# Methods

We estimated NHL risk from 1985 to 2005 under hypothetical limits on average annual exposure to soluble MWF by applying the recently-developed hazard-extended iterative conditional expectation (ICE) parametric g-formula estimator (Wen et al. 2020). We leveraged time-varying quantitative MWF exposure data in tandem with employment records to adjust for the HWSE. First, we estimated the expected number of NHL cases per 1000 workers that we would observe if there were no competing risks. Then, we contrasted this counterfactual risk to that when in addition, the hypothetical exposure limits of 0.5, 0.25, and 0.05 mg/m3 were enforced for soluble MWF over workers’ entire working lifetimes. Under this hypothetical intervention, exposures in years when the observed exposure was above the hypothetical limit of interest were set to the limit. Otherwise, exposure not intervened upon. This kind of intervention is known as a threshold longitudinal modified treatment policy (Haneuse and Rotnitzky 2013; Richardson and Robins 2013; Taubman et al. 2009; Young et al. 2014). Longitudinal modified treatment policies are intervention strategies that depend on the natural value of exposure at a particular time ie the value of exposure that would have been observed at that time if the intervention were discontinued immediately prior (Dı́az et al. 2021; Richardson and Robins 2013).

## Study population

The UAW-GM cohort includes all hourly workers at three automobile manufacturing plants in Michigan who had worked at least 3 years by 1985. Past papers provide detailed descriptions of the cohort (Eisen et al. 1992, 2001). The present study population (N = 34,738) was restricted to the autoworkers who were at work in 1941 or not yet hired, missing no more than half of their employment history, and still alive at the start of follow-up. Autoworkers in the study population were followed for NHL incidence from January 1, 1985 until NHL diagnosis, death, January 1, 2005 (10 years after the end of employment record availability), or age 108 years (the oldest observed age at death), whichever came earlier.

## Outcome and covariates

We identified incident cancers in the UAW-GM Cohort that occurred between 1985 and 2004 by linkage to the Michigan Cancer Registry (MCR). Workers at Plants 1 and 2, located in the greater Detroit metropolitan area, were also linked to the Detroit Regional Registry of the Surveillance, Epidemiology, and End Results (SEER) Program. Cancer types were distinguished using site and histology codes conforming to the International classification of Diseases for Oncology, 3rd edition (ICD-O-3). Non-Hodgkin lymphoma was defined by cancers with any of the following ICD-O-3 Histology codes: 9590-9597, 9670-9671, 9673, 9675, 9678-9680, 9684, 9687-9691, 9695, 9698-9702, 9705, 9708-9709, 9712, 9714-9719, 9724-9729, 9735, 9737-9738, 9811-9818, 9823, 9827, 9837. Details regarding cancer incidence follow-up are described elsewhere (Colbeth et al. 2022, in press). Vital status was ascertained from company records and by linkage to Social Security Administration, National Death Index, and state mortality files.

Covariates including year of hire, sex, race, and plant location were obtained from company records. Race was missing for about 16% of the cohort, most commonly among workers hired before 1960 in Plant 2. In analyses, missing race was considered a distinct category. All covariates were coded as categorical variables. Cut-points for continuous covariates were determined according to the quantiles among cases.

## Exposure

Company industrial hygienists collected several hundred personal and area samples for total particulate matter (mg/m3) composed of MWF over many decades. Research industrial hygienists collected additional air sampling data when the cohort study was launched in the mid 1980s. These data were combined with the historical data to derive quantitative 8-hour time-weighted average exposure estimates to soluble, straight, and synthetic MWFs for each combination of job, department, and plant over time. Workers’ time-weighted average annual exposure to each MWF type was determined by combining this job-exposure matrix with employment records, which recorded time-varying job type, department, and plant through 1994. For employment records that were at least half complete, gaps in the record were interpolated by carrying forward the last known job type. The exposure assessment is described in detail elsewhere (Hallock et al. 1994; Woskie et al. 1994, 2003). Previous analyses of NHL applied exposure lags of 1 to 20 years; we lagged cumulative MWF exposures by 10 years to account for disease latency and therefore ended follow-up on January 1, 2005 (Karipidis et al. 2007; Smith et al. 2007; Zhang et al. 2019). In analyses, MWF exposure history was summarized as the cumulative sum of annual exposure intensities and coded as categorical variables with cut-points determined according the quantiles of cumulative exposure among cases.

