugUR study)	14%	36%	97%
UKBB fundus images			
Average grayscale	46.4	49.0	53.6
% false positive AMD	3.0% [2.3%, 3.7%]	1.9% [1.7%, 2.2%]	1.2% [1.1%, 1.4%]
% manually ungradable with automated AMD classification	54.5% [32.2%, 75.6%]	54.8% [46.3%, 63.1%]	38.8% [32.1%, 45.9%]

FIGURE 5 Evidence for differential misclassification in automatically derived AMD with respect to the *HERC2* variant rs12913832. Shown are (a) estimated odds ratios from the naïve analysis ignoring misclassification and various characteristics per genotype group; (b) the fraction of persons with self-reported "light eye color" in the AugUR study; (c) randomly selected fundus images in UK Biobank; (d) image-lightness quantified by mean average grayscale; (e) proportion of false-positive AMD in the automated classification (1-specificity) and 95% confidence intervals estimated via MLA2; and (f) observed proportion of manually ungradable images that were deemed gradable by the algorithm and classified as "any AMD" or "AMD-free." AMD, age-related macular degeneration; GWAS, genome-wide association study; MLA, maximum likelihood approach

Such artifacts, that is, false positives, can derive from misclassification that is associated with a genetic variant. Our data and analyses provide a compelling example for such an artifact: our MLA revealed the *HERC2* signal as false-positive signal and suggested darker eye color and darker fundus images as a relevant source of misclassification for this machine learning algorithm. It is perceivable that the misclassification process of other algorithms for AMD and for other image-based diseases will depend on one or the other characteristic as well, and that such a characteristic is picked up by some genetic variants due to the abundant range of genetically pinpointed characteristics (see, e.g., NHGRI-EBI GWAS Catalog; Buniello et al., 2019), which can yield artifact signals when left unaccounted.

Our MLA, developed for bilateral diseases, does not only quantify the misclassification and the dependencies, but also guards against bias and artifacts in association analyses. Our approach has certain limitations: since we use statistical modeling for the error-prone classification, the analysis is only valid if the corresponding assumptions hold. This concerns independence of entity-specific classification given the true disease status, the correct specification of the misclassification model based on the validation data, and a neglectable error in the gold standard classification. Similar approaches are available for classic diseases (Carroll et al., 2006; Lyles et al., 2011). Thus, this concept can be generalized to other algorithms and other image-based diseases. Our work here links the theory of misclassification to machine learning-derived disease classification, which can be generalized also to measurement error and quantitative phenotypes.

We recommend a GWAS combined with a post-GWAS evaluation of emerging genetic effects for nondifferential and differential misclassification not only to search for GWAS signals on image-based, machine learning-derived disease phenotypes. We also recommend such a GWAS as a quality control for diseases like AMD, where strong genetic signals are known: a GWAS on AMD ascertained by any classification approach, manual or automatic, should be able to detect at least the two strong known signals around *ARMS2/HTRA1* and *CFH*. When a GWAS does not detect these signals, this indicates issues that can be anything from mismatched biosamples, analytical errors, or imperfect disease ascertainment—like from machine learning algorithms as highlighted here. A GWAS can be a quick guide toward phenotype classification quality when genomic data are available.

Overall, we illustrate chances and challenges of machine learning-derived disease classification in GWAS, and the applicability of our MLA to guard against bias and artifacts.

ACKNOWLEDGMENTS

This study was supported by the DFG HE 3690/5-1 (to I. M. H.) and NIH R01 EY RES 511967 (to I. M. H.), the University of Regensburg and the Ludwig Maximilians University Munich. The UK Biobank (accessed via application number 33999) was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government, and the Northwest Regional Development Agency. This study was also supported by the Welsh Assembly Government, British Heart Foundation and Diabetes UK. The authors would also like to thank the two anonymous reviewers whose comments helped improve and clarify this manuscript. Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

DATA AVAILABILITY STATEMENT

Data that support the findings of this study are available as UK Biobank resource (accessed via application number 33999). The fundus-image derived AMD classifications will be returned to UK Biobank and can be accessed by other researchers via the Data Showcase. An open source R implementation of the developed maximum likelihood approach to account for misclassification in bilateral disease is available at: https://www.genepi-regensburg.de/MLA-bilateral/. The convolutional neural network ensemble used for automated AMD classification and recommendations by the authors can be found at: https://github.com/RegensburgMedicalImageComputing/ARIANNA. IrfanView: https://www.irfanview.com/; GWAS catalogue: https://www.ebi.ac.uk/gwas/.

ORCID

Felix Guenther http://orcid.org/0000-0001-6582-1174
Caroline Brandl https://orcid.org/0000-0001-8223-6137
Thomas W. Winkler https://orcid.org/0000-0003-0292-5421
Klaus Stark https://orcid.org/0000-0002-7832-1942
Helmut Kuechenhoff https://orcid.org/0000-0002-6372-2487

REFERENCES

Brandl, C., Zimmermann, M. E., Günther, F., Barth, T., Olden, M., Schelter, S. C., … Heid, I. M. (2018). On the impact of different approaches to classify age-related macular degeneration: Results from the German AugUR study. *Scientific Reports*, *8*(1), 8675. https://doi.org/10.1038/s41598-018-26629-5

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Burlina, P. M., Joshi, N., Pekala, M., Pacheco, K. D., Freund, D. E., & Bressler, N. M. (2017). Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmology*, *135*(11), 1170. https://doi.org/10.1001/jamaophthalmol.2017.3782

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., … Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Chakravarthy, U., Wong, T. Y., Fletcher, A., Piault, E., Evans, C., Zlateva, G., … Mitchell, P. (2010). Clinical risk factors for age-related macular degeneration: A systematic review and meta-analysis. *BMC Ophthalmology*, *10*(1), 31. https://doi.org/10.1186/1471-2415-10-31

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Routledge. https://doi.org/10.4324/9780203771587

Csurka, G. (2017). A comprehensive survey on domain adaptation for visual applications, *Domain Adaptation in Computer Vision Applications* (pp. 1–35). Cham, Switzerland: Springer.

Davis, M. D., Gangnon, R. E., Lee, L.-Y., Hubbard, L. D., Klein, B. E. K., & Klein, R., … Age-Related Eye Disease Study Group. (2005). The Age-Related Eye Disease Study severity scale for age-related macular degeneration: AREDS Report No. 17. *Archives of Ophthalmology*, *123*(11), 1484–1498. https://doi.org/10.1001/archopht.123.11.1484

Devlin, A. B., Roeder, K., & Devlin, B. (2013). Genomic control for association. *Biometrics*, *55*(4), 997–1004.

Fritsche, L. G., Igl, W., Bailey, J. N. C., Grassmann, F., Sengupta, S., Bragg-Gresham, J. L., … Heid, I. M. (2016). A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics*, *48*(2), 134–143. https://doi.org/10.1038/ng.3448

Grassmann, F., Mengelkamp, J., Brandl, C., Harsch, S., Zimmermann, M. E., Linkohr, B., … Weber, B. H. F. (2018). A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology*, *125*(9), 1410–1420. https://doi.org/10.1016/j.ophtha.2018.02.037

Günther, F., Brandl, C., Heid, I. M., & Küchenhoff, H. (2019). Response misclassification in studies on bilateral diseases. *Biometrical Journal*, *61*(4), 1033–1048. https://doi.org/10.1002/bimj.201900039

Hausman, J. A., Abrevaya, J., & Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, *87*(2), 239–269. https://doi.org/10.1016/S0304-4076(98)00015-3

Heinze-Deml, C., & Meinshausen, N. (2017). Conditional variance penalties and domain shift robustness. arXiv:1710.11469 [stat.ML].

Keane, P. A., Grossi, C. M., Foster, P. J., Yang, Q., Reisman, C. A., & Chan, K., … UK Biobank Eye Vision Consortium. (2016). Optical coherence tomography in the UK Biobank study–Rapid automated analysis of retinal thickness for large population-based studies. *PLoS One*, *11*(10), e0164095. https://doi.org/10.1371/journal.pone.0164095

Klein, R., Meuer, S. M., Myers, C. E., Buitendijk, G. H. S., Rochtchina, E., Choudhury, F., … Klein, B. E. K. (2014). Harmonizing the classification of age-related macular degeneration in the three-continent AMD Consortium. *Ophthalmic Epidemiology*, *21*(1), 14–23. https://doi.org/10.3109/09286586.2013.867512

Kutalik, Z., Johnson, T., Bochud, M., Mooser, V., Vollenweider, P., Waeber, G., … Bergmann, S. (2011). Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics*, *12*(1), 1–17. https://doi.org/10.1093/biostatistics/kxq039

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*(1995), 60–88. https://doi.org/10.1016/j.media.2017.07.005

Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., & Price, A. L. (2018). Mixed-model association for biobank-scale datasets. *Nature Genetics*, *50*(7), 906–908. https://doi.org/10.1038/s41588-018-0144-6

Lyles, R. H., Tang, L., Superak, H. M., King, C. C., Celentano, D. D., Lo, Y., & Sobel, J. D. (2011). Validation data-based adjustments for outcome misclassification in logistic regression: An illustration. *Epidemiology*, *22*(4), 589–597. https://doi.org/10.1097/EDE.0b013e3182117c85

Ma, C., Blackwell, T., Boehnke, M., Scott, L. J., & GoT2D Investigators (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic Epidemiology*, *37*(6), 539–550. https://doi.org/10.1002/gepi.21742

Machiela, M. J., & Chanock, S. J. (2015). LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, *31*(21), 3555–3557. https://doi.org/10.1093/bioinformatics/btv402

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., & Teumer, A., … Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, *45*(1), 521–530. https://doi.org/10.1016/j.patcog.2011.06.019

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., … Sanseau, P. (2015). The support of human genetic evidence for approved drug indications. *Nature Genetics*, *47*(8), 856–860. https://doi.org/10.1038/ng.3314

Neuhaus, J. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, *86*(4), 843–855. https://doi.org/10.1093/biomet/86.4.843

Peng, Y., Dharssi, S., Chen, Q., Keenan, T. D., Agrón, E., Wong, W. T., … Lu, Z. (2019). DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*, *126*(4), 565–575. https://doi.org/10.1016/j.ophtha.2018.11.015

R Core Team. (2019). *R: A language and environment for statistical computing*. Retrieved from https://www.r-project.org/

Stark, K., Olden, M., Brandl, C., Dietl, A., Zimmermann, M. E., Schelter, S. C., … Heid, I. M. (2015). The German AugUR study: Study protocol of a prospective study to investigate chronic diseases in the elderly. *BMC Geriatrics*, *15*(1), 130. https://doi.org/10.1186/s12877-015-0122-0

Stephane, C. (2018). pwr: Basic functions for power analysis. Retrieved from https://cran.r-project.org/package=pwr

Sturm, R. A., Duffy, D. L., Zhao, Z. Z., Leite, F. P. N., Stark, M. S., Hayward, N. K., … Montgomery, G. W. (2008). A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *American Journal of Human Genetics*, *82*(2), 424–431. https://doi.org/10.1016/j.ajhg.2007.11.005

Ting, D. S. W., Cheung, C. Y.-L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., … Wong, T. Y. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Journal of the American Medical Association*, *318*(22), 2211–2223. https://doi.org/10.1001/jama.2017.18152

Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., … Zhang, W. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*(7571), 82–89. https://doi.org/10.1038/nature14962

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Guenther F, Brandl C, Winkler TW, et al. Chances and challenges of machine learning-based disease classification in genetic association studies illustrated on age-related macular degeneration. *Genetic Epidemiology*. 2020;44:759–777. https://doi.org/10.1002/gepi.22336

APPENDIX A: MLA TO ADJUST FOR RESPONSE MISCLASSIFICATION IN BILATERAL DISEASES

We developed a maximum likelihood approach (MLA) to adjust for response misclassification from an error-prone, entity-specific disease classification in bilateral diseases. Here, we illustrate it based on the example of age-related macular degeneration, where AMD can occur in each eye (eye-specific AMD) and the person-specific binary outcome is defined as worse eye outcome, that is, "AMD in at least one eye," and modeled using logistic regression. We assume that we have an error-prone, eye-specific AMD classification (e.g., from a machine learning-based automated classification) available for nearly all eyes and true, gold-standard classifications (e.g., manual classification) for a subset of individuals from validation data.

Let $(Z_{1i}, Z_{2i}) \in \{0, 1\}$ be the true, binary disease stages in the two eyes of study participant i, that is, $(Z_{1i} = 1, Z_{2i} = 0)$ means that participant i suffers from AMD in the left eye and is unaffected from AMD in the right. When estimating the association of person-specific risk factors with AMD, one often defines a binary person-specific disease status as worse eye AMD,$\mathrm{Y_i} := \max(\mathrm{Z_{1i}}, \mathrm{Z_{2i}})$, $Z_{1i}, Z_{2i} \in \{0, 1\}$, and uses logistic regression to estimate the association of some covariates X with AMD: the person-specific disease status Y_i equals 1, if at least one eye of individual i is classified as AMD, and Y_i equals 0, if both eyes are unaffected. As described previously (Günther et al., 2019), such a worse eye disease status can be misclassified because of two reasons: either, because of missing disease information in one of two eyes (in this case disease can be overlooked), or because of error-prone disease status for any of the two eyes. Here, we assume that we observed an error-prone, eye-specific disease status (Z^*_{1i}, Z^*_{2i}) for each of the two eyes of a "main study" participant i and additionally the true disease status in each of the two eyes (Z_{1j}, Z_{2j}) for a subset of study participants j from the "validation study." For all participants from the main study (error-prone classifications only) or the validation subset (error-prone and true classification), there is the additional issue that the disease information can be missing in one of two eyes, because of missing or ungradable fundus images. Since the automated (error-prone) and manual (gold standard, "true") classification may judge differently on whether an image is gradable or ungradable, any possible subset of $(Z_{1i}, Z_{2i}, Z^*_{1i}, Z^*_{2i})$ might be the available information for a specific study participant. To obtain valid estimates for the association of covariates with the true AMD status, we set up a likelihood based on the conditional probabilities of the observed error-prone and/or true eye-specific disease classifications given covariates. The product of these conditional probabilities over all individuals forms the likelihood, which has to be numerically optimized with respect to the regression parameters to obtain estimates. The different likelihood contributions for the individuals depend on the available AMD classifications (true and/or error-prone for one or both eyes).

The general problem of response misclassification when AMD information is missing in one of two eyes and/or the eye-specific classification suffers from misclassification with known classification probabilities has already been evaluated in a previous publication (Günther et al., 2019). There, we also derived the corresponding likelihood contributions for the different scenarios of available outcome data. Here, we add the aspect that validation data are available for some study participants or, more specifically, a collection of error-free (gold-standard) classified single eyes, and that we model the eye-specific misclassification process based on information from this validation data.

In the following, we describe the general idea and provide formulas for the respective likelihood contributions.

The assumed logistic regression model for the true worse eye disease corresponds to the assumption that $\max(Z_{1i}, Z_{2i}) = Y_i \sim \mathrm{Bernoulli}(\pi_i)$, where we model the success probability based on a linear predictor via $\pi_i = 1/(1 + \exp(-x'_i\beta)) = \mathrm{Logist}(x'_i\beta)$; x_i is a vector of observed person-specific covariates and β the vector of corresponding regression coefficients. It follows that $P(Y_i = 1|x_i) = \pi_i$. If we focus on single-eye disease classifications, there exist four different pattern of true disease classifications (Z_{1i}, Z_{2i}): $(1, 1), (1, 0)$,$(0, 1), (0, 0)$. From the assumed logistic regression model for Y_i, it follows that. $P(Z_{1i} = 0, Z_{2i} = 0|x_i) = 1 - \pi_i$ Based on the law of total probability, we can derive $P(Z_{1i} = 1, Z_{2i} = 1|x_i) = P(Z_{1i} = 1, Z_{2i} = 1|x_i, Y_i = 1) \times P(Y_i = 1|x_i)$ and we define the person-specific conditional probability of being affected by AMD in both eyes given AMD in at least one eye as $\delta_i := P(Z_{1i} = 1, Z_{2i} = 1|x_i, Y_i = 1)$. When assuming symmetric probabilities for disease in one but not the other eye for left and right eyes (i.e., same probabilities to be affected in the left but not the right eye and vice versa), the conditional probability mass function of the two-entity disease status distribution can be written concisely as

$P(\cdot,\cdot\,|x_i)$	$Z_{2i} = 1$	$Z_{2i} = 0$
$Z_{1i} = 1$	$\delta_i\pi_i$	$\frac{1-\delta_i}{2}\pi_i$
$Z_{1i} = 0$	$\frac{1-\delta_i}{2}\pi_i$	$1 - \pi_i$

(1)

which specifies the *true data model*. If we look at a single eye selected randomly from both eyes, we can derive (without loss of generality for $\mathrm{Z_{1i}}$)

$$\mathrm{P(Z_{1i} = 1|x_i) = P(Z_{1i} = 1, Z_{2i} = 1|x_i)} \\ \mathrm{+ P(Z_{1i} = 1, Z_{2i} = 0|x_i)} = \left(\frac{1}{2} + \frac{1}{2}\delta_\mathrm{i}\right)\pi_\mathrm{i}. \quad (2)$$

We now assume that we observed potentially misclassified single eye disease stages (Z^*_{1i}, Z^*_{2i}) for each participant and describe the *misclassification process* based on the sensitivity and specificity of the classification

$$\begin{aligned} P(Z^*_{li} = 1|Z_{li} = 1, x_i) &= \pi_{1i} \\ P(Z^*_{li} = 0|Z_{li} = 0, x_i) &= \pi_{0i} \end{aligned} \tag{3}$$

with $l = 1, 2$; π_{1i} and π_{0i} are the person-specific sensitivity and specificity from the eye-specific classification process. We assume that the eye-specific classification process within an individual is independent in the two eyes, that is

$$\begin{aligned} &P(Z^*_{1i} = z^*_{1i}, Z^*_{2i} = z^*_{2i}|Z_{1i} = z_{1i}, Z_{2i} = z_{2i}, x_i) \\ &\quad = P(Z^*_{1i} = z^*_{1i}|Z_{1i} = z_{1i}, x_i) \times P(Z^*_{2i} = z^*_{2i}|Z_{2i} \\ &\quad = z_{2i}, x_i). \end{aligned}$$

Based on the *true data model* and the description of the *misclassification process* via sensitivity and specificity, we can now express the conditional probabilities of all combinations of observed outcomes, by using Bayes' rule and the law of total probability. If all four AMD classifications were observed for an individual (individual with full validation data, true and error-prone disease status for each of the two eyes), we can derive the following (omitting a random variable notation and only using the small z's for the observed data):

$$\begin{aligned} P(z^*_{1i}, z^*_{2i}, z_{1i}, z_{2i}|x_i) &= P(z^*_{1i}, z^*_{2i}|z_{1i}, z_{2i}, x_i) \times P(z_{1i}, z_{2i}|x_i) \\ &= P(z^*_{1i}|z_{1i}, x_i) \times P(z^*_{2i}|z_{2i}, x_i) \\ &\quad \times P(z_{1i}, z_{2i}|x_i). \end{aligned}$$

Here, we fraction the conditional probability of the observed data into terms of the eye-specific classification process (depending on sensitivity or specificity when the observed true outcome z_{li} is 1 or 0, respectively, Equation 3) and the true data model (1). If only the two eye-specific error-prone classifications are observed (individual in the main study, not part of the validation subset), the law of total probability can be used and the conditional probability can be expressed as

$$\begin{aligned} P(z^*_{1i}, z^*_{2i}|x_i) &= \sum_{z_{1i}, z_{2i} \in \{0,1\}} P(z^*_{1i}, z^*_{2i}|z_{1i}, z_{2i}, x_i) \times P(z_{1i}, z_{2i}, |x_i) \\ &= \sum_{z_{1i}, z_{2i} \in \{0,1\}} P(z^*_{1i}|z_{1i}, x_i) \times P(z^*_{2i}|z_{2i}, x_i) \\ &\quad \times P(z_{1i}, z_{2i}|x_i). \end{aligned}$$

This again yields an expression that depends on the eye-specific classification probabilities (3) and the *true data model* (1).

