

Is Disease Intensity a Good Surrogate for Yield Loss or Toxin Contamination? A Case Study with Fusarium Head Blight of Wheat

Laurence V. Madden† and Pierce A. Paul

Department of Plant Pathology, Ohio State University, Wooster, OH 44691
Accepted for publication 3 May 2020.

ABSTRACT

Sometimes plant pathologists assess disease intensity when they are primarily interested in other response variables, such as yield loss or toxin concentration in harvested products. In these situations, disease intensity potentially could be considered a surrogate of yield or toxin. A surrogate is a variable which can be used instead of the variable of interest in the evaluation of experimental treatments or in making predictions. Surrogates can be measured earlier, more conveniently, or more cheaply than the variable of primary interest, but the reliability or validity of the surrogate must be shown. We demonstrate ways of quantifying two facets of surrogacy by using a protocol originally developed by Buysse and colleagues for medical research. Coefficient-of-determination type statistics can be used to conveniently assess the strength of surrogacy on a unitless scale. As a case study, we evaluated whether field severity of

Fusarium head blight (i.e., FHB index) can be used as a surrogate for yield loss and deoxynivalenol (DON) toxin concentration in harvested wheat grain. Bivariate mixed models and corresponding approximations were fitted to data from 82 uniform fungicide trials conducted from 2008 to 2013. Individual-level surrogacy—for predicting the variable of interest (yield or DON) from the surrogate (index) in plots with the same treatment—was very low. Trial-level surrogacy—for predicting the effect of treatment (e.g., mean difference) for the variable of interest based on the effect of the treatment on the surrogate (index)—was moderate for yield, and only low for DON. Challenges in using disease severity as a surrogate for yield and toxin are discussed.

Keywords: analytical and theoretical plant pathology

It is routine to measure disease intensity, such as incidence, severity, or lesion counts, in experiments or observational studies (Bock et al. 2010; Nutter 2002). Such measurements are used to better understand plant–microbe interactions (from the molecular to the population and landscape scales) and evaluate disease control tactics and strategies. Epidemiologists often measure intensity to characterize and predict the temporal, spatial, or spatiotemporal dynamics of disease in populations (Madden et al. 2007; Parnell et al. 2015, 2017). Sometimes, researchers measure disease intensity when they are actually more interested in another response variable that may be dependent on or directly associated with disease, such as yield loss or toxin contamination in harvested products (Madden and Paul 2009; Paul et al. 2005b). For instance, after determining a functional relationship between disease intensity and other response variables of interest, it may be possible to predict yield or toxin levels based on disease measurements, or make management, grain handling, or marketing decisions.

A case in point is Fusarium head blight (FHB), caused by *Fusarium graminearum* and other related fungal species, which is one of the most economically important diseases of wheat in the world (McMullen et al. 2012; Savary et al. 2017a). Reductions in grain yield occur as a result of death or damage to spikes that lead to smaller, lighter, and fewer grains produced per unit area.

†Corresponding author: L. V. Madden; madden.1@osu.edu

Funding: This material is based upon work supported by the U.S. Department of Agriculture, under Agreement number 59-0206-8-187. This is a cooperative project with the U.S. Wheat & Barley Scab Initiative (USWBSI). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the U.S. Department of Agriculture.

*The e-Xtra logo stands for “electronic extra” and indicates that two supplementary files are published online.

The authors declare no conflict of interest.

Importantly, there is often the accumulation of fungal-produced mammalian toxins such as deoxynivalenol (DON) in *F. graminearum*-infected grain. The concentration of DON determines, in part, the price received for grain (Salgado et al. 2014). Considerable efforts are made in many regions of the world to improve ways of reducing yield losses and DON concentrations using host plant resistance, fungicide applications, and cultural practices (Cowger et al. 2016; McMullen et al. 2012; Paul et al. 2019; Wegulo et al. 2011; Willyerd et al. 2012; Yuen and Schoneweis 2007).

It is common practice to measure the severity of disease symptoms on wheat spikes on a percentage scale (i.e., the average percent spike area with symptoms, based on a sample of spikes from a plot or field), which is generally labeled “FHB index” by FHB researchers (Paul et al. 2005a), in field, greenhouse, and growth chamber investigations. Means of multiple spikes per sampling or experimental unit are then calculated. Yield (e.g., bushels/acre or tons/hectare) and DON concentration (ppm) are also often measured, the latter based on chromatographic or immunological methods (Mirocha et al. 1998; Sinha and Savard 1996). Given that FHB can affect several yield components (Cowger et al. 2016; Salgado et al. 2015, 2017) and that DON is a virulence factor for FHB development (Möbius and Hertweck 2009), it is therefore natural to ask the following: can one use just FHB index to make predictions about the effects of treatments (e.g., fungicide application, cultivar selection) on DON or yield? In other words, can FHB index be used as a surrogate response variable for DON or yield? There are many practical aspects to this question. For instance, winter wheat growers, crop advisors and breeders in the northern US need to make decisions about which cultivars or breeding lines to plant in the autumn, less than 2 months after harvest, usually before DON results are available from test labs. Moreover, direct DON testing can be time-consuming and costly (Man et al. 2017), direct yield determination can be labor intensive and involve expensive equipment, and both responses can be affected by harvest delays due to adverse late-season weather conditions (Cowger et al. 2009; Czarnecki and Evans 1986; Edwards et al. 2018; Farrer et al. 2006).

Results from several meta-analyses show that there is general consistency between mean treatment effects across studies for FHB index, yield, and DON (Paul et al. 2008, 2010, 2018a, b; Wilyerd et al. 2012) for a range of fungicide treatments and fungicide treatment \times cultivar resistance management combinations. For instance, many treatments that produce the highest percent control of FHB index across the population of studies also produce the highest percent control of DON contamination of grain and the highest mean or percent increase in yield. But the estimated among-study variability for treatment effects is also high, overall, indicating a wide range of results could occur for individual studies in terms of the different response variables. No formal assessment of the relationships between the responses in experimental units was conducted in these meta-analyses (Paul et al. 2008, 2010, 2018a, b). In another investigation, this one involving 126 studies, Paul et al. (2005b, 2006) found a significant relationship between DON in harvested grain and FHB index (measured at or around Feekes growth stage 11.1 [Zadoks stage 75]). The slope of the linear relationship was 0.22 ppm per unit of FHB index (SE = 0.017), with a 95% confidence interval of 0.186 to 0.254. However, the variability was high, and the 95% prediction interval (Madden and Paul 2011 [pages 23-24]) for the slope was -0.128 to 0.568 (L. V. Madden, unpublished data), indicating that a positive, negative, or no relation between index and DON in an individual study is probable. Likewise, Madden and Paul (2009) and Salgado et al. (2015) found high variability across studies in the relationship between FHB index and yield. For instance, although there was a highly significantly linear relationship overall between index and yield in Madden and Paul (2009) (slope = -0.038 metric tons per hectare per unit of FHB index; SE = 0.003), a meta-analysis showed that the 95% confidence interval for the slope was -0.047 to -0.034 , but that the 95% prediction interval went from -0.071 to -0.010 (L. V. Madden, unpublished data), with the upper limit almost reaching 0.

The objective of this paper is to evaluate the extent to which FHB index can be used as a surrogate for wheat grain yield and DON contamination of grain in experiments involving fungicide applications. To do this, we present a rigorous framework for assessing different facets of surrogacy, building on the concepts and statistical methodology originally developed by Buyse and Molenberghs (1998) and Buyse et al. (2000) for continuous random variables. After first utilizing a hypothetical example to demonstrate the different aspects of surrogacy, we fit an expanded version of the model in Buyse et al. (2000) to a dataset consisting of 82 uniform fungicide trials (UFTs) supported by the U.S. Wheat and Barley Scab Initiative (USWBSI).

MATERIALS AND METHODS

Surrogacy. The concept of, and analytical methods for, assessing surrogacy were developed primarily to address medical research problems, starting with the work of Prentice (1989) and Freedman et al. (1992). Over the last quarter century, surrogate analysis has grown into an extensive field of research (Alonso et al. 2017; Bujkiewicz et al. 2019; Buyse et al. 2016; Molenberghs et al. 2008), although the ideas have not received much formal attention in plant pathology. As defined by Buyse et al. (2000), “A surrogate endpoint [response variable] is one which can be used in lieu of the endpoint [response variable] of primary interest in the evaluation of experimental treatments or other interventions.” It should be noted that the term endpoint is commonly used in medical statistics for response variable. The goal of surrogate analysis is to “look for surrogate endpoints...that can be measured earlier, more frequently, more conveniently, or more cheaply than the true... endpoint of interest” (Buyse 2017). In medical clinical trials, the true endpoint could be long-term survival time of patients after treatment for cancer (which can be measured in years), and a surrogate could be a biochemical blood assay result that could be

measured relatively soon after treatment. For FHB, disease index is certainly a *potentially* good surrogate because it can be measured earlier, more conveniently, or more cheaply than DON or yield.

There are two levels to surrogacy that can be considered, individual-level and trial-level (Alonso Abed and Van der Elst 2017b; Buyse et al. 2016). With individual-level surrogacy, one predicts the response variable of interest in individual experiment units (e.g., plants, plots, fields, depending on the research context) based on the surrogate. This is done possibly to prescribe a treatment, although with FHB, disease index measurements are made too late to make fungicide-application decisions. In particular, the latest time to apply a fungicide and still achieve acceptable levels of FHB suppression is 4 to 6 days after anthesis, which is before symptom expression (Paul et al. 2018b). However, FHB index can potentially be used as a guide for making grain marketing decisions for individual fields (e.g., harvest for grain or animal feed, harvest early to avoid late-season increase in DON, use grain-cleaning methods to reduce contaminated seeds, adjust the harvester to reduce number of scabby grains, etc.) (Cowger et al. 2016; Delwiche et al. 2005; Salgado et al. 2014; Tkachuk et al. 1991). For other pathosystems, disease severity (or a measure of inoculum density) is a potentially useful predictor variable for making control-intervention decisions (Carisse et al. 2008; Hughes 2017). Even without the possibility of a control intervention, a good individual-level surrogate can be used to predict the course of disease progression or predict the level of toxin or yield loss in plots or fields. With individual-level surrogacy, there is an individual-level association or correlation between the surrogate and true response variables, *after* accounting for the effect of any treatment.

With trial-level surrogacy, one predicts the *effect* of a treatment (e.g., mean difference between the control and treatment) on the response variable of interest based on the effect of a treatment on the surrogate. The concept applies to experiments where treatments (e.g., fungicides, cultivars, cultural practices) are being tested in randomized trials. With FHB, disease index would be a good surrogate if the mean difference in index between the control and the fungicide treatment of interest in a study can be used to (precisely) predict the mean difference in yield between the control and the fungicide treatment in that study. As explained below, assessment of trial-level surrogacy requires results from multiple trials (Buyse et al. 2000).

Depending on the objectives of the research, either trial-level, individual-level, or both levels of surrogacy may be desirable (Alonso et al. 2017; Buyse et al. 2016).

Hypothetical example. The concept of trial- and individual-level surrogacy is demonstrated in Figure 1 for three scenarios with two treatments (red for control and blue for fungicide treatment) and six replicates per treatment in a completely randomized design. We use S for the surrogate random variable and T for the (true) random variable of primary interest. Figure 1A exemplifies the situation with trial-level surrogacy but no individual-level surrogacy. The control has high levels for the surrogate (abscissa) and true response (ordinate), and the treatment has low levels for both random variables, indicating substantial treatment effects for the surrogate and the true response. Here, there is no overlap of the individual replicate values for the two treatments, but that is not a requirement for trial-level surrogacy. The key point is that the means are substantially different, with a mean difference between the control and the treatment of $\Delta_S = 19.2$ for the surrogate and $\Delta_T = 5.7$ for the true response (Table 1).

In Figure 1A there is no individual-level surrogacy because there is no relation between surrogate and true response values for replicates receiving the same treatment. The correlation coefficient *adjusted* for treatment (as explained below) is $r = -0.14$. It is important to note that the correlation has to be calculated after treatment adjustment; a simple product-moment correlation of all the points in Figure 1A (with no adjustment) would give a misleading large positive value.