## Statistical methods

We applied the hazard-extended ICE parametric g-formula with pooling over treatment history (Wen et al. 2020) to estimate four 20-year counterfactual risks of non-Hodgkin lymphoma in the absence of competing risks. The first estimand represents a baseline case with no intervention on exposure, but with the elimination of competing risks. The remaining three were the counterfactual risks under threshold interventions enforcing the following hypothetical limits: (1) the NIOSH Recommended Exposure Limit (REL) for total particulate mass (PM) from MWFs (0.5 mg/m3), (2) half the REL (0.25 mg/m3), and (3) a tenth of the REL (0.05 mg/m3) (Rosenstock 1998). Interventions were applied at hire, before the start of follow-up, through the end of follow-up. Person-years were not intervened upon if the observed average annual exposure was below the hypothetical limit.

The hazard-extended ICE parametric g-formula may be thought of as a series of model-based standardization steps, which we implemented using logistic regression. We split the 20-year follow-up period into eight time periods; the first two periods spanned four years each, and the remaining six periods spanned two years each. The first two periods are longer in length to account for the smaller number of cases in those years. Post-intervention estimates of the discrete hazard of NHL given all exposures and covariates were combined iteratively from the end of follow-up to the start. In each iteration, predicted discrete hazards were standardized over post-intervention exposure and covariate histories before combining with discrete hazards from the previous iteration. This iterative process results in sequentially standardized estimates of NHL risk over the entire follow-up period. Averaging over the baseline distribution of covariates yields the counterfactual risk estimate of NHL when the intervention of interest was enforced for the entire study population. An overview of the general steps of the estimation procedure are presented below.

1. For the hypothetical exposure limit of interest, compute the cumulative exposure each worker would have accrued by the end of each follow-up period since hire.
2. Fit a pooled logistic regression to the observed data for NHL on covariates and cumulative exposure over all at-risk person-periods, excluding those ending with a censoring event.
3. Replace the cumulative exposure vector in the observed data with the post-intervention cumulative exposure vector. Using the model fitted in the previous step, compute the post-intervention discrete hazard estimates for each at-risk person-period including those that end with a censoring event.
4. Set where is the number of follow-up periods:
   1. Among the event-free and uncensored through the th period, fit a logistic regression on the predicted post-intervention discrete hazard for the period spanning the through th periods given observed covariate and exposure history up through (and including) the th period.
   2. Replacing the cumulative exposure vector in the observed data with the post-intervention cumulative exposure vector. Using the model fitted in the previous step, obtain predicted values for all those event-free and uncensored through the th period.
   3. Compute the predicted post-intervention discrete hazard for the period spanning the through th periods by multiplying the predicted values from the previous step by 1 minus the discrete hazard estimate from Step 2 and adding the same discrete hazard estimate to the product.
   4. If , set and return to step 3.a.
5. Compute the counterfactual risk by averaging the predicted post-intervention discrete hazards for the entire follow-up period for all units.

Post-intervention exposure and exposure history were summarized as cumulative exposure. We modeled discrete hazards by fitting a pooled logistic regression for NHL over at-risk person-periods given cumulative exposure to straight, soluble, and synthetic MWFs, employment status, cumulative time off, year of hire, sex (male/female), race (Black/white/unknown), and plant (Plant 1/Plant 2/Plant 3). Cumulative exposure to MWFs, employment status, and cumulative time off were lagged 10 years. All continuous variables were represented as categorical variables with cut points determined by the tertiles of nonzero values among NHL cases. During the iterative combination of discrete hazards, we performed model-based standardization over baseline covariates and the complete set of time-varying covariate histories.

We estimated risk under the observed distribution of soluble MWF exposure (natural course) and under the six interventions. We contrasted the risk under intervention to that under the natural course by computing relative risks. Confidence intervals were computed using the nonparametric bootstrap with 1000 samples and centering on the estimate computed from observed data.

