In a recent article in the *Journal*, Ogburn et al. (1) highlighted the issues with using predicted values when estimating associations or effects. While the authors cautioned against using predicted values to estimate associations or effects, they noted that predictions can be useful for descriptive purposes. In this work, we highlight the issues with using individual-level predicted values to estimate population-level descriptive parameters.

Epidemiologists are often interested in describing some variable of interest Y in a population. Commonly used descriptive parameters are the mean $E(Y)$ and the proportion of the population in which Y falls above or below a threshold t (i.e., the cumulative distribution function (CDF), $F_Y(t) = P(Y \leq t)$, or its complement). Sometimes we are not able to directly measure Y_i (where i indexes n individuals in the population) but instead have an individual-level prediction of Y_i obtained from a model conditional on covariates $g(Z_i) = \hat{V}_i$, which estimates the conditional expectation $V_i = E(Y|Z_i)$. For simplicity of notation, we suppress the circumflex and subscript i hereafter. When we use such a model, all individuals with $Z = z$ are given the same V even though the true Y values vary, implying that $\mathrm{Var}(V) < \mathrm{Var}(Y)$. If we use V in place of Y to estimate descriptive parameters, descriptive parameters that are linear functions are unbiased but descriptive parameters that are nonlinear functions may be biased (see the Web Appendix, available at https://doi.org/10.1093/aje/kwad020). For example, the mean value, a linear function, is unbiased (i.e., $E(V) = E(E(Y|Z)) = E(Y)$). However, the standard error of the mean, a nonlinear function, is biased. Thus, we will not be able to obtain a valid confidence interval (CI) for the mean using only V. The CDF is also a nonlinear function, so $F_V(t) \neq F_Y(t)$. Intuitively, because $\mathrm{Var}(V) < \mathrm{Var}(Y)$, the 90th percentile, for example, of V will underestimate the 90th percentile of Y.

To ground ideas, say we are interested in describing gestational age at birth in a population. Common metrics are the proportions of infants born preterm and postterm (i.e., CDF and 1 − CDF for gestational age at birth evaluated at 37 weeks and 42 weeks, respectively). Gestational age often cannot be determined directly but is generated prenatally from a model (i.e., dating equation) that converts anatomical size, measured via ultrasound, to gestational age. The equation gives a single, predicted gestational age to all fetuses of the same size, even though there is biological variation in growth. This implies that the variance of the predicted gestational ages is smaller than the true gestational ages. Consequently, the proportions preterm or postterm estimated using predicted age are expected to be biased downward. In this work, we illustrate the issue with estimating descriptive parameters using V instead of Y in a simple simulation using realistic inputs. We also introduce an approach to reduce bias in estimation using V. Throughout, we use estimation of the CDF of gestational age at birth as our example.

ISSUES WITH NAIVELY USING PREDICTIONS

Simulation design

We conducted a simple simulation to illustrate the issues with estimating the CDF of gestational age at birth when gestational age is a predicted value calculated by a dating equation (software code is available on GitHub (2)). To inform this simulation, we used the equations from the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st)—one for early pregnancy using crown-rump length and one for late pregnancy using femur length and head circumference (3, 4). In addition to equations for the predicted gestational age given size, INTERGROWTH-21st provides equations for the standard deviation of gestational age given size. Gestational age dating typically occurs at the first prenatal care visit (gestational age at first visit plus time elapsed equals gestational age at birth), and the standard deviation increases with predicted gestational age (i.e., there is heteroskedasticity; see Web Figure 1). The standard deviation varies from 2.7 days at 8.4 weeks to 6.1 days at 26.0 weeks.

In our simulation, we drew conditional expectations of gestational age at birth given size (V), from a Weibull distribution with parameters (shape 31.619, scale 39.729) set so that the *observed* frequencies of preterm and postterm birth were 10% and 0.03%, respectively, as seen in the US population in 2020 (5). Next, we drew the true gestational age at birth, Y, from a normal distribution centered on V, $Y \sim N(V, \sigma)$. We set σ to 2.7 days in one scenario and 6.1 days in a second scenario, reflecting the smallest and largest standard deviations provided by INTERGROWTH-21st up to 26 weeks. We simulated data for 5,000 cohorts each with $n = 2{,}000$.

