SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(9) 2876–2891
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280218790104
journals.sagepub.com/home/smm
SAGE

Optimal Bayesian point estimates and credible intervals for ranking with application to county health indices

Patricia I Jewett,[1] Li Zhu,[2] Bin Huang,[3] Eric J Feuer[2] and Ronald E Gangnon[1,4]

Abstract

It is fairly common to rank different geographic units, e.g. counties in the USA, based on health indices. In a typical application, point estimates of the health indices are obtained for each county, and the indices are then simply ranked as if they were known constants. Several authors have considered optimal rank estimators under squared error loss on the rank scale as a default method for general purpose ranking, e.g. situations where ranking units across the full spectrum of performance (low, medium, high) is important. While computationally convenient, squared error loss on the rank scale may not represent the true inferential goals of rank consumers. We construct alternative loss functions based on three components: (1) the inferential goal (rank position or pairwise comparisons), (2) the scale (original, log-transformed or rank) and (3) the (positional or pairwise) loss function (0/1, squared error or absolute error). We can obtain optimal ranks for loss functions based on rank positions and nearly optimal ranks for loss functions based on pairwise comparisons paired with highest posterior density (HPD) credible intervals. We compare inferences produced by the various ranking methods, both optimal and heuristic, using low birth weight data for counties in the Midwestern United States, from 2006 to 2012.

Keywords

Loss functions, rankings, visualization

1 Introduction

Rankings are a common tool to assess relative performance.[1–6] In public health, ranked health indices are used to compare geographical regions such as counties with each other. However, point estimates of health rankings do not convey any information about their uncertainty from measurement error. It is crucial to evaluate the uncertainty of estimates in order to judge their reliability, and to avoid incorrect conclusions.[7,8] Therefore, ranges of health rankings may represent a region's health indices better than point estimates of rankings,[9–13] and it may be desirable to shrink observed health indices towards a regional mean if the health indices are based on few measurements only.[14,15] The plausible range of rankings for a geographical region can be large especially if populations are small and cases rare, leading to greater measurement error and more unstable estimated case rates.[16,17] Also, point estimates can be misleading because a one-rank difference does not adequately quantify underlying health status differences and can mean different things: two similar regions may be one rank apart because of minimal, and possibly random, differences in case rates, or two dissimilar regions may be one rank apart because their case rates truly stand out from the other. Although these challenges in ranking have been

[1]Department of Population Health Sciences, University of Wisconsin-Madison, Madison, WI, USA
[2]Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[3]Department of Biostatistics, University of Kentucky, Lexington, KY, USA
[4]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA

Corresponding author:
Ronald E Gangnon, Department of Population Health Sciences, University of Wisconsin-Madison, 610 Walnut St, WARF, Room 603, Madison 53726, WI, USA.
Email: ronald@biostat.wisc.edu

previously identified and discussed, there is no consensus on the best method for conveying this uncertainty to a broad audience.

In a Bayesian framework, point estimates (of rankings) should be obtained by minimizing posterior expected loss functions. Squared error loss on the rank scale has been recommended for general purpose ranking, e.g. situations where ranking units across the full spectrum of performance (low, medium, high) is important.[18] The primary advantage of squared error loss on the rank scale over other potential loss functions for general purpose ranking is the simple, closed-form solution for the point estimate of the ranking, e.g. the ranks of the posterior mean ranks. In this paper, we demonstrate that the optimal ranking for any loss function based on weighted sums of unit-specific loss functions, which includes (weighted) squared error loss on the rank scale as a special case, can be found using the Hungarian algorithm for the linear sum assignment problem (LSAP).[19] The ready availability of optimal rankings for a general class of loss functions removes a large impediment to the use of appropriate, problem-specific loss functions. In addition, it facilitates our ability to consider novel loss functions that could serve as alternative choices for general purpose ranking instead of squared error loss on the rank scale. For example, squared (or absolute) error loss on the rate (or logit transform of the rate) rather than the rank scale may better capture the clinical and public health relevance of errors in ranking, because the transformation to ranks can magnify small differences and minimize large differences in health outcomes.

In this assessment, we used County Health Rankings data to demonstrate ranking solutions based on a variety of loss functions suitable for general purpose ranking. County Health Rankings have been reported annually by The University of Wisconsin Population Health Institute since 2003.[20,21] The goals and use of annual rankings may differ depending on the stakeholder. For example, rankings may help monitor and increase the awareness of public health issues across counties and states over time, identify factors that may influence the distribution of health indicators across regions, and facilitate the allocation of funds for health improvement efforts[21] by identifying regions with elevated or low health indicators.[14] Users of the county health rankings may be interested in the rank position of a specific county, the county at a specific ranking position, or the order of two or more counties relative to each other. Loss functions for the first two goals would typically be represented as sums of unit-specific loss functions, while loss functions for the third goal would be represented as sums of comparison-specific loss functions.

We complemented the County Health Rankings point estimates by estimating and visualizing their ranges and distributions. We designed our visualizations with the goal to enhance the understanding of ranking uncertainty, making the latter more accessible for a broad audience of public health researchers and decision makers. We primarily focussed on ranking counties within one state, but our methodology is applicable to other, smaller or larger geographical contexts, depending on the research question of interest. Our analyses used low birth weight (LBW) counts in Midwestern states between 2006 and 2012 as an exemplary health indicator. We used hierarchical regression analysis to estimate rate point estimates and their standard errors. Using empirical Bayes methods, we obtained posterior samples for the county rates (and ranks) and obtained optimal point estimates for the ranks for a variety of position-based loss functions using the Hungarian algorithm[19] and comparison-based loss functions using a heuristic algorithm.

2 Point estimation of ranks

Point estimates of ranks are most frequently obtained by simply ranking optimal point estimates (least squares, maximum likelihood, or posterior means) of the underlying unit-specific parameters,[2] posterior exceedence probabilities for a fixed, application-specific threshold[4] or traditional test statistics (Z-scores or p-values). Ranks based on maximum likelihood estimates of the unit-specific parameters are maximum likelihood estimates of the ranks, but, otherwise, ranking optimal point estimates of the parameters does not produce optimal point estimates of the ranks.[11] Laird and Louis[14] proposed using ranks of the posterior mean ranks (instead of ranks of the posterior mean parameters); these estimates are optimal under squared error loss on the rank scale.[10,18] Tailored loss functions for the identification of the top (or bottom) α% of all units have been explored.[6,18] For unit-specific 0/1 loss (0 if placed correctly inside/outside of the top α%, 1 if placed incorrectly), Lin et al.[18] identified ranking the posterior probability that the unit falls in the top (or bottom) α% of all units to be a (not necessarily unique) optimal ranking; the optimal rankings for different thresholds need not (and typically will not) be consistent with each other. Henderson and Newton[6] additionally required the proposed ranking be consistent across all possible thresholds.

Ideally, point estimation of ranks should be based on a loss function that corresponds to the inferential goals of ranking the counties, not computational convenience. Three possible inferential goals of ranking are (1) identifying the rank position of a specific county, which may be of interest to county health departments, (2) identifying the county at a specific rank position, which may be of interest to state health departments, and (3) identifying the correct ordering of a pair of counties, which may be of interest to either county or state health departments. We will refer to the first two, closely related goals as *ranking for position*, while we will refer to the third goal as *ranking for comparison*. Ranking for position attempts to place each county in the correct position within the overall ranking, and the overall loss function is constructed as the (weighted) average of position-specific or unit-specific loss functions, while ranking for comparisons attempts to place each county in the correct position relative to each other county, and the overall loss function is constructed as the (weighted) average of comparison-specific loss functions.

In our setting, K units (counties) are to be ranked on the basis of their associated (unknown) probabilities (or frequency) of an adverse health event, denoted as $p_1, p_2, \ldots, p_K$. We denote the ordered probabilities by $p_{(1)} < p_{(2)} < \cdots < p_{(K)}$. Inferences will be based on independent binomial samples from each unit, $y_j \sim \text{binomial}(n_j, p_j)$. Within a Bayesian framework, we assume access to (samples from) the (joint) posterior distribution of $p_1, p_2, \ldots, p_K$.