Figure 1B exemplifies the situation with no trial-level surrogacy but substantial individual-level surrogacy. In particular, although treatment has a substantial effect on the surrogate, with a difference in treatment means of $\Delta_S = 19.2$, treatment has virtually no effect on the true response variable (mean difference of only $\Delta_T = 0.1$) (Table 1). However, there is a strong positive relation between the surrogate and true response across replicates receiving the same treatment. The correlation adjusted for treatment is $r = 0.81$ (without adjustment for treatment, the correlation would be about 0). Figure 1B demonstrates that just because there is a strong relationship between measurements of two random variables across a collection of experimental units under the same conditions, imposition of a treatment (e.g., fungicide application) that affects

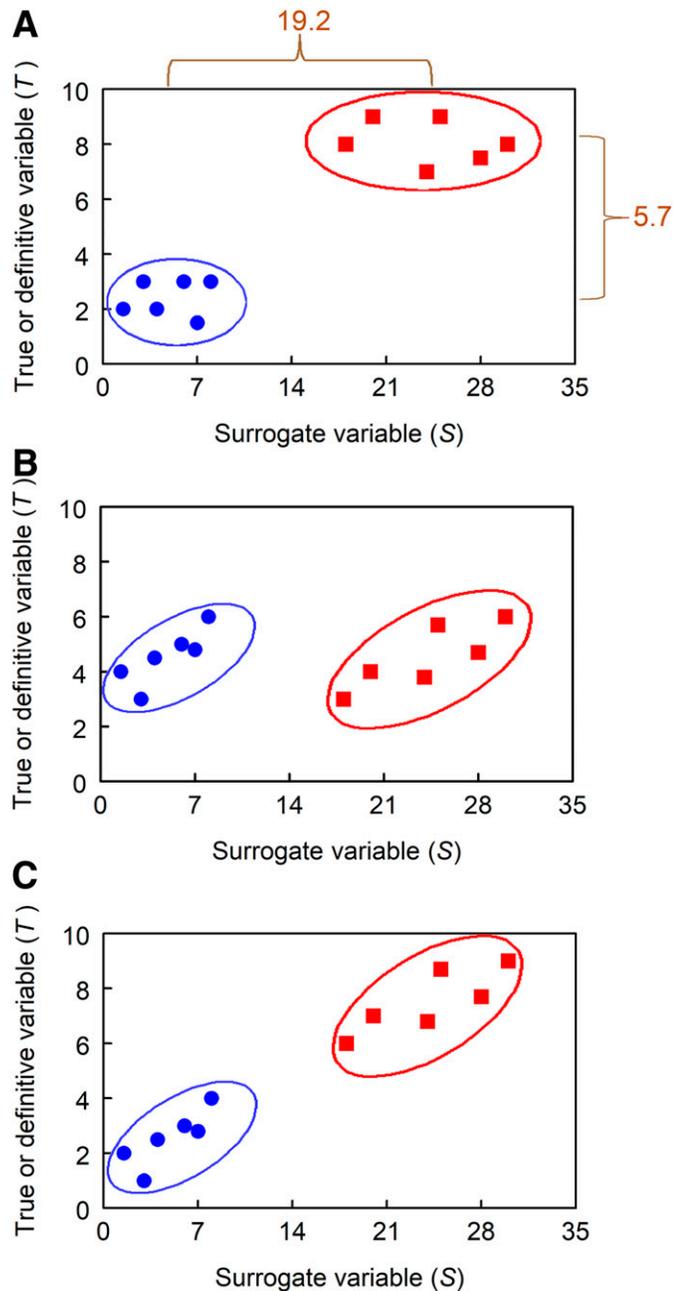


Fig. 1. Example scenarios to demonstrate trial-level and individual-level surrogacy for a single trial. Red solid squares indicate the control and blue solid circles indicate the treatment. **A.** Example of trial-level but no individual-level surrogacy. **B.** Example of individual-level surrogacy but no trial-level surrogacy. **C.** Example of both trial- and individual-level surrogacy. Results are given in Table 1. In A, mean difference between control and treatment (i.e., treatment effect) are given for the surrogate (S) and the true (definitive) response (T).

the surrogate does not automatically affect the true response (or vice versa). Failure to appreciate this fact has caused failures in selecting suitable trial-level surrogates in medical studies (Alonso Abed and Van der Elst 2017a, b).

Figure 1C exemplifies the situation with both trial-level and individual-level surrogacy. There is a substantial difference in treatment means for both the surrogate and true responses (Table 1) and the correlation coefficient adjusted for treatment is $r = 0.81$, showing the association between the two response variables for experimental units with the same treatment.

As stated by Alonso Abed and Van der Elst (2017b), “...it has become clear that the single-trial setting is too restrictive for the evaluation of surrogate endpoints.” Figure 1A and C are single-trial manifestations of trial-level surrogacy, but these results are not sufficient to predict the treatment effect (e.g., mean difference) for the true response (Δ_T) based on the treatment effect for the surrogate (Δ_S). This can be seen by plotting Δ_T versus Δ_S for the example in Figure 1A. As seen in Figure 2A, the graph consists of only one data point. One would need to use the straight line through the origin and intersecting the (Δ_S, Δ_T) point to predict the effect of treatment on the true response for any magnitude of treatment effect on the surrogate. For instance, if in another study, it was found that $\Delta_S = 6$, then one would predict that $\Delta_T = 1.8$. This is a strong and very restrictive assumption. Even if the (Δ_S, Δ_T) data point is known without error from an actual trial (unlikely, of course), there is an indefinite number of lines that can go through this point. Figure 2B demonstrates a selection of the possible lines. Plus, the true relationship does not necessarily go through this single (Δ_S, Δ_T) point. So, it is impossible to tell from a single trial what is the relationship between treatment effects for the true and surrogate variables.

One must analyze data from multiple trials to determine the relationship between Δ_T and Δ_S (Molenberghs et al. 2008). A range of bivariate and univariate meta-analytical methods have been developed to tackle the problem, based on mixed model or information theoretic principles (Alonso and Molenberghs 2007; Alonso et al. 2006; Bujkiewicz et al. 2019; Buyse et al. 2016). In order to characterize both individual- and trial-level surrogacy, one requires the replicate (experimental unit) data from each trial, not just the treatment means in each trial. The new reference book by Alonso et al. (2017) reviews many of the methods. We use an expansion of the bivariate mixed models originally developed by Buyse et al. (2000) and Tibaldi et al. (2003) for two continuous variables to evaluate FHB disease index as a surrogate for DON toxin or yield. The standard trial-level statistical methods for surrogacy focus on a single treatment (or combined treatment) versus the control, and we take this approach here to utilize previously developed statistical theory.

Data: FHB case study. There have been over 300 UFTs carried out in the United States between 1995 and 2013, conducted since 1998 under the auspices of USWBSI. Several meta-analyses have been published based on this expanding database (Madden et al. 2016; Paul et al. 2007, 2008, 2010, 2018a, b) for the purpose of determining the efficacy of different fungicide treatments for controlling FHB disease index, DON in grain, and yield loss. The trials up through 2007 primarily focused on different demethylation inhibitor (DMI) fungicides applied at wheat anthesis (Feekes 10.5.1; Zadoks stage 60); later trials focused on other fungicide chemistries (e.g., quinone outside inhibitor [QoI]), combinations of fungicides, frequency of applications, and different application timings. Overall, three different DMI fungicide products applied at or shortly after anthesis provided the highest, most consistent, and very similar levels of control across trials (Paul et al. 2018a):

- (i) metconazole (trade name Caramba; BASF Corp., Research Triangle Park, NC)
- (ii) prothioconazole (Proline; Bayer CropScience, Research Triangle Park, NC)
- (iii) prothioconazole plus tebuconazole mixture (Prosaro; Bayer CropScience)

These three treatments plus the control were used here in the surrogacy analysis, based on the premise that if a given variable is a useful trial-level surrogate then this will be manifested with the best treatment(s).

The published meta-analyses to-date from the UFT database have used the treatment means and corresponding sampling (i.e., within-trial) variances for each trial (Paul et al. 2007, 2008, 2010, 2018a, b). This is sufficient for univariate and network (multitreatment) meta-analysis (Madden et al. 2016; Whitehead 2002). In fact, the UFT database up to 2007 consists only of means and sampling variances (i.e., within-trial weights) for each trial. However, to assess individual-level surrogacy, individual replicate observations within trials are required with measurements of the surrogate and true response variables (Buyse et al. 2016). So, for the surrogacy analysis, we used the 82 trials for the 6 years from 2008 through 2013 with the available replicate data for both DON and FHB index. Randomized block designs were used in all trials. Prothioconazole plus tebuconazole (as a single premixed product of the two active ingredients; Prosaro) was used in all trials, metconazole in 81 of the 82 trials, and prothioconazole as a single active-ingredient product (Proline) in seven of the trials.

The experimental procedures, including plot size, sample size for disease index and DON, fungicide dose and timing for each of the three fungicides, and harvesting methods, generally followed a standard protocol in all trials, since they were performed using the guidelines specified by the USWBSI. The fungicides were applied at the labeled dose for the treatments considered here. Methods were all described in detail in Paul et al. (2018a, b), and are not repeated here. Critical aspects of the methods are brought up in the Discussion. These previous investigations determined the effects of many different treatments on control of FHB index, DON, and increase in yield, and dealt with effects of moderator variables (trial-level properties) on treatment effects. These past investigations dealt only with each response variable (endpoint) individually. The present investigation focuses exclusively on surrogacy, which entails quantifying the effects of treatments on pairs of endpoints.

For each trial in the current analysis, an indicator variable Z was used to represent the treatment-group designation of each experimental unit (plot). Z was coded as -1 for any control plot, and as $+1$ for any plot of the three DMI treatments listed above. Thus, these three treatments were considered as multiple replicates of a single fungicide ‘treatment’ in each trial. We use all three treatments together here in order to maximize the amount of data in the surrogacy analysis, therefore increasing statistical power and precision (Brown and Prescott 2015). We thus follow the meta-analytical approach of Caldwell and Welton (2016) and Melendez-Torres et al. (2015) of lumping treatments into “clinically meaningful units”.

Analytical methods. Let i be the index for trial ($i = 1, \dots, N$, with N representing the number of trials), j be the index for the block within the i th trial, k be the index for observation (e.g., experimental unit; plot) within block within trial, n_i be the number of blocks within the i th trial, ($j = 1, \dots, n_j$), and n_{ij} be the number of observations in the j th block of the i th trial ($k = 1, \dots, n_{ij}$). S_{ijk} is the surrogate response variable (endpoint) for the k th observation in the j th block of the i th trial, and T_{ijk} is the true (or definitive) response variable (endpoint) for the k th observation in the j th block of the i th trial. Z_{ijk} is the treatment indicator variable ($-1, +1$) for the k th

observation in the j th block of the i th trial. Although we use the designation $Z = -1$ for the control and $Z = 1$ for the treatment, other coding could be used (with corresponding modifications in the determination of treatment effects in the models below). n_{ij} could be as few as two (a control and one of the three DMI fungicides listed above), or as high as four here (control and all three fungicide treatments present in a trial). The “ ijk ” index combination uniquely identifies each experimental unit in the i th trial.

Models. In the case study, S is FHB disease index at Feekes growth stage 11.1 (Zadoks scale 75) and T is either yield (T/ha) or DON (ppm) in harvested grain at the end of the season. In other pathosystems, S could be area under the disease progress curve or measure of disease severity or incidence at particular times during the epidemics.

Expanding on the work of Buyse et al. (2000), a bivariate model for the two response variables is

$$\begin{aligned} S_{ijk} &= \eta_{Si} + B_{Sij} + \alpha_i Z_{ijk} + e_{Sijk} \\ T_{ijk} &= \eta_{Ti} + B_{Tij} + \beta_i Z_{ijk} + e_{Tijk} \end{aligned} \quad (1)$$

where η_{Si} represents the expected value of the surrogate for the i th trial, B_{Sij} is the effect of the j th block in the i th trial on the surrogate, α_i is the parameter for the effect of treatment on the surrogate in trial i , e_{Sijk} is the residual for the surrogate. Likewise, η_{Ti} represents the expected value of the true response for the i th trial, B_{Tij} is the effect of the j th block in the i th trial on the true response, β_i is the parameter for the effect of treatment on the true response in trial i , and e_{Tijk} is the residual for the surrogate. The original and subsequent surrogacy models of Buyse et al. (2000, 2016), and of most other authors (Alonso et al. 2017), are for a single binary categorization of treatment without any other within-study or study-level covariates; our analysis continues with this approach. One could consider the block effect as either fixed or random, but we consider it a random effect here. Distributional properties of the residuals and block effects are given below. In equation 1, the η parameters (η_{Si} , η_{Ti}) represent the overall level of the response variables across the control and treatment, reflecting among other things, the favorability of the environment for disease development or toxin production, or grain production, as well as local cropping practices, wheat cultivar, wheat market class, and so on.

There are four relevant expected values (means) for the i th trial based on the application of equation 1, corresponding to the control and treatment means for the surrogate ($\mu_{Si(Con)}$, $\mu_{Si(Treat)}$) and for the true response variable ($\mu_{Ti(Con)}$, $\mu_{Ti(Treat)}$) (Buyse et al. 2000; Molenberghs et al. 2008). With the $-1/1$ coding for Z in equation 1, these are given by

$$\begin{aligned} \mu_{Si(Con)} &= \eta_{Si} - \alpha_i \\ \mu_{Si(Treat)} &= \eta_{Si} + \alpha_i \\ \mu_{Ti(Con)} &= \eta_{Ti} - \beta_i \\ \mu_{Ti(Treat)} &= \eta_{Ti} + \beta_i \end{aligned} \quad (2)$$

when the block effect is random with an expected value of 0 (the standard assumption). The treatment effects (i.e., differences in

TABLE 1. Results for the hypothetical examples in Figure 1

Scenario ^a	Response variable	Treatment mean	Control mean	Mean difference ^b	Adjusted correlation
A	Surrogate	4.9	24.1	19.2	-0.14
	True	2.4	8.1	5.7	
B	Surrogate	4.9	24.1	19.2	0.81
	True	4.6	4.5	0.1	
C	Surrogate	4.9	24.1	19.2	0.81
	True	2.6	7.5	4.9	

^a Each scenario represents one trial.

^b Treatment effect for the trial.

expected values between the treatment and control) are therefore given by

$$\begin{aligned}\Delta_{Si} &= \mu_{Si(Treat)} - \mu_{Si(Con)} = 2\alpha_i \\ \Delta_{Ti} &= \mu_{Ti(Treat)} - \mu_{Ti(Con)} = 2\beta_i\end{aligned}\quad (3)$$

With three different fungicides in a trial, this treatment effect is the difference between the expected value for the combined treatment group and the control. Other coding for Z would produce alternative versions of equation 3. For instance, reversing the sign of Z (so that -1 is for the treatment) would reverse the sign of the right side of the equations. Use of $-1/2$ and $1/2$ would result in the treatment effects being simply α_i and β_i . The parameterization chosen here allows some straight-forward postmodel-fitting calculations of surrogacy as used in Alonso Abed and Van der Elst (2017b).