If only a classification for one error-prone outcome was observed (e.g., $Z^*_{1i} = z^*_{1i}$), the conditional probability is given by

$$\begin{aligned} P(z^*_{1i}|x_i) &= P(z^*_{1i}|Z_{1i} = 0, x_i) \times P(Z_{1i} = 0|x_i) \\ &\quad + P(Z^*_{1i}|Z_{1i} = 1, x_i) \times P(Z_{1i} = 1|x_i), \end{aligned}$$

where the first terms in each summand depend on the specificity and the sensitivity of the eye-specific observation process; an expression for the second was already given above (Equation 2).

When three classifications were observed, for example, $(Z_{1i} = z_{1i}, Z^*_{1i} = z^*_{1i}, Z^*_{2i} = z^*_{2i})$ or $(Z_{1i} = z_{1i}, Z_{2i} = z_{2i}, Z^*_{1i} = z^*_{1i})$, we can derive

$$\begin{aligned} P(z_{1i}, z^*_{1i}, z^*_{2i}|x_i) &= P(z^*_{1i}, z^*_{2i}|z_{1i}, Z_{2i} = 0, x_i) \\ &\quad \times P(z_{1i}, Z_{2i} = 0 \mid x_i) \\ &\quad + P(z^*_{1i}, z^*_{2i} \mid z_{1i}, Z_{2i} = 1, x_i) \\ &\quad \times P(z_{1i}, Z_{2i} = 1|x_i) \\ &= P(z^*_{1i} \mid z_{1i}, x_i) \times P(z^*_{2i}|Z_{2i} = 0, x_i) \times P(z_{1i}, Z_{2i} \\ &= 0 \mid x_i) + P(z^*_{1i}|z_{1i}, x_i) \times P(z^*_{2i}| Z_{2i} = 1, x_i) \\ &\quad \times P(z_{1i}, Z_{2i} = 1 \mid x_i), \end{aligned}$$

and

$$\begin{aligned} P(z_{1i}, z_{2i}, z^*_{1i}, |x_i) &= P(z^*_{1i}|z_{1i}, z_{2i}, x_i) \times P(z_{1i}, z_{2i}|x_i) \\ &= P(z^*_{1i}|z_{1i}, x_i) \times P(z_{1i}, z_{2i}|x_i). \end{aligned}$$

All conditional probabilities characterizing the *true data model* and the *misclassification process*, that is, (a) the probability of true worse eye AMD $P(Y_i = 1|x_i) = \pi_i$, (b) the probability of AMD in both eyes given AMD in at least one eye $P(Z_{1i} = 1, Z_{2i} = 1|Y_i = 1, x_i) = \delta_i$, (c) the eye-specific sensitivity $P(Z^*_{1i} = 1|Z_{1i} = 1, x_i) = \pi_{1i}$, and (d) the eye-specific specificity of the error-prone classification $P(Z^*_{1i} = 0|Z_{1i} = 0, x_i) = \pi_{0i}$, can potentially vary with person-specific characteristics. We therefore decided to model them based on the logistic function of a linear predictor, where relevant covariates can be specified for each probability. Combining all these expressions, we can set up the whole likelihood based on the derived conditional probabilities and numerically optimize with respect to the regression coefficients of the linear predictors for π_i, δ_i, π_{1i}, and π_{0i}. Standard errors of the maximum likelihood estimates are derived based on standard likelihood theory from the square root of the diagonal elements of the inverse of the observed Fisher information (Hessian) and used for inference. An implementation of the MLA in the statistical programming language R (R Core Team, 2019) is available.

APPENDIX B: SIMULATION STUDY TO EVALUATE CONSEQUENCES OF IGNORING MISCLASSIFICATION AND THE PERFORMANCE OF THE MLA IN CORRECTING IT

We performed a simulation study to evaluate the consequences of ignoring response misclassification and to evaluate the performance of the derived MLA in data scenarios similar to the situations in AMD studies. For each simulation scenario (data generating process), we simulated 1,000 datasets, applied different models to the sampled data, and evaluated the distribution of effect estimates, frequencies of significant statistical tests, and coverage frequencies of confidence intervals for a central covariate of interest.

To sample data mimicking studies on AMD with internal validation data, we performed the following steps.

1. We sampled the true binary “worse-eye” AMD data Y for 5,000 individuals by sampling from a Bernoulli distribution, where we modeled the success probability based on the logistic function of a linear predictor (corresponding to the assumed data generating process in logistic regression). For the linear predictor, we used an intercept of -0.25 (corresponding to an average probability of person-specific AMD of ~0.44) and a continuous standard normal covariate X. We varied the log OR of X on Y between zero (simulation under H_0 of no effect) and one.
2. To create the true eye-specific disease data (two binary observations per individual, (Z_1, Z_2)) we specified the conditional probability of being affected in both eyes given disease in at least one eye (i.e., Y = 1 based on “worse-eye definition), δ, to be (on average) $\delta = 1/(1 + \exp(-1)) = 0.73$. We assumed this probability to be either constant or varying with the continuous covariate X based on formula $\delta = 1/(1 + \exp(-(1 + 1 \times X))) = \text{Logist}(1 + 1 \times X)$. For all individuals with sampled Y = 1, we sampled a Bernoulli variable based on probability δ, to decide whether they were affected in both eyes or not. If they were affected on only one eye, we sampled randomly from the left or right.
3. To mimic the situation of missing information in one of two eyes, we sampled a Bernoulli random variable for each individual based on a fixed success probability (e.g., 0.75), to indicate whether information on both eyes was available. If not, we removed the disease information from a randomly selected eye.
4. To obtain eye-specific error-prone outcome data (Z_1^*, Z_2^*), we conditioned on the true, sampled observations (Z_1, Z_2), and sampled the error-prone outcomes based on specified classification probabilities, the sensitivity $P(Z^*{=}1|Z = 1)$ and specificity $P(Z^*{=}0|Z = 0)$. Sensitivity and specificity were either fixed (nondifferential misclassification, e.g., sens = spec= 0.9) or varying between individuals based on the formula $\text{sens} = \text{Logist}(2.20 + \beta_{\text{sens}} \times X)$ for different values of β_{sens} (analogously for the specificity, corresponding to an average sens = spec = 0.9).
5. Afterward, we split the data into two parts, the “main study” and the “validation” subset $(n^{\text{val}} = 1,000, n^{\text{main}} = 4,000)$. For the validation subset we kept both, the true and the error-prone eye-specific AMD observations (Z_1, Z_2, Z_1^*, Z_2^*); for the main study, we kept only the error-prone outcomes (Z_1^*, Z_2^*) (or only the respective information for one of the two eyes, when information in one eye was missing for an individual).
6. For the naïve analysis ignoring response misclassification, we defined an observed, binary naïve person-specific outcome Y_{obs}^* the following way: for individuals from the validation data, we used the true eye-specific disease information; for individuals from the main study data, we used the error-prone eye-specific information. When disease information was available for both eyes, we defined $Y_{\text{obs}}^* = \max(Z_1, Z_2)$ or $Y_{\text{obs}}^* = \max(Z_1^*, Z_2^*)$, respectively; for observations with information only on one eye Z_1, we used $Y_{\text{obs}}^* = Z_1$ or $Y_{\text{obs}}^* = Z_1^*$. For individuals from the validation data with information on both eyes, $Y_{\text{obs}}^* = \max(Z_1, Z_2)$ corresponds to the true Y; for all others, Y_{obs}^* might be misclassified.

For each sampled dataset we estimated three models: (a) standard logistic regression based on the error-prone naïve worse entity outcome Y_{obs}^*, (b) the derived MLA (see above) modeling the probability of person-specific AMD and the probability of AMD in both eyes given AMD in at least one eye, δ, based on covariate X, while assuming a constant eye-specific sensitivity and specificity and accounting for missing information in one of two eyes (MLA1), and (c) the derived MLA allowing for a dependency of sensitivity and specificity on X (MLA2).

APPENDIX C: POWER ANALYSIS FOR REPORTED LEAD VARIANTS BASED ON UK BIOBANK SAMPLE SIZE

We wanted to evaluate the impact of using the MLA on selected variants including the 34 reported lead variants known for their association with advanced AMD. Given reported effect sizes and EAFs, we expected the power to detect some of these 34 associations to be limited in a sample size of approximately 3,500 cases and 44,500 controls. Therefore, we aimed to assess the power to detect reported genetic associations for AMD in the available data of UK Biobank, to focus our analyses with the MLA only on adequately powered reported associations and to avoid over-interpreting noisy results from underpowered analyses. It is, however, not fully straight forward how to compute power for the scenario of "any AMD" from machine learning based disease classification, due to the power-diminishing effect of misclassification and some uncertainty of what effect size to use. We chose to use the reported (Fritsche et al., 2016) EAFs in advanced AMD cases and AMD-free controls for the established 34 lead variants and computed the power for a test on differences in (effect allele) fractions for differently sized groups (Cohen, 2013; Stephane, 2018). Group sizes correspond to the automated "any AMD" classification in UK Biobank GWAS data (Table S2). The number of observations in each group is two times the observed number of individuals, that is, $n_{\text{case}} = 2 \times 3{,}500$ and $\text{n}_{\text{contr}} = 2 \times 44{,}500$, since each individual contributes two (independent) alleles.

Based on these power calculations, we selected all lead variants with at least 80% power to yield Bonferroni-corrected ($\alpha = .05/34$) significant associations in UK Biobank. By this, we made the assumptions that EAFs in advanced AMD cases are transferable to EAFs of "any AMD" cases and that no misclassification was present in the machine learning-derived any AMD classification. Therefore, this is probably an overestimate of available power. We performed the power analysis, however, mainly to dismiss variants with an obvious lack of power.

APPENDIX D: MLA AVOIDS BIAS AND EXCESS OF TYPE I ERROR IN SIMULATION STUDIES

In our simulation study, we investigated bias and type I error of logistic regression-based association estimates for a binary worse entity outcome $Y := \max(Z_1, Z_2) \in \{0, 1\}$ and a continuous covariate X, when error-prone single-entity observations $(Z_1^*, Z_2^*) \in \{0{,}1\}$ are observed instead of the true entity-specific disease classifications $(\text{Z}_1, \text{Z}_2) \in \{0{,}1\}$. When utilizing the error-prone observations for deriving the worse entity outcomes $Y^* := \max(Z^*_1, Z^*_2)$, the entity-specific misclassification is passed on to the worse entity disease stage. We compare the performance of the naïve analysis (logistic regression ignoring misclassification) and the two versions of our MLA for different simulation scenarios.

In the naïve analysis, we found a similar pattern for bilateral disease misclassification as reported for classic diseases (Carroll et al., 2006; Neuhaus, 1999): (a) under the null hypothesis (Tables 1 and S1, $\beta_Y = 0$), we found biased estimates and a lack of type I error control (potential for false-positive association findings) for differential misclassification. With nondifferential misclassification, estimates were unbiased and type I error frequencies were at the desired levels. (b) When X was associated with true AMD (Tables 1 and S1, $\beta_Y = 1$), effect estimates were biased toward the null for nondifferential misclassification and into any direction for differential misclassification. Specific for the bilateral disease situation was (c) increasing bias with increasingly missing AMD in one of the two eyes, and (d) a larger bias by decreased specificity than by decreased sensitivity. (Tables 1 and S1).

In logistic regression, the larger the misclassification probabilities, the larger the bias of estimates (Neuhaus, 1999), with similar influence of increased probabilities for false-positive and false-negative classifications for balanced data. In the following, we provide an explanation of the findings (c) and (d) for bilateral diseases from above. Finding (c) is explained by the fact that an increased fraction of missing eyes implies a reduced sensitivity for person-specific AMD: AMD in the missing eye can be overlooked, which can lead to a false-negative person-specific AMD classification if only the missing eye of an individual is affected. Finding (d) was that decreased specificity had larger impact on bias than decreased sensitivity, for example, for (sens, spec) = (0.9, 0.9) and a fraction of 25% of individuals with "missing eyes" and a true log OR of X on Y of 1 the observed bias was −0.27. When the sensitivity was reduced to 0.8 (specificity = 0.9), the bias increased (in absolute value) to −0.32; when the specificity was reduced to 0.8 (sensitivity = 0.9), the bias increased to −0.39. This can be explained by rewriting the probability of misclassification in the worse entity outcome, $\text{P}(\text{Y}^* \neq \text{Y})$ as

$$
\begin{aligned}
\text{P}(\text{Y}^* \neq \text{Y}) &= \text{P}(\text{Y}^* = 1 | \text{Y} = 0)\text{P}(\text{Y} = 0) \\
&\quad + \text{P}(\text{Y}^* = 0 | \text{Y} = 1)\text{P}(\text{Y} = 1) \\
&= \text{P}(\max(Z_1^*, Z_2^*) = 1 | \text{Z}_1 = 0, \text{Z}_2 = 0)\text{P}(\text{Y} = 0) \\
&\quad + \text{P}(Z_1^* = 0, Z_2^* = 0 | \max(\text{Z}_1, \text{Z}_2) = 1)\text{P}(\text{Y} = 1) \\
&= (1 - \text{spec}^2)\text{P}(\text{Y} = 0) + ((1 - \text{sens})^2\delta \\
&\quad + \text{spec}(1 - \text{sens})(1 - \delta))\text{P}(\text{Y} = 1),
\end{aligned}
$$

This illustrates the dependency of $P(Y^* \neq Y)$ on entity-specific sensitivity, specificity, probability of disease in both entities given disease in one eye δ, and the fraction of truly affected individuals $P(Y = 1)$. This probability can be evaluated for different combinations of parameters: for example, in the simulation study, we assumed $P(Y = 1) = 0.44$, $\delta = 0.75$ (Appendix B), which leads to a misclassification probability of 12%, 14%, or 22% for (sens, spec) = (0.9, 0.9), (sens, spec) = (0.8, 0.9), or (sens, spec) = (0.9, 0.8), respectively, illustrating the larger impact of reducing specificity. This is even more true in scenarios with a lower fraction of affected individuals: if we assume a probability of person-specific disease of 0.10 instead of 0.44, we obtain misclassification probabilities of 17%, 18%, or 33%, for the same combinations of sensitivity and specificity. A reduced entity-specific specificity increases the probability of falsely classifying healthy entities toward disease, and falsely classifying only one of two healthy entities toward disease is sufficient to misclassify the person-specific disease status.

When applying the MLA1, we found it to effectively correct for bias and to yield the expected confidence interval coverage rates (~95%) when the misclassification was nondifferential, but we found it to still result in biased estimates and excess type I error when the misclassification was differential (Tables 1 and S1). When applying the MLA2, we found it effective in bias correction and type I error control under all misclassification scenarios, but with larger standard errors due to the larger number of parameters in the model (Tables 1 and S1). Overall, our simulation results documented substantial bias and lack of type I error control when the naïve analysis was applied to misclassified data and our MLA to effectively remove bias and keep type I error when specified correctly.

APPENDIX E: DETAILED RESULTS OF MLA FOR THE SELECTED 26 VARIANTS

For estimating sensitivity and specificity, we found the following: (a) for the three lead variants from this GWAS (*CFH*, *ARMS2/HTRA1*, or *HERC2*, respectively), the MLA1-derived sensitivity and specificity (at mean age and two copies of the noneffect allele) showed only small differences between the three variants (sensitivity = 65%, 67%, 63%; specificity = 98%, 98%, 99%, respectively, Table S6a). From a model without including a genetic covariate, we obtained an overall sensitivity of 64.5% (95% CI: 60.1%, 68.7%) and a specificity of 98.6% (98.4%, 98.8%). (b) We did not find strong evidence for associations with age using MLA1 or MLA2 based on any of the 26 selected variants, except for an association of the specificity with age based on MLA1 for the *HERC2* variant that disappeared when applying MLA2 (age*:* $p = 6.71 \times 10^{-9}$ or .70, respectively, Table S6a). (c) Applying MLA2, we found no association of the sensitivity with any selected variant ($p > .05/[23 \times 2]$), but a strong association of the specificity with the *HERC2* lead variant rs12913832 and with the reported *CFH* lead variant rs10922109 ($OR_{spec} = 0.64$, $P_{spec} = 7.38 \times 10^{-9}$ and $OR_{spec} = 1.36$, $P_{spec} = 2.29 \times 10^{-4}$, respectively; Table S6).

Second, we obtained genetic association estimates from MLA1 and MLA2 accounting for misclassification and compared these with naïve analysis estimates. We found interesting patterns: (a) when applying MLA1, we found comparable, slightly increased effect estimates for the *CFH*, *ARMS2/HTRA1,* and *HERC2* lead variant when compared to the naïve analysis (MLA1: OR = 1.23, 1.48, 1.34; $p = 1.69 \times 10^{-6}$, 8.9×10^{-18}, 1.11×10^{-12}; naïve: OR = 1.14, 1.30, 1.26, $p = 6.18 \times 10^{-7}$, 2.44×10^{-20}, 4.16×10^{-16}; Figure 2 and Table S7a). These results were as expected, since a model considering nondifferential misclassification leads in general, by assumption, to larger estimates and widened confidence intervals if any misclassification is present. (b) When applying MLA2, we found similar effect estimates for *CFH and ARMS2/HTRA1* compared to MLA1 and naïve analysis (OR = 1.19 or 1.28, respectively), which is in line with limited bias due to differential misclassification. We also found larger *p* values ($p = .02$ or 2.47×10^{-4}, respectively, which is in line with larger uncertainty when estimating more model parameters. In contrast, we found a completely vanished effect estimate for the *HERC2* variant (MLA2: OR = 1.03, $p = .76$; Figure 2 and Table S7a), indicating a bias in the naïve analysis and MLA1 when ignoring a differential misclassification. (c) Effect estimates for the three reported *CFH* variants increased when applying MLA2 compared to the naïve analysis. This was particularly interesting for the reported *CFH* lead variant rs10922109, where we now observed a nominally significant association into the reported direction (MLA2: OR = 1.15, $p = .047$; naïve: OR = 1.00, $p = .98$; Table S7c). This is in line with the observed association of the specificity with this *CFH* variant. (d) For the other 20 reported lead variants, we found many variants with increased effect estimates by MLA1 or MLA2 compared to the naïve analysis; effect estimates were mostly more comparable to reported effect sizes for advanced AMD (Fritsche et al., 2016; Figure 3c). For one variant, this MLA2 analysis yielded an effect into the opposite direction compared to the reported effect direction, which is the *C9* lead variant (OR = 0.83, $p = .59$). With an effect allele frequency of 1%, it is the rarest analyzed variant of the 26 selected variants and estimates from the reported association as well as for the MLA2 analysis have low precision (i.e., large standard errors).