2.1 Loss functions for ranking for position

We first consider ranking for position. We will define the unit-specific loss associated with assigning rank k to unit j (or the position-specific loss associated with assigning unit j to rank k) by

$$\mathscr{L}(j,k) = |g(p_j) - g(p_{(k)})|^{\rho}$$

where $g(\cdot)$ is a monotone function and $\rho \in (0, \infty)$. Natural choices for g are the identity function (probability or rate scale), the logit function (log odds scale) and the rank function (rank scale). Typical choices for ρ would be $\rho = 1$ (absolute error loss), $\rho = 2$ (squared error loss) and $\rho = 0$ (all-or-nothing or 0-1 loss). For 0-1 loss, the choice of scale (g) does not matter.

The overall loss function for a proposed ranking $R_1, R_2, \ldots, R_K$ is either the (weighted) average of the unit-specific losses

$$\mathscr{L}_u(R_1, R_2, \ldots, R_k) = \sum_{j=1}^{K} w_u(j) L(j, R_j)$$

where $w_u(\cdot)$ is a user-supplied, unit-specific weight $\left(w_u(j) > 0, j = 1, 2, \ldots, K, \sum_{j=1}^{K} w_u(j) = 1\right)$ or the weighted average of position-specific losses

$$\mathscr{L}_p(R_1, R_2, \ldots, R_k) = \sum_{j=1}^{K} w_p(R_j) L(j, R_j)$$

where $w_p(\cdot)$ is a user-supplied, position-specific weight such that $w_p(k) > 0, k = 1, 2, \ldots, K, \sum_{k=1}^{K} w_p(k) = 1$. The weights must be strictly positive to ensure a valid overall loss function (zero if and only if the ranking is correct, greater than zero if the ranking is incorrect).

The choice of weights would reflect the relative importance of assigning the correct rank to unit j or assigning the correct unit to rank k. For example, one user might be particularly interested in the ranking of a particular unit (or set of units); such a user could assign weight 1 to that unit (or units) and weight $0 < f < 1$ to the remaining units. Another user might be particularly interested in identifying the units ranked in the top 10% of all units; such a user could assign weight 1 to rank positions $1, 2, \ldots, \lfloor K/10 \rfloor$ and weight $0 < f < 1$ to rank positions $\lfloor K/10 \rfloor + 1, \ldots, N$, where $\lfloor\ \rfloor$ is the floor function.

Given (samples from) the posterior distribution, we can calculate the posterior expected loss or risk associated with assigning rank k to unit j (or the position-specific loss associated with assigning unit j to rank k)

$$\mathscr{R}(j,k) = E\left\{\mathscr{L}(j,k) | y_1, y_2, \ldots, y_K\right\} = E\left\{|g(p_j) - g(p_{(k)})|^{\rho} | y_1, y_2, \ldots, y_K\right\}$$

The overall posterior risk function for a proposed ranking $R_1, R_2, \ldots, R_K$ is either

$$\mathscr{R}_u(R_1, R_2, \ldots, R_k) = \sum_{j=1}^{K} w_u(j) \mathscr{R}(j, R_j)$$

or

$$\mathscr{R}_p(R_1, R_2, \ldots, R_k) = \sum_{j=1}^{K} w_p(R_j)\mathscr{R}(j, R_j)$$

The individually optimal rank assignment for unit j is $R_j^{opt} = \text{argmin}_k \mathscr{R}(j,k)$. If the set of individually optimally rank assignments is a ranking, e.g. each rank assignment is unique, then it is the optimal ranking. If not, the optimal ranking is the solution to the (linear sum) assignment problem[19] with cost $c_{jk} = w_u(j)\mathscr{R}(j,k)$ or $c_{jk} = w_p(k)\mathscr{R}(j,k)$ of assigning county j to rank k. In its general form, the assignment problem has K *workers* (in our case, counties) and an equal number of *jobs* (in our case, ranks). Any worker (county) can be assigned to any job (rank), incurring some *cost* (in our case, the weighted risk). It is required to perform all jobs by assigning exactly one worker (county) to each job (rank) (in our case, produce a valid ranking) in such a way that the *total cost* (in our case, the overall risk function) is minimized. Given the cost matrix $C = (c_{jk})$, the optimal assignment is the permutation of the columns (or rows) that minimizes the sum of the diagonal elements of the matrix. The solution to the (linear sum) assignment problem or optimal ranking can be found using the Hungarian algorithm.[19] In practice, we use the solve_LSAP function in the clue package in R (originally intended to calculate partition proximities for cluster ensembles) to find the optimal ranking.[22]

2.2 Loss functions for ranking for comparison

We next consider ranking for comparison. We will define the comparison-specific loss associated with placing unit i before unit j by

$$\mathscr{L}(i,j) = \left(g(p_i) - g(p_j)\right)_+^{\rho}$$

where $(\cdot)_+ = \max(0, \cdot)$, $g(\cdot)$ is a monotone function and $\rho \in (0, \infty)$. Natural choices for g are the identity function (probability or rate scale), the logit function (log odds scale) and the rank function (rank scale). Typical choices for ρ would be $\rho = 1$ (absolute error loss), $\rho = 2$ (squared error loss) and $\rho = 0$ (all-or-nothing or 0–1 loss). For 0–1 loss, the choice of scale (g) does not matter.

The overall loss function for a proposed ranking $R_1, R_2, \ldots, R_K$ is the (weighted) average of the comparison-specific losses

$$\mathscr{L}_c(R_1, R_2, \ldots, R_k) = \sum_{i,j: R_i < R_j} w_c(i,j)L(i,j)$$

where $w_c(\cdot,\cdot)$ is a user-supplied, comparison-specific weight such that $w_c(i,j) > 0, w_c(i,j) = w_c(j,i), \sum_{i,j} w_c(i,j) = 1$. The weights must be strictly positive to ensure a valid overall loss function (zero if and only if the ranking is correct, greater than zero if the ranking is incorrect). One possible choice of comparison-specific weights is the average of the corresponding position-specific weights described previously, i.e. $w_c(i,j) = (w_u(i) + w_u(j))/2$.

Given (samples from) the posterior distribution, we can calculate the posterior expected loss or risk associated with placing unit i before unit (j)

$$\mathscr{R}(i,j) = E\{\mathscr{L}(i,j) | y_1, y_2, \ldots, y_K\} = E\left\{\left(g(p_i) - g(p_j)\right)_+^{\rho} | y_1, y_2, \ldots, y_K\right\}$$

The overall posterior risk function for a proposed ranking $R_1, R_2, \ldots, R_K$ is

$$\mathscr{R}_c(R_1, R_2, \ldots, R_k) = \sum_{i,j: R_i < R_j} w_c(i,j)\mathscr{R}(i,j)$$

At the comparison level, it is optimal to place unit i before unit j if $\mathscr{R}(i,j) < \mathscr{R}(j,i)$ and to place unit j before unit i if $\mathscr{R}(j,i) < \mathscr{R}(i,j)$. If the set of optimal assignments is transitive, then it is the optimal ranking. If not, the optimal ranking is the solution to the linear ordering (or triangulation) problem[23] with weights $c_{ij} = w_c(i,j)\mathscr{R}(i,j)$.[24] For a square matrix $C = (c_{ij})$, the triangulation problem is to find a simultaneous permutation of the rows and columns (in our case, counties) of C such that the sum of elements of the lower triangle (in our case, the overall risk) is as small as possible. For a complete directed graph on K nodes (in our case, counties) with weight c_{ij} for the arc (i, j) (from node i to node j) and weight c_{ji} for the arc (j, i), the equivalent linear ordering problem is to find an *acyclic tournament* or

Table 1. Loss functions.

Inferential Goal	Type	Scale	Label
Position	Absolute Error ($\rho = 1$)	Probability	*Pos_Abs_Rate*
		Logit	*Pos_Abs_Logit*
		Rank	*Pos_Abs_Rank*
	Squared Error ($\rho = 2$)	Probability	*Pos_Sel_Rate*
		Logit	*Pos_Sel_Logit*
		Rank	*Pos_Sel_Rank*
	0/1 Loss ($\rho = 0$)	Any	*Pos_01*
Comparison	Absolute Error ($\rho = 1$)	Probability	*Comp_Abs_Rate*
		Logit	*Comp_Abs_Logit*
		Rank	*Comp_Abs_Rank*
	Squared Error ($\rho = 2$)	Probability	*Comp_Sel_Prob*
		Logit	*Comp_Sel_Logit*
		Rank	*Comp_Sel_Rank*
	0/1 Loss ($\rho = 0$)	Any	*Comp_01*

ordering of the nodes (in our case, ranking) that minimizes the sum of the weights of the arcs included in the tournament (in our case, the overall risk) is as small as possible. Several exact and heuristic algorithms for solving the linear ordering problem are available in the literature. We adopt local search based on pairwise exchanges, a simple heuristic algorithm based on pairwise swapping of elements to find a local optimum.[23,24] Briefly, starting from an initial rank ordering, swaps of rank positions for pairs of counties were repeatedly considered until a local optimum (no improvement possible with a single swap of counties) was found.

