

1
2 Introducing riskCommunicator: an R package to obtain interpretable effect estimates for public
3 health

4
5 Jessica A. Grembi¹ and Elizabeth T. Rogawski McQuade^{2*}

6
7 ¹Division of Infectious Diseases and Global Health, Department of Medicine, Stanford University School
8 of Medicine, Stanford, CA, United States of America

9
10 ²Department of Epidemiology, Emory University, Atlanta, GA, United States of America

11
12

13 *Corresponding author

14 E-mail: erogaws@emory.edu

15

16

17 **Keywords** communication, g-computation, R package, risk differences, causal inference

18 Abstract

19 Common statistical modeling methods do not necessarily produce the most relevant or
20 interpretable effect estimates to communicate risk. Overreliance on the odds ratio and relative
21 effect measures limit the potential impact of epidemiologic and public health research. We
22 created a straightforward R package, called `riskCommunicator`, to facilitate the presentation
23 of a variety of effect measures, including risk differences and ratios, number needed to treat,
24 incidence rate differences and ratios, and mean differences. The `riskCommunicator` package
25 uses g-computation with parametric regression models and bootstrapping for confidence
26 intervals to estimate effect measures in time-fixed data. We demonstrate the utility of the
27 package using data from the Framingham Heart Study to estimate the effect of prevalent
28 diabetes on the 24-year risk of cardiovascular disease or death.

29 The absolute 24-year risk of cardiovascular disease or death was 30% (95% confidence
30 interval (CI): 22, 38) higher among subjects with diabetes compared to subjects without
31 diabetes at baseline. The relative 24-year risk was 55% (95% CI: 40, 70) higher. Because the
32 outcome was common (41.8%), the odds ratio (4.55) is highly inflated compared to the risk ratio
33 (1.55). An expected 4 additional persons would need to have diabetes at baseline to observe an
34 increase in the number of cases of cardiovascular disease or death by 1 over 24 years of follow-
35 up.

36 The package promotes the communication of public-health relevant effects and is
37 accessible to a broad range of epidemiologists and health researchers with little to no expertise
38 in causal inference methods or advanced coding.

39

40 **Background**

41 The communication of disease risk and the effects of exposures and interventions on
42 that risk are core components of public health research and practice. Unfortunately, reporting of
43 results from epidemiologic studies both in the published scientific literature and to the public is
44 often confused by imprecise language, jargon, and incomplete reporting [1,2]. While it may be
45 easiest to rely on the default output from standard functions in statistical programs, common
46 statistical methods estimate parameters that are often not the most informative. Epidemiologists
47 and the larger community of public health practitioners could benefit from easy-to-use tools to
48 facilitate the presentation of relevant effects.

49 Overreliance on the odds ratio [3–6] and more broadly on relative effect measures [7,8]
50 are two examples of opportunities to improve the reporting and interpretability of epidemiologic
51 results. Efforts to increase the reporting of difference effect measures and risk ratios over odds
52 ratios are not new, and several solutions have been previously proposed, including log-binomial
53 and log-linear regression models, Poisson regression to approximate log-binomial regression
54 when the latter does not converge [9], standardization-based approaches [10], linear-expit
55 regression [11], and ordinary least-squares regression with transformed variables [12].
56 However, these models are not as efficient as logistic regression, can have convergence
57 problems, and may require robust variance estimators [9,13].

58 Parametric g-computation is an attractive alternative because of the flexibility to estimate
59 a variety of effect measures while relying on the preferable statistical properties of logistic
60 regression for the parametric modeling. G-computation is conceptually equivalent to
61 standardization, and the use of parametric models allows for highly-dimensional data and
62 continuous covariates. G-computation has been applied to estimate risk differences and risk
63 ratios from logistic regression models previously [14–16].

64 Despite the availability of g-computation-based methods, these methods are rarely used
65 to estimate risk differences and risk ratios in standard time-fixed study designs. Recent
66 applications of these methods have focused on complicated study designs, such as with
67 longitudinal data with time-varying confounding affected by prior exposure [17]. In these
68 applications, the methods are complex and difficult to understand and/or implement for the
69 average data analyst. Coding requirements and computational limitations may also dissuade
70 users from attempting these methods. Recently available R packages [18,19] and Statistical
71 Analysis System (SAS) macros [20] are geared towards estimating these more complicated
72 effects and may be overwhelming to new users.

73 We aimed to create a straightforward R package, called `riskCommunicator`, to
74 facilitate the presentation of a variety of effect measures, including risk differences and ratios,
75 number needed to treat, incidence rate differences and ratios, and mean differences, using g-
76 computation. To make the package accessible to a broad range of health researchers, our goal
77 was to design functions that were as easy to use as the standard logistic regression functions in
78 R (e.g. `glm`) and that would require little to no expertise in causal inference methods or
79 advanced coding.

80

81 **Implementation**

82 The `riskCommunicator` package uses g-computation [21,22,16,17] with standard
83 parametric regression models and bootstrapping for confidence intervals to estimate effect
84 measures in the context of time-fixed exposure and outcome data. Broadly, the effects
85 estimated are average treatment effects (ATEs), estimated for difference measures with a
86 binary exposure variable as:

87
$$\varphi_{ATE} = \sum_w [E(Y | A = 1, W = w) - E(Y | A = 0, W = w)] * Pr(W = w),$$

88 where Y is the outcome of interest, A is the exposure of interest, and W are covariates. In this
89 way, the effects are standardized to the joint distribution of covariates in the total study
90 population.