Chapter 4

Nowcasting the COVID-19 pandemic in Bavaria

Chapter 4 presents a Bayesian hierarchical nowcasting model that we proposed for real-time analysis of the Bavarian SARS-CoV-2 surveillance data.

Contributing article:
Günther, F., Bender, A., Katz, K., Küchenhoff, H., & Höhle, M. (2021). Nowcasting the COVID-19 pandemic in Bavaria. *Biometrical Journal, 63(3), 490-502.*

Supplementary material:
`https://onlinelibrary.wiley.com/doi/10.1002/bimj.202000112`

Author contributions:
Höhle, Küchenhoff, and Günther performed conception of the work. Günther developed and implemented the nowcasting model(s), performed the data analysis and simulation study, set up the webpage for an ongoing presentation of current results, and drafted a first version of the manuscript. Bender implemented the estimation of the effective reproduction number, R_t, from draws of the nowcast posterior. All authors contributed to discussions during model development, interpretation of the results, and to writing and revising the manuscript.

Received: 23 April 2020 | Revised: 23 August 2020 | Accepted: 19 October 2020

DOI: 10.1002/bimj.202000112

Biometrical Journal

RESEARCH PAPER

Nowcasting the COVID-19 pandemic in Bavaria

Felix Günther[1,2] | **Andreas Bender**[1] | **Katharina Katz**[3] | **Helmut Küchenhoff**[1] | **Michael Höhle**[4]

[1] Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Munich, Germany

[2] Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany

[3] Bavarian Health and Food Safety Authority, Oberschleißheim, Germany

[4] Department of Mathematics, Stockholm University, Stockholm, Sweden

Correspondence

Felix Günther, Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich, Germany.

Email: felix.guenther@stat.uni-muenchen.de

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 were reproduced partially for data confidentiality reasons.

Abstract

To assess the current dynamics of an epidemic, it is central to collect information on the daily number of newly diseased cases. This is especially important in real-time surveillance, where the aim is to gain situational awareness, for example, if cases are currently increasing or decreasing. Reporting delays between disease onset and case reporting hamper our ability to understand the dynamics of an epidemic close to now when looking at the number of daily reported cases only. Nowcasting can be used to adjust daily case counts for occurred-but-not-yet-reported events. Here, we present a novel application of nowcasting to data on the current COVID-19 pandemic in Bavaria. It is based on a hierarchical Bayesian model that considers changes in the reporting delay distribution over time and associated with the weekday of reporting. Furthermore, we present a way to estimate the effective time-varying case reproduction number $R_e(t)$ based on predictions of the nowcast. The approaches are based on previously published work, that we considerably extended and adapted to the current task of nowcasting COVID-19 cases. We provide methodological details of the developed approach, illustrate results based on data of the current pandemic, and evaluate the model based on synthetic and retrospective data on COVID-19 in Bavaria. Results of our nowcasting are reported to the Bavarian health authority and published on a webpage on a daily basis (https://corona.stat.uni-muenchen.de/). Code and synthetic data for the analysis are available from https://github.com/FelixGuenther/nc_covid19_bavaria and can be used for adaption of our approach to different data.

KEYWORDS

Bayesian hierarchical model, COVID-19, epidemic surveillance, infectious disease epidemiology, nowcasting

1 INTRODUCTION

Daily reported case numbers of an infectious disease outbreak do not correspond to the actual number of disease onsets on that day. Due to delays from reporting and testing, the number of newly reported cases and the actual number of newly diseased cases can substantially differ. It is the latter, however, that is of central interest when assessing the state and dynamics of an epidemic outbreak. Focusing on the daily number of reported cases hampers our ability to understand current dynamics of the outbreak close to now. This is especially problematic when one wants to gain insight about the current trend or if one wants to assess the effects of political and social interventions in real time. Knowledge of the actual number of new infections per day is highly relevant for the current COVID-19 pandemic, where far-reaching political action was taken in order to contain the epidemic outbreak in 2020.

The problem of occurred-but-not-yet-reported cases during outbreaks is well known from the HIV/AIDS outbreak and different statistical approaches have been proposed to handle delayed reporting. A standard reference is Lawless (1994). A more flexible Bayesian approach, which is the basis of the model we use here, has been developed by Höhle and an der Heiden (2014). In the following, we will refer to this delay adjustment approach as the *nowcast* and define the reporting delay as the time between *disease onset* and official case reporting by a health authority. Other authors use the term nowcasting for models that focus on adjusting the administrative delay between the first case report to a local health authority and registration (in aggregated data) at higher (e.g., state and/or federal) authorities (De Nicola, Schneble, Kauermann, & Berger, 2020), or to perform nowcasting of fatal cases between case registration and fatality date (Schneble, De Nicola, Kauermann, & Berger, 2020).

The basic idea of the nowcasting approach proposed here is to estimate the reporting delay between disease onset and reporting date based on reported cases from the past for which the date of disease onset and the reporting date are known. Given the delay distribution and the current number of case reports with reporting dates close to now, we can infer the actual number of new disease onsets at current dates. The resulting estimated epidemic curve of disease onsets per day gives a more realistic picture of the current state of the epidemic than looking at daily counts of new case reports. Furthermore, the nowcast can facilitate estimation of the time-varying effective reproduction number R_t (Wallinga & Teunis, 2004). There are other approaches including mathematical infection models (compartmental models) for the estimation of R_t, see, for example, Khailaie et al. (2020).

One complication of using nowcasting for COVID-19 reports is that reporting of symptom onset in cases is not complete: either this information could not be elicited due to difficulties getting in contact with the case or because symptoms had not manifested (yet) at the time of contact with the case. This point was first addressed in Glöckner, Krause, and Höhle (2020) and a similar approach based on Lawless (1994) is used by the Robert Koch Institute for analyzing COVID-19 in Germany (an der Heiden & Hamouda, 2020).

Using our approach, we provide nowcast estimates for the COVID-19 pandemic in Bavaria using data from the Bavarian Health and Food Safety Authority (LGL) including the estimation of R_t. The results are updated daily with recent data. In this article, we provide methodological details, show results based on data obtained from the LGL until April 9, 2020, 10 a.m., and provide results of the evaluation of the proposed nowcasting approach.

2 DATA

We use daily data on reported COVID-19 cases from Bavaria from the mandatory notification data based on the German Infection Protection Act (IfSG). The data are provided by the Bavarian Health and Food Safety Authority (LGL) on a daily basis and includes a list of all reported cases with the date of reporting to the LGL, the date of reporting to the local health authority (*Gesundheitsamt*), the date of disease onset if available, and the district of residence for the case (*Kreis*). Since we get our data from the LGL, the number of cases reported to the LGL on a specific date is complete and will not change on subsequent days. These consistent data offer a valid base for inferring the epidemic curve and the considered associated quantities.

The date of reporting to the local health authority is closer to disease onset due to a delay between reporting at the local health authority and transmission to the LGL. However, based on the data obtained from the LGL, the aggregated number of cases reported to the local health authorities on a given day may be incomplete because a case reported to the local health authority can be reported to the LGL with a delay of several days and therefore may not be included in the data yet. Therefore, we use the date of reporting to the local health authority only for the imputation of missing disease onsets, while the nowcast is based on the date a case was reported to the LGL (cf. Steps 1 and 2 in Section 3.1).

Information on disease onset stems from a retrospective collection of the day of symptom onset. However, the daily COVID-19 surveillance data of Bavaria contain about 50%–60% cases with missing information on the day of symptom onset in the weeks close to now. For a specific week, this fraction becomes lower over time since more information on the cases is collected. The missing onset information exists partly due to the heavy workload imposed on health authorities during the pandemic, but also because a certain proportion of cases have no or only very mild symptoms. However, we expect the latter explanation to be less prominent than the former.

Note also that the date of symptom onset does not correspond to the infection date due to a preceding incubation time.

3 | METHODS

In the following sections, we provide methodological details regarding the proposed nowcasting (cf. Section 3.1 as well as the estimation of the time-varying case reproduction number, Section 3.2). The nowcast itself consists of two steps: imputation of missing disease onset dates (Step 1) and Bayesian nowcasting based on the imputed data (Step 2).

3.1 | Nowcasting

Due to the many cases with a missing disease onset date, we decided to proceed with a two-step approach for nowcasting. First, we impute missing data on disease onset and, second, perform the nowcast based on the information on reporting date (available for all cases) and the date of disease onset, which is partly available and partly imputed. Imputing missing disease onset information implies that we also consider presymptomatic and asymptomatic COVID-19 cases in our analyses (to the part at which they are observed in the official COVID-19 case counts). The rationale is that this allows to compare the nowcasting results to the daily reported case numbers. In addition, it is not straightforward to limit the analysis to symptomatic cases, because in cases with missing disease onset date it is not entirely clear whether they are asymptomatic, just symptomless at the time of reporting (pre-symptomatic), or actually show symptoms, but information on the symptom onset date is missing for other reasons, for example, not yet collected.

Step 1: Imputation of disease onset
In the imputation step, we fit a flexible generalized additive model for location, scale, and shape (GAMLSS, Stasinopoulos, Rigby, Heller, Voudouris, & De Bastiani, 2017), assuming a Weibull distribution for the delay time $t_d > 0$ between disease onset and reporting date at the local health authority:

$$t_d \sim WB(\mu, \sigma), \mu > 0, \sigma > 0,$$

where μ and σ are the location and scale parameters of the Weibull distribution with density $f(t_d|\mu,\sigma) = \sigma \cdot \mu \cdot t_d^{(\sigma-1)} \exp(-\mu t_d^{\sigma})$. The same, additive predictor (1) was defined for both, μ and σ,

$$\eta_j = \beta_{0,j} + \sum_{k=1}^{6} \beta_{k,j} I(x_{weekday} = k) + f_{1,j}(x_{week}) + f_{2,j}(x_{age});\ j \in \{\mu, \sigma\}, \tag{1}$$

however, the estimated effects could differ for the two distributional parameters. In (1), parameter $\beta_{0,j}$ is the location- or scale-specific global intercept and $\beta_{k,j}$ is the effect of the weekday on which the report arrived at the local health authority. Furthermore, $f_{1,j}$ and $f_{2,j}$ are smooth effects of the calendar week (of report arrival) and age of case, respectively, both parameterized via cubic splines.

To estimate the model, we use data of all cases for which the disease onset date and thereby t_d is available. Afterward, we impute the delay time t_d, if missing, by sampling from the fitted, conditional Weibull distribution and derive the missing symptom/disease onset date. No imputation is performed for observations for which the symptom onset date is reported.

Since this imputation induces, conditional on the predictors of the GAMLSS imputation model, a missing at random assumption with respect to the time between disease onset and case reporting, we perform a sensitivity analysis, where we omit (i) all individuals where the reports say explicitly that they were symptom-free and (ii) all individuals with missing information about symptoms. This allows us to check, whether the dynamics of the daily number of individuals with available symptoms are structurally different compared to all registered cases over time.

Step 2: Bayesian nowcasting

For the nowcasting step, we use a Bayesian hierarchical model based on Höhle and an der Heiden (2014), which associated implementation in the R-package `surveillance` (Salmon, Schumacher, & Höhle, 2016). In the present work, we have extended the approach considerably, adapted it to the context of COVID-19, and provide a novel implementation in `rstan` (Stan Development Team, 2020).

Let $N_{t,d} = n_{t,d}$ be the (observed) number of cases, with disease onset on day t and reported with a delay of d days (case report arrives on day $t + d$). On day $T > t$ ("current" day, i.e., "now"), the information is available on $N(t, T) = \sum_{d=0}^{T-t} n_{t,d}$ cases that had disease onset on day t and are reported until day T. The aim of nowcasting is to predict the unobserved total number of disease onsets on day t, $N(t, \infty) = \sum_{d=0}^{\infty} N_{t,d}$, based on information available up until the current day T. For identifiability reasons, one defines a maximum relevant delay time of $d = D$ and considers each observation with an observed delay $> D$ as having a delay of D. As described in Höhle and an der Heiden (2014), the hierarchical Bayesian model for nowcasting consists essentially of two parts: a model for the expected number of disease onsets on day t, $E(N(t, \infty)) = \lambda_t$, and a model for the delay distribution at day t, specifying the probability of a reporting delay of d days for a case with disease onset at day t, $P(\text{delay} = d|\text{onset} = t) = p_{t,d}$. Both parts of the model can in general be flexibly specified. We set the maximum delay to $D = 21$ and utilize the following hierarchical model for nowcasting:

$$\log(\lambda_0) \sim N(0, 1), \log(\lambda_t)\,|\lambda_{t-1} \sim N\left(\log(\lambda_{t-1}), \sigma^2\right), \tag{2}$$

$$N_{t,d}|\lambda_t, p_{t,d} \ \sim NB\left(\lambda_t \times p_{t,d}, \phi\right),\ t = 1, \ldots, T, d = 0, \ldots, D. \tag{2}$$

The number of cases with disease onset at day t and reporting delay d days, $N_{t,d}$, is assumed to follow a negative binomial distribution with expectation $\lambda_t \times p_{t,d}$, and overdispersion parameter ϕ. For the delay distribution, we utilize a discrete time hazard model $h_{t,d} = P(\text{delay} = d|\text{delay} \geq d, W_{t,d})$ as

$$\text{logit}(h_{t,d}) = \gamma_d + W'_{t,d}\eta,\ d = 0, \ldots, D - 1;\ h_{t,D} = 1, \tag{3}$$

where $W_{t,d}$ is a vector of time- and delay-specific covariates and η the corresponding regression coefficients. In our main model, we use linear effects of time with breakpoints every 2 weeks before the current day (corresponding to a first-order spline), and a categorical weekday effect of the reporting day with a common effect for holidays and Sunday, since there are substantial differences in the reported case numbers over the week. From model (3), we can derive the probabilities of interest in (2), $p_{t,0} = h_{t,0}$ and $p_{t,d} = (1 - \sum_{d=0}^{d-1} p_{t,d}) \times h_{t,d}$. The main goal of nowcasting is to obtain inference about $N(t, \infty) = \sum_{d=0}^{D} N_{t,d}$. Based on the described Bayesian hierarchical model, this corresponds to a sum of negative binomial distributed counts and we can obtain such inference by summing up the Markov chain Monte Carlo (MCMC) samples of $N_{t,d}$ at each timepoint t. In an alternative specification of the model during evaluation (see below) we assume that $N_{t,d}|\lambda_t, p_{t,d} \sim \text{Po}(\lambda_t \times p_{t,d})$. In this case $N(t, \infty)$ is Poisson distributed as well and it is directly possible to sample from $\text{Po}(\lambda_t)$ to obtain inference about $N(t, \infty)$.

The utilization of the first-order random walk for modeling λ_t in (2) was motivated by results of McGough, Johansson, Lipsitch, and Menzies (2020). For the modeling of the delay distribution, we utilized several different approaches and covariates and evaluated them on synthetic data and retrospectively on the Bavarian COVID-19 data (see below for a description of the approaches).

3.2 | Estimation of the time-varying case reproduction number $R_e(t)$

Once a depletion of susceptibles occurs during an outbreak of a person-to-person transmitted disease or specific interventions are made, a key parameter to track is the so-called effective reproduction number (also referred to as case reproduction number). This time-varying quantity is defined as follows: consider a case with disease onset on day t—the expected number of secondary cases one such primary case generates will be denoted by $R_e(t)$. The time until these secondary cases will show symptoms is governed by the serial-interval distribution, which is defined as the time period between manifestation of symptoms in the primary case to time of symptom manifestation in the secondary case (Svensson, 2007).

We estimate the time-varying case reproduction number by the procedure of Wallinga and Teunis (2004): Consider a case j showing symptoms for the first time on day t_j. The relative likelihood that a case i (with symptom onset on day t_i) was infected by j is given by

$$p_{ij} = \frac{g(t_i - t_j)}{\sum_{k \neq i} g(t_i - t_k)},$$

where g is the probability mass function of the serial-interval distribution. For the serial interval distribution, we use a discretized version of the results from Nishiura, Linton, and Akhmetzhanov (2020), which find a log-normal distribution with mean 4.7 days and standard deviation 2.9 as the most suitable fit to data from 28 infector–infectee pairs. An estimate of the effective reproduction number at time t is now given as the average reproduction number of each case j showing first symptoms of the illness on day t:

$$\hat{R}_e(t) = \frac{1}{|j : t_j = t|} \sum_{j:t_j=t} \sum_{i \neq j} p_{ij}. \tag{4}$$

We prefer this $R_e(t)$ estimation over the method used in an der Heiden and Hamouda (2020), because it is unbiased for our generation time distribution (see the discussion in Höhle, 2020). For each imputed data set, we extract $K = 500$ time series of case counts from the posterior distribution of the nowcast and then estimate $R_e(t)$ as defined in (4) for each time series using the R-package R0 (Obadia, Haneef, & Boëlle, 2012). Furthermore, each $R_e(t)$ estimation generates $M = 100$ samples from the corresponding sampling distribution of $R_e(t)$. Altogether, we report $\hat{R}_e(t)$ as mean of these $K \times M$ samples together with the 2.5% and 97.5% quantiles to form a 95% credibility interval for $R_e(t)$. We estimate $R_e(t)$ for all t so that $t + q_g(0.95) \leq T$, where $q_g(0.95)$ is the 95% quantile of the serial interval distribution. This avoids a downward bias in the $R_e(t)$ estimation near "now." Alternatively, one could employ correction methods near T (Cauchemez et al., 2006).

3.3 | Evaluation of the methods

We perform an evaluation of the hierarchical nowcasting based on synthetic data mimicking the reported Bavarian COVID-19 data and retrospectively on the official data from the LGL that were reported until July 31. For creation of the synthetic data, we utilized a smoothed version of the observed number of reported disease onsets per day and specified a reporting delay model similar to the model described in (3) with five changepoints in the linear time effect on the hazard. This leads first to a slight increase, followed by a decrease and stabilization, and a final slight increase of the (average) reporting delay over time (see the supplemental material for a detailed description and visualization of the data generating process). The aggregated daily numbers of disease onsets and daily numbers of reported cases are similar in structure to the officially reported data. For faster computation during the evaluation, we divided the daily cases by two. For the retrospective evaluation on the official COVID-19 data, we focus on all reported cases with available disease onset and on the time period between March 17 and June 30, assuming that all cases that will be reported with disease onset until June 30 are reported on July 31.