In the remainder of the manuscript, we restrict consideration to 14 (unweighted) loss functions listed in Table 1.

While the rank scale is very useful for hypothesis testing making minimal assumptions, the rank scale is not as suited to estimating the magnitude of effects, and it is not necessarily an ideal basis for developing loss functions. To illustrate the potential advantages of a loss function based on the probability scale rather than the rank scale, we consider a small example of state-level rates of low birth weight births in 2016 for Wisconsin and its four neighboring states. The correct ranking of the states is Minnesota (6.6%), Iowa (6.8%), Wisconsin (7.4%), Illinois (8.4%), Michigan (8.5%). For any loss function based on the rank scale (position or comparison; squared error, absolute error or 0/1), the best alternatives to the correct ranking are the four rankings that swap adjacent counties. According to the loss functions, all of these rankings are considered equally good, even though the rankings that swap Iowa and Minnesota or Illinois and Michigan are intuitively better than the rankings that swap Wisconsin and Iowa or Wisconsin and Illinois. In contrast, using squared error loss on the probability scale, the best incorrect ranking is Minnesota (6.6%), Iowa (6.8%), Wisconsin (7.4%), Michigan (8.5%), and Illinois (8.4%), the second best incorrect ranking is Iowa (6.8%), Minnesota (6.6%), Wisconsin (7.4%), Illinois (8.4%), Michigan (8.5%), and the third best incorrect ranking is Iowa (6.8%), Minnesota (6.6%), Wisconsin (7.4%), Michigan (8.5%), Illinois (8.4%) ahead of the remaining two single swap rankings. The ranking Minnesota (6.6%), Wisconsin (7.4%), Iowa (6.8%), Illinois (8.4%), Michigan (8.5%) is fourth best, while the ranking Minnesota (6.6%), Iowa (6.8%), Illinois (8.4%), Wisconsin (7.4%), and Michigan (8.5%) is twelfth best. More generally, the loss functions based on the rank scale have relatively low resolution, 7–32 unique values for 120 possible rankings, compared to position-based squared error loss on the probability scale, 72 unique values, or comparison-based squared error loss on the probability scale, 118 unique values.

3 Uncertainty in ranking

To display the uncertainty and possible range of rank estimates, we used several techniques for visualizing posterior distributions and related distributional summaries.

(1) *Posterior distribution of percentile ranks within the Midwest census region*: We visualize the posterior distribution of the percentile rank for each county relative to the entire Midwest (1006 counties) using a density strip plot,[25] see Figure 1(a). The density strip plot is appropriate, because the percentile ranks are

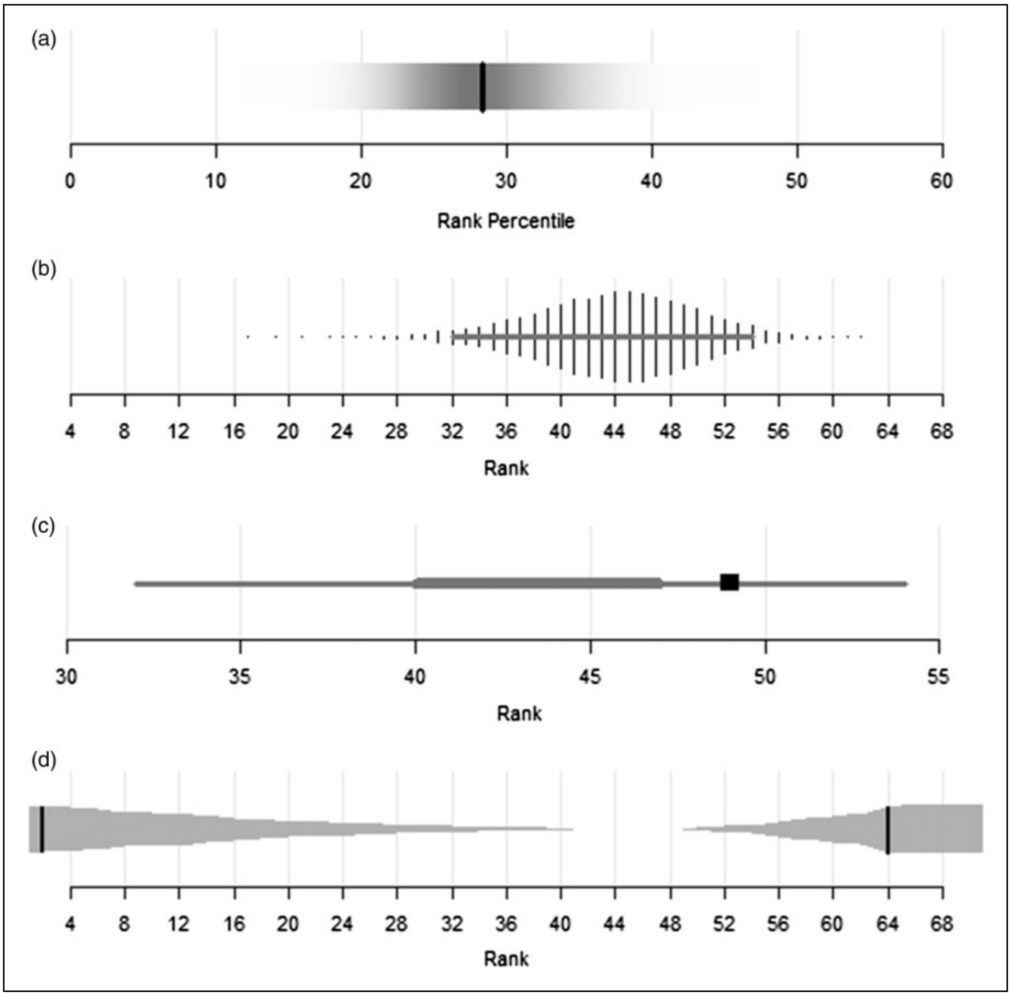

Figure 1. (a) Posterior distribution of percentile rank for Dane County, Wisconsin within the Midwest census region, (b) posterior distribution of (integer) rank for Dane County within the state of Wisconsin, (c) highest posterior density intervals (PDI) for the rank of Dane County within the state of Wisconsin, (d) simultaneous correct ranking probabilities for rank of Dane County within the state of Wisconsin based on the Pos_Sel_Logit ranking of Wisconsin counties.

essentially continuous given the large number of counties. The vertical bar represents the posterior mean percentile rank for the county within the Midwest region.

(2) *Posterior distribution of (integer) ranks within the state*: We visualize the posterior distribution of the rank for each county relative to their state (39–114 counties) using a variant of a violin plot[26] for discrete data along with the 95% highest posterior density interval (95 PDI), as shown in Figure 1(b). The height of the vertical bars represents the posterior probability of the rank for the given county. The thick horizontal bar represents the corresponding 95 PDI of the rank for the given county.

(3) *Highest posterior density intervals (PDI) for ranks within a state*: We visualize the 50% and 95% highest posterior density intervals (50 PDI, 95 PDI) of the rank of a given county using gray bars of varying thickness, as shown in Figure 1(c). The point estimate for the rank for the given county is plotted using a solid black square.

(4) *Simultaneous correct ranking probabilities for ranks within a state*: Given a point estimate of the complete ranking, for each given county, we calculate the probability that all counties identified as being ranked at or below (at or above) any lower (higher) rank position than the given county are simultaneously ranked correctly relative to the given county. We visualize these probabilities using vertical gray bars with widths proportional to the probability, as demonstrated in Figure 1(d). Vertical black lines denote the first rank positions for which the probability exceeds 97.5%. Unlike the previous displays, this display focuses on comparisons between counties rather than positions of individual counties.