91 The package contains two main functions available to end users: `gComp` (the primary
92 function) and `pointEstimate` (used internally within the `gComp` function, but provided to users
93 in case of complex dependencies among observations, e.g. nested clusters-within-clusters,
94 where a single cluster-level bootstrap resampling might not be optimal). `pointEstimate`
95 computes a point estimate by executing three steps of g-computation. First, a regression of the
96 outcome on the exposure and relevant covariates is fit using the provided dataset with a
97 generalized linear model. The underlying model distribution is based on the outcome type
98 supplied by the user (see `outcome.type` in Table 1 for details). Next, using the estimated
99 parameters from the model, counterfactuals are predicted for each observation in the data set
100 under each level of the exposure. Finally, the mean predicted value for each exposure regime
101 across all observations is calculated and used to estimate marginal difference and ratio effects.
102 The `gComp` function first estimates effects in the original data (using the `pointEstimate`
103 function). Then, bootstrap resampling of the original dataset is conducted, and the
104 `pointEstimate` function is called on each resample in order to estimate accurate standard
105 errors and provide a 95% confidence interval (CI). Confidence intervals are based on the 2.5th
106 and 97.5th percentiles of the bootstrap resampling results [23].

107

108 **Table 1. Arguments supplied to the `gComp` function in the `riskCommunicator` package.**

Argument	Description
<code>data</code>	(Required) A <code>data.frame</code> or tibble containing variables for Y, X, and Z or with variables matching the model variables specified in a user-supplied formula. Data set should also contain variables for the optional <code>subgroup</code> and <code>offset</code> , if they are specified.

outcome.type	<p>(Required) Character argument to describe the outcome type. Acceptable responses, and the corresponding error distribution and link function used in the <code>glm</code>, include:</p> <p>binary (Default) A binomial distribution with link = 'logit' is used.</p> <p>count A Poisson distribution with link = 'log' is used.</p> <p>count_nb A negative binomial distribution with link = 'log' is used, where the theta parameter is estimated internally; ideal for over-dispersed count data.</p> <p>rate A Poisson distribution with link = 'log' is used; ideal for events/person-time outcomes.</p> <p>rate_nb A negative binomial distribution with link = 'log' is used, where the theta parameter is estimated internally; ideal for over-dispersed events/person-time outcomes.</p> <p>continuous A gaussian distribution with link = 'identity' is used.</p>
formula	<p>(Optional) Default NULL. An object of class "formula" (or one that can be coerced to that class) which provides the complete model formula, similar to the formula for the <code>glm</code> function in R (e.g. 'Y ~ X + Z1 + Z2 + Z3'). Can be supplied as a character or formula object. If no formula is provided, Y and X must be provided.</p>
Y	<p>(Optional) Default NULL. Character argument which specifies the outcome variable. Can optionally provide a formula instead of Y and X variables.</p>
X	<p>(Optional) Default NULL. Character argument which specifies the exposure variable (or treatment group assignment), which can be binary, categorical, or continuous. This variable can be supplied as a factor variable (for binary or categorical exposures) or a continuous variable. For binary/categorical exposures, X should be supplied as a factor with the lowest level set to the desired referent. Numeric variables are accepted, but will be centered. Character variables are not accepted and will throw an error. Can optionally provide a formula instead of Y and X variables.</p>
Z	<p>(Optional) Default NULL. List or single character vector which specifies the names of covariates or other variables to adjust for in the <code>glm</code> function. All variables should either be factors, continuous, or coded 0/1 (i.e. not character variables). Does not allow interaction terms.</p>

subgroup	(Optional) Default NULL. Character argument that indicates subgroups for stratified analysis. Effects will be reported for each category of the subgroup variable. Variable will be automatically converted to a factor if not already.
offset	(Optional, only applicable for rate/count outcomes) Default NULL. Character argument which specifies the person-time denominator for rate outcomes to be included as an offset in the Poisson regression model. Numeric variable should be on the linear scale; function will take natural log before including in the model.
rate.multiplier	(Optional, only applicable for rate/count outcomes) Default 1. Numeric variable signifying the person-time value to use in predictions; the offset variable will be set to this when predicting under the counterfactual conditions. This value should be set to the person-time denominator desired for the rate difference measure and must be inputted in the units of the original offset variable (e.g. if the offset variable is in days and the desired rate difference is the rate per 100 person-years, <code>rate.multiplier</code> should be inputted as 365.25×100).
exposure.scalar	(Optional, only applicable for continuous exposure) Default 1. Numeric value to scale effects with a continuous exposure. This option facilitates reporting effects for an interpretable contrast (i.e. magnitude of difference) within the continuous exposure. For example, if the continuous exposure is age in years, a multiplier of 10 would result in estimates per 10-year increase in age rather than per a 1-year increase in age.
exposure.center	(Optional, only applicable for continuous exposure) Default TRUE. Logical or numeric value to center a continuous exposure. This option facilitates reporting effects at the mean value of the exposure variable, and allows for a mean value to be provided directly to the function in cases where bootstrap resampling is being conducted and a standardized centering value should be used across all bootstraps. See note below on continuous exposure variables for additional details.
R	(Optional) Default 200. The number of data resamples to be conducted to produce the bootstrap confidence interval of the estimate.
clusterID	(Optional) Default NULL. Character argument which specifies the variable name for the unique identifier for clusters. This option specifies that clustering should be accounted for in the calculation of confidence intervals. The <code>clusterID</code> will be used as the level for resampling in the bootstrap procedure.
parallel	(Optional) Default “no.” The type of parallel operation to be used. Available options (besides the default of no parallel processing) include “multicore” (not available for Windows) or “snow.” This argument is passed directly to <code>boot</code> . See note about setting seeds and parallel computing.
ncpus	(Optional, only used if parallel is set to “multicore” or “snow”) Default 1. Integer argument for the number of CPUs available for parallel processing/ number of parallel operations to be used. This argument is passed directly to <code>boot</code> .