For the evaluation of the nowcasting, we estimate several different models (Table 1). We vary the distributional assumptions of $N_{t,d}$ between Poisson and negative binomial (cf. Section 3.1, Step 2). Furthermore, we vary the specification of the model for the reporting delay distribution: first, we assume a reporting delay distribution without changes over time, second, we estimate linear effects of time on the delay distribution with changepoints every 2 weeks, and third use a different specification of the discrete time hazard model, where we model

$$\text{logit}(h_{t,d}) = \gamma_d + \alpha_t, \; d = 0, \ldots, D-1; \; h_{t,D} = 1, \tag{5}$$

with a prior on $\alpha_t | \alpha_{t-1} \sim N(\alpha_{t-1}, \sigma^2_{\alpha_t})$ and $\alpha_0 = 0$. With this model, we aim to estimate smooth daily changes in the delay distribution over time similar to the first-order random walk in the modeling of λ_t. In case of the synthetic data,

TABLE 1 Estimated hierarchical nowcast models in the evaluation on synthetic and actual Bavarian COVID-19 data

Synthetic data	
Distribution $N_{t,d}$	**Delay distribution**
Poisson	No changes
Poisson	Linear time-effect with changepoints every 2 weeks
Poisson	Linear time-effect with true changepoints
Negative binomial	Linear time-effect with changepoints every 2 weeks
Negative binomial	Linear time-effect with true changepoints
Negative binomial	Daily changes (first-order random walk)
Retrospective evaluation on Bavarian data	
Distribution $N_{t,d}$	**Delay distribution**
Poisson	No changes
Poisson	Linear time-effect with changepoints every 2 weeks
Negative binomial	Linear time-effect with changepoints every 2 weeks
Negative binomial	Daily changes (first-order random walk)
Negative binomial	Linear time-effect with changepoints every 2 weeks + reporting weekday effect
Negative binomial	Daily changes (first-order random walk) + reporting weekday effect

we additionally estimate the nowcasting with the known true changepoints in the delay distribution (that are unknown in real-world applications) and in case of the retrospective evaluation on Bavarian data, we additionally include in some scenarios dummy effects of the weekday of the reporting date.

To compare the performance of the different models, we estimate the log scoring rule (logS) and the continuous ranked probability score (CRPS) (Jordan, Krüger, & Lerch, 2019), root mean squared error (RMSE), as well as coverage frequencies of 95% prediction intervals. For all those criteria, we average over all dates and nowcast predictions 2–6 days before the current date. In addition to the quantitative measures, we visually inspect the performance of the different approaches based on the nowcasting predictions and the estimated delay distribution in comparison to the retrospective truth in order to identify potential problems of the models.

We extend the retrospective evaluation of the nowcasting on Bavarian data to the estimation of $R_e(t)$ and compare the estimated $\hat{R}_e(t)$ on the most current day $\max(t)$ s.t. $t + q_g(0.95) \leq T$ for all T to the retrospective *true* $R_e(t)$ given all available case data until July 31. This is done based on all evaluated models, and we visually inspect the estimated $\hat{R}_e(t)$ over time and compute coverage frequencies of 95% credibility intervals.

3.4 | Implementation

All calculations were done using the statistical programming environment `R` (R Core Team, 2020). Nowcasting was performed based on a custom `rstan` (Stan Development Team, 2020) implementation. Estimation of $R_e(t)$ was based on code of the `R0` package (Obadia et al., 2012) for each selected posterior sample. For computation of the proper scoring rules we used the `scoringRules` package (Jordan et al., 2019).

Code to reproduce our analysis and for adaption to other application scenarios is available at https://github.com/FelixGuenther/nc_covid19_bavaria. There, we also provide an artificial data set based on the observed reporting dates of cases but for data protection reasons featuring only artificial information on the age and disease onset dates of the cases.

4 | RESULTS

4.1 | Data

We present results based on data obtained from LGL on April 9, 2020, 10 a.m. The data contain information on 29,262 COVID-19 cases, which we restrict to 29,246 cases reported after March 1, as the first 16 COVID-19 cases reported between January 28 and February 13 (reported disease onset between January 23 and February 3, three with missing onset infor-

TABLE 2 Week-specific observed number of cases with available information on disease onset. Empirical mean, median, and 25%/75% quantile of delay distribution between disease onset and reporting at local health authority. Cases are grouped into weeks based on their reporting date at local health authority. Data from April 9, 10 a.m.

Rep. week	*n*	Delay available	% avail.	Mean	Median	25% quant.	75% quant.
10	114	77	68	5.8	5	4	8
11	1074	459	43	5.4	5	3	7
12	4660	2100	45	6.0	5	4	8
13	8858	4268	48	7.5	7	4	10
14	11,003	4800	44	8.8	8	5	12
15	3532	1335	38	8.9	7	4	12

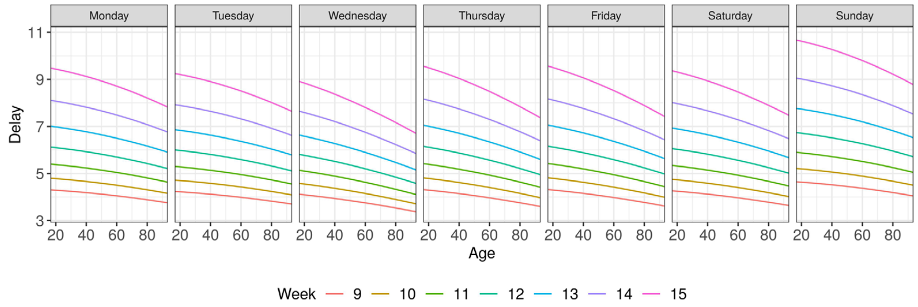


FIGURE 1 Results of the Weibull GAMLSS imputation model. Shown is the estimated median of the delay time given case-specific covariates (reporting week, weekday of reporting, age)

mation) concerned a contained outbreak (Böhmer et al., 2020) and no further cases were detected upon February 27. This outbreak can therefore be assumed to not have contributed to the later disease spread.

Information on disease onset is available for 13,137 cases, but reported disease onset was past the official reporting date for 50 cases and before January 23 for 16. We set the disease onset date for these cases as missing, yielding 13,071 cases with valid information on disease onset (44.7%). For these, the median delay between disease onset and reporting was 7 days (25% quantile: 5, 75% quantile: 11), Table 2 shows observed delay times over the observation period and reveals a considerable increase in the delay distribution over time.

4.2 | Imputation of missing disease onset

For imputation of missing disease onset dates, we estimate a Weibull GAMLSS with smooth effects of the reporting week, the cases' age, and a categorical effect of the weekday of report arrival on location and scale. We utilize the reporting date at the local health authority in the imputation model, since it is closer to the actual disease onset than the reporting date at LGL and is available for all cases contained in our data. Thereby, we do not have to deal with the additional reporting delay between the local health authorities and the LGL for the imputation, which might also change over time. Figure 1 shows the estimated association of the covariates with the median delay. All covariates turned out to be relevant: we find an increase in expected delay time over the reporting weeks, lower reporting delay for older cases, and differences over the course of a week. The estimated GAMLSS model is used to impute the date of disease onset for cases with missing onset information.

4.3 | Nowcasting

Figure 2 shows the number of daily reported cases and the number of cases with reported and imputed disease onset on a certain day over time. Furthermore, we display the estimated new cases from nowcasting based on our main model (2).

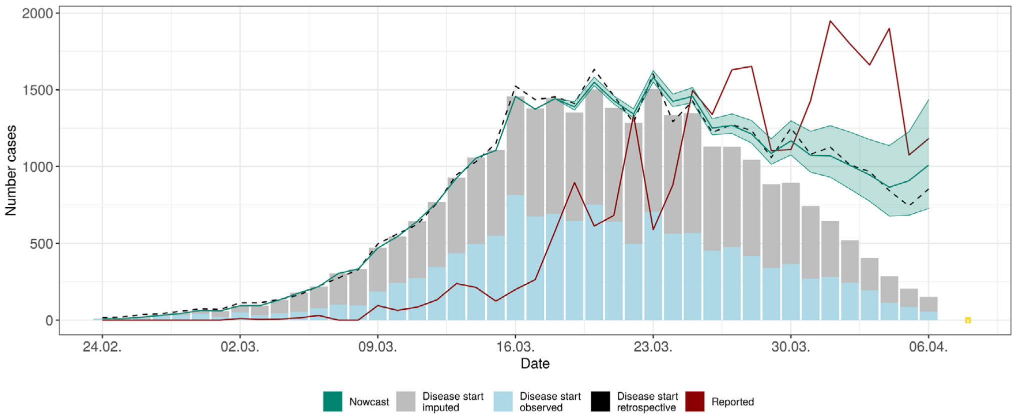


FIGURE 2 Nowcasting based on Bavarian COVID-19 data until April 8, 2020. Shown is the point estimate + 95% prediction interval of the daily number of disease onsets on a given day based on the Bayesian hierarchical nowcast. The model considers changes in the delay distribution over time based on a linear time effect with 2-week changepoints and effects of the weekday of reporting. The expected number of new disease onsets is modeled based on a first-order random walk. Additionally, we show the observed number of cases with disease onset (reported and imputed) that are known on April 8, based on daily bars, the number of newly reported cases per day (red line), and the retrospective *true* number of disease onsets known up until July 31 (black dotted line). The current day for nowcasting is April 8, and nowcasts are performed up until April 6

We observe a clear difference between the estimated new cases from the nowcast and the daily numbers of reported cases. The induced bias due to the reporting delay is obvious: the estimated daily new cases stabilize from around March 20 on and start to decrease afterward, while the reported cases still show a rapid increase. The 95% prediction interval, however, shows substantial uncertainty in estimates, especially for more recent estimates. Note that we set the current day for the nowcasts to April 8, since we only consider days with fully available reporting data. Furthermore, we set a reporting lag between the current date and reported nowcast results of 2 days due to considerable uncertainty in the nowcasts for dates with very few observations with reported or imputed disease onset.

The black dotted line in Figure 2 shows the retrospective *true* number of disease onsets (reported and imputed) based on data known on July 31. We can see that the predicted number of new cases per day and the actually observed number of cases match closely and the prediction intervals contain the actual number of onsets for most days. Note, that the imputation of missing disease onset dates was performed based on the same Weibull GAMLSS but based on different data (all data available on April 8 for nowcasting and July 31 for the *retrospective truth*), and the number of cases with (imputed)disease onset on a specific day can therefore vary slightly.

4.4 | Estimation of the time-varying case reproduction number

Figure 3 depicts the estimated $R_e(t)$ as defined in (4) for the time frame from February 24 until the March 27. This time range is defined by the time of the first secondary case observed in the data and the date of the nowcast minus the number of days it takes for 95% of secondary cases to be observed, which is determined based on 95% quantile of the assumed generation time distribution (10 days). According to the estimate, $R_e(t)$ decreased steadily since the beginning of the outbreak and is about $R_e(t) = 1$ at March 20, with $R_e(t) = 0.81$ (CI $= [0.75, 0.87]$) on March 27. However, care is required, if interpreting this result with the timing of interventions, because the $R_e(t)$ estimator is defined forward in time and describes the transmission process within the following 10 days.

4.5 | Evaluation of nowcast and estimation of $R_e(t)$

We performed an evaluation of the nowcasting approach based on synthetic data and retrospectively on the Bavarian COVID-19 data to investigate the performance of the Bayesian hierarchical model under various model specifications and gain a better understanding of important aspects of modeling. For the synthetic data, we found the following (Table 3,

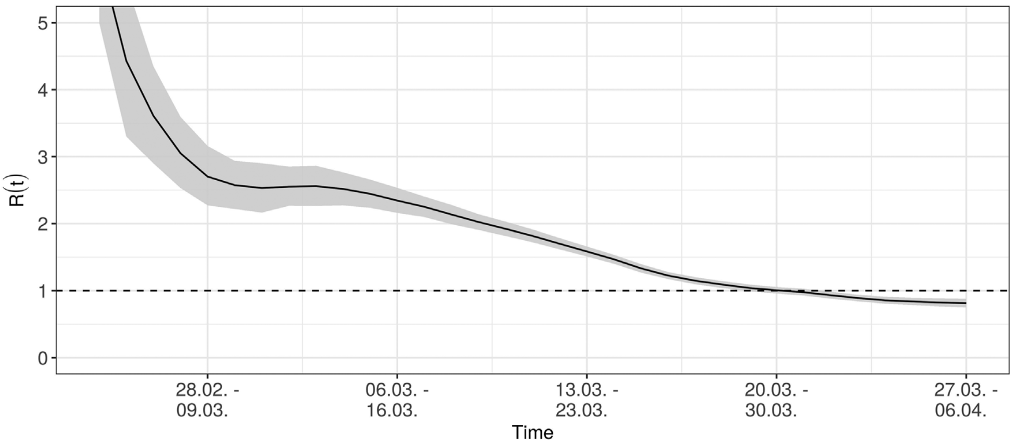


FIGURE 3 Estimated, time-varying effective case reproduction number $R_e(t)$

TABLE 3 Results of the evaluation of different nowcasting models on synthetic and actual Bavarian data (retrospectively). CRPS is the continuous ranked probability score, logS denotes the log scoring rule, RMSE denotes the root mean squared error of the posterior median. Additionally, we provide coverage frequencies of 95% prediction intervals for the number of disease onsets per day and coverage frequencies of 95% credibility intervals in the estimation of $R_e(t)$. All scores are averaged over nowcasts for $T-6, \dots, T-2$ days, with T from March 17 to June 30. For $R(t)$, we compute coverage frequencies for the estimate closest to the current date T over all T's

Synthetic data					
Model					**Cov. 95% PIs**
Distr. $N_{t,d}$	**Delay model**	**CRPS**	**logS**	**RMSE**	$\boldsymbol{N(t,\infty)}$
Poisson	No changes	46.68	13.24	89.75	0.53
Poisson	Lin. effect of time changepoints 2 weeks	12.53	3.68	36.22	0.95
Neg. binomial	Lin. effect of time changepoints 2 weeks	12.47	3.68	36.01	0.95
Neg. binomial	Daily changes (first-order RW)	28.37	3.90	92.33	0.91
Poisson	Lin. effect of time true changepoint	11.88	3.63	35.31	0.95
Neg. binomial	Lin. effect of time true changepoints	11.90	3.62	35.48	0.96

Retrospective Bavarian COVID-19 data						
Model					**Cov. 95%-PIs/CIs**	
Distr. $N_{t,d}$	**Delay model**	**CRPS**	**logS**	**RMSE**	$\boldsymbol{N(t,\infty)}$	$\boldsymbol{R_e(t)}$
Poisson	No changes	193.43	Inf	389.56	0.19	0.56
Poisson	Lin. effect of time changepoints 2 weeks	74.32	Inf	226.90	0.67	0.86
Neg. binomial	Lin. effect of time changepoints 2 weeks	61.79	5.05	205.59	0.84	0.92
Neg. binomial	Daily changes (first-order RW)	79.21	4.83	274.70	0.86	0.95
Neg. binomial	Lin. effect of time changepoints 2 weeks + weekday effect	56.63	5.22	170.59	0.82	0.92
Neg. binomial	Daily changes (first-order RW) + weekday effect	67.32	4.99	236.05	0.90	0.94

more detailed results in the supplemental material): when we supply the true, in reality unknown changepoints of the delay distribution to model fitting the nowcasting approach performs best with respect to our evaluation metrics. Averaged over all days T, and for all nowcast days $t = T-6, \dots T-2$, it shows the lowest log and CRPS score, lowest RMSE and shows the desired coverage frequencies for the 95%prediction intervals. With the models assuming changepoints in the linear time effect on the reporting delay every 2 weeks before T, we obtain similar, but slightly worse performance (see supplemental material for more details). The approach appears to be able to capture moderate changes in the delay distribution successfully. Modeling the changes on a daily basis shows a slightly worse performance with respect to the CRPS score and PI coverage frequencies. Assuming a constant reporting delay distribution over time and ignoring the changes leads to the worst performance with biggest scores and low coverage frequencies of the prediction intervals. When speci-

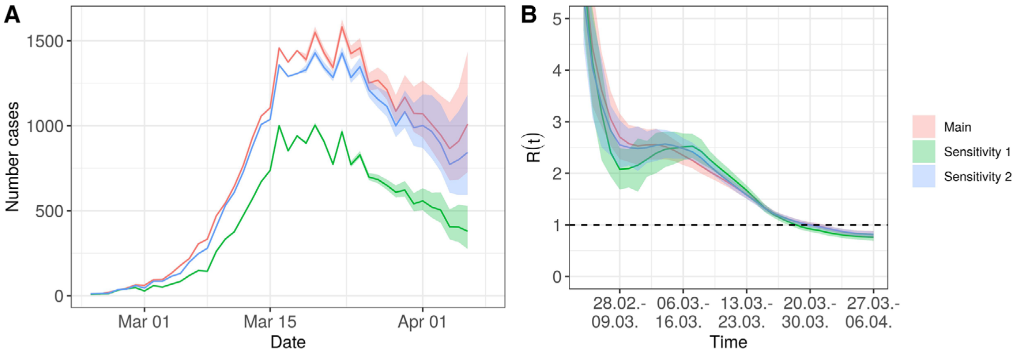


FIGURE 4 Sensitivity analysis. Estimation of nowcast and $R_e(t)$ with associated 95% prediction and credibility intervals based on three different data subsets: original analysis (red), analysis only with cases with reported COVID-19 symptoms (green), and with all cases excluded that are reported as having no symptoms (blue)

fying an adequate model for the delay distribution, the distributional assumptions regarding $N_{t,d}$ play a minor role for the evaluation based on synthetic data.

In the retrospective evaluation of the Bavarian data, the Poisson model assuming no changes in the reporting delay distribution performs badly as well. This is in line with the apparent changes in the reporting delay between disease onset and reporting at LGL over time (supplemental material). Comparing the Poisson model with 2-week changepoints with a similar model using a negative binomial distribution for $N_{t,d}$ we find the latter to perform better. Adding weekday effects to the delay distribution improves the performance of the models as well. Comparing the negative binomial model with daily changes in the delay distribution with the 2-week changepoint model, we found better coverage frequencies for the former (e.g., 90% vs. 82% when including the weekday effect) but lower CRPS score and RMSE for the latter. Comparing the estimated $\hat{R}_e(t)$ at most current t's based on the different nowcast models with the retrospective *truth* based on all reported data, we find coverage probabilities of the 95% credibility intervals bigger than 90% for all negative binomial models that consider changes in the delay distribution over time. The estimation of $R_e(t)$ is, however, biased when it is based on a biased nowcasting approach, for example, when changes in the delay distribution are ignored.

5 | DISCUSSION

Our analyses show that nowcasting is a valuable real-time tool to gain situational awareness in the middle of an outbreak. Based on our evaluation, we found several aspects to be important for the successful application of nowcasts: first and foremost, it is important to account for existing changes in delay between disease onset and case reporting over time. Ignoring such changes can severely bias the predicted number of disease onsets. In the Bavarian data, we also found evidence for changes in reporting delay associated with the weekday of reporting, which should be accounted for. Second, we found an improved performance when modeling the daily counts of disease onsets with a specific reporting delay d, $N_{t,d}$, based on a negative binomial distribution with overdispersion. In our data, the disease onset counts show bigger variability then implied by a Poisson distribution. Third, utilizing a first-order random walk for modeling the logarithmic expected daily number of new disease onsets, λ_t, as proposed by McGough et al. (2020), worked well. We also tried i.i.d. log-Gamma priors and a smooth modeling of the epidemic curve based on truncated power splines as proposed in Höhle and an der Heiden (2014), but found the first-order random walk to perform best. Altogether, we found that a negative binomial model with random-walk prior of λ_t and modeling of the delay distribution via an discrete time hazard model with linear time effects and 2-week changepoints works satisfactory. With this model, we are able to account flexibly for changes in reporting delay over time and obtain a satisfactory performance on synthetic data as well as the true retrospective Bavarian COVID-19 data. The alternative smooth modeling of the delay distribution based on daily changes using a first-order random walk also worked well for many days,but had convergence problems on some days and might be overly complex for many scenarios.

