

Summary of Work Accomplished Under TCEQ Work Order 10 -- Final Deliverable

NERA Economic Consulting Project Team

Anne E. Smith

Garrett Glasgow

Bharat Ramkrishnan

Marissa Licursi

August 31, 2018

Contents

1. Executive summary
2. Overview of the simulation method
3. Tests for thresholds with Cox proportional hazard models under measurement error
4. Estimated threshold locations with Cox proportional hazard models under measurement error
5. Estimated hazard ratios with Cox proportional hazard models under measurement error
6. Using nonparametric regressions to examine mortality data for thresholds
7. Tests for thresholds with Cox proportional hazard models under measurement error, no random variation across cities
8. Estimated threshold locations with Cox proportional hazard models under measurement error, no random variation across cities
9. Estimated hazard ratios with Cox proportional hazard models under measurement error, no random variation across cities
10. Using nonparametric regressions to examine mortality data for thresholds, no random variation across cities
11. Estimated hazard ratio using “snapshot” of PM from a single year when true PM trends downward over time
12. References



1. Executive Summary

Project Overview

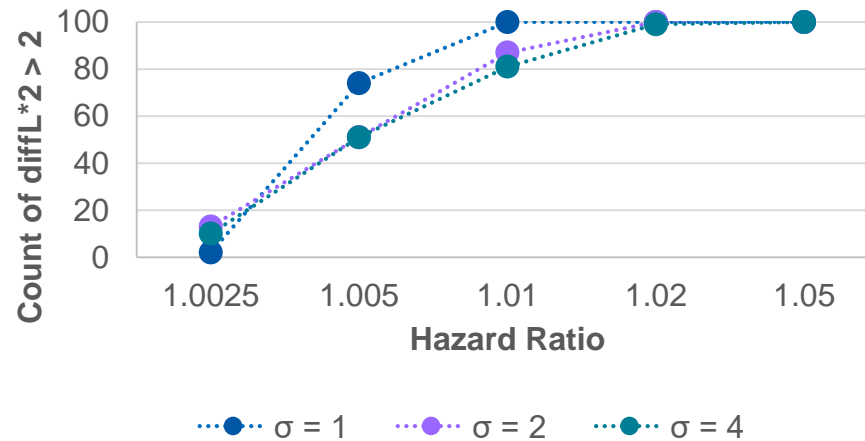
- Under this Work Order, NERA used simple prospective cohort simulations to start to explore the validity and robustness of common methods for assessing concentration-response (C-R) relationships between long-term air pollution exposures and mortality risk.
 - PM2.5 is used as the illustrative pollutant, but the study could be applicable to any of a range of criteria pollutants.
- Primary focus under this Work Order was on reliability of common statistical methods for detecting population-wide C-R thresholds in the face of inaccurate observations of population-average exposures (“measurement error”).
- Meaningful patterns were difficult to discern in initial simulation runs, requiring confirmation by studying several different types of simulations:
 - Limiting the cohort to a single age and sex stratum (Men, 60 years old at year 1 of simulation)
 - Examining unrealistically “pristine” cohorts (where there are no differences in individual mortality outcomes across cities reflecting random manifestations of the shared baseline risk).
 - Considering a very wide range of hazard ratios and levels of measurement error.
- Relationships between the detectability estimation of thresholds and measurement error have now emerged that are described in this slide deck.

Key Conclusions

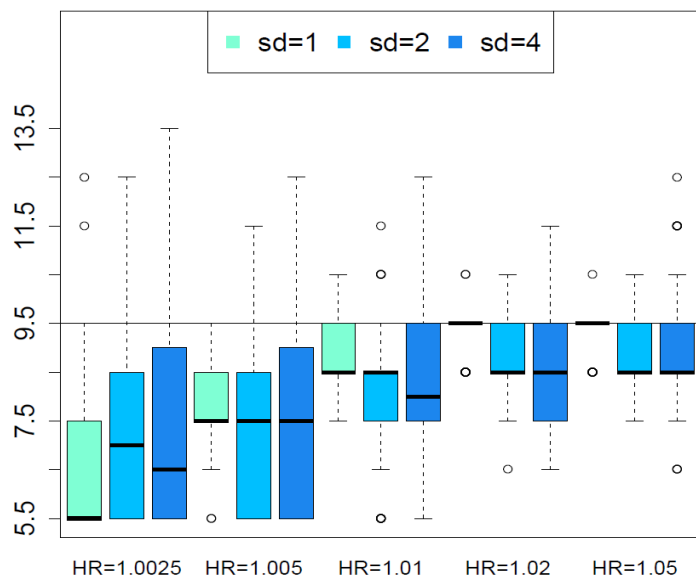
- For the type of measurement error we have simulated, as measurement error increases,
 - Ability to detect a “statistically significant” threshold in the C-R function is progressively reduced
 - When a threshold is detected, its level is progressively more likely to be underestimated.
- Even when a threshold is detected, the slope of the C-R function remains underestimated, to a degree that is also increased with increased measurement error.
- These distortions hold at policy-relevant parameter values even in a relatively non-noisy simulation (i.e., where only variability in addition to the assumed measurement error is in actual dates of death of individuals facing same mortality risk). For example:
 - We find poor detectability of a threshold of $9.5 \mu\text{g}/\text{m}^3$ (i.e., at about the mean of the PM exposures across all cities) when the hazard ratio is in the range of 1.005 per $\mu\text{g}/\text{m}^3$
 - See next slide for details of results of simulations for this case

