Web Appendix 1

To support the brief explanations in the main text, we have provided a step-by-step demonstration of the calculation of the predicted counterfactual outcomes in g-computation. These heuristic examples are intended to show, with a minimum of complexity, how the reader could implement g-computation to calculate an effect estimate. We first demonstrate two examples of how to use g-computation outside the context of MSM. We use examples with and without interaction. We also apply an MSM for a final effect estimate on the example with interaction to demonstrate the approach used in this paper. Each table below contains one observation for each of the eight combinations of W_1,W_2, A – the three variables needed to calculate a predicted value for Y. In all tables, each observation has a value for the observed Y as well as the predicted Y_0 and Y_1.

Our first example, presented in Table S1, will be based on a fit of the observed data from regression model 2 where $E(Y \mid A, W_1, W_2) = 2.87 - 0.36*A + 1.06*W_1 + 0.13*W_2$. Although it is not the correct model for the data, this main-effects only model is used for our first demonstration of the calculations when interaction is assumed to be absent.

The difference between the actual and predicted values of Y (columns (4) and (5) of Table S1, respectively) will be a function of how well the model fits the data. Table S1 also presents the calculated values of Y_0 and Y_1 for each observation (columns (6) and (7), respectively). These predicted values of Y_a are calculated by applying the regression model to each observation, holding the covariates W_1 and W_2 constant at their observed levels, while intervening on the treatment a, setting it to the appropriate value (0 and 1, respectively). The value of Y_0 is equal to the predicted value of Y when we intervene on treatment and set it at a=0 and, therefore, observations with the same value of W_1 and of W_2 have the same value for Y_0 regardless of their observed treatment, A. This equality also holds true for Y_1 when observations share values of W_1 and W_2. This is demonstrated by ID numbers 1 and 2, columns (6) and (7).

To calculate the risk difference E(Y_1 - Y_0) without specifying an MSM, we calculate Y_1 - Y_0 for each person and take the mean over the population. In Table S1, the value of Y_1 - Y_0 is presented in column (8), and is uniform across all observations: -0.36L, the same value as the coefficient for A from regression model 2 above. Recall from the main text that in the absence of covariate/exposure interaction with binary A, g-computation and traditional regression (using the same model) provide the same answer.

In Appendix 2, we provide R code demonstrating the simulation of the dataset and all steps corresponding to the eight columns of Table S1: regression, predicting counterfactual outcomes, and calculation of $Y_1 – Y_0$ for each individual.

Web Table 1: The full data with predicted outcomes based on the main-effects only regression model 2: $E(Y|A,W_1,W_2) = 2.87 - 0.36*A + 1.06*W_1 + 0.13*W_2$. Includes examples for each possible covariate pattern, with one observation per ID.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ID:				Observed	Predicted	Calculated	Calculated	
				value of	value of	value of	value of	Calculated
	W_1	W_2	A	Y	Y	Y_0	Y_1	Y_1 - Y_0
1	0	0	0	2.34	2.87	2.87	2.51	-0.36
2	0	0	1	3.06	2.51	2.87	2.51	-0.36
3	0	1	0	2.87	2.99	2.99	2.64	-0.36
4	0	1	1	2.95	2.64	2.99	2.64	-0.36
5	1	0	0	3.99	3.93	3.93	3.57	-0.36
6	1	0	1	3.76	3.57	3.93	3.57	-0.36
7	1	1	0	3.24	4.05	4.05	3.70	-0.36
8	1	1	1	3.84	3.70	4.05	3.70	-0.36

Columns 4 through 8 are in units of liters.

Observed data are in the shaded section of the table.

In contrast to the situation when there is no covariate/exposure interaction as in Table S1, the presence of an interaction term in the Q-model implies that the effect of exposure is heterogeneous and depends on the level of another covariate. We present such a case in Table S2, which like Table S1 contains observed and predicted *Y* values for one subject in each of the eight unique combinations of covariates and treatment. In Table S2, the Q-model used for prediction is regression model 4, which has a covariate/exposure interaction term.