We can assume that the trials being analyzed are a random sample from a larger population of trials, a standard assumption in meta-analysis (Madden and Paul 2011; Whitehead 2002), and re-express equation 1 as a random-coefficient mixed model. Such a model accounts for the hierarchy of the design (blocks within trials, treatments within blocks), correlations of observations within trials (because they share a random effect level), and measurement (i.e., estimation) error (Tibaldi et al. 2003). Model expansion is done by redefining the four nonblock parameters per trial (η_{Si} , η_{Ti} ,

α_i , β_i) as the sum of a (global) constant or parameter (η_S , η_T , α , β) plus a random trial effect (m_{Si} , m_{Ti} , a_i , b_i):

$$\begin{aligned}\eta_{Si} &= \eta_S + m_{Si} \\ \eta_{Ti} &= \eta_T + m_{Ti} \\ \alpha_i &= \alpha + a_i \\ \beta_i &= \beta + b_i\end{aligned}\quad (4)$$

The distributional properties are described below. Combining equations 1 and 4, one obtains:

$$\begin{aligned}S_{ijk} &= (\eta_S + m_{Si}) + B_{Sij} + (\alpha + a_i) Z_{ijk} + e_{Sijk} \\ T_{ijk} &= (\eta_T + m_{Ti}) + B_{Tij} + (\beta + b_i) Z_{ijk} + e_{Tijk}\end{aligned}\quad (5)$$

Based on equation 5, expected values for each treatment and response variable combination are given by

$$\begin{aligned}\mu_{Si(Con)} &= (\eta_S + m_{Si}) - (\alpha + a_i) \\ \mu_{Si(Treat)} &= (\eta_S + m_{Si}) + (\alpha + a_i) \\ \mu_{Ti(Con)} &= (\eta_T + m_{Ti}) - (\beta + b_i) \\ \mu_{Ti(Treat)} &= (\eta_T + m_{Ti}) + (\beta + b_i)\end{aligned}\quad (6)$$

Treatment effects per trial are then given by

$$\begin{aligned}\Delta_{Si} &= \mu_{Si(Treat)} - \mu_{Si(Con)} = 2\alpha_i = 2(\alpha + a_i) \\ \Delta_{Ti} &= \mu_{Ti(Treat)} - \mu_{Ti(Con)} = 2\beta_i = 2(\beta + b_i)\end{aligned}\quad (7)$$

Because the expressions involve random effects, equations 6 and 7 defines BLUPs rather than the fixed-effect parameters of equations 2 and 3 (Littell et al. 2006); these can be considered trial-specific predictions. The expected values of the treatment effects *across* the population of trials (with $-1/1$ coding) are given by

$$\begin{aligned}\Delta_S &= 2\alpha \\ \Delta_T &= 2\beta\end{aligned}\quad (8)$$

which are fixed-effect parameters.

Note that equations 1 and 5 are generalizations of the models proposed by Buyse et al. (2000) because they did not deal with blocking within trials. For randomized trials without blocks, j is the index for observation (experimental unit such as a plot) within the trial (rather than the index for block), and n_i refer to number of observations per trial. The k subscript is then dropped.

Distributions. Equation 5 (or equation 1) can be fitted as univariate models, that is, with a separate fit for S and T , or as bivariate models, that is, with a combined model fit for S and T . The latter takes into account the correlations of the two response variables and the random effects in the model (Alonso Abed and Van der Elst 2017b; Tibaldi et al. 2003). The bivariate approach provides one valuable method for quantifying individual- and trial-level surrogacy (Buyse et al. 2016; Molenberghs et al. 2008). For the bivariate fit of equation 5, we need to define the distributions of the random variables and random effects on the right side. The residuals are assumed to have a bivariate normal distribution with nonzero covariance. Defining \mathbf{e}_{ijk} as the matrix of the two residuals for S and T , $\mathbf{e}_{ijk} = (e_{Sijk}, e_{Tijk})^T$, where the T superscript means matrix transpose, the residual distribution is given by $\mathbf{e}_{ijk} \sim N(\mathbf{0}, \Sigma)$, where $\mathbf{0}$ is the vector of expected values (0) and Σ is the residual variance-covariance matrix written as

$$\Sigma = \begin{pmatrix} \sigma_S^2 & \sigma_{ST} \\ \sigma_{ST} & \sigma_T^2 \end{pmatrix}\quad (9)$$

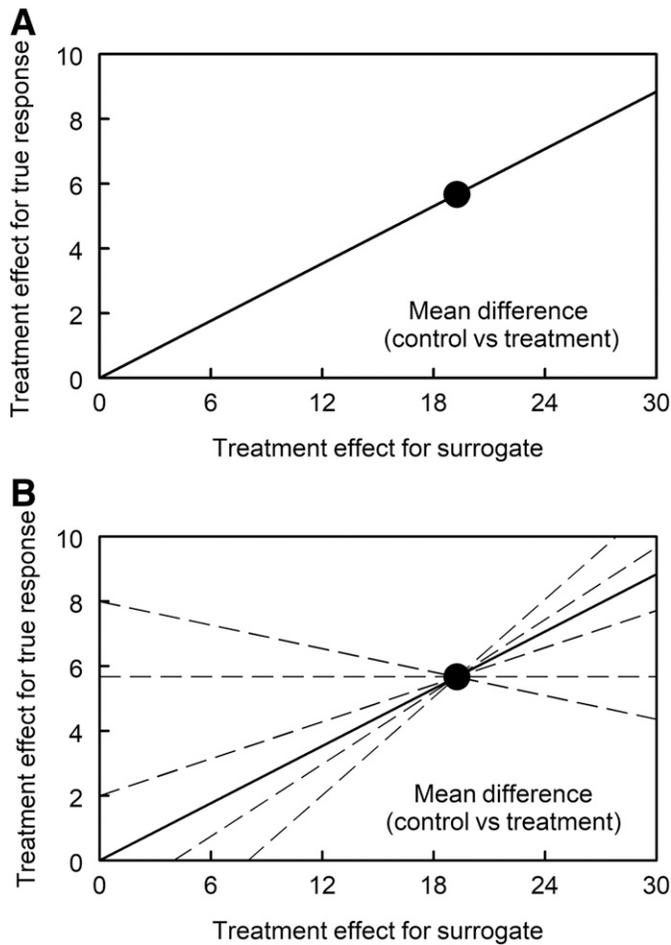


Fig. 2. Conceptualization of trial-level surrogacy. **A**, Treatment effect (e.g., mean difference) for the true (definitive) response variable versus the treatment effect for the surrogate response variable. Large solid circle: result for the single trial in scenario A in Figure 1 ($\Delta_S = 19.2$, $\Delta_T = 5.7$; see means in Table 1). The line connects this point to the origin. **B**, Same single-trial treatment-effect observation as in part A, but with several lines (out of an infinite number of possible lines) intersecting the treatment-effect point. Without multiple trials, one does not know the relationship between treatment effects for the true and response variables.

Equation 9 indicates a separate within-trial variance for S and T (σ_S^2, σ_T^2) and a covariance that can be positive, negative, or 0 (σ_{ST}). Note that the variance-covariance matrices (here and below) are symmetrical and only the upper diagonal elements are usually shown for ease of reading.

For the block effects, we assume a bivariate distribution. Defining the matrix $\mathbf{B}_{ij} = (B_{Sij}, B_{Tij})^T$ for the j th block effect of the i th trial for S and T , we write its distribution as $\mathbf{B}_{ij} \sim N(0, \Theta)$, where the variance-covariance matrix is written as

$$\Theta = \begin{pmatrix} \varsigma_S^2 & \varsigma_{ST} \\ & \varsigma_T^2 \end{pmatrix} \quad (10)$$

Equation 10 allows for different block-effect variances for the two response variables. The appropriateness of a block-effect term in equation 5 and the structure of the variance-covariance matrix in equation 10 (i.e., whether it was possible to allow for a nonzero value of the covariance ς_{ST}) was evaluated under model diagnostics (see below).

The four random effects of equation 4 that are components of equation 5 are assumed to follow a multivariate normal distribution. Defining the matrix $\mathbf{\Omega}_i = (m_{Si}, m_{Ti}, a_i, b_i)^T$ for these four random effects of the i th trial, we represent the distribution as $\mathbf{\Omega}_i \sim N(0, \mathbf{D})$, and write the variance-covariance matrix as

$$\mathbf{D} = \begin{pmatrix} d_S^2 & d_{ST} & d_{Sa} & d_{Sb} \\ & d_T^2 & d_{Ta} & d_{Tb} \\ & & d_a^2 & d_{ab} \\ & & & d_b^2 \end{pmatrix} \quad (11)$$

In equation 11, the diagonal d terms are the variances *among* trials for each random effect and the off-diagonal d elements are the covariances. The elements of the \mathbf{D} matrix characterize the variability in the treatment effects and overall level of the responses among trials (as further explained below). We assume that $\mathbf{\Omega}_i, \mathbf{B}_{ij}$, and \mathbf{e}_{ijk} are independent.

Surrogacy determination. Based on equation 5 as a bivariate model, with equations 9-11 for the variances and covariances, one can define coefficient-of-determination-type (R^2) statistics for individual- and trial-level surrogacy. Individual-level surrogacy is easily defined from the bivariate model based on equation 9 for variance-covariance matrix of the residuals. As shown by Buyse et al. (2000),

$$R_{indiv}^2 = \frac{(\sigma_{ST})^2}{\sigma_S^2 \sigma_T^2} \quad (12)$$

which is simply the square of the correlation coefficient between the residuals after adjusting for other terms in the model (e.g., adjusting for treatment effects, block effects, overall trial effects). Equation 12 can also be used if equation 1 (with fixed rather than random effects for $\alpha_i, \beta_i, \eta_{Si}$, and η_{Ti}) is directly fitted to the data as a bivariate model, as long as the residuals are modeled as correlated random variables (equation 9).

Determining trial-level surrogacy based on the bivariate mixed model is also possible. Consider a hypothetical new trial, $i = new$, with data for the surrogate but not for the true response variable. This can also be a randomly selected trial out of the original N trials. The question to be answered is: What is the effect of treatment (e.g., mean difference) on the true response given the effect of treatment on the surrogate? The calculations are for a completely randomized (without blocks) new trial or a trial in which block has no effect (note that, on average, block does not have an effect, by definition). This allows us to use the results in Buyse et al. (2000).

Fitting the single-trial version of equation 5 for just the surrogate response variable provides estimates of η_{Snew} (i.e., η_{Si} with $i = new$) and α_{new} (i.e., α_i with $i = new$). Based on equation 4, we can then define the random effects for this trial:

$$\begin{aligned} m_{Snew} &= \eta_{Snew} - \eta_S \\ a_{new} &= \alpha_{new} - \alpha \end{aligned} \quad (13)$$

where the last terms on the right side (η_S and α) are the global parameters for the population of studies (equations 4, 5, 6, and 7), and m_{Snew} and a_{new} represent the random effects m_{Si} and a_i for the $i = new$ trial. For the definition of trial-level surrogacy given above, we want to predict the effect of treatment on the true response variable, β_{new} (i.e., $\beta_{new} = \beta + b_{new}$), based on knowledge of η_{Snew} and α_{new} .

Buyse et al. (2000) derived expressions for the expected value and variance of β_{new} conditional on the results for the surrogate.

$$\begin{aligned} E(\beta_{new} | \eta_{Snew}, \alpha_{new}) &= \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_S^2 & d_{Sa} \\ d_{Sa} & d_a^2 \end{pmatrix}^{-1} \begin{pmatrix} \eta_{Snew} - \eta_S \\ \alpha_{new} - \alpha \end{pmatrix} \\ \text{var}(\beta_{new} | \eta_{Snew}, \alpha_{new}) &= d_b^2 - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_S^2 & d_{Sa} \\ d_{Sa} & d_a^2 \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix} \end{aligned} \quad (14)$$

Importantly, these terms for β_{new} on the left side depend only on the (bivariate) results for the sampled population of studies and on the results for the surrogate in the new study. A surrogate is perfect at the trial level if the conditional variance [$\text{var}(\beta_{new} | \eta_{Snew}, \alpha_{new})$] is 0. Thus, one can measure how close this conditional variance is to 0 on a relative scale (i.e., relative to the variance for b , i.e., d_b^2) with a trial-level coefficient of determination-type statistic:

$$R_{trial}^2 = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_S^2 & d_{Sa} \\ d_{Sa} & d_a^2 \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_b^2} \quad (15)$$

Equation 15 represents the proportion of the variance in the treatment effect on the true response variable (e.g., mean difference) that is explained by the treatment effect on the surrogate.

R_{trial}^2 and R_{indiv}^2 are estimated after fitting equation 5 to the data using estimates of the variances and covariances in equations 11 and 9. Together, these two coefficients of determination characterize the two aspects of surrogacy on a relative scale. Confidence intervals for these coefficients of determination can be estimated using the methods in Cortiñas et al. (2008).

Model fitting. SAS macros, an R package, and a (R-based) Shiny App for cloud-computing have been developed to fit several surrogacy models to data from multiple trials and calculate statistics such as the R^2 metrics (Alonso et al. 2017; Tilahun et al. 2007). Although different distributions for the response variables (endpoints) are accommodated (continuous normal, binomial, survival distributions), with several model choices and link-function choices being possible, the available models are not adequate for common agricultural experiments. For instance, these specialized programs cannot directly handle blocking within trials, multiple treatments, additional random effects, and covariables, and cannot deal with different experimental designs in the different trials. Of these, the need to incorporate blocking is important for our investigation. So, we fitted all models using the MIXED procedure in SAS, expanding on code provided in Buyse et al. (2000), Tibaldi et al. (2003), and Bigirimurame et al. (2017). Additional code was written in SAS data steps and the IML procedure for postmodel-fitting processing to calculate R^2 and other statistics. Example code for fitting example 5 is given in Supplementary File S1.