We derived the true frequencies of preterm and postterm birth using the sample proportions from the combined population of all 5,000 simulated cohorts (10 million). Then, within each simulated cohort, we estimated the proportions of infants born preterm and postterm using Y and V. Across the 5,000 cohorts, we calculated the average of the proportions preterm and postterm using Y and V and the bias, the empirical standard error, and 95% CI coverage for those proportions (6).

Results

As expected, the mean values of Y and V were equal (39.0 weeks), and the standard deviation of Y was larger than that of V (e.g., when σ = 6.1 days, the standard devi-

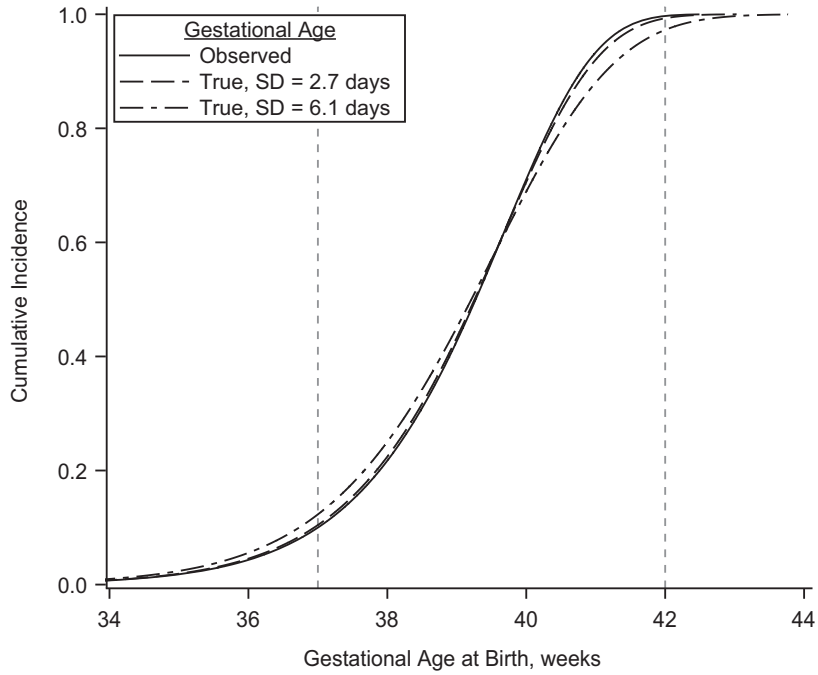

Figure 1. Cumulative incidence of birth in the simulated data using the predicted gestational age at birth ("observed" curve) and true gestational age at birth when the observed distribution mimics estimates from the United States. Dashed vertical lines mark cutoffs for preterm and postterm birth, respectively. SD, standard deviation.

ations were 1.8 and 1.5, respectively). Figure 1 plots the cumulative incidence of births by gestational age. There was some separation in the curves such that the proportions of infants born preterm and postterm are underestimated when using predicted age, V. Table 1 includes estimates of the proportions of infants born preterm and postterm, averaged across the 5,000 simulated cohorts. The observed proportion of infants born preterm, 10%, was biased downward between 0.4 and 2.3 percentage points, depending on the value of σ. The observed proportion of infants born postterm, 0.3%, had a similar absolute amount of downward bias. With σ set at 6.1 days, the true proportion of infants born postterm was more than 9 times the observed proportion. The empirical standard error was smaller when using V than when using Y, and there was poor 95% CI coverage when using V. Coverage was 89% and 8% for preterm birth and 15% and 0% for postterm birth.