However, there are important limitations of any nowcasting estimation: (i) we correct for a bias due to delays between disease onset and case reporting, but provide no correction for possible cases in the population that were not tested. This is a big issue in understanding COVID-19 spread, since there are possibly many undetected cases. Assuming a constant factor of underreporting, we can analyze the dynamics of the outbreak in a more reliable way by our nowcasting method compared to focusing on daily counts of newly reported cases. Furthermore, $R(t)$ estimates would be invariant to such constant underreporting. However, if the proportion of undetected cases varies over time, then the dynamics of the pandemic is not described adequately by our approach as well. (ii) We model the temporal variation in the delay distribution in a flexible way. However, short-term changes, especially in the time close to the current day can lead to a bias, because it is particularly hard to distinguish between developments in the epidemic curve and changes in the reporting delay with no or very less data. (iii) Our imputation method includes a missing at random assumption, which implies that the time between disease onset and reporting is the same for individuals with and without available symptom onset date. This could be violated due to many asymptomatic and presymptomatic among the reported COVID-19 cases. However, the sensitivity analyses in the Appendix show that our results are relatively stable to variations of this definition.

Comparing our approach to the one used by the Robert Koch Institute an der Heiden and Hamouda (2020), we use a more detailed modeling of the delay distribution for the nowcast, for example, including the day of the week in our model, which turned out to be relevant in our data. Furthermore, we observed and modeled a dependence of the delay time on calendar time as part of the nowcast. This was not originally taken into account by an der Heiden and Hamouda (2020). When calculating the effective reproduction number $R_e(t)$, an der Heiden and Hamouda (2020) used a constant generation time of 4 days, while our approach includes a more realistic assumption of an individually varying time originating from a lognormal distribution, which also provides a smoother estimate over time.

The approach to estimate $R(t)$ proposed by Khailaie et al. (2020) includes a complex compartmental model with many assumptions about the other model parameters, which in part can only be guesstimated from literature sources. Their procedure of estimating $R(t)$ is only partly data driven and mainly relies on cumulative reported cases in the federal states of Germany. Confidence intervals are generated by the variation of the other model parameters. This highlights the problems of the approach: while compartmental models can be useful for forecasting, its value for real-time estimation of $R(t)$ hinges on it being a realistic model with a well-calibrated parameter estimated. Instead, we prefer the more statistically driven transmission-tree–based estimates, which rely less on model assumptions and more on a statistically sound analysis of the available data.

In our retrospective evaluation of the Bavarian COVID-19 data we found, that the estimation of $R_e(t)$ based on the predicted daily counts of disease onsets from nowcasting performs well if the nowcast model is adequately specified. Coverage frequencies of the 95% credibility intervals were as desired compared to a calculation of $R_e(t)$ based on all retrospectively available disease onset data. The utilization of the predictive distribution from the Bayesian nowcast for the estimation of $R_e(t)$ helps therefore successfully to avoid a bias close to the current date due to diseased-but-not-yet reported cases. For interpretation of the estimated $\hat{R}_e(t)$ over time, similar limitations arise as in the interpretation of the estimated epidemic curve from the nowcast. If the fraction of undetected cases compared to all cases changes strongly over time, for example, due to changes in testing strategy, this can bias the estimated reproduction number. Compared to the interpretation of the estimated epidemic curve, the time-varying reproduction number might, however, have the advantage that it only requires stable conditions within a short time window, since it compares the estimated and reported number of disease onsets to the situation at time points close by, instead of looking at the absolute numbers over a longer period of time.

Summarizing, we believe that our results give a much more reliable picture of the course of the pandemic than the mostly used time series of reported cases. For the interpretation, it has to be emphasized, that we estimate the number of persons with disease onset on a certain day.

Our proposed nowcasting model can be applied to other data, when sufficient information about disease onset dates is available and the numbers are large enough for reliable modeling. On our webpage (corona.stat.uni-muenchen.de), we present daily results of the nowcasting for Bavaria and, in addition, for the city of Munich.

Since we introduce no correction for cases, which are never detected, our estimated epidemic curve should be related to other data sources, like hospital admission, ICU admission, or death numbers. This aspect highlights the need for the collection and combination of many different data sources—each bringing challenges of its own.

ACKNOWLEDGMENTS

We thank Titus Laska for initial code on the $R_e(t)$ estimation. An early version of the imputation modeling, but without any methodological extensions of the nowcast, was proposed in the now disused work by Glöckner et al. (2020). We thank two anonymous referees for constructive and helpful comments.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

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 were reproduced partially for data confidentiality reasons.

ORCID

Felix Günther https://orcid.org/0000-0001-6582-1174
Andreas Bender https://orcid.org/0000-0001-5628-8611
Helmut Küchenhoff https://orcid.org/0000-0002-6372-2487
Michael Höhle https://orcid.org/0000-0002-0423-6702

REFERENCES

an der Heiden, M., & Hamouda, O. (2020). Schätzung der aktuellen Entwicklung der SARS-CoV-2-Epidemie in Deutschland: Nowcasting. *Epidemiologisches Bulletin*, *17*, 10–15.

Böhmer, M. M., Buchholz, U., Corman, V. M., Hoch, M., Katz, K., Marosevic, D. V., … Zapf, A. (2020). Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: A case series. *The Lancet Infectious Diseases*, *20*(8), 920–928.

Cauchemez, S., Boelle, P. Y., Donnelly, C. A., Ferguson, N. M., Thomas, G., Leung, G. M., … Valleron, A. J. (2006). Real-time estimates in early detection of SARS. *Emerging Infectious Diseases*, *12*(1), 110–113.

De Nicola, G., Schneble, M., Kauermann, G., & Berger, U. (2020). Regional now- and forecasting for data reported with delay: A case study in COVID-19 infections. Retrieved from https://arxiv.org/abs/2007.16058.

Glöckner, S., Krause, G., & Höhle, M. (2020). Now-casting the COVID-19 epidemic: The use case of Japan, March 2020. Retrieved from https://www.medrxiv.org/content/early/2020/03/23/2020.03.18.20037473

Höhle, M. (2020). Effective reproduction number estimation. Retrieved from https://staff.math.su.se/hoehle/blog/2020/04/15/effectiveR0.html

Höhle, M., & an der Heiden, M. (2014). Bayesian nowcasting during the STEC O104:H4 Outbreak in Germany, 2011. *Biometrics*, *70*(4), 993–1002.

Jordan, A., Krüger, F., & Lerch, S. (2019). Evaluating probabilistic forecasts with scoring rules. *Journal of Statistical Software*, *90*(12), 1–37.

Khailaie, S., Mitra, T., Bandyopadhyay, A., Schips1, M., Mascheroni, P., Vanella, P., … Meyer-Hermann, M. (2020). Development of the reproduction number from coronavirus SARS-CoV-2 case data in Germany and implications for political measures. Retrieved from https://www.medrxiv.org/content/10.1101/2020.04.04.20053637v1

Lawless, J. F. (1994). Adjustments for reporting delays and the prediction of occurred but not reported events. *The Canadian Journal of Statistics*, *22*(1), 15–31.

McGough, S. F., Johansson, M. A., Lipsitch, M., & Menzies, N. A. (2020). Nowcasting by Bayesian smoothing: A flexible, generalizable model for real-time epidemic tracking. *PLoS Computational Biology*, *16*(4), e1007735.

Nishiura, H., Linton, N. M., & Akhmetzhanov, A. R. (2020). Serial interval of novel coronavirus (COVID-19) infections. *International Journal of Infectious Diseases*, *93*, 284–286.

Obadia, T., Haneef, R., & Boëlle, P.-Y. (2012). The R0 package: A toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Medical Informatics and Decision Making*, *12*(1), 147.

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Salmon, M., Schumacher, D., & Höhle, M. (2016). Monitoring count time series in R: Aberration detection in public health surveillance. *Journal of Statistical Software*, *70*(10), 1–35.

Schneble, M., De Nicola, G., Kauermann, G., Berger, U. (2020). Nowcasting fatal COVID-19 infections on a regional level in Germany. *Biometrical Journal. 2020*, 1–19. https://doi.org/10.1002/bimj.202000143.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Stasinopoulos, M., Rigby, R., Heller, G., Voudouris, V., & De Bastiani, F. (2017). *Flexible regression and smoothing using GAMLSS in R*. Boca Raton. FL: Chapman and Hall/CRC.

Svensson, Å. (2007). A note on generation times in epidemic models. *Mathematical Biosciences*, *208*(1), 300–311.

Wallinga, J., & Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, *160*(6), 509–516.

SUPPORTING INFORMATION

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

How to cite this article: Günther F, Bender A, Katz K, Küchenhoff H, Höhle M. Nowcasting the COVID-19 pandemic in Bavaria. *Biometrical Journal*. 2021;63:490–502. https://doi.org/10.1002/bimj.202000112

APPENDIX

Sensitivity analysis

Since the amount of data with missing information on disease/symptom onset date are rather high, we perform two sensitivity analyses. The symptom onset date can either be missing because the reporting local health authorities were not able to provide any information or because a case did not experience any symptoms until case reporting. Based on available information, we can distinguish between cases for which COVID-19 symptoms are documented at time of reporting (17,723, 61%; 73% with available onset), cases explicitly without symptoms (2,221, 8%), and cases without any information on symptoms (9,302, 32%). In the first sensitivity analysis, we focus only on the cases with reported symptoms. In a second analysis, we exclude all cases that have been explicitly reported to have no symptoms.

Figure 4(A) indicates that the estimated structure of the epidemic curve is very similar to the main analysis when excluding asymptomatic cases and cases without known symptom status in the sensitivity analyses. This also applies to the estimated $R_e(t)$ as show in Figure 4(B). Since the sensitivity analyses consider fewer reported cases, the actual estimated number of disease onsets per day is lower as well. However, the interpretation regarding the dynamics of the COVID-19 pandemic in Bavaria is similar based on all three analyses.

Chapter 5

Analysis of COVID-19 case numbers: adjustment for diagnostic misclassification on the example of German case reporting data

Chapter 5 discusses the implications of erroneous diagnostic test results for SARS-CoV-2 infection for disease spread surveillance based on case report data. We show how reported case counts can be adjusted for misclassification given known misclassification probabilities. This forms the basis for real-time estimation of the misclassification-adjusted epidemic curve . Based on this work, we discuss in detail the potential magnitude of bias due to erroneous tests in German surveillance data utilizing plausible assumptions about misclassification probabilities.

Contributing article:
Günther, F., Berger, U., Höhle, M., Bender, A., Wildner, M., Heid, I. M., & Küchenhoff, H. (2021). Analysis of COVID-19 case numbers: adjustment for diagnostic misclassification on the example of German case reporting data. *medRxiv*

Supplementary material:
`https://www.medrxiv.org/content/10.1101/2020.09.29.20203877v2`

Author contributions:
Küchenhoff, Günther, Berger, and Wildner devised the research question and work. Gün-

ther and Küchenhoff derived the approach to account for misclassification in the case reporting data. Günther and Berger performed data analyses. Günther created the first draft of the manuscript. All authors contributed to the interpretation of the results and to writing and revising the manuscript.

Analysis of COVID-19 case numbers: adjustment for diagnostic misclassification on the example of German case reporting data

Felix Günther*[1], Ursula Berger[2], Michael Höhle[3], Andreas Bender[1], Manfred Wildner[4], Iris M. Heid[5], and Helmut Küchenhoff[1]

[1]Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Germany
[2]Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich, Germany
[3]Department of Mathematics, Stockholm University, Sweden
[4]Bavarian Health and Food Safety Authority / Pettenkofer School of Public Health, Oberschleißheim, Germany
[5]Department of Genetic Epidemiology, University of Regensburg, Germany

March 4, 2021

Abstract

Background Reported COVID-19 case numbers are key to monitoring pandemic spread and decision-making on policy measures but require careful interpretation as they depend substantially on testing strategy. A high and targeted testing activity is essential for a successful Test-Trace-Isolate strategy. However, it also leads to increased numbers of false-positives and can foster a debate on the actual pandemic state, which can slow down action and acceptance of containment measures.

Aim We evaluate the impact of misclassification in COVID-19 diagnostics on reported case numbers and estimated numbers of disease onsets (epidemic curve).

Methods We developed a statistical adjustment of reported case numbers for erroneous diagnostic results that facilitates a misclassification-adjusted real-time estimation of the epidemic curve based on nowcasting. Under realistic misclassification scenarios, we provide adjusted case numbers for Germany and illustrate misclassification-adjusted nowcasting for Bavarian data.

Results We quantify the impact of diagnostic misclassification on time-series of reported case numbers, highlighting the relevance of a specificity smaller than one when test activity changes over time. Adjusting for misclassification, we find that the increase of cases starting in July might have been smaller than indicated by raw case counts, but cannot be fully explained by increasing numbers of false-positives due to increased testing. The effect of misclassification becomes negligible when true incidence is high.

Conclusions Adjusting case numbers for misclassification can improve this important measure on short-term dynamics of the pandemic and should be considered in data-based surveillance. Further limitations of case reporting data exist and have to be considered.

1 Introduction

In the acute COVID-19 pandemic, politics as well as public health and academic institutions worldwide are faced with the challenge of evaluating existing surveillance data like time series of reported case counts in real time. It is important to analyze and interpret this data taking into account all potential limitations and uncertainties, in order to maintain the highest possible confidence in generated insights.

*felix.guenther@stat.uni-muenchen.de

This is particularly important the longer the pandemic lasts and ongoing restrictions in public life also foster a growing wariness of people.

Interpreting daily reported COVID-19 case numbers is pivotal to gain insights into the state and dynamics of the current pandemic situation in different regions, but has several drawbacks. Problems can especially occur if the number of performed tests or the testing strategy change over time. The number of conducted PCR-tests has increased substantially in many European countries in summer 2020 [1], which coincided with an increase in observed case counts in July and August 2020, for example in Germany. In retrospect, this development of increasing case counts can be seen as a precursor of the strong second COVID-19 wave in Germany. At that time, however, it led in parts of society to increased skepticism whether the increasing case numbers were only false-positive test results and whether implemented measures to control the pandemic situation were superfluous.

There are three main challenges for the analysis and interpretation of reported case counts: first, the temporal assignment of the reported cases, second, misclassification in diagnostic tests, and third, a time-varying case detection ratio (dark figure). In this work we focus on the first two problems and refer to the third in the discussion.

To assess the short-term dynamics of an epidemic, it is common to look at the epidemic curve, defined as the number of disease onsets per day. Due to reporting delay, there are differences between date of disease onset and the date of case reporting and the time-series of newly reported cases can give a lagged and also in its structure incorrect impression of the acute pandemic situation. If data is collected on the day of disease onset, the epidemic curve can be constructed from this information. However, the reporting delay gives rise to occurred-but-not-yet-reported cases leading in real-time surveillance to a downward bias for days close to the current date. Utilizing individual-specific data on both disease onset and reporting date, it is possible to adjust for the reporting delay and to obtain an estimate of the epidemic curve based on nowcasting [2, 3, 4]. For SARS-CoV-2 and other pathogens, there exists the additional complexity that not all infected cases develop disease. Since the epidemic curve is a proxy for the number of exposed individuals over time (with a small time lag), it is still reasonable to assign those cases a synthetic disease onset date based on adequate assumptions.

Misclassification in COVID-19 diagnostics manifests in two different ways: infected persons who receive negative test results (false-negatives) and persons that are not infected but receive positive results (false-positives). One problem of false-negatives on the individual level is that infected persons are not aware of infection and not quarantined and can transmit the disease. On the population level, false-negatives lead to underestimating the number of infected individuals [5, 6]. False-positives lead on the individual level to superfluous quarantining and contact tracing, wasting time and resources. On the population level, false-positives lead to overestimating the number of infected individuals and could be the cause of intervention measures stricter than necessary [7]. In infectious disease surveillance, we are mainly interested in the population-level effects of misclassification. Since the impact of diagnostic misclassification depends on the number of tested persons and the true incidence, both changing over time, the apparent dynamics of reported case numbers can be misleading [8] and it is important to quantify the potential amount of distortion and adjust for it.

In this work, we provide an approach to adjust reported COVID-19 case counts for diagnostic misclassification based on the matrix method [9]. We illustrate the impact of diagnostic misclassification on reported case counts on the example of Germany and the federal state of Bavaria under realistic assumptions for sensitivity and specificity of person-specific diagnostics and show that the approach can also be used to establish a lower-bound for the person-specific specificity. Furthermore, we illustrate how to use adjusted case numbers in downstream analyses like nowcasting or the estimation of the time-varying reproduction number. By this, it is possible to integrate solutions for both the problem of diagnostic misclassification as well as the problem of temporal assignment in the real-time analysis of case reporting data.

2 Methods

2.1 Data

Data on reported COVID-19 cases in Germany and Bavaria are collected based on the German Infection Protection Act (IfSG). For our analyses, we use daily German case numbers published by the Robert-Koch-Institute (RKI) [10], and, for Bavaria, the daily case numbers by the Bavarian Health and Food Safety Authority (LGL) [11]. In addition, the LGL provided us with person-specific case reporting data, including information on age, gender, and date of symptom onset if available (i.e., considered date of disease onset).

Furthermore, we use data on the number of SARS-CoV-2 laboratory tests that are directly reported by German and Bavarian testing facilities (university hospitals, research institutions, laboratories) to the respective health authorities. The testing facilities report the number of performed tests (analysed specimens) and the number of positive tests. The RKI publishes weekly data on the reported number of tests by German testing facilities [12, 13]. The LGL publishes daily numbers of laboratory tests and number of positive tests as reported by Bavarian testing facilities [11]. Because the data on case numbers come from a different source than the data on test numbers, the numbers of positive tests do not directly match to the number of reported cases and the reported number of performed tests does not directly correspond to the number of tested persons. Reasons for differences are diverse: (i) multiple testing of some individuals, (ii) reporting delay between the testing and the reporting of positive results from laboratories to local, regional and federal health authorities, and (iii) inconsistent reporting of test numbers by laboratories to the health authorities [14]. Furthermore, for the Bavarian data, persons whose tests are performed in Bavarian laboratories may not reside in Bavaria or vice versa.

Due to the weekly reporting of test numbers by the RKI, we focus our analysis on aggregated weekly case numbers for the German data. For Bavaria, we perform analyses based on daily data and make use of the person-specific information on disease onsets for estimating the epidemic curve.

In our main analysis, we focus on the time between May, 1 and Mid-September (utilizing data published on Sept 23, 2020, for Germany and data as per Sept 21, 2020, for Bavaria) as this was the most interesting phase of the pandemic in Germany with respect to changes in testing activity and their potential effects of misclassification on the epidemic curve, so far. Additionally, we provide current results for the analysis of the Bavarian data in the Supplemental Material and regular updates of the analysis on a public webpage (`https://corona.stat.uni-muenchen.de/nowcast/`).