The use of equation 5 assumes a linear relationship between S and T , which serves as the basis of the individual-level surrogacy (equation 12). Assessment of trial-level surrogacy based on the use of equation 14 assumes a linear relationship between β_i and α_i . We assessed linearity graphically, by plotting the estimated residuals (from equation 5) versus predicted values and estimated residuals for S versus estimated residuals for T . Furthermore, in anticipation of the approximations (below), we plotted estimates of $2\beta_i$ versus estimates of $2\alpha_i$ to assess the linearity of the trial-level surrogacy relationship.

Equation 1 can also be directly fitted to data from multiple trials, with α_i and β_i as fixed-effect parameters. There will be a wider range of estimates of these effects with the fixed-effect approach than with the random-coefficient approach (Madden and Paul 2009) because of the shrinkage effect obtained with BLUPs (Littell et al. 2006). This was not pursued here.

Approximations, alternatives, and simplifications. As pointed out by Tibaldi et al. (2003), fitting a bivariate random-coefficient mixed model such as the one in equation 5 can be quite challenging, especially with a limited number of trials. Even with a successful model fit (i.e., convergence of the optimization algorithm), the estimated \mathbf{D} matrix of equation 11 for variances and covariances may be nonpositive definite or ill-conditioned, leading to uncertain R^2_{trial} estimates. In our case, the full impact of the block effect in the original trials also is uncertain in the R^2_{trial} calculations. Although not of concern with our investigation, experimental designs may vary among trials in other situations, or one may want to determine surrogacy with multiple treatments per trial. Thus, alternatives are desirable. A wide range of approximations have been developed as alternatives to the direct use of the full mixed model (Buyse et al. 2016; Tibaldi et al. 2003). Even when equation 5 (or similar bivariate mixed model) is successfully fitted, some of these approximations lead to valuable ways of visualizing potential individual- and trial-level surrogates, leading to improved insight (see below).

The approximations are multistage processes. For individual-level surrogacy, one first fits the two components of equation 5 (random trial effects) or equation 1 (fixed trial effects) *separately* to the surrogate and true response variables (Alonso Abed and Van der Elst 2017a; Tibaldi et al. 2003). That is, univariate models are fitted to S and to T (with $e_{Sijk} \sim N(0, \sigma_S^2)$, $e_{Tijk} \sim N(0, \sigma_T^2)$), which is equivalent to forcing the σ_{ST} covariance in equation 9 to be 0 in the bivariate case. With random trial effects, the variance-covariance for trials is

$$\mathbf{D}_S = \begin{pmatrix} d_S^2 & d_{Sa} \\ & d_a^2 \end{pmatrix}$$

for the surrogate and

$$\mathbf{D}_T = \begin{pmatrix} d_T^2 & d_{Tb} \\ & d_b^2 \end{pmatrix}$$

for the true response variable.

The residuals are then estimated for each response variable (\hat{e}_{Sijk} , \hat{e}_{Tijk}), and R^2_{indiv} is estimated as the square of the correlation coefficient between these two residual estimates:

$$R^2_{indiv} = (\text{corr}(\hat{e}_{Sijk}, \hat{e}_{Tijk}))^2 \quad (16)$$

It is critical that the univariate models be fitted to S and T for the calculation in equation 16 to work, because if the full bivariate model is used (with equation 9 for the variance-covariance matrix), then the correlation is removed from the residuals and incorporated

directly in the model parameters (σ_{ST}). A plot of \hat{e}_{Tijk} versus \hat{e}_{Sijk} from the separate univariate fits is a useful graphic way to visualize the individual-level surrogacy, if there is any, which also serves as a diagnostic check on the linearity of the relationship between the two endpoints.

For trial-level surrogacy, one first fits either the bivariate or two separate univariate models (equations 5 or 1) to the data across trials, resulting in the following estimates or empirical BLUPs:

$$\begin{aligned} \hat{\eta}_{Si} & (= \hat{\eta}_S + \hat{m}_{Si} \text{ for the random-effects approach}) \\ \hat{\alpha}_i & (= \alpha + \hat{a}_i) \\ \hat{\eta}_{Ti} & (= \hat{\eta}_T + \hat{m}_{Ti}) \\ \hat{\beta}_i & (= \hat{\beta} + \hat{b}_i) \end{aligned}$$

for each trial. One then fits the following least squares regression equation to the $\hat{\beta}_i$, $\hat{\eta}_{Si}$, and $\hat{\alpha}_i$ parameter estimates from each trial:

$$\hat{\beta}_i = \lambda_0 + \lambda_1 \hat{\eta}_{Si} + \lambda_2 \hat{\alpha}_i + \xi_i \quad (17)$$

where the λ terms are new parameters to be estimated and ξ is a new residual term (Tibaldi et al. 2003). The regular R^2 from this regression is an estimate of R^2_{trial} from equation 15. A plot of $\hat{\beta}_i$ versus $\hat{\alpha}_i$ (or $2\hat{\beta}_i$ versus $2\hat{\alpha}_i$ for the mean differences per trial; equations 3 and 7) shows how the treatment effect for the surrogate affects the treatment effect for the true response variable. A plot of $\hat{\beta}_i$ versus $\hat{\eta}_{Si}$ shows how the overall magnitude of the surrogate in a trial affects the treatment effect for the true response; this graph also serves as a diagnostic check on the linearity of the relationship between treatment effects for the two endpoints. Weighted regression can be used also (if trials vary substantially in size), or a multivariate meta-analysis of these parameter estimates can be used to more fully capture the hierarchical nature of the data and the intercorrelations of the estimates (Tibaldi et al. 2003).

An advantage of equation 17 is that it can be used when there are multiple treatments ($m = 1, \dots, M$ treatments), with multiple β parameters, so that one could do a separate regression for each β_{mi} to estimate the trial-level surrogacy for each treatment.

FHB analyses. Analyses were performed for FHB index as a surrogate for yield and for DON in grain. For the latter, analysis was done with untransformed index and DON measurements, and also for log-transformed index and DON. Past work based on results from multiple trials has shown a linear relationship between index and yield (Madden and Paul 2009) and between index and DON (Paul et al. 2006), so we continue with analyses of these response variables (endpoints) here. For comparison purposes, to assess the sensitivity of the approach to the choice of endpoint scale, we also considered the angular transformation of index. The analysis based on logs of the response variables is needed for comparisons with previously published meta-analyses of the relative effects of several different fungicide treatments on index and DON (Paul et al. 2008, 2010, 2018a, b). With log-transformed values, a treatment effect (difference in log-means between a treatment and a control) is the same as the log-response ratio (Paul et al. 2008); back-transformation of the log-response ratio leads to an estimate of the percent control (i.e., percent reduction in index or DON by the treatment relative to the control) (Madden and Paul 2011; Paul et al. 2008).

Three sets of analyses were done for each of the surrogate: true response situations: (a) bivariate model with random trial effects (equations 5, 9, 10, and 11); (b) univariate model with random trial effects (equations 5, 16, and 17); and (c) bivariate model with fixed trial effects (equations 1, 10, and 12). Methods b and c can be considered approximations (Buyse et al. 2016).

For the full bivariate model (approach a), we first considered different versions of equation 5 before deciding on the final model form for characterizing surrogacy. In particular, we considered

three versions of the block effect ($\mathbf{B}_{ij} = (B_{Sij}, B_{Tij})^T$): no block effect (so that equation 5 reduces to the simpler original model in Buyse et al. (2000)), a block effect with a covariance of 0 between the effects for S and T (i.e., $\varsigma_{ST} = 0$ in the Θ variance-covariance matrix of equation 10), and a block effect with no constraints on the variances and covariance of the Θ matrix. We also considered the even simpler bivariate mixed model where there is no block effect and no random effect of trial (i.e., no m_{Si} term); this latter model was suggested by Buyse et al. (2000) as potentially desirable for interpretation purposes, if there is no strong effect of trial on the surrogate. The Akaike information criterion (AIC) statistic was calculated for the fit of each model form, and we chose the model with the smallest AIC (Brown and Prescott 2015).

RESULTS

Form of bivariate random-trial-effects model. Versions of equation 5 without a block effect or a random trial effect gave poorer fits to the data than the versions with a block effect (Table 2). For the log-transformed response variables, the model with a zero covariance for the block effects ($\varsigma_{ST} = 0$ in equation 10 for Θ) had the best fit based on AIC. For the yield:index and DON:index response-variable models, the estimated Θ variance-covariance matrix for the block effects (equation 10) was singular when there were no constraints on ς_{ST} . For instance, the absolute value of the correlation was estimated to be 1 for the block effects for S and T ; this suggests an overparameterized model when no constraint is placed on ς_{ST} ; using a model with this property can substantially inflate the estimates of variances (Riley et al. 2007). For quantifying surrogacy, therefore, we chose equation 5, using 0 for ς_{ST} in equation 10 for Θ .

Disease index and yield. The bivariate random-trial-effects model (equation 5; with $\varsigma_{ST} = 0$ in equation 10) was successfully fitted to the FHB index and yield data. Residual plots revealed no evidence to reject linearity between the response variables, and the plots of $2\hat{\beta}_i$ versus $2\hat{\alpha}_i$ also did not suggest nonlinearity for the relationship between treatment effects for the endpoints. Based on the estimated parameters, fungicide treatment resulted in an average reduction of 8.3 units of percent field severity for index, the surrogate ($\hat{\Delta}_S = 2\hat{\alpha} = -8.3$), and an average increase of 0.44 T/ha for yield, the response of interest in this case ($\hat{\Delta}_T = 2\hat{\beta} = 0.44$) (Table 3). Both were highly significant ($P < 0.001$). The estimated treatment effects for index across the 82 studies ($\hat{\Delta}_{Si} = 2\hat{\alpha}_i = 2(\hat{\alpha} + \hat{a}_i)$; equation 7) ranged from -35.7 to 0.24 ; the estimated treatment effect for yield ($\hat{\Delta}_{Ti} = 2\hat{\beta}_i = 2(\hat{\beta} + \hat{b}_i)$) ranged from 0.01 to 1.19 (Fig. 3). The estimated \mathbf{D} variance-covariance matrix for trial and treatment effects (equation 11), for yield and the other true response variables (below), is given in Supplementary File S2.

The estimated treatment effect for yield decreased with the estimated treatment effect for index and increased with the estimated expected value for index in each trial ($\hat{\eta}_{Si} = \hat{\eta}_S + \hat{m}_{Si}$) (Fig. 3A and B). The former indicates that the largest increases in yield from fungicide treatment tended to occur with the largest decreases in FHB index from the treatment. The latter indicates that

the increase in expected yield after treatment tends to be larger in trials when FHB index is larger than when smaller; this is consistent with a previous meta-analysis of treatment means per trial, involving trials between 1995 and 2007 (Paul et al. 2010). Based on the estimated terms in equation 15, the trial-level coefficient of determination was $\hat{R}_{trial}^2 = 0.42$, indicating a moderate level of trial-level surrogacy of index for wheat yield (Table 3). Based on equation 12, there was no evidence of individual-level surrogacy ($\hat{R}_{indiv}^2 = 0.002$). This is confirmed by a plot of estimated residuals obtained from fitting the univariate random-effect models to the data (Fig. 3C). Presence of individual-level surrogacy would have been manifested by a steeper slope and less variation around the line in Figure 3C.

Lower trial-level surrogacy was estimated using two univariate random-trial-effect models (equation 5) or with a bivariate fixed-trial-effect model (equation 1). Based on the fit of equation 17 to the estimated treatment effects and surrogate expected values per trial, \hat{R}_{trial}^2 equaled 0.22 (Table 3), about half the value found with the full bivariate random-trial-effect model. This lower coefficient of determination was close to the lower limit of the 95% confidence interval for R_{trial}^2 that was found with the full model (0.25). The estimated individual-level surrogacy, \hat{R}_{indiv}^2 , was unchanged when using the univariate random-effect model (equation 16) or the bivariate fixed-effect model (equation 12).

Disease index and DON. The bivariate random-trial-effects model (equation 5; $\varsigma_{ST} = 0$ in equation 10) was also successfully fitted to the (untransformed) FHB and DON data. As with yield, residual plots revealed no evidence to reject linearity between the response variables, and the plots of $2\hat{\beta}_i$ versus $2\hat{\alpha}_i$ also did not suggest nonlinearity. As anticipated, the same estimated expected reduction in index from the fungicide treatment was found as with the surrogacy analysis of index and yield ($\hat{\Delta}_S = 2\hat{\alpha} = -8.3$). Treatment resulted in an estimated expected reduction in DON of 1.2 ppm ($\hat{\Delta}_T = 2\hat{\beta} = -1.2$) (Table 3). The range of estimated treatment effects for index across the trials was about the same as with the analysis of index and yield (-36.1 to 0.06). Estimated treatment effects for DON ranged from -8.1 to $+3.7$, with 7% of the trials exhibiting an actual increase in mean DON when treated with fungicide (Fig. 4).