Simulation Results for Threshold = 9.5

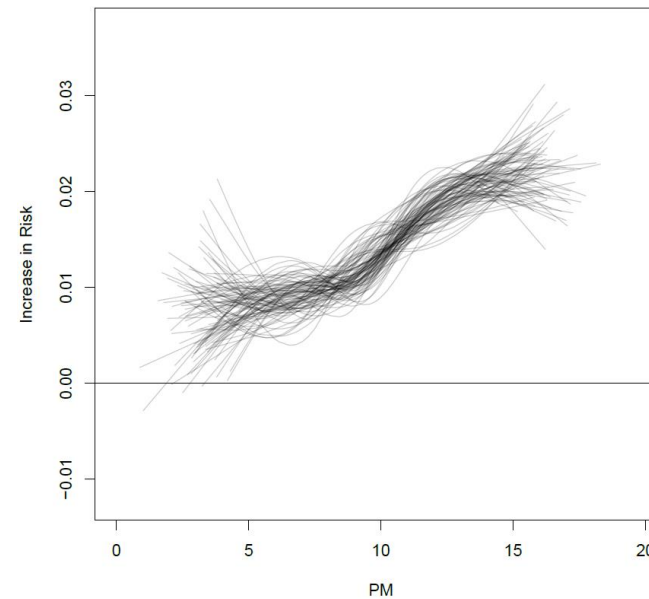
**% of simulations where Cox PH detects a threshold with statistical significance
(by level of true HR and degree of measurement error)**