2.2 Statistical approach to adjust reported case numbers for diagnostic misclassification

We present an approach to adjust reported case numbers in a given period (e.g., per day or week) for misclassification in the COVID-19 laboratory diagnostics (called examination in the following). Let NT_t be the number of examined persons whose test results would be reported to the health authorities on a given day (or within a given period) t in the event of a positive test. We denote the number of persons with a positive examination reported at time t as T_t^+(observed cases). We now assume that the examination results might be error-prone and the actual number of cases at time t (of all examined persons during that period) is D_t^+. In practice, D_t^+ is unknown and we want to estimate it based on the observed case numbers T_t^+, the number of examined persons NT_t, and assumptions regarding the sensitivity and specificity of the person-specific examination. Based on elementary probability calculations, the (expected) number of observed cases can be expressed in terms of the true case numbers by

$$E(T_t^+|D_t^+, NT_t) = D_t^+ \cdot \text{sens} + (NT_t - D_t^+) \cdot (1 - \text{spec}). \tag{1}$$

Equation (1) shows that the effects of a reduced sensitivity and specificity on the relation of observed case counts, T_t^+, and true case counts, D_t^+, differ structurally: while a reduced sensitivity leads to an under-estimation of the true number of cases by the factor of the sensitivity, the effect of the specificity is additive and depends on the number of examined persons, NT_t. To adjust the reported case numbers for misclassification we can re-order equation (1) and estimate the true number of cases at time t based on the reported case number, the number of examined persons, and the sensitivity and specificity:

$$\hat{D}_t^+ = \frac{T_t^+ - NT_t \cdot (1 - \text{spec})}{\text{sens} + \text{spec} - 1}. \tag{2}$$

This estimator relates to the well known matrix method, see e.g., [9]. As described above, the number of examined persons, NT_t, is not directly available in (German) case reporting data. We therefore approximate it using a statistical model based on the number of performed and the number of positive COVID-19 tests reported by laboratories and the number reported cases by the health authorities (cf. Supplemental Note 1).

2.3 Conceptualizing the sources of misclassification and deriving realistic person-specific misclassification probabilities

We conceptualize the sources of misclassification per PCR-test and per person-specific examination. For this, we searched for evidence on the extent of misclassification from literature and by consultation of experts from public health and virology. Besides the misclassification of PCR-tests under controlled laboratory conditions, we also consider misclassification from the following: (i) collecting and handling of specimens, (ii) additional aspects of uncertainty in PCR-tests under realistic conditions, e.g., due to timing of the test after infection or repeated testing in case of unclear results, (iii) varying testing strategy (physician-initiated/symptom-based vs. screening).

2.4 Adjusting reported case numbers for diagnostic misclassification

From the observed number of reported cases as published by the RKI (weekly numbers, Germany) or the LGL (daily numbers, Bavaria) and the estimated numbers of examined individuals, we estimate the true number of cases per time unit from equation (2) under various realistic misclassification scenarios. We compare results to reported case counts.

2.5 Estimation of the epidemic curve and the time-varying reproduction number based on adjusted case numbers

Daily reported case numbers are the basis for more complex downstream analyses, e.g., the real-time estimation of the epidemic curve (nowcasting) and estimation of the time-varying reproduction number, R(t). Those analyses provide a better characterization of the current state of the pandemic than reported case counts but require information on disease onset and reporting dates of cases [2]. To perform such analyses adjusted for diagnostic misclassification, we propose this general approach: (i) derive the adjusted number of reported cases per time unit, (ii) remove a randomly selected number of surplus reported (false-positive) cases per time unit from the data, and (iii) to estimate the epidemic curve and the reproduction number on the reduced data. Under the assumption of a low sensitivity, the number of misclassification-adjusted cases per time unit can be higher than the number of reported cases. We propose the following two-stage approach for the steps (ii) and (iii) from above and show the analytical validity (cf. Supplemental Note 1): (1) remove a given number of false-positive cases from the data based on the assumed specificity smaller than one and a sensitivity of one and estimate the epidemic curve based on these case counts and, (ii) adjust the resulting estimated epidemic curve for false-negatives based on an assumed sensitivity smaller than one. This procedure avoids the otherwise necessary up-sampling of data and cuts down considerably on computational resources.

We exemplify the misclassification-adjusted estimation of the epidemic curve and R(t) based on Bavarian case reporting data, since we were able to obtain person-specific information on disease onset for this German federal state.

2.6 Code and data

We provide R-code to reproduce our analyses on Github (`https://github.com/FelixGuenther/mc_covid_cases_public`) and update the results of the estimation of the misclassification-adjusted epidemic curve on our web page regularly (`https://corona.stat.uni-muenchen.de/nowcast/`). The proposed analysis can thus easily be extended to other countries or regions.

3 Results

3.1 Number of reported tests and reported cases in Germany and Bavaria

In the following, we refer to the time period between start of the laboratory reporting of conducted SARS-CoV-2 tests (calendar week 11 for Germany overall, March 16th for Bavaria) and end of our work's observational period (as per calendar week 38 for Germany, Sept 20th for Bavaria), which is chosen to capture the low-level incidence in summer 2020. Based on the number of tests reported by laboratories, the testing activity in Germany started increasing in July and stabilized in September (Figure 1A) and this increase was more pronounced in Bavaria. Altogether, 15.7 million PCR-tests were reported by laboratories for Germany (i.e. 19 per 100 inhabitants) and 3.5 million by Bavarian laboratories (26 per 100). During spring/summer 2020, the number of reported positive tests (by laboratories) and reported cases (by health authorities) were the highest in the beginning of April, then decreasing to a low level in Mid-May and started rising in June/July (Figure 1B). The summer rise of case counts coincides with the increase of the test activity.

Note that the number of positive tests reported by laboratories and the number of COVID-19 cases reported by health authorities are not equal due to different reporting institutions. In Germany, there were 310,630 reported positive tests at the time of our analysis and 272,664 reported cases, in Bavaria 67,214 positive tests and 63,857 cases. Differences in time-series are particularly apparent in the daily data from Bavaria (Figure 1B): at the beginning of the first wave in February, more cases were reported by health authorities than positive tests by laboratories, most likely due to incompleteness of data reported from laboratories and incomplete coverage of reporting laboratories at this early phase. In our subsequent analyses, we focus on the time-period from May 1st to Sept 20th, 2020, where reporting of test numbers by laboratories was established and changes in the testing activity combined with low-level incidence in Germany is an ideal situation studying the impact of diagnostic misclassification.

3.2 Evidence on the performance of COVID-19 diagnostics

Routine COVID-19 examination in a screening or hospital setting is currently mostly done based on the detection of unique sequences of virus RNA using PCR-tests on clinical respiratory tract specimens of examined individuals [15]. Laboratories use different PCR-tests targeting different viral genes. The analytic sensitivity and specificity of PCR-tests applied to an adequately collected and handled specimen are generally reported to be very high [16]. In a proficiency test of German laboratories from April 2020, the authors found an average target-specific specificity between 97.8% and 98.6% and a sensitivity between 98.9% and 99.7% [17]. However, such numbers on the analytic performance of PCR-tests in laboratory settings do not directly relate to the person-specific performance of COVID-19 examinations in a hospital or screening setting, which is relevant for adjusting the reported case numbers in real-time surveillance. We systematically document sources of error in the person-specific

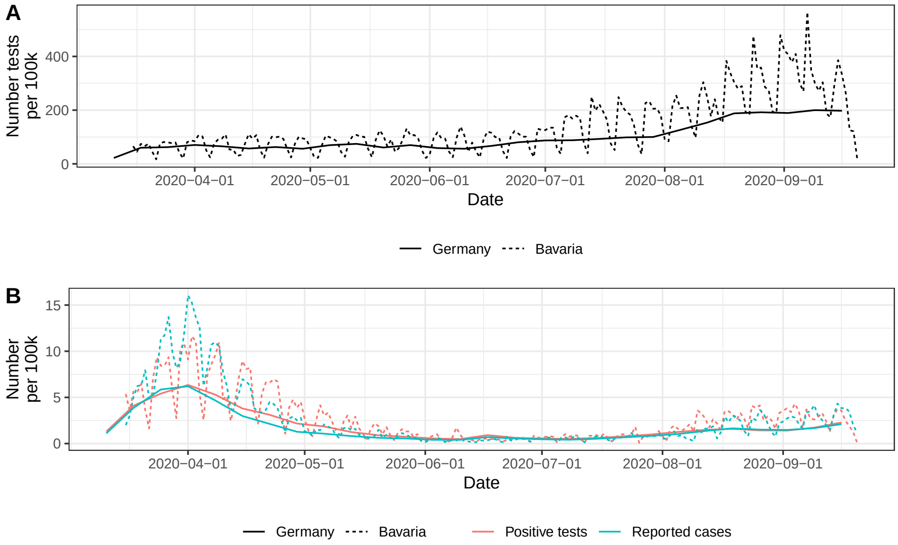


Figure 1: Reported number of performed PCR-tests and COVID-19 cases in Germany and Bavaria. Panel A shows the daily number of performed COVID-19 PCR-tests per 100,000 inhabitants in Germany and Bavaria as reported by the laboratories. Panel B shows the daily number of positive PCR-tests as reported by the laboratories and the number of COVID-19 cases as reported by the health authorities per 100,000 inhabitants. The daily numbers are average numbers from weekly reported data for Germany and daily reported numbers for Bavaria (with an obvious weekly cycle).

COVID-19 examination (Table 1). While it is a substantial challenge to cover all aspects, some aspects should be noted in detail:

With respect to the sensitivity, there are two factors that increase the probability of a false-negative examination result: first, inappropriate pre-analytical collection or handling of specimens, e.g., during transportation to laboratories, can lead to false-negative results. Second, the performance of PCR-tests in infected persons is reported to vary strongly depending on the time-point of the test after infection due to quantitatively insufficient viral RNA in the early pre-symptomatic phase [18]. The authors report a sensitivity close to zero directly after infection, an increase to 80% on day 8 (i.e., three days after typical symptom onset), and a decrease afterwards, all with high uncertainty. Similar results on the time-varying performance indicate a sensitivity of bigger 90% at the day of symptom onset with following decline [19]. Therefore, fast symptom-based testing should have a considerably higher sensitivity than testing in a screening setting. Multiple tests at different time-points (e.g., after symptom onset for a case with initial negative finding in a pre-symptomatic screening) can increase the person-specific sensitivity. It is difficult to quantify the overall sensitivity of the person-specific COVID-19 examination in the general testing regime, but it is certainly much lower than the high analytic sensitivity reported for tests in a controlled laboratory experiment. We perform our analyses based on an assumed sensitivity of 70% and 90% as a range of realistic values that also correspond to previously reported findings [20, 21, 22].

With respect to the specificity of the person-specific examination, chance or cross-contamination between specimens or swaps between infected and not infected individuals can lead to false-positive test results. It is recommended that tests are repeated to ascertain unclear results and to use tests that target two viral genes, which decreases the probability of false-positives considerably compared to a single PCR-test and/or single target tests [23]. However, the specific approach might depend on

Drivers of performance	**Comment**	**Quantification, implication**
Sensitivity		
Analytical	Very small error, depends on test	Average analytical sensitivity >98%
Specimens	Errors in collection and/or handling	Reduction due to damaged specimen, extent unclear
Real-world application	Test result depends on viral load	Reduction of sensitivity due to wrong timing of test: close to zero directly after infection, biggest ∼8 days after infection (or shortly after/around symptom onset)
Testing strategy	Infection might be overlooked in screening due to wrong timing, individuals might get tested a second time after disease onset	Sensitivity higher in case of targeted testing or physician-initiated (e.g., due to symptoms): reduced sensitivity in screening application or testing of asymptomatic, sensitivity might change over time; infected individuals might get detected after repeated testing
Specificity		
Analytical	Very small error, depends on test	Average analytical specificity >98%
Specimens	False-positives due to swaps or cross-contamination	Reduction due to cross-contamination, frequency unclear, rather low
Real-world application	Testing on dual targets and repeated testing in case of unclear results	Increase of specificity due to repeated testing and expert evaluation of results, extent might vary between laboratories
Testing strategy	Increased testing activity might reduce quality of testing due to limited resources	High workload in laboratories might decrease specificity of tests due to less time for validation of unclear results; might imply changes in specificity over time, extent of effect unclear

Table 1: Identified drivers of the performance of the person-specific COVID-19 examinations based on PCR-tests and their impact on sensitivity and specificity.

the laboratory and the overall situation (e.g., workload). Based on the Bavarian data, we show that a specificity for the person-specific examination lower than 99.5% is not empirically supported: a lower specificity would imply more false-positive cases than cases reported in the low-incidence phase in June and beginning of July, given the estimated number of examined persons in the respective time period (cf. section Adjusted case counts). For our analyses, we thus apply a specificity of 99.9%, 99.7% or 99.5%.

3.3 Number of examined persons in Germany and Bavaria

To adjust reported case counts for misclassification, we first derive the number of examined individuals per time unit (weekly or daily) based on the best performing varying-coefficient regression model to relate the number of reported cases to the number of reported positive PCR-tests over time (Supplemental Note 2, Supplemental Figure 1).

We find that, for the German data, the number of examined persons increases over time, most notably in the first half of August, and reaches a plateau in September (Figure 2A). The number of reported PCR-tests (derived from weekly data) matches the estimated number of examined persons

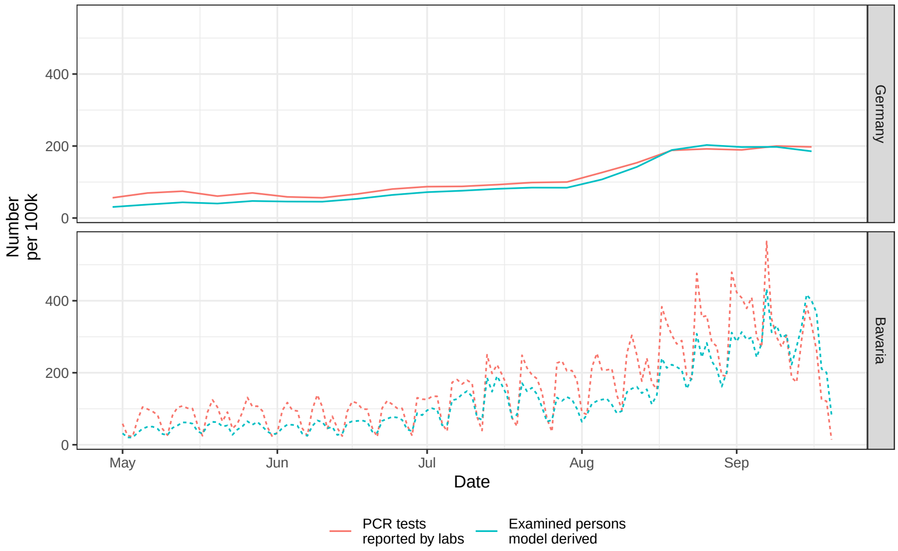


Figure 2: Reported number of PCR-tests and model-based number of examined persons in Germany and Bavaria. Shown are the daily numbers of performed PCR-tests per 100,000 inhabitants as well as the estimated number of examined persons. The estimated numbers of examined persons result from a varying-coefficient model fitted based on the association of the number of reported positive PCR-tests and the number of reported COVID-19 cases. Results for Germany stem from modeling weekly data and dividing by seven to illustrate results on a daily scale; the results for Bavaria are based on daily data (with an obvious weekly cycle).

quite closely. For the daily Bavarian data, we find a similar increase in examined persons, but larger differences in the number of reported PCR-tests by Bavarian laboratories and examined persons residing in Bavaria (Figure 2B). This difference is most pronounced at the end of German summer holidays (end of July to early September, where many inhabitants from other federal states were tested for free in Bavarian testing facilities upon travelling back home from the South through Bavaria). In the last days of our observation period here, reported numbers of PCR-tests decrease, which reflects the reporting delays between the time of the tests and the reporting of tests to health authorities. This reporting delay is also present in the case reporting data with respect to the number of reported cases per day by the local health authorities and we adjust for the reporting delay based on the nowcasting described below (cf. section Estimation of the adjusted epidemic curve).

3.4 Misclassification-adjusted case counts

Based on the derived numbers of examined persons in Germany and Bavaria, we adjust observed case numbers for misclassification in the person-specific diagnostic examination for SARS-CoV-2 infection under the assumptions of a sensitivity of 90% or 70% and a specificity of 99.9%, 99.7%, or 99.5% (Figure 3). Adjusting for a specificity less than one leads to a reduction of case numbers and adjusting for a sensitivity less than one increases the case numbers, as expected from theory. The impact of sensitivity and specificity on adjusted case counts is structurally different: while adjusting for imperfect sensitivity corresponds roughly to a time-constant upscaling of case numbers by the factor 1/sensitivity, the impact of the specificity varies over time due to varying numbers of examined individuals. The relative effect of adjusting for false-positives is the largest during July, when reported case numbers were low and testing activity started rising (Figure 4). The relative effect diminishes in August/September despite

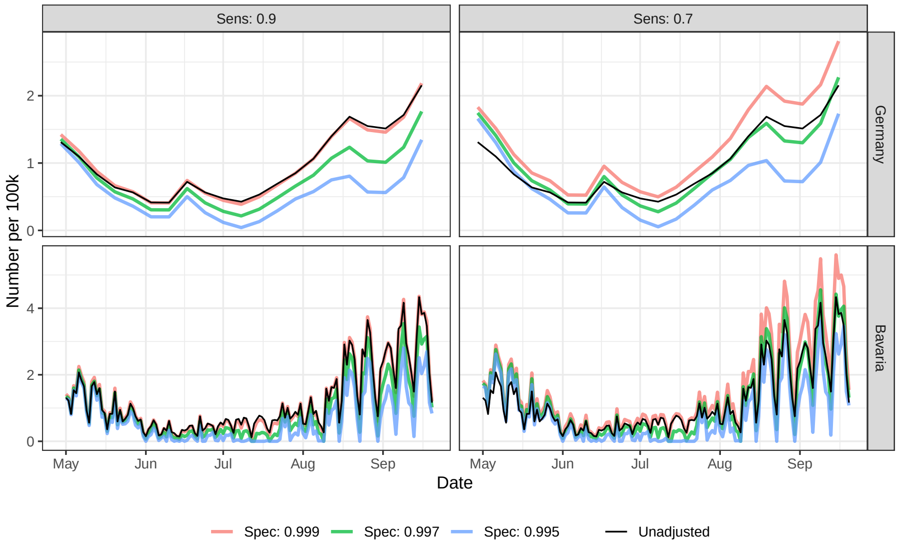


Figure 3: Misclassification adjusted case numbers for Germany and Bavaria. Shown are daily misclassification adjusted case numbers per 100,000 inhabitants for different assumptions regarding the sensitivity and specificity of the person-specific COVID-19 examination and the unadjusted reported case numbers. The smaller the assumed specificity, the higher the number of false-positives in the reported cases and the bigger the reduction of adjusted case numbers. The smaller the sensitivity, the bigger the probability of false-negative examinations and the higher the adjusted case numbers. Results for Germany stem from weekly reported data, the results for Bavaria are based on daily data (with an obvious weekly cycle). Note the different scales on the y-axis for German and Bavarian data for better visibility.

the increased testing activity during that time period due to an increased incidence.

Adjusted case numbers in Germany suggest that, dependíng on the extent of the specificity, the increase in case counts in Beginning of July is indeed partly due to false-positives: adjusted case counts are lower than the observed under the assumption of a sensitivity of 90%. However, the increase of cases is still apparent even after the adjustment, which indicates that not the full increase was induced by false-positives. Based on the daily reported Bavarian data, we find a similar pattern. This daily data also prompted us to assume a specificity of no lower than 99.5%. Even under the extreme assumption that there were no true cases in June and July and all observed cases were false-positives, the false-positive proportion might not plausibly be larger than 0.5%. More precisely, we would have expected more false-positive cases than actually reported in 20 of the 61 days of June and July given the estimated number of examined persons and an assumption of 0.5% false-positive examination results. Of all reported PCR-tests by the Bavarian laboratories in June and July, only 0.56% were reported as positive.