The magnitude of the estimated treatment effect for DON generally increased with estimated treatment effect for FHB index (Fig. 4A) and decreased with the estimated expected value for index in each trial (Fig. 4B). The former indicates that the largest reductions in DON from fungicide treatment tended to occur with the largest reductions in FHB index from treatment. Trial-level surrogacy of index for DON was weak, however, with \hat{R}_{trial}^2 calculated from equation 15 equal to 0.18 (Table 3). Although significant (the 95% confidence interval for R_{trial}^2 did not include 0), there was high variability in the treatment effect for DON in relation to the treatment effect for index. There was some very slight evidence of individual surrogacy of index for DON, with $\hat{R}_{indiv}^2 = 0.04$ (equation 12), and a confidence interval that did not

TABLE 2. Akaike information criterion statistic for the fit of the bivariate linear mixed model (equation 5) to the Fusarium head blight data, with different specifications of the covariance matrix for the block effect (equation 10), and the presence or not of a random surrogate effect of trial (m_{Si}) when there is no block effect

Model specification	Response variables in equation 5		
	Yield:index	DON:index	log(DON):log(index)
No block, no random intercept (m_{Si})	10,699.6	14,222.5	7,826.4
No block effect	9,191.7	12,477.5	6,410.5
Block effect (with 0 covariance)	9,171.0	12,463.9	6,371.6
Block effect (no constraints)	— ^a	— ^a	6,372.1

^a Singular covariance matrix for block. Model form rejected.

quite reach 0 (Table 3). The estimated residuals from the univariate analyses showed a slight positive relationship, but with high scatter (Fig. 4C).

Similar weak trial-level surrogacy was found with the univariate random-trial-effect analysis and with the bivariate fixed-trial-effect analysis. The \hat{R}_{trial}^2 values were slightly lower than those found with the full bivariate random-effect model, 0.13-0.15 versus 0.18 (Table 3). Individual-level surrogacy estimates were also similar between these two approximate methods ($\hat{R}_{indiv}^2 = 0.05$), confirming the very weak relationship of DON and index at the individual plot level.

Logarithm of index and DON. As with the previous analyses, the bivariate random-trial-effects model (equation 5, with $s_{ST} = 0$ in equation 10) was successfully fitted to the log-transformed data. Estimated expected log ratios of the treatment versus the control were -1.08 for index ($\hat{\Delta}_S = 2\hat{\alpha}$) and -0.61 for DON ($\hat{\Delta}_T = 2\hat{\beta}$) (Table 3). Large negative values indicate large reductions in the response variable. Back-transformation leads to an estimate of percent control (Madden and Paul 2011; Paul et al. 2007) of 66% for FHB index and 46% for DON. The estimated treatment effect for $\log(\text{DON})$ [i.e., $\hat{\Delta}_{Ti} = 2\hat{\beta}_i = 2(\hat{\beta} + \hat{b}_i)$] slightly increased with the estimated treatment effect for $\log(\text{index})$ [i.e., $\hat{\Delta}_{Si} = 2\hat{\alpha}_i = 2(\hat{\alpha} + \hat{a}_i)$] and with the estimated expected value of $\log(\text{index})$ for each trial ($\hat{\eta}_{Si} = \hat{\eta}_S + \hat{m}_{Si}$) (Fig. 5A and B). However, there was a great deal of scatter, translating to a very low value of 0.06 for \hat{R}_{trial}^2 (Table 3). Back-transformation of the individual $\hat{\Delta}_{Si}$ and $\hat{\Delta}_{Ti}$ values also showed a high degree of scatter (Fig. 5C), although with a slight upward trend in the percent control for DON in relation to percent control for index. There also was little evidence of individual-level surrogacy (Table 3).

The two approximations (univariate random-trial and bivariate fixed-trial) resulted in very similar results to that found for the full bivariate random-trial model (Table 3). A plot of the estimated residuals from the two univariate-model fits demonstrates the very low individual-level surrogacy (Fig. 5D).

Checks on the analyses. Although there was no evidence to reject a linear relationship between FHB index and yield or DON based on residual plots, the analyses were repeated using the angular transformation of index because variability of a percent or proportion depends on the mean value (Littell et al. 2006). R^2 -based surrogacy results were little affected by use of this transformation. For yield and angular-transformed index, \hat{R}_{trial}^2 was 0.42, 0.21, and 0.21 for the full bivariate model, bivariate model with fixed trial effects (approximation), and univariate model with random trial effects (approximation), respectively. \hat{R}_{indiv}^2 values were all 0.01 or less. These are almost the same as shown in Table 3 for the analysis on index directly. For DON and angular-transformed FHB index,

\hat{R}_{trial}^2 was 0.19, 0.14, and 0.16, respectively, for the same three analytic approaches, and \hat{R}_{indiv}^2 values were all 0.04 or less. These results are all close to the values shown in Table 3 for untransformed index. This shows that the results for these data were not sensitive to possible variance heterogeneity of index.

DISCUSSION

Epidemiologists have invested considerable effort into characterizing relationships between disease intensity and other epidemic components (e.g., AUDPC, time to a fixed level of disease, epidemic rate parameter, healthy area duration) and resulting yield or yield loss (Madden et al. 2007; Savary et al. 2006, 2017b). Although to a lesser extent, researchers have also explored the relationship between disease and toxin concentration in harvested products (Okoth et al. 2017; Reid et al. 1996; Williams et al. 2011). Even when the primary interest is in crop yield or toxin, measurements of disease intensity (or other estimated epidemic components) may shed light on how plant diseases directly or indirectly affect yield and toxin. For instance, one may investigate whether declines in yield are due to the effects of disease on radiation interception or radiation use efficiency (Savary et al. 2017b). Or, one may wish to determine the biological and environmental conditions that result in increased (or decreased) levels of toxin for different pathogens (Andersen et al. 2015; Cowger et al. 2009; Culler et al. 2007; Gautam and Dill-Macky 2012; Ojiambo et al. 2018; Okoth et al. 2017; Xu et al. 2014). In this regard, one could determine if pathogen biomass (based on DNA content) could be used as a surrogate for the toxin it produces (Xu and Nicholson 2009). In other circumstances, measurement of disease early in an epidemic could be considered a surrogate for disease intensity later in an epidemic (say, for a long-season crop or for a perennial), or for area under the disease progress curve based on the entire epidemic (Madden et al. 2007). Also, easy-to-measure disease incidence (proportion of individuals diseased) could be considered as a surrogate for the more-difficult-to-measure disease severity (area of diseased tissue), after a suitable linearizing transformation was used (McRoberts et al. 2003). Likewise, severity of above-ground symptoms (e.g., wilting) could be considered a surrogate or root disease severity for infection by soilborne pathogens (Campbell and Neher 1994). If a random variable is to be used *instead* of the random variable of ultimate (primary) interest, then it should ideally be formally evaluated as a surrogate.

In principle, it may be possible to predict yield loss, toxin concentration, or other response variables from measurements of disease or estimates of epidemic components (Madden and Paul 2009; Madden et al. 2007; Paul et al. 2005b,2006). Since disease

TABLE 3. Trial-level and individual-level surrogacy results for wheat yield (T/ha) and deoxynivalenol (DON) in harvested grain (ppm) as the true response variable, with Fusarium head blight index (field severity) as the surrogate, and for $\log(\text{DON})$ as the true response and $\log(\text{index})$ as the surrogate, with expected value and coefficient of determination (R^2)^a estimates given for the fit of the full random-trial-effects bivariate model, and R^2 estimates for two approximations (a fixed-trial-effects bivariate model and combined use of two random trial-effects univariate models)

Response variable	Estimate ^b	Estimated CI	Random, bivariate				Fixed, bivariate		Random, univariate	
			R_{trial}^2	R_{trial}^2 CI	R_{indiv}^2	R_{indiv}^2 CI	R_{trial}^2	R_{indiv}^2	R_{trial}^2	R_{indiv}^2
Yield:index										
Surrogate (index) treatment effect (Δ_S)	-8.29 (1.08)	-10.45 ↔ -6.13	0.42	0.25 ↔ 0.59	0.002	0 ↔ 0.01	0.22	0.002	0.22	0.002
True response (yield) treatment effect (Δ_T)	0.44 (0.04)	0.36 ↔ 0.52								
DON:index										
Surrogate (index) treatment effect (Δ_S)	-8.28 (1.08)	-10.44 ↔ -6.12	0.18	0.03 ↔ 0.33	0.04	0.02 ↔ 0.06	0.13	0.05	0.15	0.05
True response (DON) treatment effect (Δ_T)	-1.20 (0.24)	-1.68 ↔ -0.72								
$\log(\text{DON}):\log(\text{index})^c$										
Surrogate ($\log(\text{index})$) treatment effect (Δ_S)	-1.08 (0.10)	-1.27 ↔ -0.87	0.06	0 ↔ 0.16	0.02	0.01 ↔ 0.03	0.05	0.02	0.07	0.02
True response ($\log(\text{DON})$) treatment effect (Δ_T)	-0.61 (0.08)	-0.77 ↔ -0.45								

^a R_{trial}^2 : trial-level surrogacy, estimated with equation 15 for the full model, and with equation 17 for the two approximations. R_{indiv}^2 : individual-level surrogacy, estimated with equation 12 for the full model and for the fixed-trial bivariate model, and with equation 16 for the univariate models. If a confidence interval (CI) for R_{trial}^2 or R_{indiv}^2 extends below 0, it is truncated at 0.

^b Estimates of expected value for the population of trials (equation 8). CI at 95% is shown.

^c Back-transformation of values correspond to a percent control of 66% for index and 46% for DON.

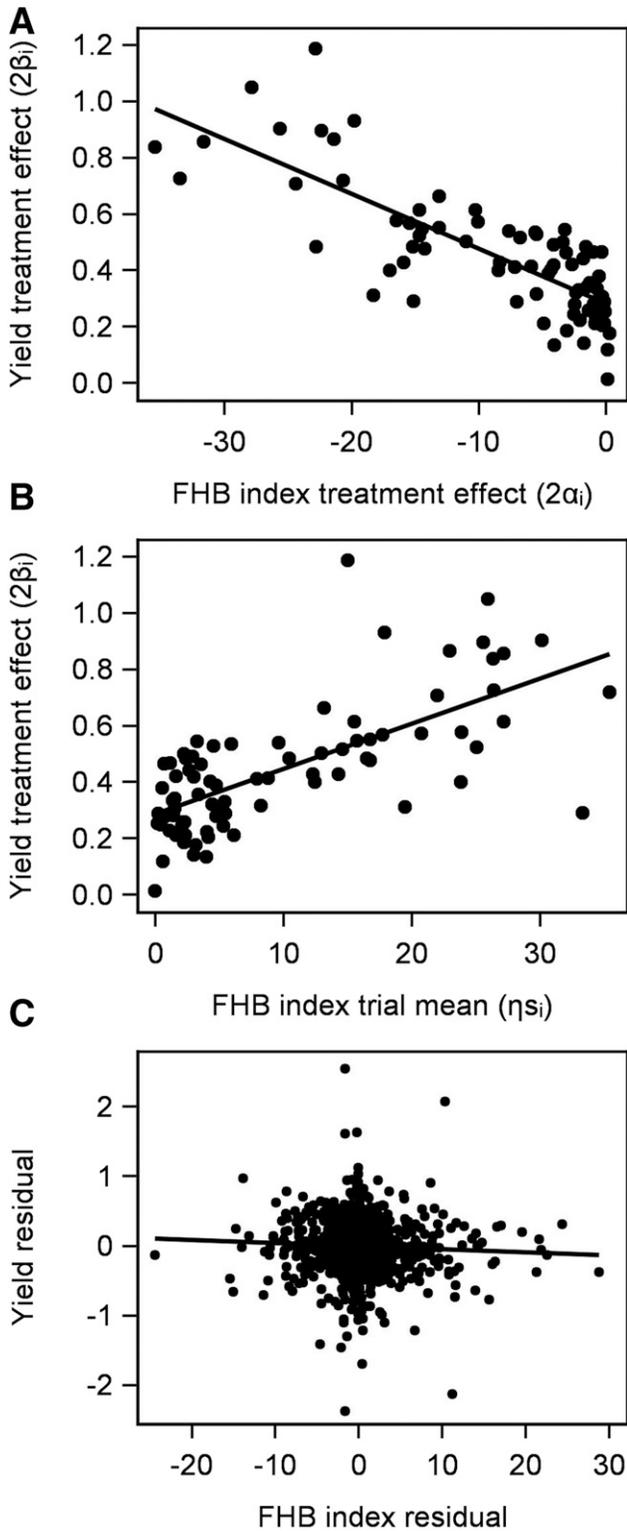


Fig. 3. Trial-level and individual-level surrogacy results for wheat yield (T/ha) in relation to Fusarium head blight (FHB) index (field severity) across 82 trials. **A**, Estimated treatment effect (mean difference between fungicide treatment and control) for wheat yield versus estimated treatment effect for index. **B**, Estimated treatment effect for yield versus estimated expected value of index per trial. Results in A and B based on the fit of the bivariate random-trial-effects model (equation 5). **C**, Estimated residuals for yield for each experimental unit versus estimated residuals for index. Residuals in C estimated from the fit of two univariate random-trial-effects models (each component of equation 5 fitted separately), followed by a linear regression of the estimated residuals. Lines in A and B are best fits from ordinary least squares regression with a single predictor variable, shown to improve visual presentation. These are not the predictions from the use of equation 5 or from either of the approximations (which are based on a multiple-regression [two predictor-variable] model).

measurements occur earlier than measurements of yield or toxin, and may be obtained more economically and with fewer specialized equipment requirements, disease intensity could be then used as a surrogate for the response variables of primary interest, if the relationships between the variables were strong enough, at the individual and/or trial level. FHB index and other measures of disease intensity data (e.g., incidence) are routinely collected in fungicide and integrated management trials (Paul et al. 2019), as well as in surveys of grower fields before crop maturity (often for fields where no toxin or yield data are obtained) (Kriss et al. 2012). Even when precision is not high enough for prediction, differences in treatment effects between the potential surrogates and true response variable can shed light on the mechanisms underlying treatment effects (Savary et al. 2006). Formal concepts of surrogacy have become well established in medicine and pharmacology (Alonso et al. 2017; Buyse et al. 2016), but these ideas have received little recognition in plant pathology. It is critical to understand that one can consider the strength of the relationship between the surrogate and the true response variable in individual experimental units (after adjusting for any treatment effects), known as individual-level surrogacy, and the strength of the relationship between the treatment effect size (e.g., mean difference between a treatment and a control) in the true response variable and treatment effect size in the surrogate, known as trial-level surrogacy (Alonso Abed and Van der Elst 2017b; Molenberghs et al. 2008). The latter mandates that data from multiple trials be analyzed.

