Regulatory Toxicology and Pharmacology 65 (2013) 251–258

Contents lists available at SciVerse ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Bayesian statistical approaches to compositional analyses of transgenic crops 2. Application and validation of informative prior distributions

Jay M. Harrison [a,*], Matthew L. Breeze [b], Kristina H. Berman [b], George G. Harrigan [b]

[a] Maritz Holdings Inc., 1395 North Highway Drive, Fenton, MO 63099, USA
[b] Regulatory Product Characterization and Safety Center, Monsanto Company, 800 North Lindbergh Blvd., St. Louis, MO 63167, USA

ARTICLE INFO

Article history:
Received 15 June 2012
Available online 20 December 2012

Keywords:
Bayesian statistics
Compositional analyses
GM soybean
Transgenic
Informative prior distributions

ABSTRACT

Bayesian approaches to evaluation of crop composition data allow simpler interpretations than traditional statistical significance tests. An important advantage of Bayesian approaches is that they allow formal incorporation of previously generated data through prior distributions in the analysis steps. This manuscript describes key steps to ensure meaningful and transparent selection and application of informative prior distributions. These include (i) review of previous data in the scientific literature to form the prior distributions, (ii) proper statistical model specification and documentation, (iii) graphical analyses to evaluate the fit of the statistical model to new study data, and (iv) sensitivity analyses to evaluate the robustness of results to the choice of prior distribution. The validity of the prior distribution for any crop component is critical to acceptance of Bayesian approaches to compositional analyses and would be essential for studies conducted in a regulatory setting. Selection and validation of prior distributions for three soybean isoflavones (daidzein, genistein, and glycitein) and two oligosaccharides (raffinose and stachyose) are illustrated in a comparative assessment of data obtained on GM and non-GM soybean seed harvested from replicated field sites at multiple locations in the US during the 2009 growing season.

1. Introduction

Advantages of Bayesian statistical methodology in the compositional assessment of genetically modified (GM) crops were first described by Harrison et al. (2011). Two key elements characterized this original paper: (1) establishment of quality expectations for model validation and documentation following the precedent set by the US Food and Drug Administration in its guidelines for Bayesian analyses for clinical trials of medical devices (US FDA, 2010) and (2) the use of *low-information* prior distributions in assessing crop component values. Here, we reiterate the importance of determining quality expectations in Bayesian analyses but now further expand on considerations in the selection and validation of *informative* prior distributions.

The prior distribution is a fundamental component of Bayesian analysis. It is a mathematical function that quantifies the degree of belief that a parameter in a statistical model can assume certain values. Prior distributions can be based on results from earlier experiments, they can be defined by physical constraints, or they may reflect the investigator's experiences and intuition. The prior distribution and the newly acquired data are combined to form a posterior distribution that is then used to make inferences about the parameters. See the Supplementary information section for further discussion on this topic.

Prior distributions can vary in the amount of information about a model parameter that is to be incorporated into the analyses. The most general description is given by a *noninformative* prior distribution, which does not favor any particular value of the parameter (Berry et al., 2011). Next, a *low-information* prior distribution provides some information about the likely values of a parameter in a model. For example, the mean of the prior distribution may correspond to a mean value that was previously observed, or bounds on the possible values may be established. A low-information prior distribution has a large variance to allow a wide range of possible parameter values, and such prior distributions contribute negligible information to the analyses (Ntzoufras, 2009). The most specific type of prior distribution is an *informative* prior distribution, which assigns more probability to values that are believed to be likely to occur. The variance of an informative prior distribution is large enough to allow plausible values from previous experience, but the probabilities that are assigned to extreme parameter values are very low.

The use of noninformative prior distributions for the means of GM and conventional varieties may seem appropriate, since such prior distributions seem to convey impartiality and objectivity. However, there are several arguments in favor of using informative prior distributions,

---

* Corresponding author.
E-mail address: jay.harrison@maritz.com (J.M. Harrison).

0273-2300/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.yrtph.2012.12.002

under the condition that those prior distributions are supported by previous findings, are derived in an appropriate and transparent manner, and are not selected solely for mathematical convenience. Fundamentally, Kruschke (2011) argued that Bayesian analyses with informative prior distributions that are based on previous research represent the potential for cumulative scientific progress. In a practical sense, informative prior distributions may be needed in order to perform the computations that are involved in a Bayesian analysis. According to Carlin and Louis (2009), in models where the number of parameters increases with the sample size, at least some of the prior distributions on the individual parameters must be informative. The addition of extra varieties or extra locations also increases the number of parameters in the statistical model, thus necessitating some prior information to be applied. Gelman et al. (2004) wrote, "We should never find ourselves seeking a noninformative prior distribution for large number of parameters." Also, computational difficulties may be encountered with large numbers of parameters. In fact, Spiegelhalter et al. (2007) suggested the use of informative prior distributions as a solution if a numerical overflow error occurs when fitting Markov Chain Monte Carlo (MCMC) models in WinBUGS software, which is often used for Bayesian analysis.