**Distributions of Cox PH threshold estimates
(by level of true HR and degree of measurement error)**



**Results of 100 nonlinear spline simulations
(HR=1.005 and $\sigma=2$)**





2. Overview of the Simulation Method

Setting Up the Simulated Cohorts

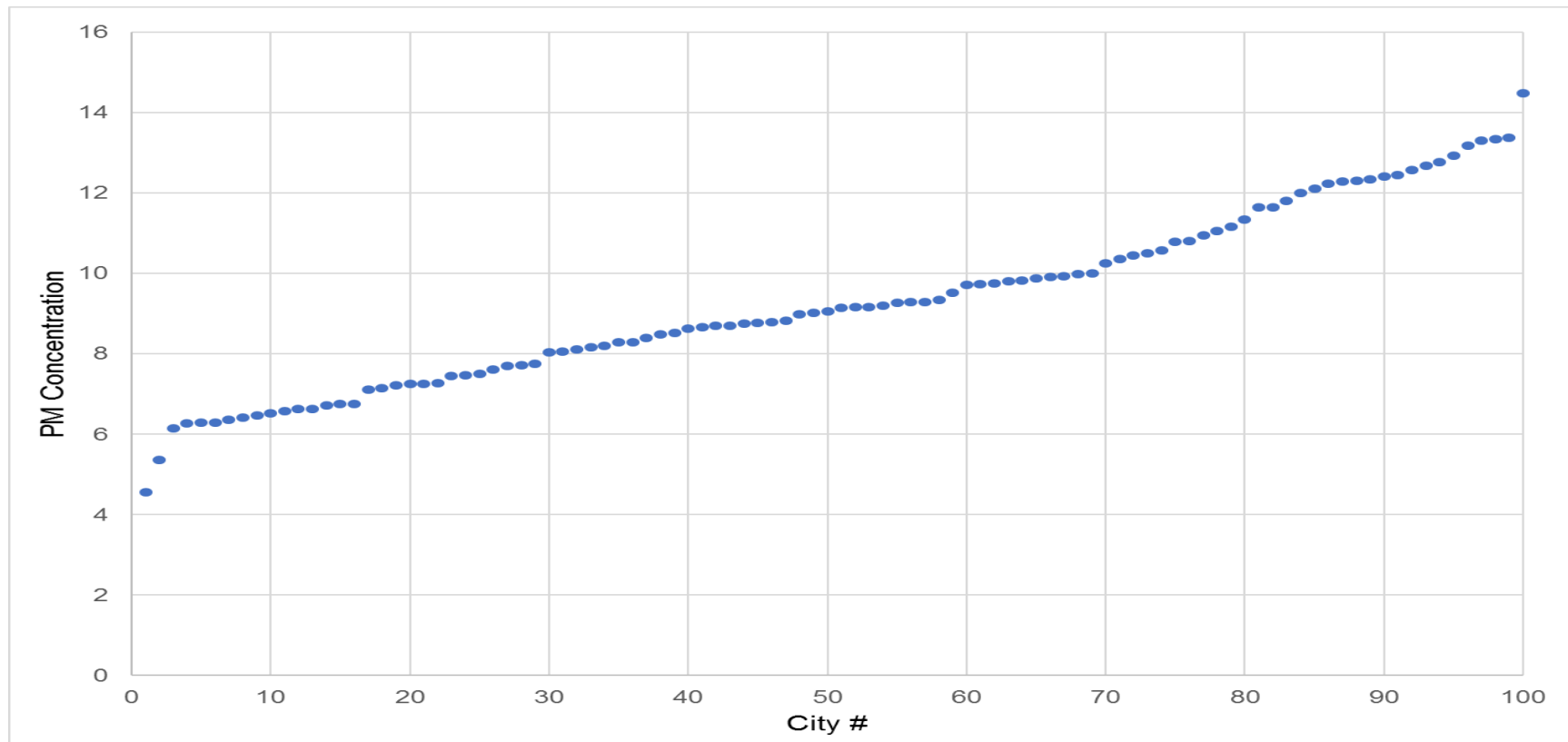
- Large numbers of hypothetical individuals for each city are generated, and their survival over time is simulated to create a “cohort” database for statistical study.
 - Results presented here are from cohorts of men aged 60 at the start of the study (for purposes of understanding underlying reasons for unusual results)
 - Patterns in our results would also hold for cohorts with more varied age & sex mix

- 100 cities, each with 20,000 simulated individuals. (2 million total in cohort, ~900,000 deaths observed after 20 years follow up).
 - Assumes cohort first forms in year 2000, is followed for 20 years.
 - Baseline mortality based on the US Census life tables for all-cause mortality.
 - “Non-noisy” cohort simulation:
 - Same baseline mortality risk is applied in every city
 - Same sensitivity to PM is applied to every individual in every city (no variability in C-R)
 - As in the real world, actual dates of death of individuals facing same mortality risk can vary randomly across cities

- We also studied a set of unrealistically “pristine” cohorts to better understand dynamics underlying our findings
 - Pristine simulations eliminated even the random variation in actual dates of death across cities for a given level of mortality risk – differences in mortality across cities are due to PM only
 - Results using these “pristine” simulations are presented in sections 7-10