Model 4: $E(Y|A,W_1,W_2) = 2.95 - 0.49 *A + 1.05*W_1 + 0.31*A*W_2$. As a result, the exposure effect is different among those observations where $W_2 = 0$, as compared to those for whom $W_2 = 1$. This is demonstrated by the two unique values in Table S2, column (8): for some individuals, $Y_1 – Y_0$ equals -0.49L, while for others this value is -0.18L. The difference is 0.31L, the coefficient for $A*W_2$. In this scenario, the g-computation approach and the traditional approach estimate different treatment effects, because the g-computation estimator averages the effect across all individuals into a single, marginal treatment effect of -0.34L (calculated from the full data).

In Appendix 2, we provide R code demonstrating the simulation of the dataset and all steps corresponding to the eight columns of Table S2: regression, predicting counterfactual outcomes, and calculation of $Y_1 – Y_0$ for each individual.

Web Table 2: The full data with predicted outcomes based on regression model 4 with interaction: $E(Y|A,W_1,W_2) = 2.95 - 0.49 *A + 1.05*W_1 + 0.31*A*W_2.$ Includes examples for each possible covariate pattern, with one observation per ID.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Observed	Predicted	Calculated	Calculated	
ID:				value of	value of	value of	value of	Calculated
	W_1	W_2	A	Y	Y	Y_0	Y_1	Y_1-Y_0
1	0	0	0	2.34	2.95	2.95	2.46	-0.49
2	0	0	1	3.06	2.46	2.95	2.46	-0.49
3	0	1	0	2.87	2.95	2.95	2.77	-0.18
4	0	1	1	2.95	2.77	2.95	2.77	-0.18
5	1	0	0	3.99	4.00	4.00	3.51	-0.49
6	1	0	1	3.76	3.51	4.00	3.51	-0.49
7	1	1	0	3.24	4.00	4.00	3.82	-0.18
8	1	1	1	3.84	3.82	4.00	3.82	-0.18

Columns 4 through 8 are in units of liters.

Observed data are in the shaded section of the table.

To use an MSM to calculate an effect estimate with the same Q-model as in Table S2, the dataset could be structured as follows in Table S3. Note that there are now two records per ID: the observed values are included in one record, and now each ID has a second record where the value of the treatment has been reversed (i.e. if observed $A = 0$, then $a = 1$ in the new record). This setup highlights the missing data structure inherent in observed data: by necessity, only observed values of A and Y (columns (3) and (4), respectively), are included. By intervening on a (column (5)) and using g-computation to predict values of Y for all treatments, for all records—even those that were unobserved—we simulate the full data.

Thus, a general Y_a variable for all the counterfactual outcomes is created (column (6)). The values from columns (6) and (7) in Table S2 are replicated Table S3, column (6) below. Once this structure is created for all observations in the original dataset, regress Y_a (column (6)) on a (column (5)) to get the marginal effect of interest. From Table 3 of the paper, we can see when model 4 is used as the Q-model, the g-computation estimate for the marginal effect of a equals -0.34L, which is very close to the truth of the simulation protocol, where β=-0.35L.

Web Table 3: The full data with predicted outcomes based on Regression Model 4 with interaction:

E($Y|A,W_1,W_2$) = 2.95 - 0.49*A + 1.05*W_1 + 0.31*A*W_2. Includes examples for each possible covariate pattern, with two records per ID (corresponding to a=1 and a=0).

	(1)	(2)	(3)		(5)	(6)
				(4)		Calculated
				Observed		value of
ID:	W_1	W_2	A	value of Y	a	Y_a
1	0	0	0	2.34	0	2.95
1	0	0	unobserved	unobserved	1	2.46
2	0	0	unobserved	unobserved	0	2.95
2	0	0	1	3.06	1	2.46
3	0	1	0	2.87	0	2.95
3	0	1	unobserved	unobserved	1	2.77
4	0	1	unobserved	unobserved	0	2.95
4	0	1	1	2.95	1	2.77
5	1	0	Unobserved	unobserved	0	4.00
5	1	0	1	3.99	1	3.51
6	1	0	0	3.76	0	4.00
6	1	0	Unobserved	unobserved	1	3.51
7	1	1	0	3.24	0	4.00
7	1	1	unobserved	unobserved	1	3.82
8	1	1	unobserved	unobserved	0	4.00
8	1	1	1	3.84	1	3.82

Columns 4 and 6 are in units of liters.