In Figure 1 we demonstrate these visualizations for Dane County, Wisconsin; point estimates of the rankings are based on the Pos_Sel_Logit loss function. Dane County is at the 28th percentile (95% PDI 20–37) of Midwestern counties (Figure 1(a)); Dane County is in rank position 49 (95% PDI 32–54) within Wisconsin (Figure 1(b)). With at least 95% posterior probability, Dane County ranks below the county ranked first (Vernon County), which is the top ranked county according to the Pos_Sel_Logit ranking; with at least 95% posterior probability, it ranks above all counties ranked 64 or worse (Manitowoc, Outagamie, Waupaca, Rock, Winnebago, Kenosha, Racine, and Milwaukee) according to the Pos_Sel_Logit ranking (Figure 1(d)).

4 Data example: Midwest low birth weight data

We ran our analyses on low birth weight data in the Midwest. Using the 14 loss functions described above, we calculated alternative plausible rankings for the entire Midwest region, and state by state. For this presentation, we focused on Wisconsin counties. Findings for other Midwestern states and the whole Midwestern region can be seen in an online shiny app available at https://pijewett.shinyapps.io/shiny_main/.

4.1 Data sources

2006–2012 Data on low birth weight counts per county were available at the County Health Rankings website at http://www.countyhealthrankings.org/rankings/data. Data were available nation-wide; we used only counties from Midwestern states (Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin). In some counties, low birth weight counts were missing for data privacy protection reasons because their number was small. We left these data gaps as missing and only used the data from counties where low birth weight counts were available. To calculate case rates from the low birth weight count data, we used 2010 Census county population estimates available at http://www.census.gov/geo/reference/centersofpop.html.

4.2 Hierarchical logistic regression model

We used the gam( ) function in the R package mgcv to fit a hierarchical logistic regression model to the low birthweight data. The model included fixed state effects, a spatial trend surface (thin plate regression spline based on longitude and latitude), and random county effects to obtain empirical Bayes estimates for low birth weight rates for each county in order to account for and adjust unreliable raw rate measurements in small counties.[15] When running the estimation for a larger area than a state, the use of a spatial trend surface as a regional model improves the estimates within each state by taking into account counties outside the state borders, i.e. by accounting for spatial patterns that are not confined to individual states. Using 10,000 simulations from the joint posterior distribution of the county-level rates, we obtained posterior distributions of the rankings (and other parameters) by direct transformation. We repeated this procedure for all Midwestern states simultaneously, and each state individually.

4.3 Clustering of rankings

We used complete-linkage hierarchical cluster analysis to group the loss functions based on the similarity (Spearman squared correlation) of the optimal rankings within each state and across the entire Midwest using the varclus( ) function from the R Hmisc package (Harrell,[27] p.81). We used clustering to identify loss functions which produced similar rankings and could hence be plausibly viewed as substitutes for each other. Figure 2 shows the clustering of ranking methods for Illinois, Minnesota, Wisconsin and the entire Midwest region. The clustering results in each region are quite similar, so we will primarily discuss the findings using the Wisconsin plot in the bottom-right corner. We observe four major groupings of the ranking methods. Starting from the top, the Pos_01L ranking differs most dramatically from the other rankings and stands out as a very unique ranking (relatively low correlation with all of the other rankings). Next, the commonly used ranking using position-based squared error loss on the rank scale (Pos_Sel_Rank) is clustered with comparison-based loss functions using the rank scale (Comp_Sel_Rank and Comp_Abs_Rank). Third, rankings using position-based absolute error loss functions on all three scales (Pos_Abs_Rank, Pos_Abs_Rate and Pos_Abs_Logit) form a grouping. Rankings based on the remaining loss functions including comparison-based all-or-nothing loss (Comp_01L) and position-based squared error loss on the rate or logit scales (Pos_Sel_Rate or Pos_Sel_Logit) form a final grouping.

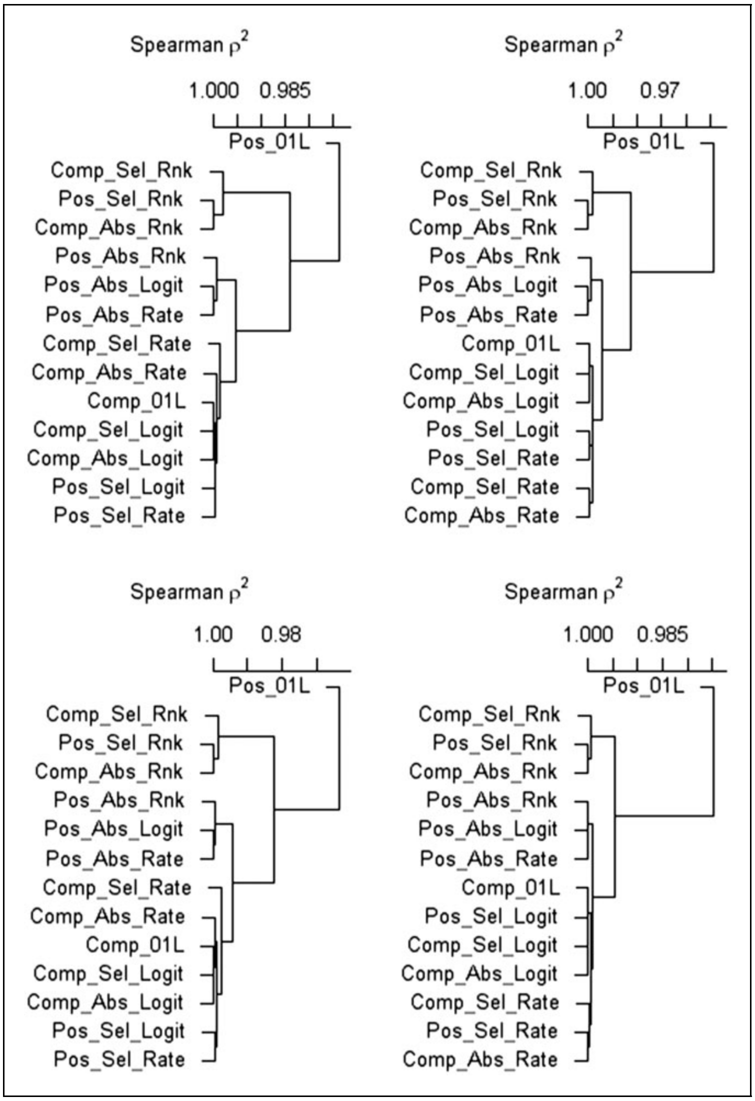

Figure 2. Clustering of correlations between different rankings across different regions, from top left clockwise: Illinois, Minnesota, Midwest region, Wisconsin.

These findings suggest that Comp_01L or Pos_Sel_Logit rankings may provide distinct inferences from the commonly used Pos_Sel_Rank ranking.

4.4 Results

Figure 3 shows a map of the rankings distribution in Wisconsin counties according to the Pos_Sel_Rnk optimized ranking. The lighter the shading of a county, the better was the ranking, i.e. the lower was the low birth weight rate in that county. Overall, the rural counties in the West of the state tended to have better rankings than the more urban counties in the East, with Milwaukee County having the worst ranking in the state. Iron County was not included in this analysis because of missing data on low birth weight births.

Table 2 shows rankings of the 71 Wisconsin counties across multiple loss functions. We chose to contrast the Pos_Sel_Rank with the Pos_Sel_Logit and Comp_01L rankings because our cluster analysis had found the latter two rankings to differ from the commonly used Pos_Sel_Rank ranking. Rankings based on the logit(rate) scale, or on all-or-nothing losses may not be as intuitive as losses on the rank scale. However, their distinctness from the Pos_Sel_Rank emphasizes their validity in their own right. Also, losses calculated on the (logit) scale use the scale of the original hierarchical logistic regression simulations, and from that point of view might be prefered depending on the application. Milwaukee County was ranked last by all loss functions, and its 95 PDI was 71,