110 Most users will only need to call the `gComp` function to estimate the effects of interest.
111 Arguments to be supplied are listed in Table 1 (and examples of how to call the function are
112 provided below in the Results section and S1 Appendix). Users can supply individual variable
113 names for the exposure, outcome, and covariates, or can provide a model formula. The `gComp`
114 function (and also `pointEstimate`) does not allow for interaction terms, however subgroup
115 analysis is possible by specifying the variable name in the dataset corresponding to the
116 subgroup classification, which automatically adds an interaction term between the subgroup
117 variable and the exposure to the model formula. Both functions also allow for the specification of
118 a categorical (in addition to binary) exposure. In cases of single-level clustered data, the `gComp`
119 function can conduct bootstrap resampling at the cluster, instead of individual sample, level by
120 specifying the variable identifying the cluster in the `clusterID` argument.

121 Output of the `gComp` function is a list with several pieces of data, including parameter
122 estimates and 95% confidence intervals for the effect measures (e.g. for a binary outcome, this
123 would include risk difference, risk ratio, odds ratio, and number needed to treat). Confidence
124 intervals are not reported for the number needed to treat given the primary utility of the number
125 needed to treat for communication and the challenges in construction and interpretation of the
126 number needed to treat confidence interval when the confidence interval for the risk difference
127 crosses the null [24,25]. Additional output includes marginal mean predicted outcomes for each
128 exposure level. Users can visualize the distribution of parameter estimates over all bootstrap
129 resamples of the data by plotting the resulting data with the base R `plot()` call to the output of
130 the `gComp` function, which provides a quantile-quantile plot [26] and histogram of all parameter
131 estimates (see S1 Appendix).

132 Bootstrap resampling is necessary to estimate accurate 95% confidence intervals for the
133 population-standardized marginal effects obtained with g-computation, since the standard errors
134 for the coefficients from the underlying parametric model (covariate-conditional effects) are no

135 longer applicable [16,22]. We recommend setting the number of bootstrap resamples (R) to
136 1000 for the final analysis. However, this can result in potentially long runtimes, depending on
137 the computing power of the user's computer (>30min). Thus, exploratory analyses can be
138 conducted with a lower number of bootstraps (default is $R = 200$, which should compute on
139 datasets of 5000-10000 observations in <60s).

140 Package code was written in R version 4.1.2 [27], and the package was built in RStudio
141 [28] using `devtools` and `roxygen2` to generate and populate the package documentation
142 [29,30]. `riskCommunicator` is open-source and freely available on GitHub
143 (<https://github.com/jgrembi/riskCommunicator>) and Comprehensive R Archive Network
144 (<https://CRAN.R-project.org/package=riskCommunicator>).

145

146 Results

147 We demonstrate the utility of `riskCommunicator` using the teaching data set from the
148 Framingham Heart Study [31], a prospective cohort study of cardiovascular disease conducted
149 in Framingham, Massachusetts. The use of these data for the purposes of this package were
150 approved on 11 March 2019 (request #7161) by National Institutes of Health/National Heart,
151 Lung, and Blood Institute. These data were altered prior to receipt by the authors to ensure an
152 anonymous dataset that protects patient confidentiality. This project was deemed by the
153 Institutional Review Board at Emory University to not be research with human subjects and
154 therefore did not require IRB review or consent from participants. The following analysis was
155 conducted among 4,240 participants who conducted a baseline exam and were free of
156 prevalent coronary heart disease when they entered the study in 1956. Participants were
157 followed for 24 years for the combined outcome of cardiovascular disease or death due to any
158 cause. A complete vignette highlighting the full range of analyses that are available with
159 `riskCommunicator` is available on Comprehensive R Archive Network (CRAN).

160 A relatively straightforward research aim for these data would be to estimate the effect of
 161 having prevalent diabetes at the beginning of the study on the 24-year risk of cardiovascular
 162 disease or death, adjusting for the potential confounders, including patient's age, sex, body
 163 mass index, smoking status (current smoker or not), and prevalence of hypertension. For a
 164 binary outcome, `riskCommunicator` estimates the risk difference, risk ratio, odds ratio, and
 165 number needed to treat. The output of the `gComp` function for this analysis as follows reports the
 166 strong effect of diabetes on cardiovascular disease and mortality (Table 2):

```
167 binary.res <- gComp(data = cvdd, Y = "cvd_dth", X =
168 "DIABETES", Z = c("AGE", "SEX", "BMI", "CURSMOKE", "PREVHYP"),
169 outcome.type = "binary", R = 1000)
```

172 **Table 2. Effect of prevalent diabetes at the beginning of the study on the 24-year risk of**
 173 **cardiovascular disease or death among 4,240 participants in the Framingham Heart**
 174 **Study.**

	<code>riskCommunicator</code>	Standard regression models*
	Marginal effect	Covariate-conditional effect
Effect measure	estimate [†] (95% CI)	estimate [†] (95% CI)
Risk difference	0.29 (0.20, 0.39)	N/A [‡]
Risk ratio	1.70 (1.48, 1.97)	1.49 (1.33, 1.66)
Odds ratio	4.55 (2.77, 9.09)	4.55 (2.66, 7.78)
Number needed to treat	3.48	N/A [‡]

175 *Log-linear regression for the risk difference, Poisson approximation of log-binomial regression with robust variance
 176 for the risk ratio, logistic regression for the odds ratio.