PM Exposure Levels

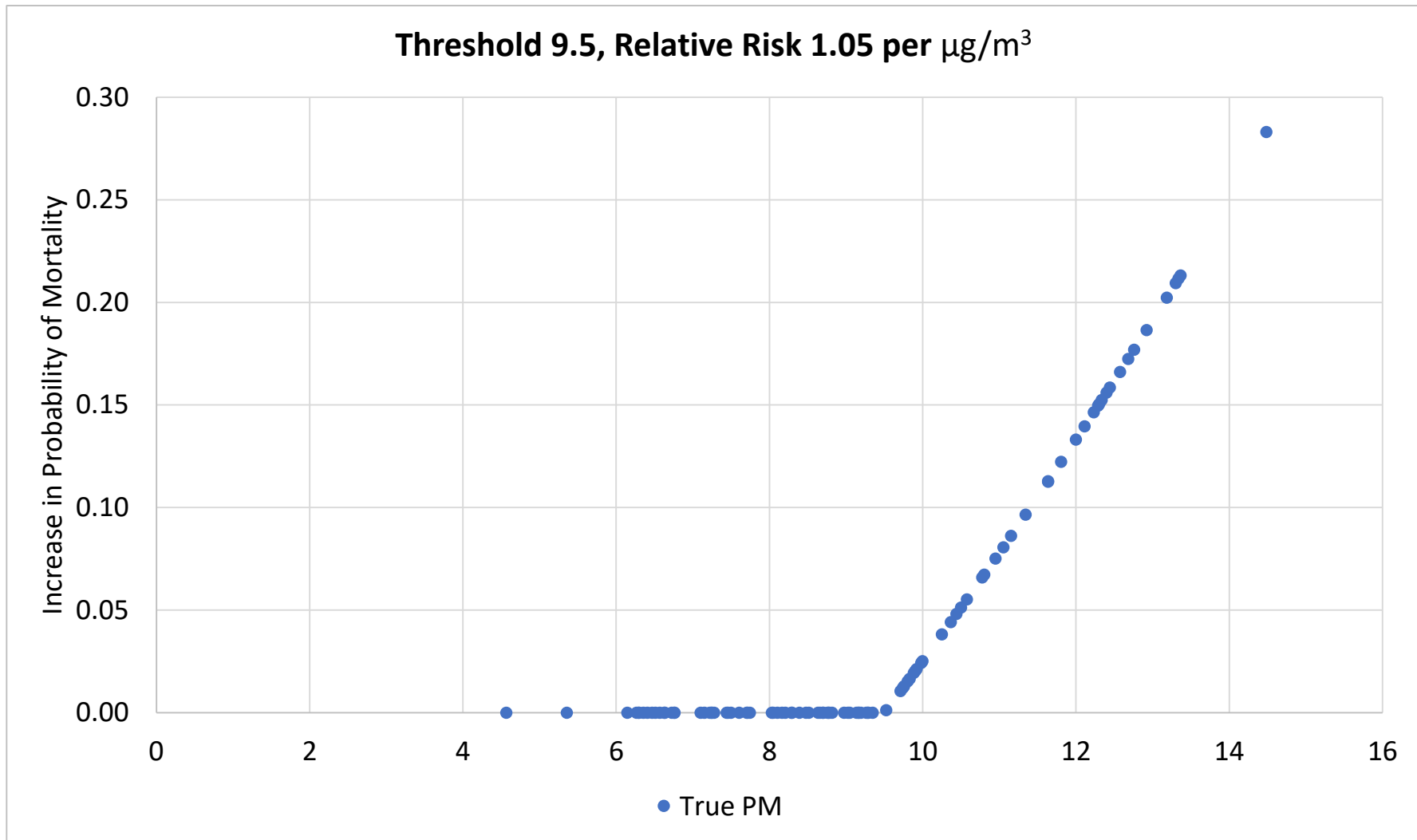
- Except for one analysis presented below, “true” PM in each of 100 generic cities was assumed to be constant over time.
 - The distribution was consistent with the US-wide distribution from 2000 to 2017 (see figure below)
 - The mean PM across all cities was approximately $9.3 \mu\text{g}/\text{m}^3$.



The Concentration-Response Function

- The “true” PM level in each city is used to alter the observed survival outcomes in the cohort according to an assumed “true” hazard ratio (HR)
 - For example, a hazard ratio of 1.005 means that for every $\mu\text{g}/\text{m}^3$ of true $\text{PM}_{2.5}$ in a given city, we multiply the baseline mortality probability of each individual in that city by 1.005.
 - We assume risk is a function of total $\text{PM}_{2.5}$ mass, with no differences due to mix of $\text{PM}_{2.5}$ constituents.
- The C-R functions we focused on in the Work Order have well-defined, population-wide thresholds, such that PM below the threshold has no effect on mortality, and PM above the threshold has a linear effect as before.
- We ran simulations for a wide range of true C-Rs:
 - Three alternative “true” threshold levels: $7 \mu\text{g}/\text{m}^3$, $8.5 \mu\text{g}/\text{m}^3$, and $9.5 \mu\text{g}/\text{m}^3$.
 - Five alternative “true” HRs above the threshold: 1.0025, 1.005, 1.01, 1.02, 1.05