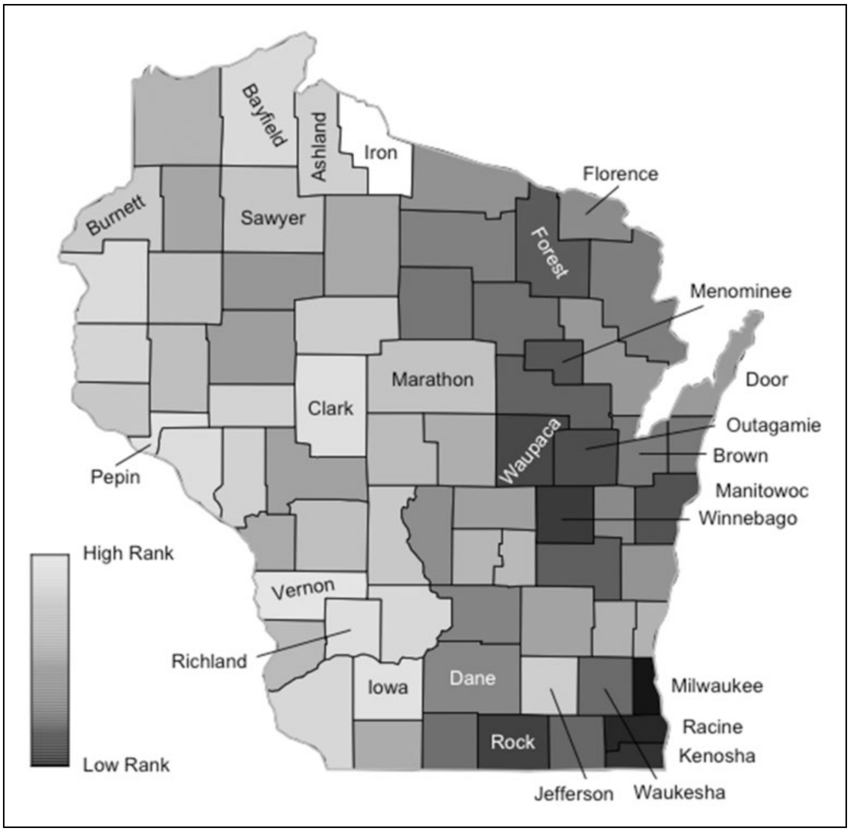

Figure 3. Map of Wisconsin low birth weight rankings according to the Pos_Sel_Logit ranking. Lower ranked counties have a darker shade. Iron county (white) was missing low birth weight data. Counties discussed in text are labeled.

unequivocally identifying Milwaukee County as the worst performing county in Wisconsin. The next four lowest ranked counties were also consistently ranked by all loss functions: Racine County second worst (95 PDI 69–70), Kenosha County third worst (95 PDI 68–70), Winnebago County fourth worst (95 PDI 62–69), and Rock County fifth worst (95 PDI 59–68). At the top of the rankings, Vernon County (95 PDI 1–31) was consistently ranked first and Richland County (95 PDI 1–41) was consistently ranked fifth; Pepin County (95 PDI 1–43), Iowa County (95 PDI 1–35), and Clark County (95 PDI 1–35) were ranked second through fourth in differing orders. Loss function-based rankings for these smaller counties differed from their observed rankings, where Vernon County would rank sixth, Pepin County first, Iowa county third, Clark County eighth, and Richland County seventh. More striking, Florence County would rank second based on its observed rate, but ranks 45th or 47th based on the loss functions. In terms of consistency of the rankings across the loss functions, we observe that the Pos_Sel_Rank ranking differs from the other two rankings for 38 of the 71 counties, while the other two rankings show more agreement (Pos_Sel_Logit differs from the other two rankings for 12 counties, Comp_01L differs from the other two rankings for 11 counties).

Figures 4 and 5 visualize the uncertainty in Wisconsin county rankings. Counties were ordered by the optimized rankings according to the Pos_Sel_Logit loss function. Panel a) displays the posterior distribution of LBW rates for each Wisconsin county. Panel b) shows the posterior distribution of percentile ranks within the entire Midwest region for each Wisconsin county; the bar marks the posterior mean percentile rank. There are counties within Wisconsin across the entire spectrum of rates of low birth weight, but the majority of Wisconsin counties had posterior distributions centered below the 50th percentile for the Midwest, i.e. Wisconsin counties typically ranked above than the median Midwestern county. The posterior distributions of the lowest ranked (and highly populated) counties, covering mostly the 75th to 95th percentile, were narrower than those of better ranked counties. Panel (c) shows the posterior distribution of the integer rank for each county within Wisconsin. With the exception of urban counties assigned the worst ranks, the posterior distributions for most counties spanned a large range of possible ranks, indicating a large degree of uncertainty regarding the ranking of many Wisconsin counties. Counties with adjacent point estimates of ranks tended to have similar posterior distributions.

Table 2. Wisconsin county rankings based on the Pos_Sel_Logit, Pos_Sel_Rank, and Comp_01L loss functions.

Rank	Pos_Sel_Logit	Pos_Sel_Rank	Comp_01L
1	Vernon	Vernon	Vernon
2	*Pepin*	Iowa	Iowa
3	*Iowa*	*Clark*	*Pepin*
4	Clark	*Pepin*	Clark
5	Richland	Richland	Richland
6	Buffalo	*Polk*	Buffalo
7	Polk	*Buffalo*	Polk
8	Bayfield	*Sauk*	Bayfield
9	Sauk	*St. Croix*	Sauk
10	*Grant*	*Eau Claire*	*St. Croix*
11	*St. Croix*	Grant	Grant
12	Eau Claire	*Bayfield*	Eau Claire
13	Trempealeau	Trempealeau	Trempealeau
14	Ashland	*Jefferson*	Ashland
15	Jefferson	*Ashland*	Jefferson
16	Taylor	*Marathon*	Taylor
17	*Burnett*	*Taylor*	*Juneau*
18	Juneau	Juneau	*Burnett*
19	Pierce	Pierce	Pierce
20	*Sawyer*	*Barron*	*Marathon*
21	*Marathon*	*Burnett*	*Sawyer*
22	Barron	*Monroe*	Barron
23	Monroe	*Sawyer*	Monroe
24	Dunn	*Wood*	Dunn
25	Wood	*Dunn*	Wood
26	Crawford	*Green Lake*	Crawford
27	Green Lake	*Crawford*	Green Lake
28	Marquette	Marquette	Marquette
29	Douglas	Douglas	Douglas
30	Ozaukee	Ozaukee	Ozaukee
31	Washington	Washington	Washington
32	Portage	Portage	*Lafayette*
33	Lafayette	Lafayette	*Portage*
34	La Crosse	La Crosse	*Price*
35	Price	Price	*La Crosse*
36	Washburn	Washburn	Washburn
37	Dodge	Dodge	Dodge
38	*Jackson*	Waushara	Waushara
39	*Waushara*	Jackson	Jackson
40	Chippewa	*Door*	Chippewa
41	*Door*	Oconto	Oconto
42	Oconto	*Chippewa*	Door
43	Rusk	Rusk	Rusk
44	Sheboygan	*Vilas*	Sheboygan
45	Vilas	*Florence*	Vilas
46	Adams	*Sheboygan*	Adams
47	Florence	*Adams*	Florence
48	Calumet	Calumet	Calumet
49	Dane	*Columbia*	Dane
50	Columbia	*Dane*	Columbia
51	Oneida	Oneida	Oneida
52	Marinette	Marinette	Marinette
53	Brown	*Kewaunee*	Brown
54	Kewaunee	*Langlade*	Kewaunee

(continued)

Table 2. Continued

Rank	Pos_Sel_Logit	Pos_Sel_Rank	Comp_01L
55	Lincoln	Lincoln	Lincoln
56	Langlade	*Brown*	Langlade
57	Green	Green	Green
58	Waukesha	*Shawano*	Waukesha
59	*Walworth*	*Forest*	*Shawano*
60	*Shawano*	Walworth	Walworth
61	Fond du Lac	*Menominee*	Fond du Lac
62	Forest	*Waukesha*	Forest
63	Menominee	*Fond du Lac*	Menominee
64	Manitowoc	Manitowoc	Manitowoc
65	Outagamie	*Waupaca*	Outagamie
66	Waupaca	*Outagamie*	Waupaca
67	Rock	Rock	Rock
68	Winnebago	Winnebago	Winnebago
69	Kenosha	Kenosha	Kenosha
70	Racine	Racine	Racine
71	Milwaukee	Milwaukee	Milwaukee

Note: Rankings that differ from the other two rankings are in *italics*.

Among counties with mid-level rankings, 95% PDIs covered virtually the entire range of possible ranks within Wisconsin. Panel (d) shows, for each county in Wisconsin, the probability that all counties identified as being ranked at or below (at or above) any lower (higher) rank position than the given county are simultaneously ranked correctly relative to the given county; the bars mark where the probability exceeds 97.5%. We observe that Milwaukee County is worse than all other Wisconsin counties, Racine County is worse than all other Wisconsin counties except for Milwaukee County, and Kenosha County is worse than counties ranked 55 (Lincoln County) or better. For the vast majority of Wisconsin counties, we are unable to pinpoint, with few exceptions, specific counties with better rates of low birth weight births.