177 †Adjusted for patient's age, sex, body mass index (BMI), smoking status (current smoker or not), and prevalence of
 178 hypertension.

179 ‡Log-linear model did not converge

180

181 The absolute 24-year risk of cardiovascular disease or death due to any cause was 29%
182 (95% CI: 20, 40) higher among subjects with diabetes at baseline compared to subjects without
183 diabetes at baseline. The relative 24-year risk was 70% (95% CI: 48, 97) higher. Because the
184 outcome was common (41.8%), the odds ratio (4.55) is highly inflated compared to the risk ratio
185 (1.70). This is a clear example where the odds ratio may be misleading since the odds ratio is
186 commonly misinterpreted as a risk ratio. Furthermore, the relative effect may be interpreted as
187 much larger than the absolute effect, even though the absolute risk difference more closely
188 corresponds to the expected additional number of cases due to diabetes. For public health
189 communication, the number needed to treat derived from the risk difference (1/risk difference)
190 provides an easily interpreted estimate of the magnitude of effect. We would expect that only 4
191 additional persons would need to have diabetes at baseline to observe an increase in the
192 number of cases of cardiovascular disease or death by 1 over 24 years of follow-up.

193 We may also be interested in the effect of diabetes on the rate of cardiovascular disease
194 or death, incorporating person-time at risk. If the Framingham Heart Study were an open cohort
195 with variable follow-up time, rate-based effects would be more appropriate than risk-based
196 measures, which assume a constant follow-up period. In addition, we may be interested in
197 effects stratified by a potential effect measure modifier, such as participant sex.

198 `riskCommunicator` can estimate the incidence rate difference and incidence rate ratio by sex
199 for this analysis (Fig 1). As the person-time variable has units of days, rates are reported per
200 100 person-years by using the `rate.multiplier` option.

```
201   rate.res.subgroup <- gComp(data = cvdd.t, Y = "cvd_dth", X =  
202     "DIABETES", Z = c("AGE", "SEX", "BMI", "CURSMOKE", "PREVHYP"),  
203     subgroup = "SEX", outcome.type = "rate", rate.multiplier =  
204     365.25*100, offset = "timeout", R = 1000)
```

207 **Fig 1. Effect of having prevalent diabetes at the beginning of the study on the 24-year risk**
208 **of cardiovascular disease or death overall and stratified by sex among 4,240 participants**

209 **in the Framingham Heart Study.** A) Incidence rate ratio. B) Incidence rate difference.
210 `riskCommunicator` was used to obtain marginal effect estimates (purple) and Poisson
211 regression was used to obtain covariate-conditional estimates (green; not available for
212 incidence rate difference). All models were adjusted for patient's age, sex, body mass index,
213 smoking status (current smoker or not), and prevalence of hypertension. Each point represents
214 the point estimate and error bars show the 95% CI.

215
216 There is evidence for effect modification on the additive scale. The absolute rate of
217 cardiovascular disease or death due to any cause is 2.49 cases/100 person-years (95% CI:
218 1.34, 4.06) higher among males with diabetes compared to males without diabetes. In contrast,
219 the effect among women is smaller: the absolute rate of cardiovascular disease or death is 1.92
220 cases/100 person-years (95% CI: 1.00, 3.32) higher among females with diabetes compared to
221 females without diabetes. The relative effects suggest effect modification in the opposite
222 direction on the multiplicative scale, such that the effect of diabetes is stronger among females
223 compared to males. This difference is observed because the baseline rate of cardiovascular
224 disease and death without diabetes is higher among males (2.77 cases/100 person-years) than
225 females (1.64 cases/100 person-years), such that with the relative effect, the greater absolute
226 effect among males is diluted by their higher baseline risk.

227 The overall incidence rate ratio in the total study population (1.91, 95% CI: 1.60, 2.29)
228 can be estimated using the same code as above without the subgroup option. As expected, the
229 incidence rate ratio is further from the null than the risk ratio, but closer to the null than the odds
230 ratio (Table 1). This relationship among the magnitudes of these effect measures is expected
231 due to their mathematical properties, and specifically the differences in the denominators of risk
232 (total population), rates (person-time at risk), and odds (non-cases at the end of follow-up).
233 The estimation of these effects with standard regression models is not trivial. To estimate the
234 risk difference and risk ratio, we used log-binomial and log-linear regression, respectively.

235 However, in these data, both models fail to converge, and the Poisson approximation with
236 robust variance was necessary to estimate the risk ratio. The risk ratio estimate from g-
237 computation (confidence limit ratio: 1.33) had slightly lower precision compared to the estimate
238 from Poisson regression with robust variance (confidence limit ratio: 1.25). Minor differences in
239 the magnitude of the estimates can be attributed to the difference between the covariate-
240 conditional effects (as estimated by Poisson regression) and the marginal effects (as estimated
241 by `riskCommunicator`; Table 2). Poisson regression could also be used to estimate the
242 incidence rate ratios, resulting in equivalent magnitudes of estimates as those from
243 `riskCommunicator`, but slightly less precision (Table 3). Adjusted incidence rate differences
244 are not easy to estimate using standard regression models, but are readily available from
245 `riskCommunicator`.