3.5 Misclassification-adjusted epidemic curve and time-varying reproduction number in Bavaria

We use misclassification-adjusted case numbers to estimate the epidemic curve or the reproduction number R(t), which provide key information for real-time surveillance. When comparing the epidemic curve adjusted for a diagnostic specificity of 99.7% or 99.5% to the unadjusted epidemic curve, we find structural differences in June/July: while the unadjusted epidemic curve started to rise slowly

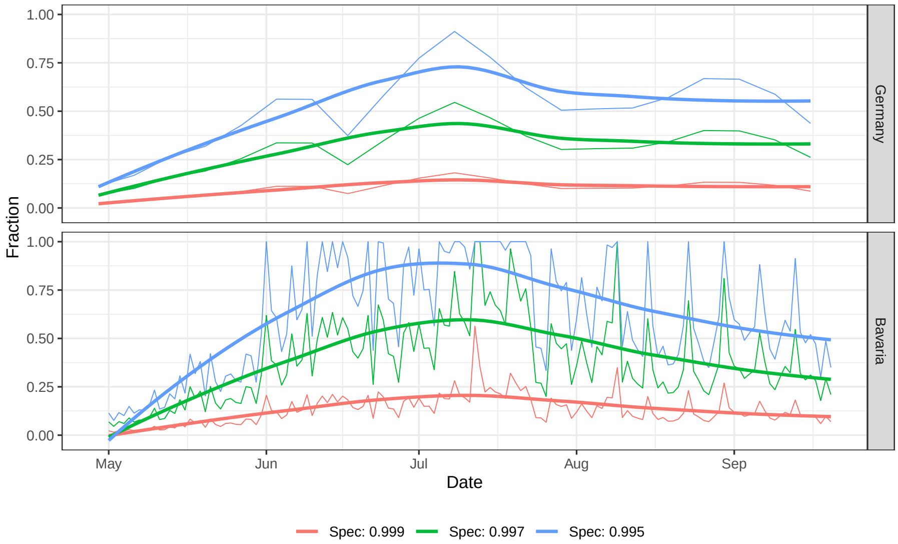


Figure 4: Fraction of false-positives in Germany and Bavaria over time for different assumptions on the specificity. Shown is the fraction of all cases that is removed in the misclassification adjusted case counts due to false-positive cases over time assuming different values for the (time-constant) specificity and a constant sensitivity. For both, German and Bavarian data, the relative effect of the false-positive cases is biggest during July, i.e., in the time of low reported case numbers and rising testing activity and is less important with fewer testing and/or higher case numbers. This fraction is independent of the assumed (constant) sensitivity due to the multiplicative nature of the adjustment for a sensitivity smaller than 100%.

around Mid-June, the adjusted epidemic curves remained on a very low level until Mid-July (Figure 5A). When additionally accounting for a sensitivity of 90% or 70%, we find the same (Figure 5B); in fact, the impact of different values for the sensitivity was low during this time due to the low incidence. After Mid-July, the increase of the epidemic curve was genuine: in all scenarios, we observe an increase of case counts. In August/September, the reduction of case counts when accounting for false-positives based on a specificity of 99.7% is neutralized by the increase of the case count when considering also false-negatives based on a sensitivity of 70%: the unadjusted and adjusted epidemic curve are basically the same.

The misclassification-adjusted estimation of the epidemic curve facilitates estimation of an adjusted time-varying effective reproduction number $R(t)$. It is an estimate of the average number of individuals that are infected by an individual with disease onset on a given day t. If this factor is smaller than one, case numbers are decreasing, if it is bigger than one, case numbers are increasing within the following days. At the beginning of July, adjusted $R(t)$, is smaller than one when assuming a specificity of 99.5% or 99.7%, while the unadjusted $R(t)$ was slightly larger than one (Figure 5C). This is in line with the observation that, at that time, true case numbers might have been close to zero or decreasing, but reported numbers were slightly increasing due to false-positives from increased testing. However, shortly thereafter in Mid-July, the adjusted $R(t)$ exceeds one for all considered values of the specificity, which is in line with increasing case numbers in August/September and an actual true increase in the case counts. The (relative) increase in case numbers in July is bigger when accounting for a relatively high fraction of false-positives in June, yielding bigger estimates of $R(t)$ the lower the assumed specificity.

4 Discussion

In this work, we conceptualize sources of uncertainty in person-specific PCR-based COVID-19 diagnostics and quantify realistic extents of misclassification in terms of plausible values for sensitivity and specificity. We provide an approach to adjust reported case counts for the diagnostic misclassification and extend this approach to a misclassification-adjusted real-time estimation of the epidemic curve and reproduction number R(t). This helps to solve two important problems of case reporting data in real-time surveillance: temporal assignment of reported cases and accounting for misclassification in COVID-19 diagnostics. On the example of data from Germany and Bavaria, we quantify the impact of diagnostic misclassification on the time series of reported case counts and the real-time estimation of the epidemic curve.

Sensitivity and specificity of the diagnostics have a structurally different impact on case counts: a sensitivity smaller than one leads to an underestimation of the true number of cases by a factor independent of the number persons tested, a specificity is smaller than one leads to an overestimation of case numbers, the extent of which depends on the number of tested persons and corresponds to the number of false-positive cases. When the number of tested persons changes over time, the number of false-positives changes as well. This can distort the estimated epidemic curve and thus the apparent dynamics of the epidemic.

For the German case counts during summer 2020, we find that the reported case numbers during the low incidence phase in June and July were to a relevant part false-positives, but that the observed increase of cases since Mid-July was not entirely driven by false-positives. When accounting for false-positive and false-negative test results, the adjusted case numbers in August and September were on a similar level than the unadjusted. Therefore, the increase seen in German case numbers from July until September, which was debated as being predominantly driven by false-positives due to increased testing activity, was a genuine increase of infections. Based on our developed analysis approach such questions can also be answered in future real-time surveillance.

Since the relative impact of misclassification on reported case counts depends on the number of examined persons and the current incidence, we update our analyses on a regular basis and pro-

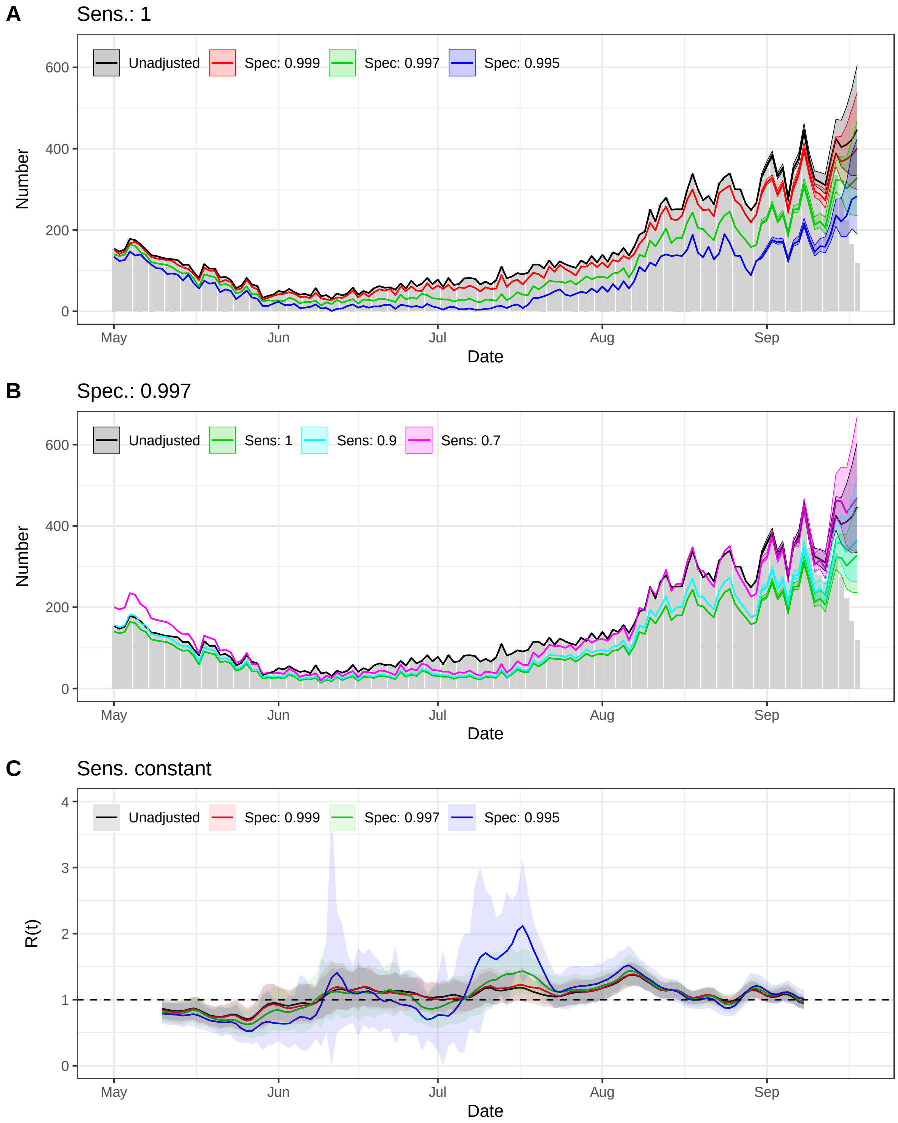


Figure 5: Misclassification adjusted epidemic curve and time-varying reproduction number $R(t)$. Panel A shows the estimated epidemic curve (number of disease onsets per day) for different assumptions regarding the specificity (assuming a perfect sensitivity as in the unadjusted analysis). Panel B shows the results for different assumptions regarding the sensitivity assuming a specificity of 99.7% and additionally the unadjusted curve. The grey bars show the number of cases with disease onset on a specific day (or imputed if missing) as reported to the LGL until Sept, 21. Panel C shows the misclassification adjusted time-varying reproduction number for different assumptions regarding the specificity. Results are independent of the assumed (time-constant) sensitivity. The epidemic curve is estimated based on a Bayesian hierarchical nowcasting model considering misclassification adjusted case counts, the time-varying reproduction number is estimated based on the estimated epidemic curves.

vide the results adjusted for various scenarios of sensitivity and specificity on our website (`https://corona.stat.uni-muenchen.de/nowcast/`). Based on current results from January, 2021, it becomes obvious that the true number of cases during the strong second wave in Bavaria might be considerably underestimated due to false-negative results (Supplemental Figure 2). The general structure of the epidemic curve is, however, not changed by taking into account errors in COVID-19 examinations. The distortion due to false-positive tests hardly plays a role in times of high incidence.

Our analyses have some assumptions and limitations. The adjustment for misclassification in COVID-19 diagnostics depends on accurate information with respect to the number of examined individuals. Such information is not directly available in German surveillance data and we rely on a model-based approach to estimate this number. The results appear plausible, but cannot be directly validated. Furthermore, we assume constant misclassification probabilities for the COVID-19 diagnostics over time. This assumption is likely violated by changes in the diagnostic procedures, changes in workload for the laboratories, and changes or improvements in standard operating procedures. We believe, however, that our calculations over a range of plausible assumptions for sensitivity and specificity can give a realistic overview about potential biases due to misclassification.

Misclassification in COVID-19 diagnostics is not the only problem for the interpretation of reported case numbers. It is well known that not all infected persons are captured and examined. This leads to a relevant difference between the number of infected individuals and the number of reported infections. This problem is not COVID-19-specific, but also occurs with other diseases when they only cause mild symptoms in some cases - as typically described by the so-called surveillance pyramid [24]. The case detection ratio for COVID-19 was estimated in Germany from antibody prevalence studies to be as low as 0.2 to 0.4 for the first phase of the pandemic in spring 2020 [25, 26, 27]. If the case detection ratio is constant over time, the structure of the epidemic curve remains unchanged. However, expanding the testing activity increases the detection ratio. Therefore, one has to be careful when comparing absolute numbers of cases over longer periods of time. More specifically, the absolute number of reported cases in Germany during March-April 2020 is not comparable to the case numbers reported during August and September (and today) due to increased testing. In our analysis and figures, we focused on the epidemic curve starting in May 2020, i.e., after the first phase of the pandemic. Nevertheless, the remaining increase of the misclassification-adjusted case numbers during August-September might still partly be driven by an increasing case detection ratio. This question cannot be answered directly from reported case numbers, but requires additional information, e.g. numbers of hospital admissions, deaths or longitudinal data on antibody prevalence. Recent modelling approaches [28] might also be extended to account for misclassification in COVID-19 testing.

Overall, a thorough analysis of case reporting data adjusting for misclassification is important to improve monitoring of short-term changes in the pandemic situation. Given the high relevance of reported case counts for the surveillance of an acute pandemic, we recommend to analyse case counts adjusted for plausible assumptions regarding diagnostic misclassification in the corresponding testing regimes to understand potential distortions due to misclassification and avoid the appearance of false precision. This remains especially important as containment and surveillance of the current pandemic remains a central task worldwide. Consideration of misclassification in tests will also be relevant with the increasing use of alternative test methods, especially if their results are not confirmed by the established PCR-tests. In this work, we have shown that a thorough analysis of surveillance data can help to capture current trends more reliably and reduce the ambiguity of the available information, which can ultimately support public confidence in the available evidence. This will also be important in the event of a future emergence of new virus strains or other pandemic pathogens.

References

[1] ECDC. Data on testing for COVID-19 by week and country, 2020. URL `https://www.ecdc.europa.eu/en/publications-data/covid-19-testing`.

[2] Felix Günther, Andreas Bender, Katharina Katz, Helmut Küchenhoff, and Michael Höhle. Nowcasting the COVID-19 Pandemic in Bavaria. *Biometrical Journal*, 2020. doi: https://doi.org/10.1002/bimj.202000112.

[3] Michael Höhle and Matthias an der Heiden. Bayesian Nowcasting during the STEC O104:H4 Outbreak in Germany, 2011. *Biometrics*, 70(4):993–1002, December 2014. doi: 10.1111/biom.12194.

[4] Sarah F. McGough, Michael A. Johansson, Marc Lipsitch, and Nicolas A. Menzies. Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking. *PLoS Comput Biol*, 16(4): e1007735, April 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007735. URL `https://dx.plos.org/10.1371/journal.pcbi.1007735`.

[5] Steven Woloshin, Neeraj Patel, and Aaron S. Kesselheim. False Negative Tests for SARS-CoV-2 Infection — Challenges and Implications. *New England Journal of Medicine*, 383(6):e38, 2020. doi: 10.1056/NEJMp2015897. URL `https://doi.org/10.1056/NEJMp2015897`.

[6] M. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2004.

[7] Elena Surkova, Vladyslav Nikolayevskyy, and Francis Drobniewski. False-positive COVID-19 results: hidden problems and costs. *The Lancet Respiratory Medicine*, 2020. Publisher: Elsevier.

[8] Igor Burstyn, Neal D. Goldstein, and Paul Gustafson. Towards reduction in bias in epidemic curves due to outcome misclassification through Bayesian analysis of time-series of laboratory test results: case study of COVID-19 in Alberta, Canada and Philadelphia, USA. *BMC Med Res Methodol*, 20(1):146, December 2020. ISSN 1471-2288. doi: 10.1186/s12874-020-01037-4. URL `https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01037-4`.

[9] WJ Rogan and B Gladen. Estimating prevalence from results of a screening-test. *American Journal of Epidemiology*, 107(1):71–76, 1978. ISSN 0002-9262. doi: 10.1093/oxfordjournals.aje.a112510.

[10] Robert Koch Institut (RKI). COVID-19: Fallzahlen in Deutschland, . URL `https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Fallzahlen.html`.

[11] Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit (LGL). Übersicht der Fallzahlen von Coronavirusinfektionen in Bayern. URL `https://www.lgl.bayern.de/gesundheit/infektionsschutz/infektionskrankheiten_a_z/coronavirus/karte_coronavirus/index.htm`.

[12] Robert Koch Institut (RKI). Daily Situation Report, . URL `https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Gesamt.html`.

[13] Robert Koch Institut. Erfassung der SARS-CoV-2-Testzahlen in Deutschland. URL `https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Testzahl.html`.

[14] D Staat, D Stern, J Seifried, S Böttcher, S Albrecht, N Willrich, B Zacher, M Mielke, U Rexroth, and O Hamouda. Erfassung der SARS-CoV-2-Testzahlen in Deutschland (Stand 9.9.2020). *Epid Bull*, 45:16–19, 2020. doi: 10.25646/7202.

[15] World Health Organization. Laboratory testing for coronavirus disease (COVID-19) in suspected human cases: interim guidance, 19 March 2020. Technical report, World Health Organization, 2020.

[16] European Commission. COVID-19 In Vitro Diagnostic Devices and Test Methods Database. URL `https://covid-19-diagnostics.jrc.ec.europa.eu/`.

[17] Heinz Zeichhardt and Martin Kammel. Kommentar zum Extra Ringversuch Gruppe 340 Virusgenom Nachweis SARS-CoV-2. Technical report, INSTAND, 2020.

[18] Lauren M Kucirka, Stephen A Lauer, Oliver Laeyendecker, Denali Boon, and Justin Lessler. Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based SARS-CoV-2 tests by time since exposure. *Annals of Internal Medicine*, 2020. Publisher: American College of Physicians.

[19] Paul S Wikramaratna, Robert S Paton, Mahan Ghafari, and José Lourenço. Estimating the false-negative test probability of SARS-CoV-2 by RT-PCR. *Eurosurveillance*, 25(50), December 2020. ISSN 1560-7917. doi: 10.2807/1560-7917.ES.2020.25.50.2000568. URL https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.50.2000568.

[20] Jessica Watson, Penny F Whiting, and John E Brush. Interpreting a covid-19 test result. *British Medical Journal*, 369, 2020. Publisher: British Medical Journal Publishing Group.

[21] Yang Yang, Minghui Yang, Jing Yuan, Fuxiang Wang, Zhaoqin Wang, Jinxiu Li, Mingxia Zhang, Li Xing, Jinli Wei, Ling Peng, Gary Wong, Haixia Zheng, Weibo Wu, Chenguang Shen, Mingfeng Liao, Kai Feng, Jianming Li, Qianting Yang, Juanjuan Zhao, Lei Liu, and Yingxia Liu. Laboratory Diagnosis and Monitoring the Viral Shedding of SARS-CoV-2 Infection. *The Innovation*, 1(3):100061, November 2020. ISSN 26666758. doi: 10.1016/j.xinn.2020.100061. URL https://linkinghub.elsevier.com/retrieve/pii/S2666675820300643.

[22] Nikhil S Padhye. Reconstructed diagnostic sensitivity and specificity of the RT-PCR test for COVID-19. *medRxiv*, 2020. Publisher: Cold Spring Harbor Laboratory Press.

[23] Robert Koch Institut. Hinweise zur Testung von Patienten auf Infektion mit dem neuartigen Coronavirus SARS-CoV-2: Direkter Erregernachweis durch RT-PCR (Stand: 25.01.2021), 2021. URL https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Vorl_Testung_nCoV.html.