VALIDLY USING PREDICTIONS

Multiple imputation can be used to validly estimate descriptive parameters using predictions V in place of Y. We impute values for Y multiple times by repeatedly drawing values for each individual from the distribution $N(V_i, \sigma_i)$. Subsequently, the descriptive parameter is estimated in each imputed data set and then the estimates from each imputed data set are combined using Rubin's rule (7). The approach requires σ_i, the standard deviation of $Y_i \mid Z_i = z$. Results from a simulation show that the approach is unbiased with appropriate standard error estimates when σ_i is known (Web Table 1).

We implemented this approach in an applied example. We estimated the proportions of infants born preterm and postterm using data from the Zambian Preterm Birth Prevention Study (ZAPPS), an observational prospective cohort study of pregnant people recruited at prenatal care initiation in Lusaka, Zambia, in 2015–2017 (8, 9). Our analysis included 1,169 people with a singleton pregnancy whose gestational age was predicted using INTERGROWTH-21st equations. Using the predicted gestational age, V, the estimated proportion of infants born preterm was 13.6% (95% CI: 11.6, 15.6) and the proportion born postterm was 2.0% (95% CI: 1.2, 2.8). To account for the use of predictions, we multiply imputed values for Y where σ_i was obtained from the INTERGROWTH-21st equations for the standard deviation. Using the imputed values, the estimated proportion preterm was 14.6% (95% CI: 12.6, 16.6) and the proportion postterm was 3.0% (95% CI: 2.1, 4.0). Web Figure 2 shows the cumulative incidence of birth using the predicted and imputed gestational age.

DISCUSSION

We have demonstrated that using predicted values of a variable can bias descriptive parameters commonly used in epidemiology. Our illustration specifically showed that naively using predicted values underestimates the proportion of a population in the tails of a distribution. In our simulation,

Table 1. Proportions of Infants Born Preterm and Postterm Using Predicted Gestational Age at Birth, *V*, or True Gestational Age at Birth, *Y* (Average of Estimates From 5,000 Simulated Cohorts of *n* = 2,000)

Parameter	Proportion of Infants Born	Bias		ESE	95% CI Coverage, proportion
		Absolute	Relative[a]		
Preterm					
σ = 2.7 days					
V	0.100	−0.004	−0.04	0.0066	0.89
Y	0.104	0.0	0.0	0.0068	0.95
σ = 6.1 days					
V	0.100	−0.023	−0.19	0.0066	0.08
Y	0.123	0.0	0.0	0.0073	0.95
Postterm					
σ = 2.7 days					
V	0.003	−0.004	−0.57	0.0012	0.15
Y	0.007	0.0	0.0	0.0019	0.93
σ = 6.1 days					
V	0.003	−0.024	−0.89	0.0012	0.00
Y	0.027	0.0	0.0	0.0037	0.94

Abbreviations: CI, confidence interval; ESE, empirical standard error.
[a] Absolute bias/truth.

which mimicked the observed US distribution of gestational age at birth and used realistic inputs for the standard deviation provided by INTERGROWTH-21st, we found that the true proportions of infants born preterm and postterm in the United States may be as high as 12.3% and 2.7% (compared with the 10% and 0.03% observed), respectively. Our simulations also showed that the standard error estimated using predicted values is inappropriately small. The combination of bias and a too-small standard error results in poor CI coverage, meaning that estimated intervals rarely contained the true value. Our illustration focused on the CDF; however, there may also be bias and poor coverage in any descriptive parameter that is a nonlinear function. Additionally, even though descriptive parameters that are linear functions, such as the mean, would not be biased, standard errors of the mean would be biased, resulting in misleading CIs.

We demonstrated an approach for estimating descriptive parameters using predicted values. The approach leverages resampling that treats the prediction as a mismeasured version of *Y*. In our example application, we assumed that $Y \mid V$ was normally distributed; however, assuming a parametric distribution is not necessary. When data used to fit the prediction model (i.e., training data) are available, this approach could be made nonparametric by, for example, resampling from the residuals of the fitted prediction model. Such an approach has been previously proposed (10), where the issue of using predicted values was described as Berkson measurement error (11, 12). The approach also has similarities to multiple overimputation (13).