246 Finally, an additional useful output of the package is the estimation of marginal mean
247 predicted outcomes for each exposure level. These predicted means are standardized over the
248 observed values of covariates included in the model, and therefore are not specific to set values
249 of the covariates. This difference is a major advantage over the usual `predict` function in R,
250 and similar functions in other statistical programs such as the `lsmeans` statement in Statistical
251 Analysis System (SAS), which can only predict outcomes at specific values of the other
252 covariates.

253

254 **Conclusions**

255 `riskCommunicator` facilitates the presentation of a wide range of effect measures with
256 a simple user experience, similar to running a linear regression model in R. For binary
257 outcomes, effects are modeled using logistic regression, which preserves the preferable
258 statistical qualities usually associated with odds ratios and applies them to the estimation of risk
259 ratios and risk differences. The package also facilitates the presentation of incidence rate

260 differences, which are difficult to obtain with standard generalized linear models. Finally, the
261 package supports assessment of additive effect measure modification by reporting difference
262 effects, which is important since contradictory evidence for effect modification between the
263 additive and multiplicative scales is common. While effect modification on the additive scale can
264 be more relevant to public health [32,33], it is often harder to estimate with standard regression
265 models [34,35].

266 It is important to highlight that the g-computation approach produces marginal rather
267 than covariate-conditional effect estimates. In a multivariable model, the effect estimates
268 derived directly from the covariate coefficients are covariate-conditional, interpreted as the
269 associations given constant values of the other variables (or informally, “holding all other
270 variables constant”) [32]. Covariate-conditional effects are difficult to interpret for non-collapsible
271 effect measures like the odds ratio [36]. Therefore, the reporting of marginal effects, in which the
272 effect is standardized over the covariate distribution of the total study population, may be
273 preferable in many cases. The marginal effect is interpreted as the average treatment effect in
274 the total population and is the primary effect of interest in randomized trials and in many
275 observational settings where causal inference is the goal [37].

276 One potential limitation to the g-computation approach is the use of bootstrap for the
277 confidence intervals. Bootstrapping is conservative compared to closed form solutions for the
278 variance (e.g. those used to estimate Wald confidence intervals), such that the confidence
279 intervals from bootstrapping can be slightly wider than alternatives. However, in the examples
280 above, precision improved for the rate ratios. In addition, the precision loss is often not extreme
281 when it occurs, and bootstrapped confidence intervals are more appropriate when the
282 distributional assumptions or approximations of the parameter may not be valid [32]. By using
283 percentiles of the simulated distribution of estimates from the bootstrap, one can avoid the need
284 to calculate the standard deviation of estimates under the normal distribution assumption [38].

285 The use of bootstrap allows for flexibility to estimate many effects with the same framework,
286 including allowing for clustering with bootstrap at the cluster level.

287 The g-computation approach can also be limited in settings with a continuous exposure
288 variable. For example, for a binary outcome, because the underlying parametric model is logistic
289 regression, the risks will be estimated to be linear on the log-odds (logit) scale, such that the
290 odds ratio for any one unit increase in the continuous variable is constant. However, the risks
291 will not be linear on the linear (risk difference) or log (risk ratio) scales, such that these
292 parameters will not be constant across the range of the continuous exposure. The g-
293 computation approach requires setting one specific exposure contrast within the range of the
294 continuous exposure. Therefore, users should be aware that the risk difference, risk ratio,
295 number needed to treat (for a binary outcome) and the incidence rate difference (for a
296 rate/count outcome) reported do not necessarily apply across the entire range of the continuous
297 exposure. We mitigate this issue by reporting the estimates for a relevant contrast within the
298 exposure variable by first centering the variable at the mean and allowing users to specify a
299 scaling factor for the contrast.

300 While other software packages are available to conduct more complex analyses with the
301 g-computation approach, `riskCommunicator` has been designed to be more accessible to the
302 average data analyst. For example, the GFORMULA macro for SAS [20] and the `gfoRmula`
303 package in R [18] are targeted to longitudinal data with time-varying covariates. The `qgcomp`
304 package combines g-computation with weighted quantile sum regression to estimate the effects
305 of mixtures [39]. The `tmle3` package in R includes g-computation but is designed to enable a
306 more comprehensive set of analyses to estimate Targeted Minimum Loss-Based Estimation
307 (TMLE) parameters [19], which requires advanced training even for doctoral-level
308 epidemiologists. The focus of `riskCommunicator` alternatively is on facilitating the
309 presentation of relevant and interpretable effect measures in relatively simple time-fixed

310 settings. The application of g-computation in these more traditional settings can help overcome
311 the gap for less experienced users between traditional regression modeling-based methods and
312 the g-methods, which are at the vanguard of epidemiologic methods development [40]. More
313 importantly, `riskCommunicator` can facilitate the communication of effects of exposures and
314 interventions and ultimately further the public health impact of epidemiologic and statistical
315 research.

316

317 **Acknowledgements**

318 We thank Dr. Charlie Poole for his feedback on key functions of the package. We thank Helen
319 Pitchik and Solis Winters for beta testing and useful suggestions on formatting results for
320 downstream plotting functionality.