Harrison et al. (2011) focused on the use of low-information prior distributions in their study on protein and fat levels in seed harvested from GM soybean (see also Harrigan and Harrison, 2012). However, the arguments in favor of informative prior distributions are relevant to the comparative assessments of GM crop compositions. The use of informative prior distributions could be especially useful for crop components whose levels are well-established in the scientific literature to be within certain ranges and are known to be influenced by specific factors such as genotype or environment.

We apply these considerations to a comparative assessment of isoflavones and oligosaccharide levels in seed harvested from GM and non-GM soybean grown in multiple replicated field sites in the US during the 2009 growing season. Isoflavones and oligosaccharides are key components recommended by the Organisation of Economic Cooperation and Development (OECD) for analysis in comparative assessments of new soybean varieties. Isoflavones are considered to have "biochemical activity, including estrogenic, anti-estrogenic, and hypocholesterolemic effects" (OECD, 2001). The oligosaccharides raffinose and stachyose are considered antinutrients due to flatulence caused by their consumption (OECD, 2001). Both types of analytes offer good opportunities to demonstrate the utility of informative prior distributions. Unlike the percentages of protein and fat content that appeared previously in Harrison et al. (2011), levels of isoflavones do not have a known upper bound. Even though the oligosaccharides are measured as percentages, they are present in small amounts, and a low-information uniform prior distribution from (0%, 100%) unrealistically assigns high probability to large percentages that cannot occur in soybeans. Informative prior distributions that take advantage of the knowledge about levels of these important metabolites may allow more rigorous analyses of comparative assessments of new GM soybean varieties and their traditionally bred comparators. The objective and transparent construction of the informative prior distributions, as well as additional model validation to determine whether the chosen prior distribution was appropriate and robust, are detailed below.

2. Materials and methods

2.1. Biological material

Seed samples were collected from a new GM soybean variety grown in a total of eight replicated field sites in the major growing regions of the US in 2009. The field productions included the new GM soybean variety, the variety A3525 as a near-isogenic conventional control, and a total of sixteen conventional reference varieties to provide more information on compositional variability inherent to soybean. The eight sites for the production were planted in a randomized complete block design, with four replicates per site containing the new GM variety, the conventional control, and three reference varieties. The field productions were conducted according to normal agronomic practices and seed samples were harvested at physiological maturity.

2.2. Compositional analysis

2.2.1. Isoflavones

Isoflavone levels were determined using a method based upon Seo and Morr (1984) and Pettersson and Kiessling (1984) as reported in numerous previous compositional reports (e.g. Berman et al., 2009; Lundry et al., 2008).

2.2.2. Oligosaccharides

The raffinose and stachyose levels were determined using a method based upon Mason and Slover (1971), and Brobst (1972) as reported in numerous compositional reports (e.g. Berman et al., 2009; Lundry et al., 2008).

2.3. Criteria for establishing informative distributions

An overview of the establishment, testing and application of the informative prior distributions is presented in schematic form in Fig. 1.

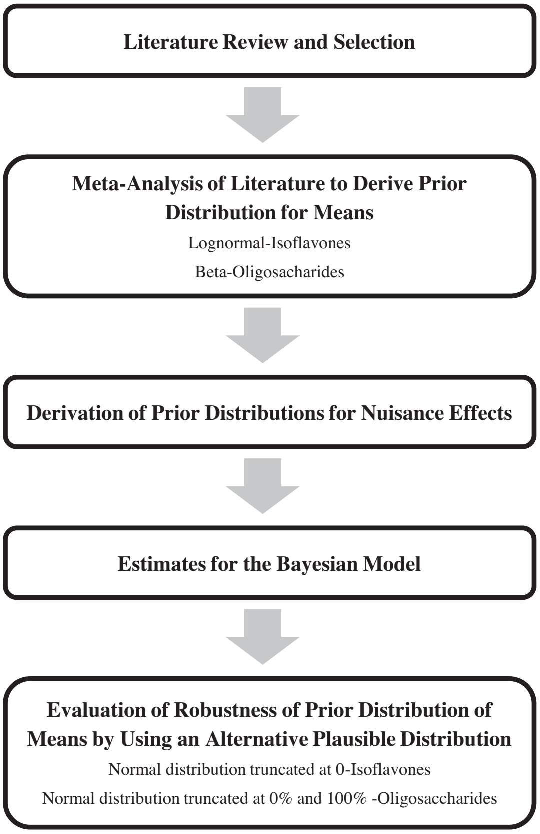

Fig. 1. Overview of selection and evaluation of informative prior distributions based on compositional data in the peer-reviewed scientific literature.