Importantly, our approach requires a measure of the standard deviation of $Y_i \mid Z_i$. In our example application, we did not have training data, so we used equations for the standard deviation available from INTERGROWTH-21st. We made the strong assumption that these equations were transportable to our study population. If this assumption does not hold (i.e., there are covariates that are related to the standard deviation and the distributions of these covariates differ between our study population and the population in which these equations were fit), then our results may be biased. In general, obtaining measures of the standard deviation is challenging. A validation sample could be used if measuring *Y* is feasible. If no measure for the standard deviation is available, a range of plausible values could be examined to provide information on the direction and potential magnitude of bias.

In general, we must be cautious when using predicted values. It can be easy to overlook that we are using predictions instead of *Y*, such as in the case of gestational age. Predictive approaches such as machine learning, which are increasingly used in epidemiology and public health (14), yield predictions, rather than *Y*. It is important to remember that we cannot naively replace an unmeasured variable with predicted values to estimate our parameters of interest, even descriptive ones.

Acknowledgments

R.K.R. was supported by a training grant from the National Institute of Child Health and Human Development (grant T32 HD52468).

The software code used in this article is available on GitHub (https://github.com/rachael-k-ross/DescriptionWithPredictions).

Conflict of interest: none declared.

References

1. Ogburn EL, Rudolph KE, Morello-Frosch R, et al. A warning about using predicted values from regression models for epidemiologic inquiry. *Am J Epidemiol.* 2021;190(6):1142–1147.
2. Ross RK. DescriptionWithPredictions. https://github.com/rachael-k-ross/DescriptionWithPredictions. Published January 11, 2023. Accessed January 11, 2023.
3. Papageorghiou AT, Kennedy SH, Salomon LJ, et al. International standards for early fetal size and pregnancy dating based on ultrasound measurement of crown-rump length in the first trimester of pregnancy. *Ultrasound Obstet Gynecol.* 2014;44(6):641–648.
4. Papageorghiou AT, Kemp B, Stones W, et al. Ultrasound-based gestational-age estimation in late pregnancy. *Ultrasound Obstet Gynecol.* 2016;48(6):719–726.
5. Osterman MJK, Hamilton BE, Martin JA, et al. Births: final data for 2020. *Natl Vital Stat Rep.* 2022;70(17):1–50.
6. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med.* 2019;38(11):2074–2102.
7. Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J Am Stat Assoc.* 1986;81(394):366–374.
8. Castillo MC, Fuseini NM, Rittenhouse KJ, et al. Zambian Preterm Birth Prevention Study (ZAPPS): cohort characteristics at enrollment. *Gates Open Res.* 2019;2(25):1–18.
9. Price JT, Vwalika B, Rittenhouse KJ, et al. Adverse birth outcomes and their clinical phenotypes in an urban Zambian cohort. *Gates Open Res.* 2020;3(1533):1–26.
10. Baldoni PL, Sotres-Alvarez D, Lumley T, et al. On the use of regression calibration in a complex sampling design with application to the Hispanic Community Health Study/Study of Latinos. *Am J Epidemiol.* 2021;190(7):1366–1376.
11. Keogh RH, Shaw PA, Gustafson P, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: part 1—basic theory and simple methods of adjustment. *Stat Med.* 2020;39(16):2197–2231.
12. Haber G, Sampson J, Graubard B. Bias due to Berkson error: issues when using predicted values in place of observed covariates. *Biostatistics.* 2021;22(4):858–872.
13. Blackwell M, Honaker J, King G. A unified approach to measurement error and missing data: overview and applications. *Sociol Methods Res.* 2017;46(3):303–341.
14. Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. *Annu Rev Public Health.* 2020;41(1):21–36.

Rachael K. Ross[1], Alexander P. Keil[1], Stephen R. Cole[1], Jessie K. Edwards[1], and Jeffrey S. A. Stringer[1,2]
(e-mail: rkross@unc.edu)

[1] *Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States*

[2] *Department of Obstetrics and Gynecology, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States*

https://doi.org/10.1093/aje/kwad020; Advance Access publication: January 27, 2023