321

322 **References**

- 323 1. Pocock SJ, Collier TJ, Dandreo KJ, de Stavola BL, Goldman MB, Kalish LA, et al. Issues in
324 the reporting of epidemiological studies: a survey of recent practice. *BMJ*. 2004;329: 883.
325 doi:10.1136/bmj.38250.571088.55
- 326 2. Vandembroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al.
327 Strengthening the Reporting of Observational Studies in Epidemiology (STROBE):
328 explanation and elaboration. *Epidemiology*. 2007;18: 805–835.
329 doi:10.1097/EDE.0b013e3181577511
- 330 3. Knol MJ, Le Cessie S, Algra A, Vandembroucke JP, Groenwold RHH. Overestimation of
331 risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression.
332 *Canadian Medical Association Journal*. 2012;184: 895–899. doi:10.1503/cmaj.101715
- 333 4. Holcomb WL, Chaiworapongsa T, Luke DA, Burgdorf KD. An odd measure of risk: use and
334 misuse of the odds ratio. *Obstet Gynecol*. 2001;98: 685–688. doi:10.1016/s0029-
335 7844(01)01488-0
- 336 5. Persoskie A, Ferrer RA. A Most Odd Ratio. *Am J Prev Med*. 2017;52: 224–228.
337 doi:10.1016/j.amepre.2016.07.030
- 338 6. Katz KA. The (Relative) Risks of Using Odds Ratios. *Arch Dermatol*. 2006;142: 761–764.
339 doi:10.1001/archderm.142.6.761

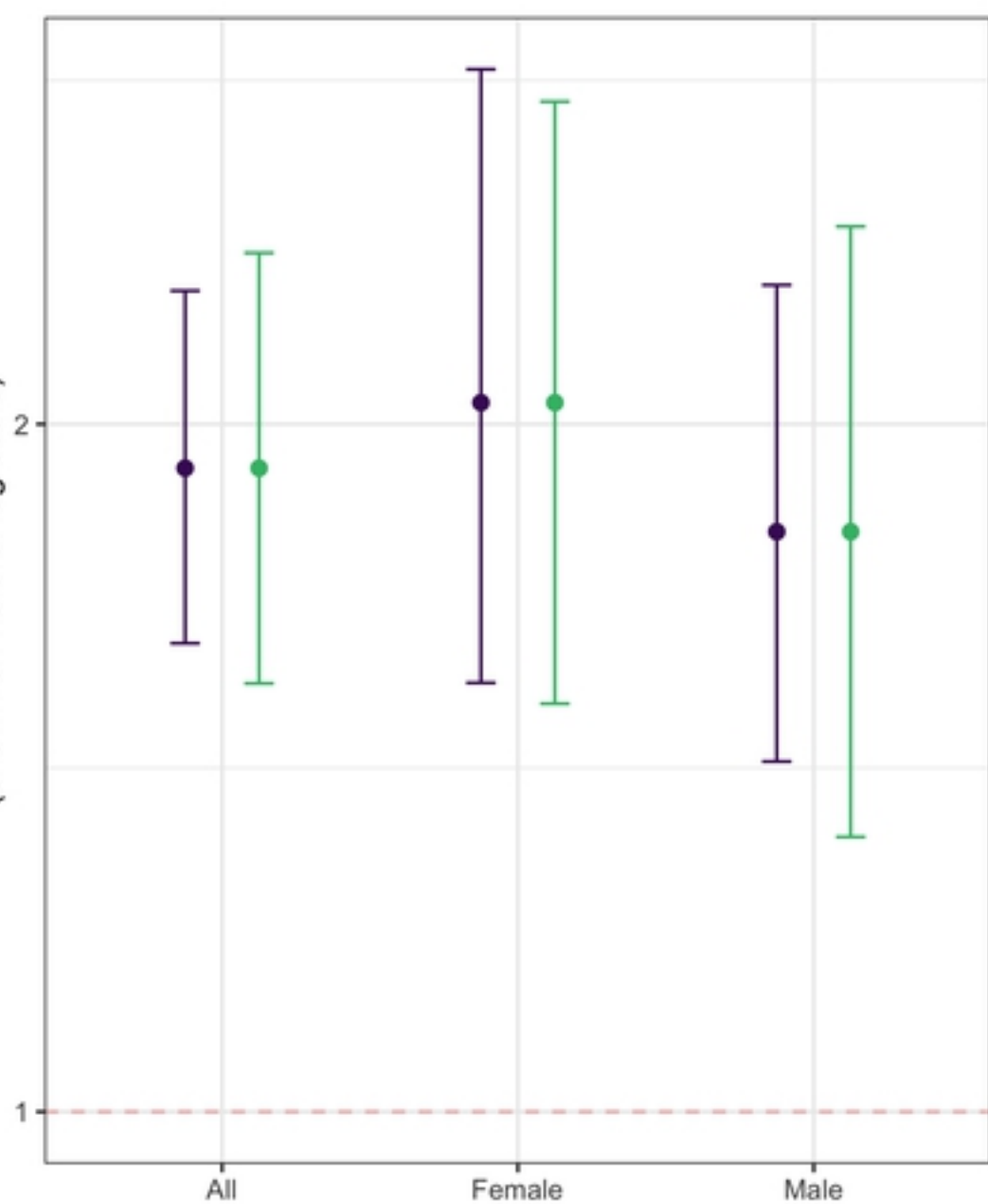
- 340 7. Schwartz LM, Woloshin S, Dvorin EL, Welch HG. Ratio measures in leading medical
341 journals: structured review of accessibility of underlying absolute risks. *BMJ*. 2006;333:
342 1248. doi:10.1136/bmj.38985.564317.7C
- 343 8. Noordzij M, van Diepen M, Caskey FC, Jager KJ. Relative risk versus absolute risk: one
344 cannot be interpreted without the other. *Nephrol Dial Transplant*. 2017;32: ii13–ii18.
345 doi:10.1093/ndt/gfw465
- 346 9. Zou G. A modified poisson regression approach to prospective studies with binary data.
347 *Am J Epidemiol*. 2004;159: 702–706.
- 348 10. Greenland S. Estimating standardized parameters from generalized linear models. *Stat*
349 *Med*. 1991;10: 1069–1074. doi:10.1002/sim.4780100707
- 350 11. Kovalchik SA, Varadhan R, Fetterman B, Poitras NE, Wacholder S, Katki HA. A general
351 binomial regression model to estimate standardized risk differences from binary response
352 data. *Stat Med*. 2013;32: 808–821. doi:10.1002/sim.5553
- 353 12. Xu Y, Cheung YB, Lam KF, Tan SH, Milligan P. A Simple Approach to the Estimation of
354 Incidence Rate Difference. *Am J Epidemiol*. 2010;172: 334–343. doi:10.1093/aje/kwq099
- 355 13. Yelland LN, Salter AB, Ryan P. Relative Risk Estimation in Randomized Controlled Trials:
356 A Comparison of Methods for Independent Observations. *The International Journal of*
357 *Biostatistics*. 2011;7: 1–31. doi:10.2202/1557-4679.1278
- 358 14. Localio AR, Margolis DJ, Berlin JA. Relative risks and confidence intervals were easily
359 computed indirectly from multivariable logistic regression. *J Clin Epidemiol*. 2007;60: 874–
360 882. doi:10.1016/j.jclinepi.2006.12.001
- 361 15. Bieler GS, Brown GG, Williams RL, Brogan DJ. Estimating model-adjusted risks, risk
362 differences, and risk ratios from complex survey data. *Am J Epidemiol*. 2010;171: 618–
363 623. doi:10.1093/aje/kwp440
- 364 16. Ahern J, Hubbard A, Galea S. Estimating the effects of potential public health interventions
365 on population disease burden: a step-by-step illustration of causal inference methods. *Am*
366 *J Epidemiol*. 2009;169: 1140–1147. doi:10.1093/aje/kwp015
- 367 17. Westreich D, Cole SR, Young JG, Palella F, Tien PC, Kingsley L, et al. The parametric g-
368 formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or
369 death. *Stat Med*. 2012;31: 2000–2009. doi:10.1002/sim.5316
- 370 18. Lin V, McGrath S, Zhang Z, Logan RW, Petito LC, Young JC, et al. gfoRmula: Parametric
371 G-formula. 30 Jan 2020 [cited 7 Feb 2020]. Available: [https://cran.r-](https://cran.r-project.org/web/packages/gfoRmula/index.html)
372 [project.org/web/packages/gfoRmula/index.html](https://cran.r-project.org/web/packages/gfoRmula/index.html)
- 373 19. Coyle J. R/tmle: The Extensible TMLE framework. [cited 7 Feb 2020]. Available:
374 <https://tlverse.org/tmle3/>
- 375 20. Harvard Program on Causal Inference: Software. [cited 7 Feb 2020]. Available:
376 <https://www.hsph.harvard.edu/causal/software/>