Occasionally, estimated optimal ranks fell outside the corresponding 95% PDI. For rankings across the entire Midwest region, the optimal point estimate fell outside the 95% CI for nine counties using three loss functions (Pos_Sel_Rank, Comp_Abs_Rank and Comp_Sel_Rank) and for five additional counties using Comp_Sel_Rank alone; this phenomenon did not occur for any of the remaining 11 loss functions. The involved counties were generally highly populated counties in large metropolitan areas, e.g. Cook, Lake and Will County, Illinois (Chicago); Allen County, Indiana (Ft. Wayne); Marion County, Indiana (Indianapolis); Sedgwick County, Kansas (Witchita); Kent County, Michigan (Grand Rapids); Macomb and Oakland County, Michigan (Detroit); Hennepin County, Minnesota (Minneapolis); Jackson County, Missouri (Kansas City); St. Louis County, Missouri (St. Louis); Franklin County, Ohio (Columbus); and Milwaukee County, Wisconsin (Milwaukee).

5 Discussion

In our analyses, we estimated county health rankings based on multiple loss functions, and visualized their distributions. Estimated rankings and their distributions differed from observed ranks since differences in measurement error informed the hierarchical regression outcomes and simulations, while unaccounted for in the observed ranks. Across Midwestern states, metropolitan counties ranked the worst and had the narrowest rank distributions. Our visualizations showed substantial overlaps in ranking distributions, demonstrating the uncertainty of point estimates.

5.1 Loss functions

Differences in optimal point estimates of the rankings across the various loss functions demonstrate that there is no universally optimal ranking. Ideally, the choice of loss function, and therefore the point estimate of the ranking, should be based on the inferential goals of the end user of the ranking, and different users would use

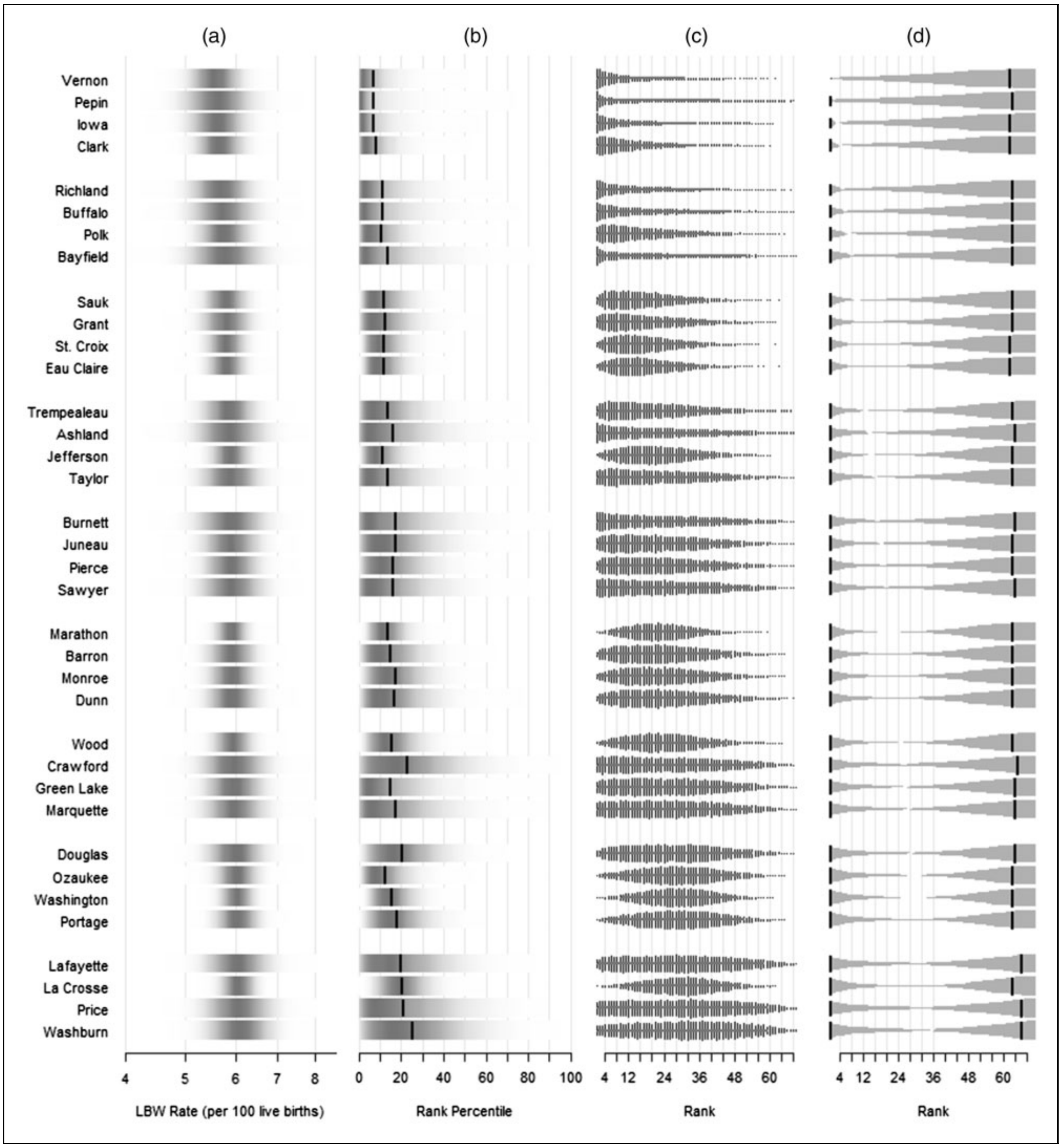

Figure 4. Highest ranked Wisconsin county results ordered according to Pos_Sel_Logit ranking.

different rankings. In situations with a single, primary consumer of the ranking, a problem-specific loss function should be used. In practice, practitioners will typically generate a single ranking that will be used by a variety of consumers with different inferential goals, requiring some consideration of appropriate choices for general purpose ranking. Squared error loss on the rank scale has previously been recommended as a good default choice for general purpose ranking,[18] but in our clustering analysis, ranks based on squared error loss on the rank scale are part of a cluster that is fairly distinct from most other ranking alternatives, which makes the choice of squared error loss rankings as a default appear somewhat arbitrary. Ranks based on comparison-wise 0/1 loss, squared error loss on the logit scale, squared error loss on the rate scale, and absolute error loss on the rank scale are distinct from ranks based on squared error loss on the rank scale, suggesting that if any of these ranks were chosen as a default, that choice would appear to be at least as justified as choosing the commonly used Pos_Sel_Rank ranking as a default. It is not the aim of this article to choose one ranking over another or to disparage the Pos_Sel_Rank per se, but rather to highlight a range of alternative ranking possibilities that have been previously neglected.