Example of True Relative Risk Function With a Threshold

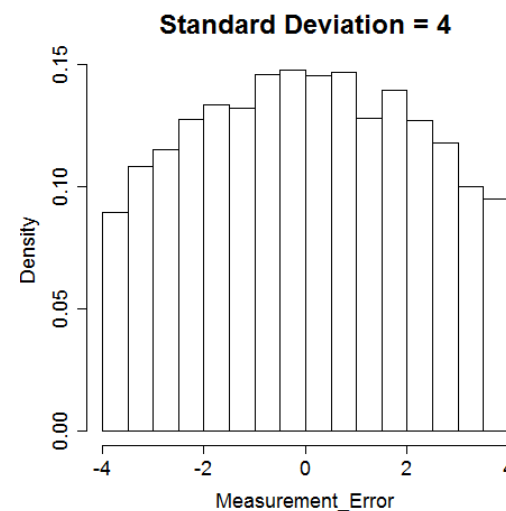
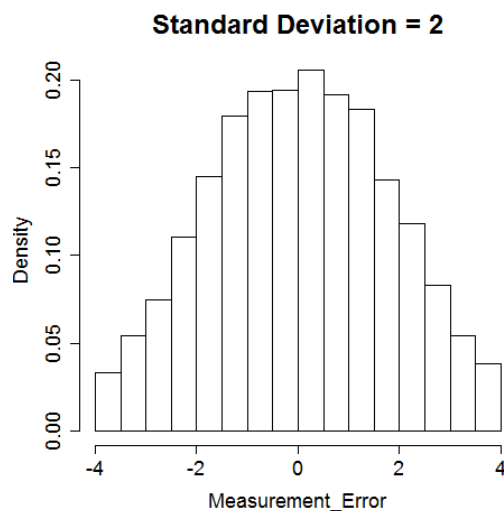
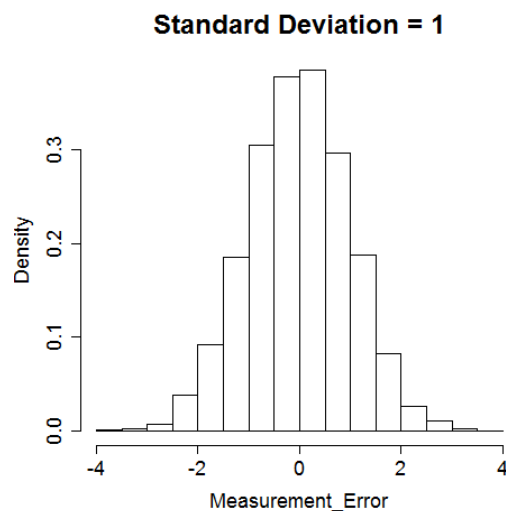


A Note on Hazard Ratios versus Relative Risks

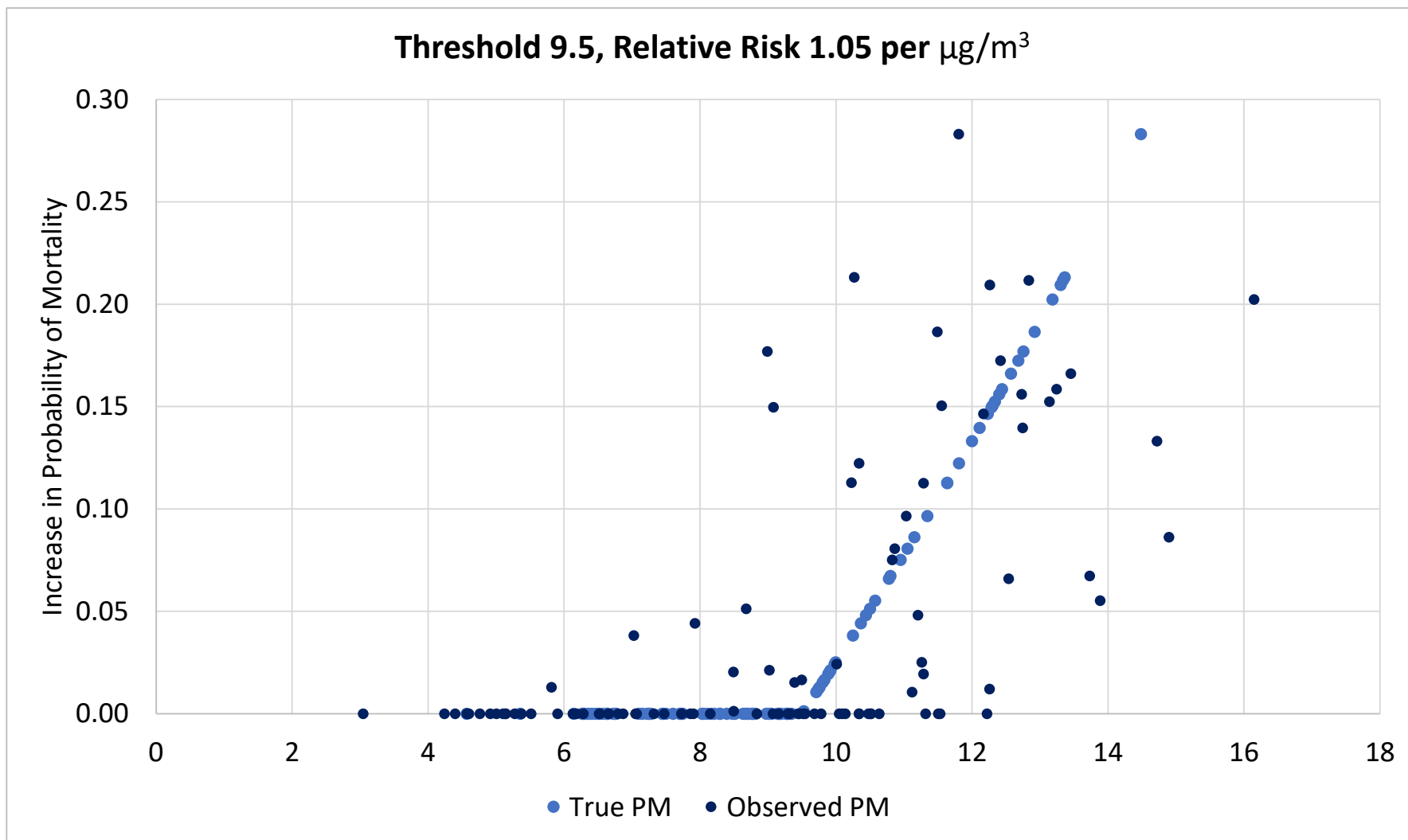
- Note that hazard ratios (HR) and relative risks (RR) are not the same thing. Hazard ratios are instantaneous risk, relative risk is cumulative.
- The RR observed between two populations will usually be lower than the HR experienced by one of those populations.
 - For example, for our illustrative cohort of 60 year old men, a HR of 1.05, comparing two cities with a 1 unit difference in $PM_{2.5}$, leads to a RR of about 1.03 after 20 years in this population.
- We suspect that RR and HR may have become inappropriately conflated in the long-term risk epidemiology literature.
 - The estimate of the β coefficient in a Cox PH model is an estimate of the HR, not of the RR. However, authors of Cox PH studies are describing their β coefficient estimates as “relative risks” (See for example Pope et al. 2002, Table 2)
 - The risk analysis literature computes the “attributable fraction” of deaths, which is a function of RR, not the HR in a survival curve analysis.
- If true, the risk analysis profession has been overestimating long-term premature deaths from the results of Cox PH studies.
 - This possibility needs further study to confirm or refute.