[24] Cheryl L Gibbons, Marie-Josée J Mangen, Dietrich Plass, Arie H Havelaar, Russell John Brooke, Piotr Kramarz, Karen L Peterson, Anke L Stuurman, Alessandro Cassini, Eric M Fèvre, and others. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC public health*, 14(1):147, 2014. Publisher: Springer.

[25] Claudia Santos-Hövener, Markus A Busch, Carmen Koschollek, Martin Schlaud, Jens Hoebel, Robert Hoffmann, Hendrik Wilking, Sebastian Haller, Jennifer Allen, Jörg Wernitz, and others. Seroepidemiologische Studie zur Verbreitung von SARS-CoV-2 in der Bevölkerung an besonders betroffenen Orten in Deutschland–Studienprotokoll von CORONA-MONITORING lokal. *Journal of Health Monitoring*, 5, 2020. Publisher: Robert Koch-Institut.

[26] Hendrik Streeck, Bianca Schulte, Beate Kuemmerer, Enrico Richter, Tobias Hoeller, Christine Fuhrmann, Eva Bartok, Ramona Dolscheid, Moritz Berger, Lukas Wessendorf, Monika Eschbach-Bludau, Angelika Kellings, Astrid Schwaiger, Martin Coenen, Per Hoffmann, Markus Noethen, Anna-Maria Eis-Huebinger, Martin Exner, Ricarda Schmithausen, Matthias Schmid, and Gunther Hartmann. Infection fatality rate of SARS-CoV-2 infection in a German community with a super-spreading event. *medRxiv*, 2020. doi: 10.1101/2020.05.04.20090076. URL https://www.medrxiv.org/content/early/2020/06/02/2020.05.04.20090076. Publisher: Cold Spring Harbor Laboratory Press _eprint: https://www.medrxiv.org/content/early/2020/06/02/2020.05.04.20090076.full.pdf.

[27] Michael Pritsch, Katja Radon, Abhishek Bakuli, Ronan Le Gleut, Laura Olbrich, Jessica Michelle Guggenbuehl Noller, Elmar Saathoff, Noemi Castelletti, Mercè Garí, Peter Puetz, Yannik Schaelte, Turid Frahnow, Roman Wölfel, Camilla Rothe, Michel Pletschette, Dafni Metaxa, Felix Forster, Verena Thiel, Friedrich Riess, Maximilian Nikolaus Diefenbach, Guenter Froeschl, Jan Bruger, Simon Winter, Jonathan Frese, Kerstin Puchinger, Isabel Brand, Inge Kroidl, Jan Hasenauer, Christiane Fuchs, Andreas Wieser, Michael Hoelscher, and KoCo19 Study Group. Prevalence and Risk Factors of Infection in the Representative COVID-19 Cohort Munich. SSRN Scholarly Paper ID 3745128, Social Science Research Network, Rochester, NY, January 2021. URL https://papers.ssrn.com/abstract=3745128.

[28] Marc Schneble, Giacomo De Nicola, Göran Kauermann, and Ursula Berger. Spotlight on the dark figure: Exhibiting dynamics in the case detection ratio of COVID-19 infections in Germany. preprint, medRxiv, December 2020. URL http://medrxiv.org/lookup/doi/10.1101/2020.12.23.20248763.

Supplementary material: Analysis of COVID-19 case numbers: adjustment for diagnostic misclassification on the example of German case reporting data

Felix Günther*[1], Ursula Berger[2], Michael Höhle[3], Andreas Bender[1], Manfred Wildner[4], Iris M. Heid[5], and Helmut Küchenhoff[1]

[1]Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Germany
[2]Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich, Germany
[3]Department of Mathematics, Stockholm University, Sweden
[4]Bavarian Health and Food Safety Authority / Pettenkofer School of Public Health, Oberschleißheim, Germany
[5]Department of Genetic Epidemiology, University of Regensburg, Germany

February 2, 2021

Supplemental Note 1 - Misclassification adjustment for case numbers

We denote a binary indicator variable D_i for the event that a person i is currently infected. $D_i = 1$ means that a person is currently infected by COVID-19 and $D_i = 0$ that a person is not infected. The result of the person's COVID-19 examination (e.g., one or multiple sequential PCR-tests) is denoted by T_i. $T_i = 1$ corresponds to a classification as infected (positive examination) and $T_i = 0$ to a negative COVID-19 examination. Furthermore, the sensitivity (i.e., the probability of a true positive examination) is defined by $\text{sens} = P(T_i = 1|D_i = 1)$ and the specificity (probability of true negative examination) is denoted by $\text{spec} = P(T_i = 0|D_i = 0)$. Then the probability of a positive examination is given by

$$P(T_i = 1) = P(D_i = 1) \cdot \text{sens} + P(D_i = 0) \cdot (1 - \text{spec}). \tag{1}$$

We denote the (unknown) number of examined persons at a certain time t by NT_t and the number of positive diagnostic procedures (i.e., reported cases) by $T_t^+ = \sum_{i=1}^{NT_t} T_i$. We can rewrite the expected number of reported cases on day t as

$$E(T_t^+|D_t^+, NT_t) = D_t^+ \cdot \text{sens} + (NT_t - D_t^+) \cdot (1 - \text{spec}), \tag{2}$$

where D_t^+ denotes the (unknown) number of actually infected persons that are examined on day t. Equation (2) is derived from equation (1) by replacing the probabilities $P(\cdot)$ by the corresponding relative frequencies, $P(T_i = 1) = T_t^+/NT_t$, $P(D_i = 1) = D_t^+/NT_t$, and $P(D_i = 0) = (NT_t - D_t^+)/NT_t = D_t^-/NT_t$, and multiplying both sides with the number of examined persons, NT_t.

Equation (2) shows that the effects of sensitivity and specificity on the observed case counts are different. While a low sensitivity leads to an underestimation of the number of cases by the factor sens, the effect of specificity is additive and depends on the number of examined persons.

If all information, i.e., the number of reported cases T_t^+, the number of examined persons NT_t, and the sensitivity and specificity of the person-specific examination are known, we can rewrite equation (2) and estimate the number of true cases D_t^+ based on

$$\hat{D}_t^+ = \frac{T_t^+ - NT_t \cdot (1 - \text{spec})}{\text{sens} + \text{spec} - 1}. \tag{3}$$

[*] felix.guenther@stat.uni-muenchen.de

This estimator relates to the well known matrix method, see e.g., Rogan and Gladen (1978).

Since only positive COVID-19 examinations are directly reported to German health authorities, only the overall number of reported cases T_t^+ is directly available, but the corresponding number of examined individuals, that would be reported as case on day t in the event of a positive test, NT_t, is unknown. However, the overall number of performed tests is separately reported by the laboratories as well as the number of positive tests. This data can have a difference in temporal allocation and, furthermore, we expect more reported tests than examined persons due to multiple tests for single persons and - in case of the Bavarian data - tests of individuals living outside of Bavaria.

To establish the relationship between the two quantities, we utilize the positive test results and model the number of reported cases at the health authorities based on the number of positive tests reported by the laboratories from the current and previous time points (i.e., based on lagged time series of reported positive tests from the laboratories). We utilize two different models and consider different degrees for the lag-number of positive tests. The first model corresponds to standard linear regression with linear effects of the (lagged) number of positive tests from the same and previous time points:

$$E(T_t^+) = \sum_k \alpha_k \cdot \text{Test}_{t-k}^+, \quad t = \text{May 1st}, \ \ldots, \ T. \tag{4}$$

Here, the number of positive tests reported by the laboratories at time point t is denoted by Test_t^+ and T corresponds to the most current time point. We estimate the model based on all possible subsets of lags $k \subset \{0, \ldots, 7\}$. As a second model, we consider a varying-coefficient model (Hastie and Tibshirani, 1993), in which the linear effect f_k of the (lagged) number of positive tests, PCR_{t-k}^+, varies smoothly over time:

$$E(T_t^+) = \sum_k f_k(t) \cdot \text{Test}_{t-l}^+, \quad t = \text{May 1st}, \ \ldots, \ T. \tag{5}$$

This model is estimated for all different subsets of lags $k \subset \{0, \ldots, 7\}$, as well. From all estimated models, we select the best performing model based on the Bayesian information criterion (BIC).

Assuming a similar relation between the number of (positively) reported tests for individuals reported as cases to the Bavarian/German health authorities as for the number of reported tests per examined persons, we estimate the number of examined persons NT_t by using the estimated parameters of the selected model (4) or (5) and plug-in the total number of reported tests:

$$\widehat{NT}_t = \sum_k \hat{f}_k(t) \cdot \text{Test}_{t-k}, \tag{6}$$

where Test_t denotes the total number of test reported by the laboratories on day t and $\hat{f}_k(t)$ are the estimated associations for the (lagged) number of tests reported by the laboratories. Depending on whether a model of type (4) or (5) is selected, they correspond to time-constant (linear) effects $f_k(t) = \alpha_k$ or (linear) effects that vary smoothly over time.

Plugging the observed numbers of positively examined persons (new cases), T_t^+, and the derived number of relevant examinations, $\widehat{NT}_t$ from (6), into (3), we obtain misclassification adjusted estimates for the number of new COVID-19 cases per time point, $\hat{D}_t^+$.

In the nowcasting, we estimate the number of cases with disease onset on a specific day based on a complex Bayesian hierarchical model using individual-specific data on the reporting and disease onset date (cf. Günther et al. (2020)). To apply the proposed adjustment for misclassification, we first focus on the scenario of no false-negative examinations (sensitivity equals one) and a reduced specificity smaller than one. We then calculate the expected number of false positives reported to the health authorities on a certain day based on the difference of

$$\hat{D}_t^+ = \frac{T_t^+ - NT_t \cdot (1 - \text{spec})}{\text{spec}} \tag{7}$$

and the reported number of new cases T_t^+. Then, we remove this number of randomly selected observations from our data and apply the nowcasting to the reduced data to estimate the *false-positive adjusted* epidemic curve. Based on the estimated epidemic curve it is possible to estimate the effective time-varying reproduction number $R_e(t)$ as also described in Günther et al. (2020).

To take possible false negatives into account, we can rewrite formula (3) by

$$\hat{D}_t^+ = \frac{T_t^+ - NT_t \cdot (1 - \text{spec})}{\text{spec}} \cdot \frac{\text{spec}}{\text{sens} + \text{spec} - 1} \tag{8}$$

and plug in different values for the sensitivity, sens < 1. The first term of (8) corresponds to the false positive adjusted estimate from (7) and the second part is a constant factor independent of the number of examinations per day.

Since the *false-negative* adjustment relies on a constant factor which is independent of the number of tested individuals, the factor $\text{spec}/(\text{sens} + \text{spec} - 1)$ can be directly applied to the result of the false positive adjusted nowcasting procedure, which reduces the computational effort considerably compared to a repeated application of the nowcasting to (upsampled) data. Note, that $\text{spec}/(\text{sens} + \text{spec} - 1) \approx 1/\text{sens}$ for a specificity close to one. The false negative adjustment corresponds therefore roughly to a point-wise up-scaling of the estimated epidemic curve by the reciprocal sensitivity.

Supplemental Note 2 - Results of model for approximating the number of examined persons over time

To approximate the number of examined persons per time interval (week or day) based on the reported number of performed PCR-tests by the laboratories, we estimate a model that establishes a relation between the number of reported cases by the health authorities and the number of reported positive PCR-tests by the laboratories also allowing for lagged associations. We select the best-performing model out of a set of candidate models with time-varying or time-constant linear associations between the number of reported cases and the (lagged) number of positive PCR-tests based on the Bayesian information criterion (BIC) (cf. Supplemental Note 1). With this approach, we aim at capturing potentially time-varying associations between the available information in a flexible way to obtain a plausible approximation for the number of examined individuals over time. Nevertheless, a closer look at the estimated associations might help to understand the structure in the available data better.

For the weekly German data, a varying coefficient model was selected with time-varying effects of the number of reported PCR-tests in the same week and the previous week. For the daily Bavarian data, the selected model considers time-varying effects of the number of positive PCR-tests on the respective day and additionally of the number of positive PCR-tests two and six days before.

At a given time t, the predicted number of reported cases at the health authorities corresponds to a weighted sum of reported positive PCR-tests at the same time and previous time points (i.e., previous week for the German data, and day $t - 2$ and $t - 6$ for the Bavarian data). The weights are given by the estimated effect for time point t. Analogously, our predicted number of examined persons corresponds to the weighted sum of reported PCR-test counts (positive and negative).

For the German model, we find a positive association between the number of reported cases by the health authorities and the number of reported positive PCR-tests in the same week that increases over time towards one. Additionally, we estimated a small positive contribution (around 10%) of the reported positive PCR-tests in the previous week. The rather low weight of the lag-0 PCR-tests in the beginning of the observation period corresponds to the observation that the German laboratories reported more positive PC-tests to the health authorities during May and June than new cases were reported by the RKI (cf. Figure 1 in the manuscript). This indicates that several laboratories reported tests that were already performed at in previous weeks (i.e., during the first wave) in this time period. During the end of the observation period PCR-test reporting by the laboratories was better established and the weekly number of reported positive tests corresponds closely to the number of reported cases by the RKI. In fact, for the German data the misclassification adjustment in August/September is very similar when utilizing the number of reported tests instead of the predicted number of examined persons (not shown).

For the Bavarian model estimated based on daily data, we obtain somewhat similar results. The association of the number of reported positive PCR-tests and the number of reported cases at the health authorities per day increases over time towards around one. Additionally we have a small contribution of the number of positive PCR-tests two and six days before the current day that remains rather constant over time. The lag-0 association (i.e., the association of the number of reported positive tests and the number of reported cases at the same day) shows, however, an additional change in July and August. At this time, there were more positive tests reported by the Bavarian laboratories than cases reported by the Bavarian health authorities and the lag-0 association decreases slightly. This pattern could be related to the German holidays, when many travellers from other federal states were tested when passing through Bavaria. At the end of the observation period the effect of the number of positive tests reported by the laboratories is slightly bigger than one: there are (already) more cases reported by the health authorities than positive tests reported by the laboratories. This might be induced by a faster reporting of person-specific positive test results to the local health authorities (leading to a case registration) compared to the reporting of aggregated (positive and overall) test-numbers by the laboratories. This indicates that the approximation of the number of examined persons based on the flexible model is an important step for a valid misclassification adjustment, especially when analysing daily data in close to real-time.

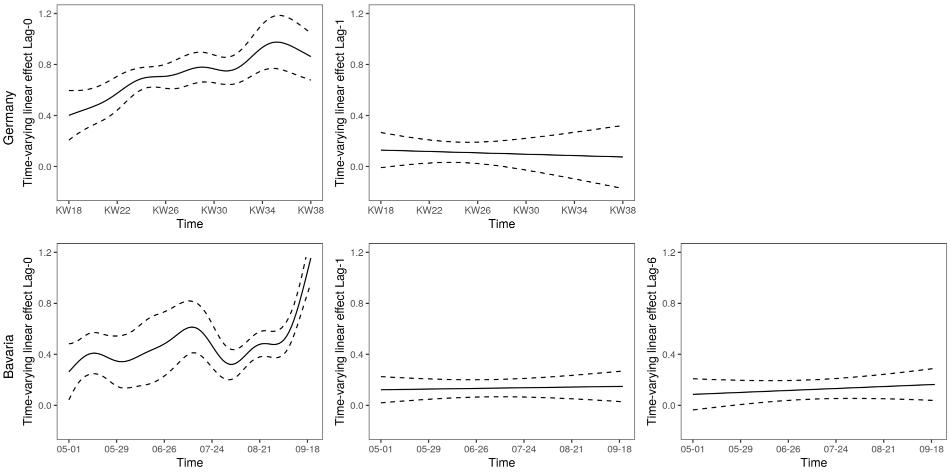
References

Günther, F., Bender, A., Katz, K., Küchenhoff, H., and Höhle, M. (2020). Nowcasting the COVID-19 pandemic in Bavaria. *Biometrical Journal*.

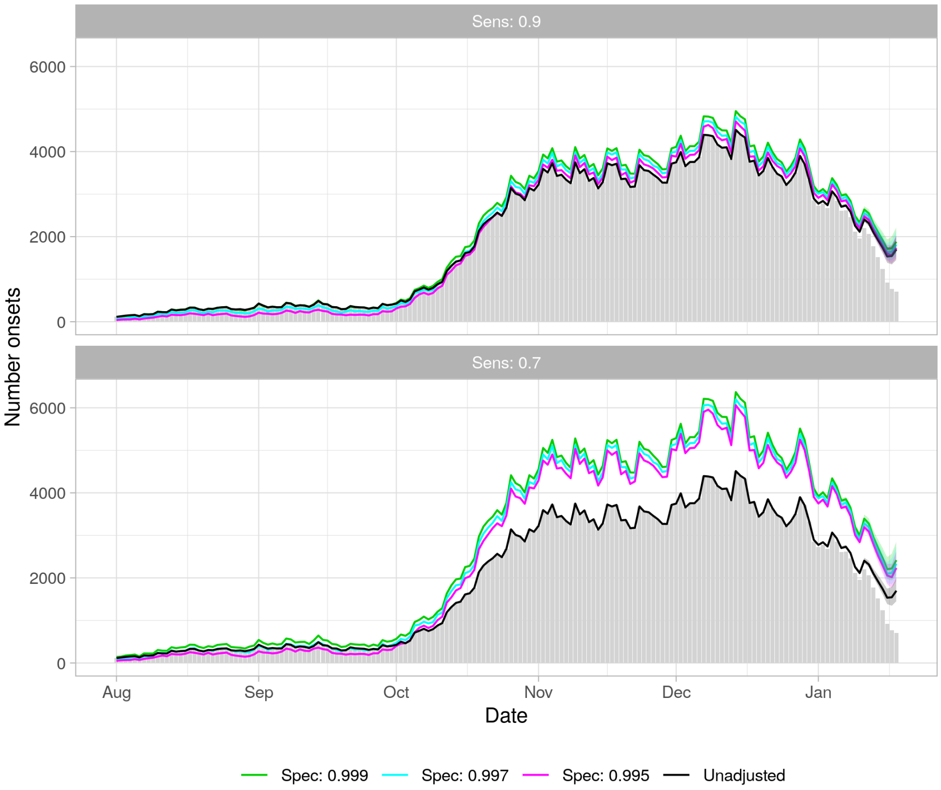
Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.

Rogan, W. and Gladen, B. (1978). Estimating prevalence from results of a screening-test. *American Journal of Epidemiology*, 107(1):71–76.

Supplemental Figures



Supplemental Figure 1: Estimated time-varying association between (lagged) numbers of positive tests reported by the laboratories and the number of reported cases at the health authorities. The model for Germany (top row) is based on weekly information and contains time-varying effects of the number of reported positive tests in the current and the previous week, the model for Bavaria (bottom row)is based on daily information and contains time-varying effects of the number of reported positive tests at the current day, and two and six days before.



Supplemental Figure 2: Estimated misclassification-adjusted epidemic curve for Bavaria based on data available an January, 21, 2020. They grey bars show the number of disease onsets per day reported until January, 21. The black line represents the estimated epidemic curve from nowcasting without adjustment for misclassification. The colored lines show the estimated epidemic curve misclassification adjustment based on different assumptions regarding the sensitivity and specificity of person-specific COVID-19 examination. While the true number of cases might be considerably underestimated due to false-negative results (Supplemental Figure 2), the general structure of the epidemic curve is not changed by taking misclassification into account.

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.

Felix Günther
München, 10.04.2021