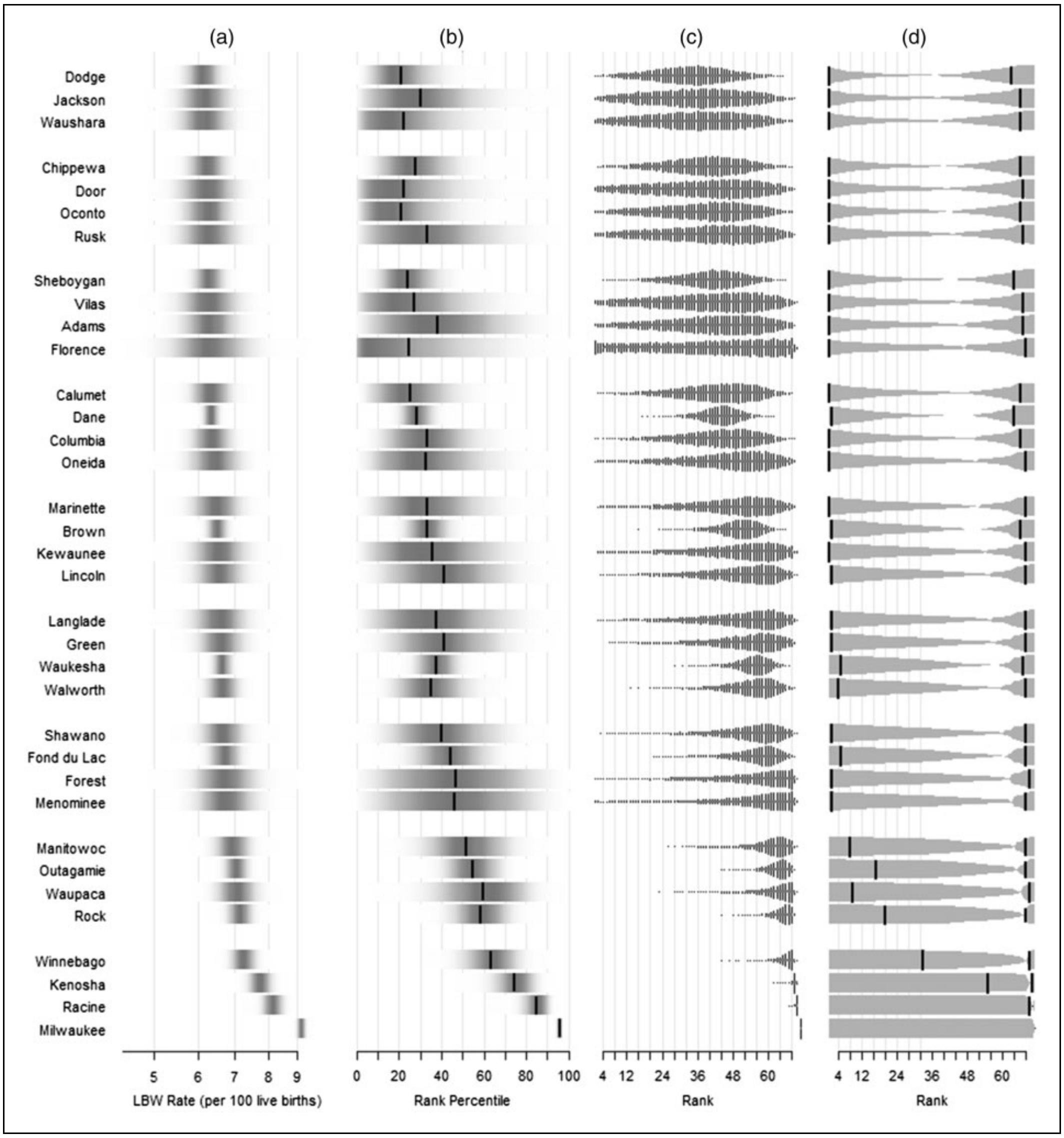

Figure 5. Lowest ranked Wisconsin county results ordered according to Pos_Sel_Logit ranking.

5.2 Visualizing uncertainty and ranking patterns

The visualizations that we used served different purposes. First, they emphasized the uncertainty in the estimates, for example in the rate and rank distribution plots by county. We used gradient and density plots which have been shown to be effective communicators of uncertainty.[28,29] Distribution similarities between counties uncovered by those plots are usually masked by cleanly separated rank point estimates. Second, our visualizations highlighted characteristic patterns: distribution plots contrasted counties with narrow vs. wide ranking distributions; and our percentile plots could be adopted to compare general ranking patterns between states. For example, Figure 6 shows the percentile plots for Wisconsin, Ohio, and Minnesota next to each other. Wisconsin counties were spread out over almost the entire Midwest percentile spectrum, but many county percentile distributions concentrated below the 50th Midwest percentile, creating a concavely shaped overall plot. Ohio had a convex pattern complementary to the concave Wisconsin pattern: most Ohio counties had narrower percentile distributions centered above the 50th Midwest percentile. Conversely, in Minnesota, almost all counties were not only centered below the 50th Midwest percentile, but the tails of many Minnesota counties hardly crossed the median line, and Hennepin as the worst ranked county exhibited a better percentile distribution than, for example, Milwaukee in Wisconsin. These overall patterns in comparison are effective in demonstrating that most Minnesota counties performed better than most Wisconsin counties with regards to low birth weight rates, while Ohio performed worst among these three states.

5.3 Variability in the estimates

Some counties displayed more variability in rankings than others. This was especially the case for counties with midlevel rankings. The reason is that while the data clearly indicated the top and bottom ranked counties, mid-level counties had more exchangeable posterior distributions for their rates (and therefore their ranks). The Pos_01L rankings differed the most from other loss function rankings, see Figure 2. Counties that still displayed noticeable variability in loss function rankings after excluding the Pos_01L or Comp_01L rankings were either large counties with rate and ranking distributions that were narrower than counties with comparable

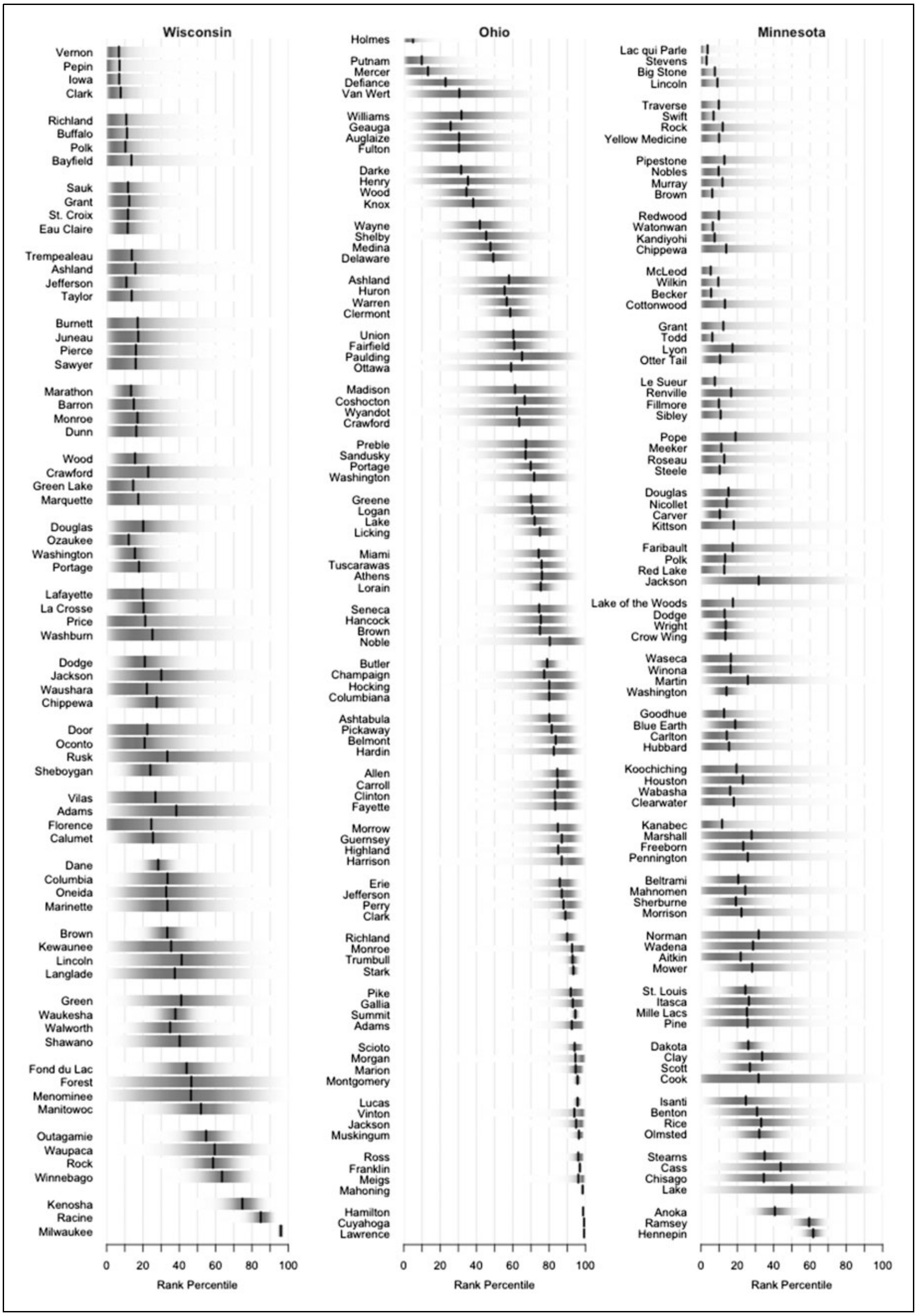

Figure 6. Rank percentile plots for Wisconsin, Ohio, and Minnesota in comparison, ordered according to the Pos_Sel_Logit ranking.

ranks (e.g. Brown, Dane, Jefferson, Marathon, Waukesha County in Wisconsin), or small counties with large measurement errors (e.g. Ashland, Bayfield, Burnett, Florence, and Sawyer County in Wisconsin).