Application of Measurement Error

- We studied impacts of a type of measurement error that is interpreted as the potential that the PM exposure assigned to a group of people (“city” in this case) deviates from the true population-weighted average experienced by that group of people.
 - This is consistent with the concept of “classical” error
- “Observed” PM measures for each city were simulated by adding a random draw to the “true” PM value.
 - The random draws came from a truncated normal with bounds at +/- 4 $\mu\text{g}/\text{m}^3$.
 - We considered impacts of standard deviations of 1, 2, and 4.
- For each SD, 100 sets of observed PM values were generated, and then used for all simulations with the same assumed true C-R parameters.



Example of Relative Risk Evidence When Observed PM Contains Measurement Error





3. Tests for Thresholds With Cox Proportional Hazard Models Under Measurement Error

Detecting Thresholds with the Cox Proportional Hazard Tests: Summary of Results

- With moderate amounts of measurement error ($sd = 2$), and a threshold higher than the mean PM level (9.5), there was only a 50% chance of detecting the threshold at $HR=1.005$.
- The ability to detect the threshold increases as the threshold increases.
- The ability to detect the threshold increases as the hazard ratio increases.
- The ability to detect the threshold decreases as measurement error increases.

Cox Proportional Hazard Threshold Tests

- This test searches for thresholds using a “grid search” type method.
 1. Examine a range of alternative threshold estimates incremented by $1 \mu\text{g}/\text{m}^3$ around the true threshold, over a range of $\pm 4 \mu\text{g}/\text{m}^3$.
 2. For each potential threshold, subtract it from the PM measure to create a new PM measure that should capture a E-R curve with that threshold.
 3. Estimate a Cox proportional hazards model with the new PM measure.
 4. Select the best fitting model across the range of thresholds as the “threshold model.”
 5. Compare the fit of the threshold model to the fit of a no-threshold model.