Observed data are in the shaded section of the table.

Web Appendix 2

By following this annotated R code, the reader is able to perform all steps required to generate Tables S1 and S2. The code demonstrates the simulation of the dataset, the regressions implemented (corresponding to regression models 2 and 4), the prediction of counterfactual outcomes, and the calculation of $Y_1 – Y_0$ for each individual. The tables that this code generates show columns 1 – 8 as represented in Tables S1 and S2, with rows for all 300 observations rather than the 8 example rows shown in the preceding tables.

```
##Generate simulated data set
n<-300
set.seed(285)
simdata<-data.frame(W1 = rbinom(n,1,0.4), W2=rbinom(n,1,0.5))
simdata<-transform(simdata, # add A
       A = rbinom(n,1,(0.5+0.2*W1-0.3*W2)))
simdata<-transform(simdata, # add Y
       Y = rnorm(n,(3-0.5*A+W1+0.3*A*W2),.4))

##S1
## Perform regression with main effects of W1,W2
## Corresponds to Table S1, which uses regression model 2
reg2<-glm(Y~A+W1+W2, data=simdata, family=gaussian)
summary(reg2)

## Create predicted Y_A for all observations
Ypred.S1<-predict(reg2)

## Generate data sets where a is set to 0 and 1
simdata.A0<-transform(simdata, A=0)
simdata.A1<-transform(simdata, A=1)

## Create Y_0 for all observations, sets a=0
Y0.S1<-predict(reg2,newdata=simdata.A0)

## Create Y_1 for all observations, sets a=1
Y1.S1<-predict(reg2,newdata=simdata.A1)

## Calculate (Y_1 - Y_0) for each individual
difference.S1<-Y1.S1-Y0.S1

## Create summary table
table1<-cbind(simdata[,c("W1", "W2", "A")], "Y"=round(simdata[,"Y"],2),
"Ypred"=round(Ypred.S1,2), "Y_0"=round(Y0.S1,2), "Y_1"=round(Y1.S1,2), "Y_1-
Y_0"=round(difference.S1,2))

## Show all variables (W1, W2, A, Y, Y_pred, Y_0, Y_1) and (Y_1 - Y_0) for each of the
300 observations
## This corresponds to the 8 columns in Table S1, for all 300 observations
table1
```

```
##S2
## Perform regression on original data, now including W1 and interaction between A*W2
## Corresponds to Table S2, which uses regression model 4
reg4<-glm(Y~A+W1+A:W2, data=simdata, family=gaussian)
summary(reg4)

## Create predicted Y_A for all observations, sets A=a
Ypred.S2<-predict(reg4)

## Create Y_0 for all observations, sets a=0
Y0.S2<-predict(reg4,newdata=simdata.A0)

## Create Y_1 for all observations, sets a=1
Y1.S2<-predict(reg4,newdata=simdata.A1)

## Calculate (Y_1 - Y_0) for each individual
difference.S2<-Y1.S2-Y0.S2

## Create summary table
table2<-cbind(simdata[,c("W1", "W2", "A")], "Y"=round(simdata[,"Y"],2),
"Ypred"=round(Ypred.S2,2), "Y_0"=round(Y0.S2,2), "Y_1"=round(Y1.S2,2), "Y_1-
Y_0"=round(difference.S2,2))

## Show all variables (W1, W2, A, Y, Y_pred, Y_0, Y_1) and (Y_1 - Y_0) for each of the
300 observations
## This corresponds to the 8 columns in Table S2, for all 300 observations
table2
```