Estimated ranks for some large counties were in the tails of their posterior distributions and occasionally even outside their 95% PDI. These discrepancies were often driven by numbers of small counties with roughly similar (posterior mean) rates to each other and to the larger county; the high posterior uncertainty regarding the underlying rates for the small counties can result in 'misplacement' of the larger county, if the small counties fall disproportionately to one side of its relatively narrow posterior distribution. Another cause of 'rank displacement' is 'competition' for the same rank. In Figure 1(c), we observe this phenomenon: the point estimate of the rank for Dane County under the Pos_Sel_Logit (or Pos_Seq_Rank) loss is outside its 50% PDI (albeit not outside its 95% PDI). While ranked 49 using the overall loss function based on Pos_Sel_Logit, the individually optimal rank for Dane County is 44. Overall, there are nine counties with individually optimal ranks between positions 38 through 43, forcing at least three of these counties to be assigned other nearby ranks in the overall ranking. The compromises required by a complete ranking make conflicts between some individual posterior distributions and their assigned rankings inevitable. The simplest way to avoid the point estimates falling outside of the corresponding credible intervals is to introduce an additional constraint, requiring the point estimate fall inside the interval, into the construction of the credible interval.

5.4 Applicability in cancer control

When assessing cancer rates in counties, rate point estimates may mask true similarities or differences.[30] Counties with different ranks may seem to imply an increased cancer problem in the lower ranked counties, but when taking ranking distributions and variability into account in addition to the point estimate, differences may virtually disappear. This may be relevant for the distribution of limited resources.[31] If ranking differences are random rather than attributable to real differences in underlying cancer rates, there is no evidence why resources should be allocated differently. Conversely, counties with similar ranks may mask a true increase in cancer rate which would merit further analysis. Thus, by simulating posterior distributions instead of using just the raw point estimates of cancer rates, it may become easier to truly identify cancer patterns, similarities, clusters, and outliers within a region of interest, which would help identify public health priorities for detailed investigations in cancer control.

6 Conclusion

Our analyses and visualizations highlight the uncertainty in rankings which is usually masked when only point rank estimates are presented. When comparing counties within a geographical region of interest, it may be known which counties perform best or worst, but everything in between may not be as clear. Hierarchical modeling and simulating posterior rate and ranking distributions takes into account that health indices in counties with small population counts are measured with greater error than in counties with larger populations, whereas differences in measurement error are disregarded in unadjusted observed rankings. Visualizations help compare distributions of health indices by region, and make overall patterns visible. Different loss functions produce rankings that are similar to each other, but vary in the specifics. Although in some case less intuitive than the commonly used Pos_Sel_Rank ranking, rankings based on the scale of the original data or the scale of the logistic regression come to different conclusions and therefore constitute ranking alternatives in their own right. Like the ranking distributions of each county which emphasize the uncertainty in the estimates, considering the range of possible optimal point rank estimates helps to bring the message across that there is no such thing as one best ranking. In future work, we will be collaborating with staff from the County Health Rankings and Roadmaps to develop a consensus on the inferential goals of their end users, to translate the inferential goals into a specific loss function (or loss functions), and to understand the implications of these choices for the ultimate rankings.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclose receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by contract HHSN261201500333P from the National Cancer Institute.

ORCID iD

Ronald E Gangnon http://orcid.org/0000-0003-2587-6714

References

1. Gibbons JD, Olkin I and Sobel M. An introduction to ranking and selection. *Am Stat* 1979; **33**: 185–195.
2. Aitkin M and Longford N. Statistical modelling issues in school effectiveness studies. *J Royal Stat Soc Ser A (General)* 1986; **149**: 1–43.
3. Goldstein H and Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J Royal Stat Soc Ser A (Stat Soc)* 1996; **159**: 385–443.
4. Normand SLT, Glickman ME and Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc* 1997; **92**: 803–814.
5. Paddock SM and Louis TA. Percentile-based empirical distribution function estimates for performance evaluation of healthcare providers. *J Royal Stat Soc: Ser C (Appl Stat)* 2011; **60**: 575–589.
6. Henderson NC and Newton MA. Making the cut: improved ranking and selection for large-scale inference. *J Royal Stat Soc: Ser B (Stat Methodol)* 2015; **78**: 781–804.
7. Potter K, Kniss J, Riesenfeld R, et al. Visualizing summary statistics and uncertainty. *Computer Graphics Forum* 2010; **29**(3): 823–832.
8. Shahian DM, Normand SL, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thoracic Surg* 2001; **72**: 2155–2168.
9. Zhang S, Luo J, Zhu L, et al. Confidence intervals for ranks of age-adjusted rates across states or counties. *Stat Med* 2014; **33**: 1853–1866, http://dx.doi.org/10.1002/sim.6071.
10. Shen W and Louis TA. Triple-goal estimates for disease mapping. *Stat Med* 2000; **19**: 2295–2308. http://dx.doi.org/10.1002/1097-0258(20000915/30)19:17/18&lt;2295::AID-SIM570&gt;3.0.CO;2-Q
11. Gelman A and Price PN. All maps of parameter estimates are misleading. *Stat Med* 1999; **18**: 3221–3234.
12. Xie M, Singh K and Zhang CH. Confidence intervals for population ranks in the presence of ties and near ties. *J Am Stat Assoc* 2009; **104**: 775–788.
13. Hall P and Miller H. Modeling the variability of rankings. *Ann Stat* 2010; **38**: 2652–2677.
14. Laird NM and Louis TA. Empirical bayes ranking methods. *J Educ Behav Stat* 1989; **14**: 29–46.
15. Berger JO and Deely J. A Bayesian approach to ranking and selection of related means with alternatives to analysis-of-variance methodology. *J Am Stat Assoc* 1988; **83**: 364–373.
16. Athens JK, Remington PL and Gangnon RE. Improving the rank precision of population health measures for small areas with longitudinal and joint outcome models. *PLoS ONE* 2015; **10**: 1–20, http://dx.doi.org/10.1371%2Fjournal.pone.0130027
17. Devine OJ and Louis TA. A constrained empirical Bayes estimator for incidence rates in areas with small populations. *Stat Med* 1994; **13**: 1119–1133.
18. Lin R, Louis TA, Paddock SM, et al. Loss function based ranking in two-stage, hierarchical models. *Bayes Analys* 2006; **1**: 915.
19. Kuhn HW. The Hungarian method for the assignment problem. *Naval Res Logistics Quarter* 1955; **2**: 83–97.
20. Peppard PE, Kindig DA, Dranger E, et al. Ranking community health status to stimulate discussion of local public health issues: the Wisconsin county health rankings. *Am J Public Health* 2008; **98**: 209–212.
21. Remington PL and Booske BC. Measuring the health of communities—how and why? *J Public Health Manage Practice* 2011; **17**: 397–400.
22. Hornik K. A CLUE for CLUster Ensembles. *J Stat Software* 2005; **14**(12): 1–25.
23. Laguna M, Martí R and Campos V. Intensification and diversification with elite tabu search solutions for the linear ordering problem. *Comput Operat Res* 1999; **26**: 1217–1230.
24. Martí R, Reinelt R and Duarte A. A benchmark library and a comparison of heuristic methods for the linear ordering problem. *Computat Optimiz Applicat* 2012; **51**: 1297–1317.
25. Jackson CH. Displaying uncertainty with shading. *The American Statistician* 2008; **62**(4): 340–347.
26. Hintze JL and Nelson RD. Violin plots: a box plot-density trace synergism. *Am Stat* 1998; **52**: 181–184.
27. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, 2nd ed. New York, NY: Springer, 2015.
28. Correll M and Gleicher M. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transact Visualizat Computer Graphics* 2014; **20**: 2142–2151.
29. Jackson CH. Displaying uncertainty with shading. *Am Stat* 2008; **62**: 340–347.
30. Devine OJ, Louis TA and Halloran ME. Empirical Bayes methods for stabilizing incidence rates before mapping. *Epidemiology* 1994; **5**: 622–630.
31. Brijs T, Karlis D, Van den Bossche F, et al. A Bayesian model for ranking hazardous road sites. *J Royal Stat Soc: Ser A (Stat Soc)* 2007; **170**: 1001–1017.