Testing for the Statistical Significance of the Threshold

- We compare the fit of the threshold model to the fit of a no-threshold model.
- The test statistic is 2 times the difference in log-likelihoods between the threshold model and the no-threshold model ($2 \times \Delta LL$). Larger differences indicate a relatively better fit for the threshold model.
- Three standards have been proposed for concluding the threshold model is a better fit. Conclude the threshold model is a better fit if $2 \times \Delta LL$ is greater than:
 - 2
 - The natural log of the number of deaths, or $\ln(\text{events})$.
 - The natural log of the number of individuals, or $\ln(n)$.

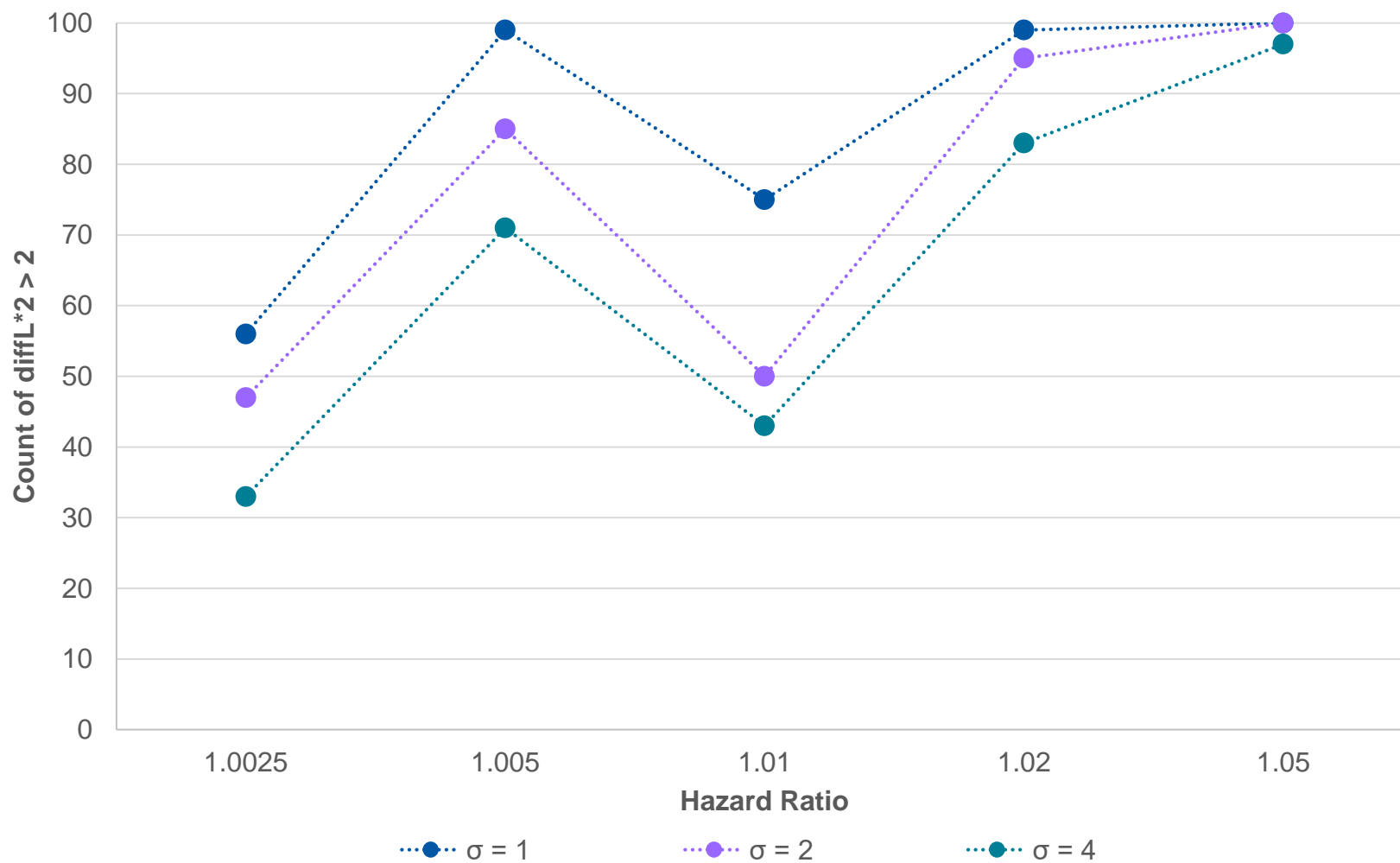
The Specific Threshold Tests Undertaken

- We examined all 45 combinations of:
 - Hazard ratios: 1.0025, 1.005, 1.01, 1.02, and 1.05.
 - Thresholds: 7, 8.5, and 9.5.
 - Measurement error: 1, 2, and 4 standard deviations.
- We ran 100 simulations for each test, each with a different set of values for observed PM. The same set of 100 observed PM values was used for all simulations involving the same level of measurement error.