Colbeth, H. L., Chen, K. T., Picciotto, S., Costello, S., and Eisen, E. A. (2022, in press), “Metalworking fluids and cancer incidence in the United Autoworkers-General Motors Cohort,” *American Journal of Epidemiology*.

Dı́az, I., Williams, N., Hoffman, K. L., and Schenck, E. J. (2021), “Nonparametric causal effects based on longitudinal modified treatment policies,” *Journal of the American Statistical Association*, Taylor & Francis, 1–16.

Eisen, E. A., Bardin, J., Gore, R., Woskie, S. R., Hallock, M. F., and Monson, R. R. (2001), “Exposure-response models based on extended follow-up of a cohort mortality study in the automobile industry,” *Scandinavian journal of work, environment & health*, 27, 240–249.

Eisen, E. A., Tolbert, P. E., Monson, R. R., and Smith, T. J. (1992), “Mortality studies of machining fluid exposure in the automobile industry I: A standardized mortality ratio analysis,” *American journal of industrial medicine*, Wiley Online Library, 22, 809–824.

Hallock, M. F., Smith, T. J., Woskie, S. R., and Hammond, S. K. (1994), “Estimation of historical exposures to machining fluids in the automotive industry,” *American Journal of Industrial Medicine*, Wiley Subscription Services, Inc., A Wiley Company, 26, 621–634. <https://doi.org/10.1002/ajim.4700260505>.

Haneuse, S., and Rotnitzky, A. (2013), “Estimation of the effect of interventions that modify the received treatment,” *Statistics in medicine*, Wiley Online Library, 32, 5260–5277.

Karipidis, K. K., Benke, G., Sim, M. R., Kauppinen, T., Kricker, A., Hughes, A. M., Grulich, A. E., Vajdic, C. M., Kaldor, J., Armstrong, B., and others (2007), “Occupational exposure to ionizing and non-ionizing radiation and risk of non-Hodgkin lymphoma,” *International archives of occupational and environmental health*, Springer, 80, 663–670.

Richardson, T. S., and Robins, J. M. (2013), “Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality,” *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, Citeseer, 128, 2013.

Rosenstock, L. (ed.) (1998), *What you need to know about occupational exposure to metalworking fluids*, Department of Health; Human Services (NIOSH).

Smith, M. T., Jones, R. M., and Smith, A. H. (2007), “Benzene exposure and risk of non-Hodgkin lymphoma,” *Cancer epidemiology, biomarkers & prevention*, AACR, 16, 385–391.

Taubman, S. L., Robins, J. M., Mittleman, M. A., and Hernán, M. A. (2009), “Intervening on risk factors for coronary heart disease: An application of the parametric g-formula,” *International journal of epidemiology*, Oxford University Press, 38, 1599–1611.

Wen, L., Young, J. G., Robins, J. M., and Hernán, M. A. (2020), “Parametric g-formula implementations for causal survival analyses,” *Biometrics*, Wiley Online Library.

Woskie, S. R., Smith, T. J., Hallock, M. F., Hammond, S. K., Rosenthal, F., Eisen, E. A., Kriebel, D., and Greaves, I. A. (1994), “Size-selective pulmonary dose indices for metal-working fluid aerosols in machining and grinding operations in the automobile manufacturing industry,” *American Industrial Hygiene Association Journal*, Taylor & Francis, 55, 20–29.

Woskie, S. R., Virji, M. A., Hallock, M., Smith, T. J., and Hammond, S. K. (2003), “Summary of the findings from the exposure assessments for metalworking fluid mortality and morbidity studies,” *Applied occupational and environmental hygiene*, Taylor & Francis, 18, 855–864.

Young, J. G., Hernán, M. A., and Robins, J. M. (2014), “Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data,” *Epidemiologic methods*, De Gruyter, 3, 1–19.

Zhang, L., Rana, I., Shaffer, R. M., Taioli, E., and Sheppard, L. (2019), “Exposure to glyphosate-based herbicides and risk for non-Hodgkin lymphoma: A meta-analysis and supporting evidence,” *Mutation Research/Reviews in Mutation Research*, Elsevier, 781, 186–206.