We successfully applied the general statistical methodology developed by (Buyse et al. 2000; 2016) for continuous random variables and a binary treatment classification to quantify the extent to which FHB field severity (FHB index) can be used as a surrogate for wheat grain yield and DON toxin contamination. One advantage of this approach is that relatively easy-to-interpret coefficient-of-determination-type statistics are obtained for both types of surrogacy. Using data from the USWBSI uniform fungicide trials (for susceptible cultivars), and model expanded to include block effects, we found that FHB index is a poor surrogate for yield or DON at the *individual* plot level. That is, knowing the (estimated) mean FHB index in a plot (the experimental unit) did not provide reliable information on mean yield or DON [or $\ln(\text{DON})$] for the same plot, after adjusting for treatment effects.

FHB index did provide a higher level of trial-level surrogacy for yield, in that there was a generally increasing *positive* yield differences with increasing *negative* index differences between the fungicide treatment and the control. Nevertheless, the R^2_{trial} value (0.4) was far from the informal threshold of 0.8 for high surrogacy suggested by Buyse et al. (2000). Moreover, the R^2_{trial} statistic from the full bivariate model (equation 15) was about double that found with the approximations (equation 17). The lower estimates from the approximate models were reasonable based on the confidence interval for the full bivariate model. As shown by Tibaldi et al. (2003), the full bivariate model can be difficult to fit and could produce biased results if the estimated among-trial covariance matrix (equation 11) is ill conditioned or not positive definite, the former being difficult to detect. All variance-covariance parameter estimates in equation 11 were nonzero, and all eigenvalues of the matrix were positive, indicating that the estimated matrix for each analysis was positive definite. Although there are different measures of an ill-conditioned variance-covariance matrix (Goodfellow et al. 2016; Tabcart et al. 2019), such as the ratio of the largest to smallest eigenvalues, there is no agreed upon threshold for this ratio in identifying problem situations, nor are there other unambiguous way of determining when the full bivariate model (equation 5) should be abandoned based on properties of $\hat{\mathbf{D}}$ (Burzykowski 2017; Tibaldi et al. 2003). These uncertainties certainly justify the use of the two approximations.

With the large number of trials in the data set (82), challenges in fitting the full bivariate equation 5 are considerably diminished, but

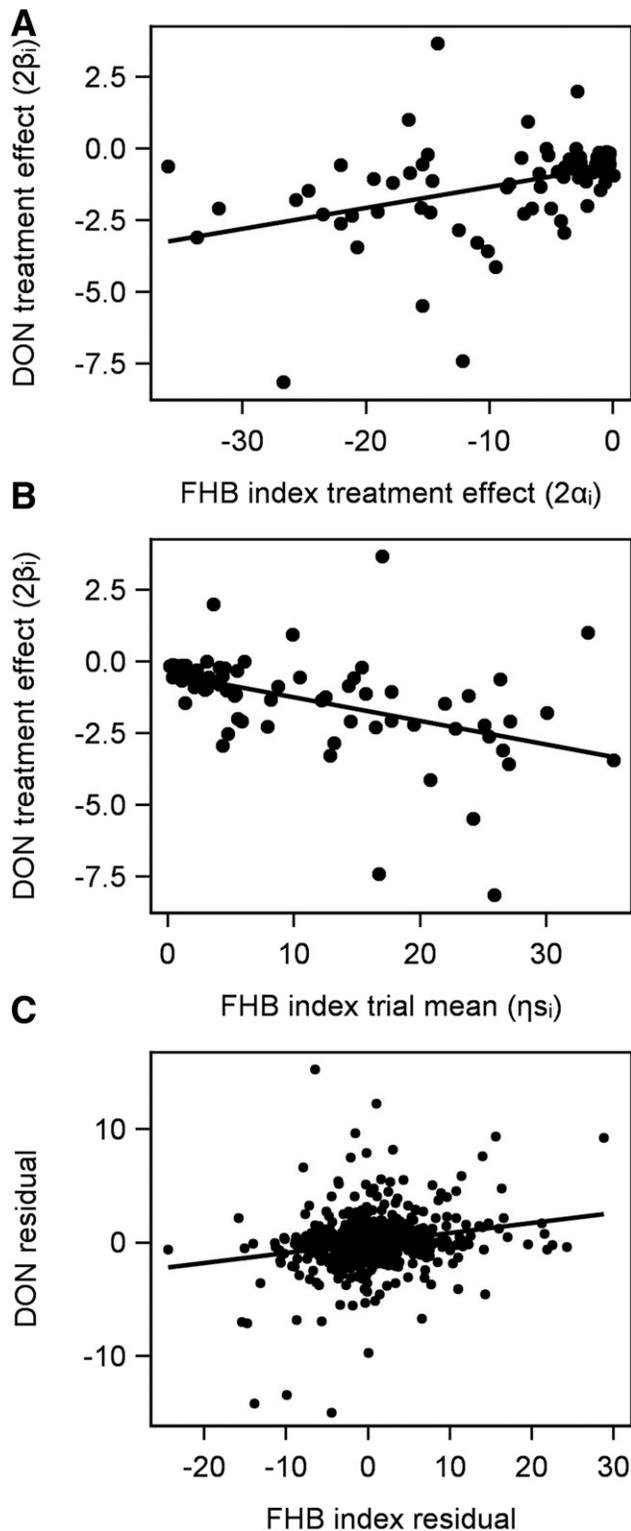


Fig. 4. Trial-level and individual-level surrogacy results for deoxynivalenol (DON) concentration in harvested wheat grain (ppm) in relation to Fusarium head blight (FHB) index (field severity) across 82 trials. **A**, Estimated treatment effect (mean difference between fungicide treatment and control) for DON versus estimated treatment effect for index. **B**, Estimated treatment effect for DON versus estimated expected value of index per trial. Results in A and B based on the fit of the bivariate random-trial-effects model (equation 5). **C**, Estimated residuals for DON for each experimental unit versus estimated residuals for index. Residuals in C estimated from the fit of two univariate random-trial-effects models (each component of equation 5 fitted separately), followed by a linear regression of the estimated residuals. Lines in A and B are best fits from ordinary least squares regression with a single predictor variable, shown to improve visual presentation. These are not the predictions from the use of equation 5 or from either of the approximations (which are based on a multiple-regression [two predictor-variable] model).

there is still the possibility that incorporation of a block effect into the model had an unexplained impact on the estimate of **D** (equation 11) and, thus, an influence on the R^2_{trial} calculation. Thus, approximation methods are further justified here (Tibaldi et al. 2003). With the approximations, R^2_{trial} was based solely on a simple two-variable linear regression equation (equation 17) for treatment effects, calculated after the bivariate or univariate mixed models were fitted to *S* and *T*. Nevertheless, it should be emphasized that the results were very similar among the different methods utilized (Table 3), with the biggest variation in results for the yield:index data. Overall, based on the results from the full bivariate model results and the two approximations, trial-level surrogacy was moderate at best for yield, suggesting that one cannot readily make strong conclusions in individual trials about magnitude of treatment effects on yield based on treatment effects on FHB index.

FHB index was found to be even less satisfactory as a trial-level surrogate for DON toxin in harvested wheat grain. In terms of mean differences between the treatment and control, there was only a weak relation in treatment effects for index and DON (with a general trend of increasing negative treatment differences for DON with increasing negative treatment differences for index). This was generally consistent when using the full model and the approximations. There was even less of a relation when treatment effects were based on differences of mean log-transformed data (and thus, also for percent control).

The low surrogacy of FHB index for the endpoints of interest may seem surprising at first, given that there are overall relationships between index and DON and index and yield (Madden and Paul 2009; Paul et al. 2005b, 2006) across a large number of independent trials. Moreover, in terms of averages (estimated expected values) across many trials, treatments that work best for reducing DON or increasing yield generally work best for decreasing FHB index (Paul et al. 2018a, b, 2019). However, as mentioned in the Introduction, those papers also showed the substantial variability in the relationship among the trials, possibly suggesting that relationships for a *given* trial, or even within a trial for individual observations (plots), was not predictable. The surrogacy analysis here builds on this past work and shows that evidence of a treatment effect on FHB index is not sufficient to predict the magnitude of a treatment effect on DON in a given trial; even for yield, the predictability of the magnitude of a treatment effect based on index is weak. This low surrogacy can in part explain the challenges in breeding and developing cultivars with high levels of DON resistance when researchers make decisions on genotype selections based primarily on FHB index (often because selections must be made before DON results are known). Although resistance to infection and fungal colonization of the wheat spike naturally leads to lower DON contamination (in the sense that DON is produced by the pathogen and is needed for colonization), the specific reaction of cultivars to the latter is fairly complex, not clearly understood, and is influenced by environmental and host-related factors that may be independent of index or operate well after visual symptoms of index are estimated (Cowger et al. 2009, 2016; Culler et al. 2007; Gautam and Dill-Macky 2012; Sneller et al. 2012). Low FHB index-DON surrogacy can also explain the fact that DON sometimes exceeds critical thresholds in a commercial field even if a low risk of index is predicted by the National FHB Prediction Center (the prediction center is based on a model for FHB index, not DON) (Moraes et al. 2018; Shah et al. 2019; L. V. Madden, *unpublished data*).

There are several possible reasons for the low surrogacy of FHB index. Two of the most likely ones involve sampling precision and observation timing. For instance, there is a great deal of variation among cooperators with the USWBSI uniform fungicide trials in the number of samples used for assessing index and how samples are collected and processed for DON analysis. A typical sample size for index is 30 to 50 spikes per 100 ft² (9.3 m²) of plot, but sometimes even lower numbers have been used (Paul et al. 2008; L.

V. Madden, unpublished data). However, new work by Moraes et al. (2017) has shown that precision of estimates of mean index per plot can be very low with under 70 samples. For DON, plot-level estimates are commonly made from a single sample of 450 to 900 g of grain from which a ground sample of less than 100 g is sent for DON extraction and quantification (Paul et al. 2008). In some cases, samples are pulled directly from the grain stream during harvest, whereas in other cases they are drawn after harvest from stored grain lots. Work by Champeil et al. (2004) and Hart and Schabenberger (2001) has shown that the precision of plot- or field-level DON estimates depends on how samples are collected and processed and when they are collected relative to harvest. Imprecise estimates of index and DON due to poor sampling would affect the variability in the observed bivariate relationship between the mean response variables at the individual plot level, which would carry over to the trial level. This sample-size issue is not typically of direct relevance in the surrogacy studies in medicine because the individual observation is a single person (with measurements of S and T),

rather than an individual plot with a sample of S and T measurements that are averaged. However, the sampling imprecision issue here is analogous to measurement error of covariates (predictor variables) in data analysis within medical and other disciplines (Li and Wang 2012).

One strong reason to use a surrogate, in general, is because the variable can be observed earlier than the variable of interest (Alonso Abed and Van der Elst 2017a; Buyse 2017), sometimes years earlier for some systems. But the length of time between the measurements potentially can also reduce the value of a variable as a surrogate. This is particularly important in agricultural production systems that are strongly influenced by weather conditions. In our situation, the time between assessment of FHB index (at Feekes 11.1, reflecting earlier infection and colonization of the spikes) and DON and yield measurements at harvest (a duration of more than 50 days for winter wheat grown in the northern half of the U.S. wheat belt [D'Angelo et al. 2014]) likely contributes to reduced surrogacy. As discussed and demonstrated by others (Andersen et al. 2015;