- 377 21. Robins J. A new approach to causal inference in mortality studies with a sustained
378 exposure period—application to control of the healthy worker survivor effect. *Mathematical*
379 *Modelling*. 1986;7: 1393–1512. doi:10.1016/0270-0255(86)90088-6
- 380 22. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data
381 set: demonstration of a causal inference technique. *Am J Epidemiol*. 2011;173: 731–738.
382 doi:10.1093/aje/kwq472
- 383 23. Efron B, Tibshirani R. *Bootstrap Methods for Standard Errors, Confidence Intervals, and*
384 *Other Measures of Statistical Accuracy*. *Statistical Science*. 1986;1: 54–75.
- 385 24. Hutton JL. Number needed to treat: properties and problems. *Journal of the Royal*
386 *Statistical Society: Series A (Statistics in Society)*. 2000;163: 381–402. doi:10.1111/1467-
387 985X.00175
- 388 25. Stang A, Poole C, Bender R. Common problems related to the use of number needed to
389 treat. *Journal of Clinical Epidemiology*. 2010;63: 820–825.
390 doi:10.1016/j.jclinepi.2009.08.006
- 391 26. Wilk MB, Gnanadesikan R. Probability Plotting Methods for the Analysis of Data.
392 *Biometrika*. 1968;55: 1–17. doi:10.2307/2334448
- 393 27. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria:
394 R Foundation for Statistical Computing; 2013. Available: <http://www.R-project.org/>
- 395 28. RStudio Team. *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc.; 2015.
396 Available: <http://www.rstudio.com/>
- 397 29. Wickham H, Chang W. *devtools: Tools to Make Developing R Packages Easier*. 2017.
398 Available: <https://CRAN.R-project.org/package=devtools>
- 399 30. Wickham H, Danenberg P, Eugster M. *roxygen2: In-Line Documentation for R*. Available:
400 <https://CRAN.R-project.org/package=roxygen2>
- 401 31. National Heart, Lung, and Blood Institute. *Teaching Datasets*. [cited 11 Mar 2019].
402 Available: <https://biolincc.nhlbi.nih.gov/teaching/>
- 403 32. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA:
404 Lippincott William & Wilkins; 2008.
- 405 33. VanderWeele TJ. *Explanation in Causal Inference: Methods for Mediation and Interaction*.
406 New York, NY: Oxford University Press; 2015.
- 407 34. Hosmer DW, Lemeshow S. Confidence Interval Estimation of Interaction. *Epidemiology*.
408 1992;3: 452.
- 409 35. Richardson DB, Kaufman JS. Estimation of the Relative Excess Risk Due to Interaction
410 and Associated Confidence Bounds. *Am J Epidemiol*. 2009;169: 756–760.
411 doi:10.1093/aje/kwn411