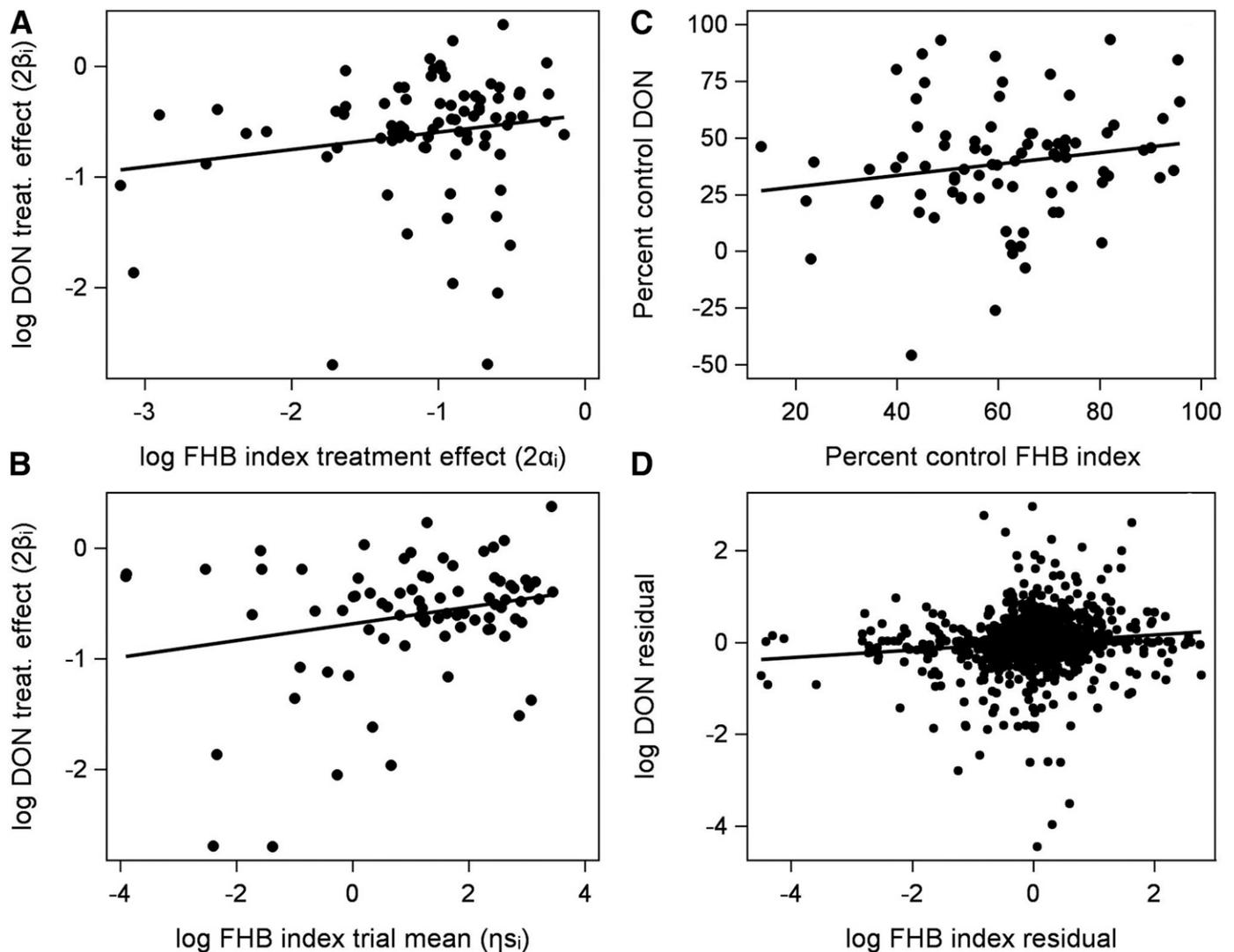


Fig. 5. Trial-level and individual-level surrogacy results for logarithm of deoxynivalenol (DON) concentration in harvested wheat grain (ppm) in relation to logarithm of Fusarium head blight (FHB) index (field severity) across 82 trials. **A**, Estimated treatment effect (mean difference between fungicide treatment and control) for log(DON) versus estimated treatment effect for log(index). **B**, Estimated treatment effect for log(DON) versus estimated expected value of log(index) per trial. **C**, Estimated percent control (percent reduction in DON by the treatment relative to the control) for DON versus percent control of index, based on back-transformation of the treatment effects in part A. Results in A to C based on the fit of the bivariate random-trial-effects model (equation 5). **D**, Estimated residuals for log(DON) for each experimental unit versus estimated residuals for log(index). Residuals in D estimated from the fit of two univariate random-trial-effects models (each component of equation 5 fitted separately), followed by a linear regression of the estimated residuals. Lines in A to C are best fits from ordinary least squares regression with a single predictor variable, shown to improve visual presentation. These are not the predictions from the use of equation 5 or from either of the approximations (which are based on a multiple-regression [two predictor-variable] model).

Cowger et al. 2009; Culler et al. 2007; Edwards et al. 2018; Gautam and Dill-Macky 2012; Moraes et al. 2018), late-season environmental conditions (especially moisture) can affect DON production, conversion of DON to DON-3-glucoside, movement of DON into grain, and retention of DON in grain for a given level of FHB index. Likewise, many factors such as postanthesis temperature and rainfall and late-season development of diseases like *Stagonospora* and leaf rust can influence grain yield after the assessment of FHB index by virtue of their effects on grain fill (Czarnecki and Evans 1986; Salgado et al. 2015, 2017). In addition, studies primarily from Europe show that the presence of other *Fusarium* species in wheat spikes can affect the production of DON by *F. graminearum* (Xu and Nicholson 2009); this is probably not an issue here since *F. graminearum* is the dominant FHB pathogen in the United States (McMullen et al. 2012). For some other pathosystems, toxins are commonly produced by *Fusarium* species without any visible disease symptoms at all (Xu et al. 2014).

In conclusion, we have introduced an expansion of the continuous-variable surrogacy models by Buyse et al. (2000) and found moderate or no surrogacy of FHB index for other response variables of interest. We are not aware of other attempts to take this type of formal and strict approach to quantify individual- and trial-level surrogacy in plant disease field studies. The question remains for future work: Are there plant pathosystems where disease severity measurement at a particular time (or possibly area under the disease progress curve or time to a particular level of disease) provides a high level of surrogacy for other response variables of interest? The full bivariate method of Buyse et al. (2000), suitably expanded for complexities of experimental designs in plant protection provides one approach for evaluating surrogacy. Theoretical statistical research is still needed, however, to evaluate the best way to account for blocking in the derivation of trial-level surrogacy statistics, and to accommodate multiple individual treatments in the model. The approximate two-stage methods described herein (both univariate and bivariate), as well as other approximations not utilized by us (Tibaldi et al. 2003), may be more practical in handling the many variations in experimental designs, including multiple treatments in each trial. As long as one can estimate for each trial the treatment effects for the true and surrogate variables ($\hat{\Delta}_T$ and $\hat{\Delta}_S$, respectively), as well as the trial-level expected value for the surrogate variable, one can estimate trial-level surrogacy using equation 17. If bivariate mixed models are too difficult to fit, correlations of estimated residuals from univariate models are sufficient to approximate individual-level surrogacy. Alternatively, new bivariate (Bayesian) network-meta-analysis methods (Bujkiewicz et al. 2019) or an information-theoretic modeling approach to surrogacy (Alonso and Molenberghs 2007) may give great flexibility in dealing with multiple treatments, blocking, trial-level and individual-level covariates, and other issues related to experimental designs common in plant pathology. This latter approach gave very similar results to those shown here when applied to the FHB:DON:yield dataset (*unpublished data*).

ACKNOWLEDGMENTS

We thank the U.S. Wheat & Barley Scab Initiative (USWBSI) cooperators who conducted field trials over multiple years and locations.

LITERATURE CITED

- Alonso, A., Bigirimurame, T., Burzykowski, T., Buyse, M., Molenberghs, G., Muchene, L., Perualila, N. J., Shkedy, Z., and Van der Elst, W. 2017. Applied Surrogate Endpoint Evaluation Methods with SAS and R. CRC Press, Boca Raton, FL.
- Alonso, A., and Molenberghs, G. 2007. Surrogate marker evaluation from an information theoretical perspective. *Biometrics* 63:180-186.
- Alonso, A., Molenberghs, G., Geys, H., Buyse, M., and Vangeneugden, T. 2006. A unifying approach for surrogate marker validation based on Prentice's criteria. *Stat. Med.* 25:205-221.
- Alonso Abed, A., and Van der Elst, W. 2017a. The history of surrogate endpoint evaluation: Single-trial methods. Pages 35-46 in: *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. Alonso, T. Bigirimurame, T. Burzykowski, M. Buyse, G. Molenberghs, L. Muchene, N. J. Perualila, Z. Shkedy, W. Van der Elst, eds. CRC Press, Boca Raton, FL.
- Alonso Abed, A., and Van der Elst, W. 2017b. Two continuous outcomes. Pages 49-66 in: *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. Alonso, T. Bigirimurame, T. Burzykowski, M. Buyse, G. Molenberghs, L. Muchene, N. J. Perualila, Z. Shkedy, W. Van der Elst, eds. CRC Press, Boca Raton, FL.
- Andersen, K. F., Madden, L. V., and Paul, P. A. 2015. Fusarium head blight development and deoxynivalenol accumulation in wheat as influenced by post-anthesis moisture patterns. *Phytopathology* 105:210-219.
- Bigirimurame, T., Shkedy, Z., and Burzykowski, T. 2017. SAS software. Pages 159-214 in: *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. Alonso, T. Bigirimurame, T. Burzykowski, M. Buyse, G. Molenberghs, L. Muchene, N. J. Perualila, Z. Shkedy, W. Van der Elst, eds. CRC Press, Boca Raton, FL.
- Bock, C. H., Poole, G. H., Parker, P. E., and Gottwald, T. R. 2010. Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. *Crit. Rev. Plant Sci.* 29:59-107.
- Brown, H., and Prescott, R. 2015. *Applied Mixed Models in Medicine*, 3rd Ed. Wiley, New York.
- Bujkiewicz, S., Jackson, D., Thompson, J. R., Turner, R. M., Städler, N., Abrams, K. R., and White, I. R. 2019. Bivariate network meta-analysis for surrogate endpoint evaluation. *Stat. Med.* 38:3322-3341.
- Burzykowski, T. 2017. Two failure-time endpoints. Pages 67-79 in: *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. Alonso, T. Bigirimurame, T. Burzykowski, M. Buyse, G. Molenberghs, L. Muchene, N. J. Perualila, Z. Shkedy, W. Van der Elst, eds. CRC Press, Boca Raton, FL.
- Buyse, M. 2017. Introduction. Pages 3-14 in: *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. Alonso, T. Bigirimurame, T. Burzykowski, M. Buyse, G. Molenberghs, L. Muchene, N. J. Perualila, Z. Shkedy, W. Van der Elst, eds. CRC Press, Boca Raton, FL.
- Buyse, M., and Molenberghs, G. 1998. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 54:1014-1029.
- Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. 2000. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 1:49-67.
- Buyse, M., Molenberghs, G., Paoletti, Z., Oba, K., Alonso, A., Van der Elst, W., and Burzykowski, T. 2016. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biom. J.* 58:104-132.
- Caldwell, D. M., and Welton, N. J. 2016. Approaches for synthesizing complex mental health interventions in meta-analysis. *Evid. Based Ment. Health* 19:16-21.
- Campbell, C. L., and Neher, D. 1994. Estimating disease severity and incidence. Pages 117-147 in: *Epidemiology and Management of Root Diseases*. C. L. Campbell and D. M. Benson, eds. Springer-Verlag, Berlin.
- Carisse, O., McRoberts, N., and Brodeur, L. 2008. Comparison of monitoring- and weather-based risk indicators of Botrytis leaf blight of onion and determination of action thresholds. *Can. J. Plant Pathol.* 30:442-456.
- Champeil, A., Fourbet, J.-F., and Dore, T. 2004. Effects of grain sampling procedures on *Fusarium* mycotoxin assays in wheat grains. *J. Agric. Food Chem.* 52:6049-6054.
- Cortiñas, J., Shkedy, Z., and Molenberghs, G. 2008. Alternative methods to evaluate trial level surrogacy. *Clin. Trials* 5:194-208.
- Cowger, C., Patton-Özkurt, J., Brown-Guedira, G., and Perugini, L. 2009. Post-anthesis moisture increased Fusarium head blight and deoxynivalenol levels in North Carolina winter wheat. *Phytopathology* 99:320-327.
- Cowger, C., Weisz, R., Arellano, C., and Murphy, P. 2016. Profitability of integrated management of Fusarium head blight in North Carolina winter wheat. *Phytopathology* 106:814-823.
- Culler, M. D., Miller-Garvin, J. E., and Dill-Macky, R. 2007. Effect of extended irrigation and host resistance on deoxynivalenol accumulation in *Fusarium*-infected wheat. *Plant Dis.* 91:1464-1472.
- Czarnecki, E., and Evans, L. E. 1986. Effect of weathering during delayed harvest on test weight, seed size, and grain hardness of wheat. *Can. J. Plant Sci.* 66:473-482.
- D'Angelo, D. L., Bradley, C. A., Ames, K. A., Willyerd, K. T., Madden, L. V., and Paul, P. A. 2014. Efficacy of fungicide applications during and after anthesis against Fusarium head blight and deoxynivalenol in soft red winter wheat. *Plant Dis.* 98:1387-1397.
- Delwiche, S. R., Pearson, T. C., and Brabec, D. L. 2005. High-speed optical sorting of soft wheat for reduction of deoxynivalenol. *Plant Dis.* 89:1214-1219.
- Edwards, S. G., Kharbikar, L. L., Dickin, E. T., MacDonald, S., and Scudamore, K. A. 2018. Impact of pre-harvest rainfall on the distribution of *Fusarium* mycotoxins in wheat mill fractions. *Food Control* 89:150-156.

- Farrer, D., Weisz, R., Heiniger, R., Murphy, J. P., and Pate, M. H. 2006. Delayed harvest effect on soft red winter wheat in the Southeastern USA. *Agron. J.* 98:588-595.
- Freedman, L. S., Graubard, B. I., and Schatzkin, A. 1992. Statistical validation of intermediate endpoints for chronic diseases. *Stat. Med.* 11:167-178.
- Gautam, P., and Dill-Macky, R. 2012. Impact of moisture, host genetics and *Fusarium graminearum* isolates on Fusarium head blight development and trichothecene accumulation in spring wheat. *Mycotoxin Res.* 28:45-58.
- Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*. MIT Press, Cambridge, MA.
- Hart, P., and Schabenberger, O. 2001. Early detection of deoxynivalenol in wheat grain. Pages 164-167 in: *Proceedings of the National Fusarium Head Blight Forum*. S. M. Canty, J. Lewis, L. Siler, and R. W. Ward, eds. University of Michigan, Flint, MI.
- Hughes, G. 2017. The evidential basis of decision making in plant disease management. *Annu. Rev. Phytopathol.* 55:41-59.
- Kriss, A. B., Paul, P. A., and Madden, L. V. 2012. Characterizing heterogeneity of disease incidence in a spatial hierarchy: A case study from a decade of observations of Fusarium head blight of wheat. *Phytopathology* 102:867-877.
- Li, H., and Wang, L. 2012. Consistent estimation in generalized linear mixed models with measurement error. *J. Biomet. Biostat.* 2012:S7.
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. 2006. *SAS for Mixed Models*, 2nd ed. SAS Institute, Cary, NC.
- Madden, L. V., Hughes, G., and van den Bosch, F. 2007. *The Study of Plant Disease Epidemics*. American Phytopathological Society, St. Paul, MN.
- Madden, L. V., and Paul, P. A. 2009. Assessing heterogeneity in the relationship between wheat yield and Fusarium head blight intensity using random-coefficient mixed models. *Phytopathology* 99:850-860.
- Madden, L. V., and Paul, P. A. 2011. Meta-analysis for evidence synthesis in plant pathology: An overview. *Phytopathology* 101:16-30.
- Madden, L. V., Piepho, H.-P., and Paul, P. A. 2016. Statistical models and methods for network meta-analysis. *Phytopathology* 106:792-806.
- Man, Y., Liang, G., Li, A., and Pan, L. 2017. Recent advances in mycotoxin determination for food monitoring via microchip. *Toxins (Basel)* 9:324.
- McMullen, M., Bergstrom, G., De Wolf, E., Dill-Macky, R., Hershman, D., Shaner, G., and Van Sanford, D. 2012. A unified effort to fight an enemy of wheat and barley: Fusarium head blight. *Plant Dis.* 96:1712-1728.
- McRoberts, N., Hughes, G., and Madden, L. V. 2003. The theoretical basis and practical application of relationships between different disease intensity measurements in plants. *Ann. Appl. Biol.* 142:191-211.
- Melendez-Torres, G. J., Bonell, C., and Thomas, J. 2015. Emergent approaches to the meta-analysis of multiple heterogeneous complex interventions. *BMC Med. Res. Methodol.* 15:47.
- Mirocha, C. J., Kolaczowski, E., Xie, W., Yu, H., and Jelen, H. 1998. Analysis of deoxynivalenol and its derivatives (batch and single kernels) using gas chromatography/mass spectrometry. *J. Agric. Food Chem.* 46:1414-1418.
- Möbius, N., and Hertweck, C. 2009. Fungal phytotoxins as mediators of virulence. *Curr. Opin. Plant Biol.* 12:390-398.
- Molenberghs, G., Burzykowski, T., Alonso, A., Assam, P., Tilahun, A., and Buyse, M. 2008. The meta-analytic framework for the evaluation of surrogate endpoints in clinical trials. *J. Stat. Plan. Inference* 138:432-449.
- Moraes, W. B., De Wolf, E. D., Shah, D. A., Salgado, J. D., Madden, L. V., and Paul, P. A. 2018. Using predictions from a Fusarium head blight risk assessment tool as predictors of the risk of deoxynivalenol contamination of wheat grain. (Abstr.) *Phytopathology* 108(suppl.):S1.298.
- Moraes, W. B., Madden, L. V., and Paul, P. A. 2017. Effects of sample size on Fusarium head blight index estimation and its relationship with deoxynivalenol accumulation in wheat. In *2017 National Fusarium Head Blight Forum*, Milwaukee, WI.
- Nutter, F. W., Jr. 2002. Disease assessment. Pages 312-323 in: *Encyclopedia of Plant Pathology*. O. C. Maloy and T. D. Murray, eds. John Wiley & Sons, New York.
- Ojiambo, P. S., Battilani, P., Cary, J. W., Blum, B. H., and Carbone, I. 2018. Cultural and genetic approaches to manage aflatoxin contamination: Recent insights provide opportunities for improved control. *Phytopathology* 108:1024-1037.
- Okoth, S., Rose, L. J., Ouko, A., Netshifhefe, N. E. I., Sila, H., and Viljoen, A. 2017. Assessing genotype-by-environment interactions in *Aspergillus ear rot* and pre-harvest aflatoxin accumulation in maize inbred lines. *Agronomy (Basel)* 7:86.
- Parnell, S., Gottwald, T. R., Cunniffe, N. J., Chavez, V. A., and van den Bosch, F. 2015. Early detection surveillance for an emerging plant pathogen: A rule of thumb to predict prevalence at first discovery. *Proc. Roy. Soc. B.* 282:47-55.
- Parnell, S., van den Bosch, F., Gottwald, T., and Gilligan, C. A. 2017. Surveillance to inform control of emerging plant diseases: An epidemiological perspective. *Annu. Rev. Phytopathol.* 55:591-610.
- Paul, P. A., Bradley, C. A., Madden, L. V., Dalla Lana, F., Bergstrom, G. C., Dill-Macky, R., Esker, P. D., Wise, K. A., McMullen, M., Grybauskas, A., Kirk, W. W., Milus, E., and Ruden, K. 2018a. Meta-analysis of the effects of QoI and DMI fungicide combinations on Fusarium head blight and deoxynivalenol in wheat. *Plant Dis.* 102:2602-2615.
- Paul, P. A., Bradley, C. A., Madden, L. V., Dalla Lana, F., Bergstrom, G. C., Dill-Macky, R., Wise, K. A., Esker, P. D., McMullen, M., Grybauskas, A., Kirk, W. W., Milus, E., and Ruden, K. 2018b. Effects of pre- and post-anthesis applications of demethylation inhibitor fungicides on Fusarium head blight and deoxynivalenol in spring and winter wheat. *Plant Dis.* 102:2500-2510.
- Paul, P. A., Lipps, P. E., Hershman, D. E., McMullen, M. P., Draper, M. A., and Madden, L. V. 2007. A quantitative review of tebuconazole effect on Fusarium head blight and deoxynivalenol content in wheat. *Phytopathology* 97:211-220.
- Paul, P. A., Lipps, P. E., Hershman, D. E., McMullen, M. P., Draper, M. A., and Madden, L. V. 2008. Efficacy of triazole-based fungicides for Fusarium head blight and deoxynivalenol control in wheat: A multivariate meta-analysis. *Phytopathology* 98:999-1011.
- Paul, P. A., Lipps, P. E., and Madden, L. V. 2005b. Relationship between visual estimates of Fusarium head blight intensity and deoxynivalenol accumulation in harvested wheat grain: A meta-analysis. *Phytopathology* 95:1225-1236.
- Paul, P. A., Lipps, P. E., and Madden, L. V. 2006. Meta-analysis of regression coefficients for the relationship between Fusarium head blight and deoxynivalenol content of wheat. *Phytopathology* 96:951-961.
- Paul, P. A., McMullen, M. P., Hershman, D. E., and Madden, L. V. 2010. Meta-analysis of the effects of triazole-based fungicides on wheat yield and test weight as influenced by Fusarium head blight intensity. *Phytopathology* 100:160-171.
- Paul, P. A., Salgado, J. D., Bergstrom, G. C., Bradley, C., Byamukama, E., Byrne, A. M., Chapara, V., Cummings, J. A., Chilvers, M. I., Dill-Macky, R., Friskop, A. J., Kleczewski, N. M., Madden, L. V., Nagelkirk, M., Stevens, J., Smith, M., Wegulo, S. N., Wise, K. A., and Yabwalo, D. 2019. Integrated effects of genetic resistance and prothioconazole + tebuconazole application timing on Fusarium head blight in wheat. *Plant Dis.* 103:223-237.
- Paul, P. M., El-Allaf, S. M., Lipps, P. E., and Madden, L. V. 2005a. Relationships between incidence and severity of Fusarium head blight on winter wheat in Ohio. *Phytopathology* 95:1049-1060.
- Prentice, R. L. 1989. Surrogate endpoints in clinical trials: Definitions and operational criteria. *Stat. Med.* 8:431-440.
- Reid, L. M., Stewart, D. W., and Hamilton, R. I. 1996. A 4-year study of the association between Gibberella ear rot severity and deoxynivalenol concentration. *J. Phytopathol.* 144:431-436.
- Riley, R. D., Abrams, K. R., Sutton, A. J., Lambert, P. C., and Thompson, J. R. 2007. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Med. Res. Methodol.* 7:3.
- Salgado, J. D., Lindsey, L. E., and Paul, P. A. 2017. Effects of row spacing and nitrogen rate on wheat grain yield and profitability as influenced by diseases. *Plant Dis.* 101:1998-2011.
- Salgado, J. D., Madden, L. V., and Paul, P. A. 2014. Efficacy and economics of integrating in-field and harvesting strategies to manage Fusarium head blight of wheat. *Plant Dis.* 98:1407-1421.
- Salgado, J. D., Madden, L. V., and Paul, P. A. 2015. Quantifying the effects of Fusarium head blight on grain yield and test weight in soft red winter wheat. *Phytopathology* 105:295-306.
- Savary, S., Djurle, A., Yuen, J., Ficke, A., Rossi, V., Esker, P. D., Fernandes, M. C., Del Ponte, E. M., Kumar, J., Madden, L. V., Paul, P., McRoberts, N., Singh, P. K., Huber, L., Pope de Vallavielle, C., Saint-Jean, S., and Willocquet, L. 2017a. A white paper on global wheat health based on scenario development and analysis. *Phytopathology* 107:1109-1122.
- Savary, S., Nelson, A. D., Djurle, A., Esker, P., Sparks, A., Amorim, L., Bergamin Filho, A., Caffi, T., Castilla, N., Garrett, K., McRoberts, N., Rossi, V., Yuen, J., and Willocquet, L. 2017b. Concepts, approaches, and avenues for modelling crop health and crop losses. *Eur. J. Agron.* 100:4-18.
- Savary, S., Teng, P. S., Willocquet, L., and Nutter, F. W., Jr. 2006. Quantification and modeling of crop losses: a review of purposes. *Annu. Rev. Phytopathol.* 44:89-112.
- Shah, D. A., Paul, P. A., De Wolf, E. D., and Madden, L. V. 2019. Predicting plant disease epidemics from functionally-represented weather series. *Philos. Tran. Roy. Soc. B* 374:20180273.
- Sinha, R. C., and Savard, M. E. 1996. Comparison of immunoassay and gas chromatography methods for the detection of the mycotoxin deoxynivalenol in grain samples. *Can. J. Plant Pathol.* 18:233-236.
- Sneller, C., Guttieri, M., Paul, P., Costa, J., and Jackwood, R. 2012. Variation for resistance to kernel infection and toxin accumulation in

- winter wheat infected with *Fusarium graminearum*. *Phytopathology* 102:306-314.
- Tabcart, J. M., Dance, S. L., Lawless, A. S., Nichols, N. K., and Waller, J. A. 2019. Improving the condition number of estimated covariance matrices. arXiv 1810.10984v4
- Tibaldi, F. S., Cortiñas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. 2003. Simplified hierarchical linear models for the evaluation of surrogate endpoints. *J. Stat. Comput. Simul.* 73:643-658.
- Tilahun, A., Pryseley, A., Alonso, A., and Molenberghs, G. 2007. Flexible surrogate marker evaluation from several randomized clinical trials with continuous endpoints, using R and SAS. *Comput. Stat. Data Anal.* 51: 4152-4163.
- Tkachuk, R., Dexter, J. E., Tipples, K. H., and Nowicki, T. W. 1991. Removal by specific gravity table of tombstone kernels and associated trichothecene from wheat infected with *Fusarium* head blight. *Cereal Chem.* 68:428-431.
- Wegulo, S. N., Bockus, W. W., Nopsa, J. H., De Wolf, E. D., Eskridge, K. M., Peiris, K. H. S., and Dowell, F. E. 2011. Effects of integrating cultivar resistance and fungicide application on *Fusarium* head blight and deoxynivalenol in winter wheat. *Plant Dis.* 95:554-560.
- Whitehead, A. 2002. *Meta-Analysis of Controlled Clinical Trials*. John Wiley & Sons, West Sussex, England.
- Williams, W. P., Ozkan, S., Ankala, A., and Windham, G. L. 2011. Ear rot, aflatoxin accumulation, and fungal biomass in maize after inoculation with *Aspergillus flavus*. *Field Crops Res.* 120:196-200.
- Willyerd, K. T., Li, C., Madden, L. V., Bradley, C. A., Bergstrom, G. C., Sweets, L. E., McMullen, M., Ransom, J. K., Grybauskas, A., Osborne, L., Wegulo, S. N., Hershman, D. E., Wise, K., Bockus, W. W., Groth, D., Dill-Macky, R., Milus, E., Esker, P. D., Waxman, K. D., Adee, E. A., Ebelhar, S. E., Young, B. G., and Paul, P. A. 2012. Efficacy and stability of integrating fungicide and cultivar resistance to manage *Fusarium* head blight and deoxynivalenol in wheat. *Plant Dis.* 96:957-967.
- Xu, X., Madden, L. V., and Edwards, S. G. 2014. Modeling the effects of environmental conditions on HT2 and T2 toxin accumulation in field oat grains. *Phytopathology* 104:57-66.
- Xu, X., and Nicholson, P. 2009. Community ecology of fungal pathogens causing wheat head blight. *Annu. Rev. Phytopathol.* 47:83-103.
- Yuen, G. Y., and Schoneweis, S. D. 2007. Strategies for managing *Fusarium* head blight and deoxynivalenol accumulation in wheat. *Int. J. Food Microbiol.* 119:126-130.