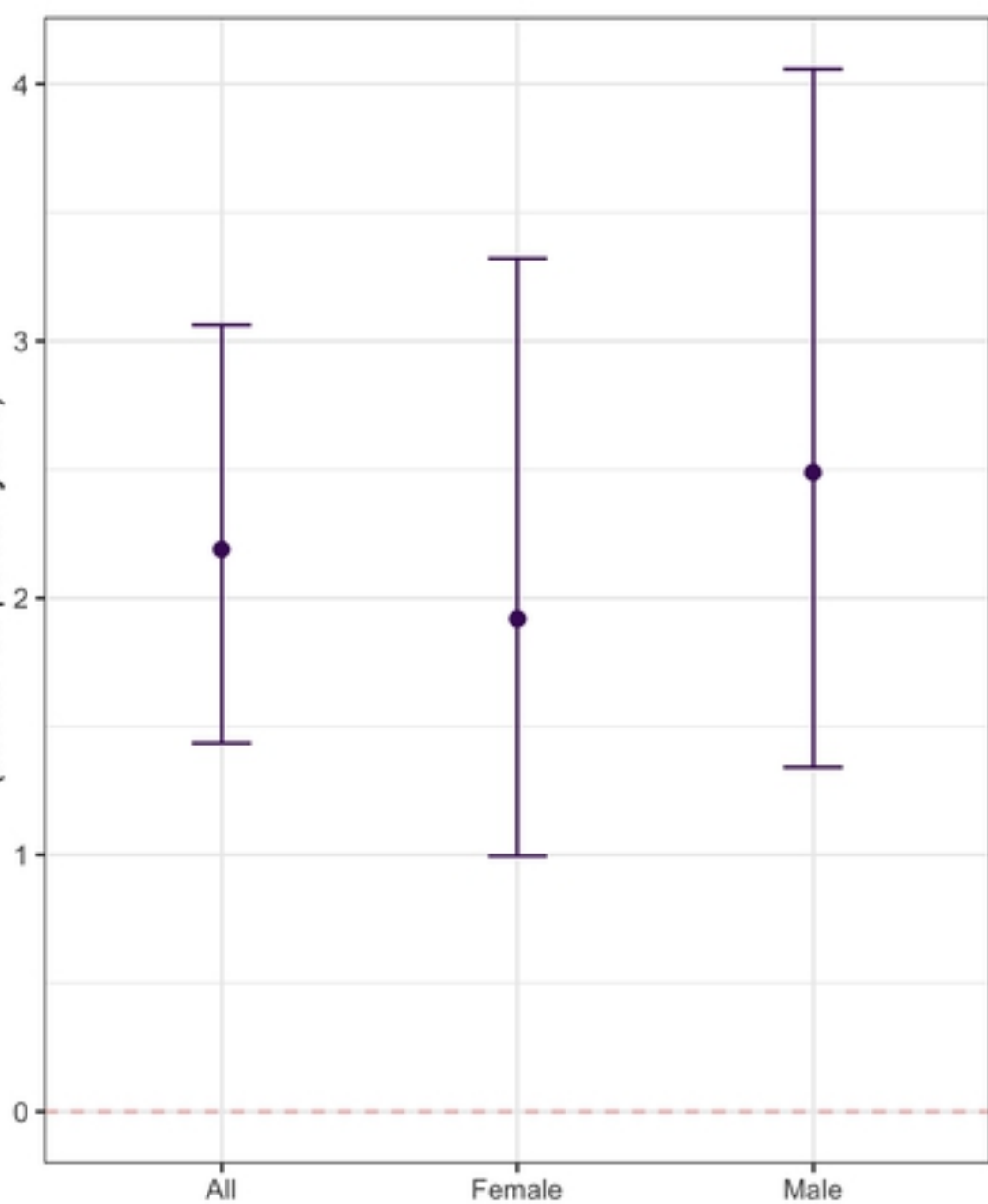
- 412 36. Pang M, Kaufman JS, Platt RW. Mixing of confounding and non-collapsibility: a notable
413 deficiency of the odds ratio. *Am J Cardiol.* 2013;111: 302–303.
414 doi:10.1016/j.amjcard.2012.09.002
- 415 37. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly Robust
416 Estimation of Causal Effects. *Am J Epidemiol.* 2011;173: 761–767.
417 doi:10.1093/aje/kwq439
- 418 38. Greenland S. Interval estimation by simulation as an alternative to and extension of
419 confidence intervals. *Int J Epidemiol.* 2004;33: 1389–1397. doi:10.1093/ije/dyh276
- 420 39. Keil Alexander P., Buckley Jessie P., O'Brien Katie M., Ferguson Kelly K., Zhao Shanshan,
421 White Alexandra J. A Quantile-Based g-Computation Approach to Addressing the Effects
422 of Exposure Mixtures. *Environmental Health Perspectives.* 128: 047004.
423 doi:10.1289/EHP5838
- 424 40. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. *Int J Epidemiol.* 2017;46:
425 756–762. doi:10.1093/ije/dyw323
- 426

A

Incidence rate ratio of
cardiovascular disease or death
(shown on natural log scale)

**B**

Incidence rate difference of
cardiovascular disease or death
(cases/100 person-years)



● riskCommunicator ● Standard Poisson Regression

Figure 1