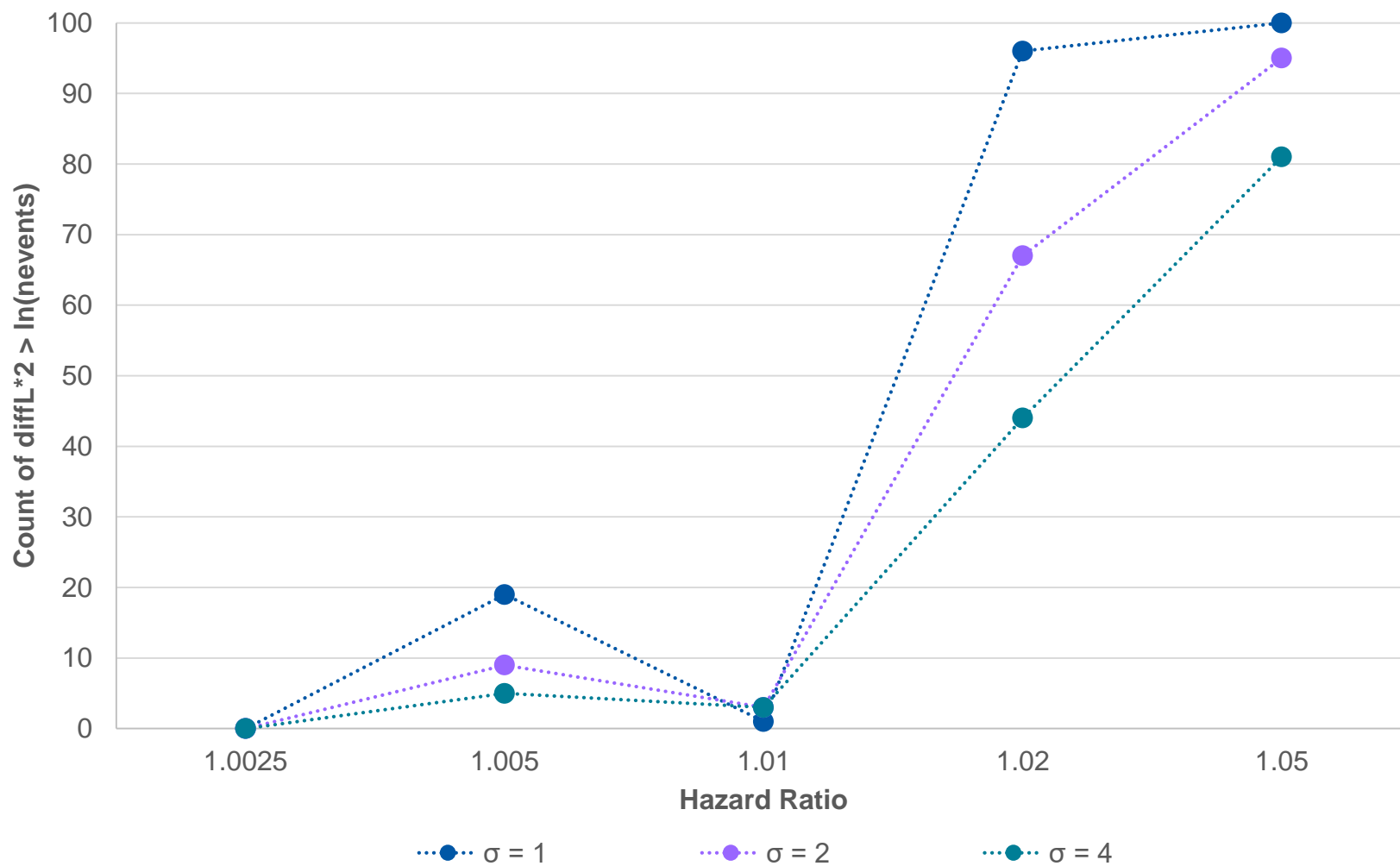
Threshold Test Results

- The results of these threshold tests across 100 simulations are presented in the 9 plots below:
 - One plot for each combination of threshold and standard for significance.
 - The vertical axis indicates the number of simulations (out of 100) in which we would conclude there is a threshold.
 - The horizontal axis indicates different hazard ratios.
 - Each line indicates a different level of measurement error.
- Points to keep in mind:
 - Due to random variation across cohorts, results may not be monotonically increasing or decreasing when the relationship is weak.
 - Only about 15% of cities have PM less than 7.

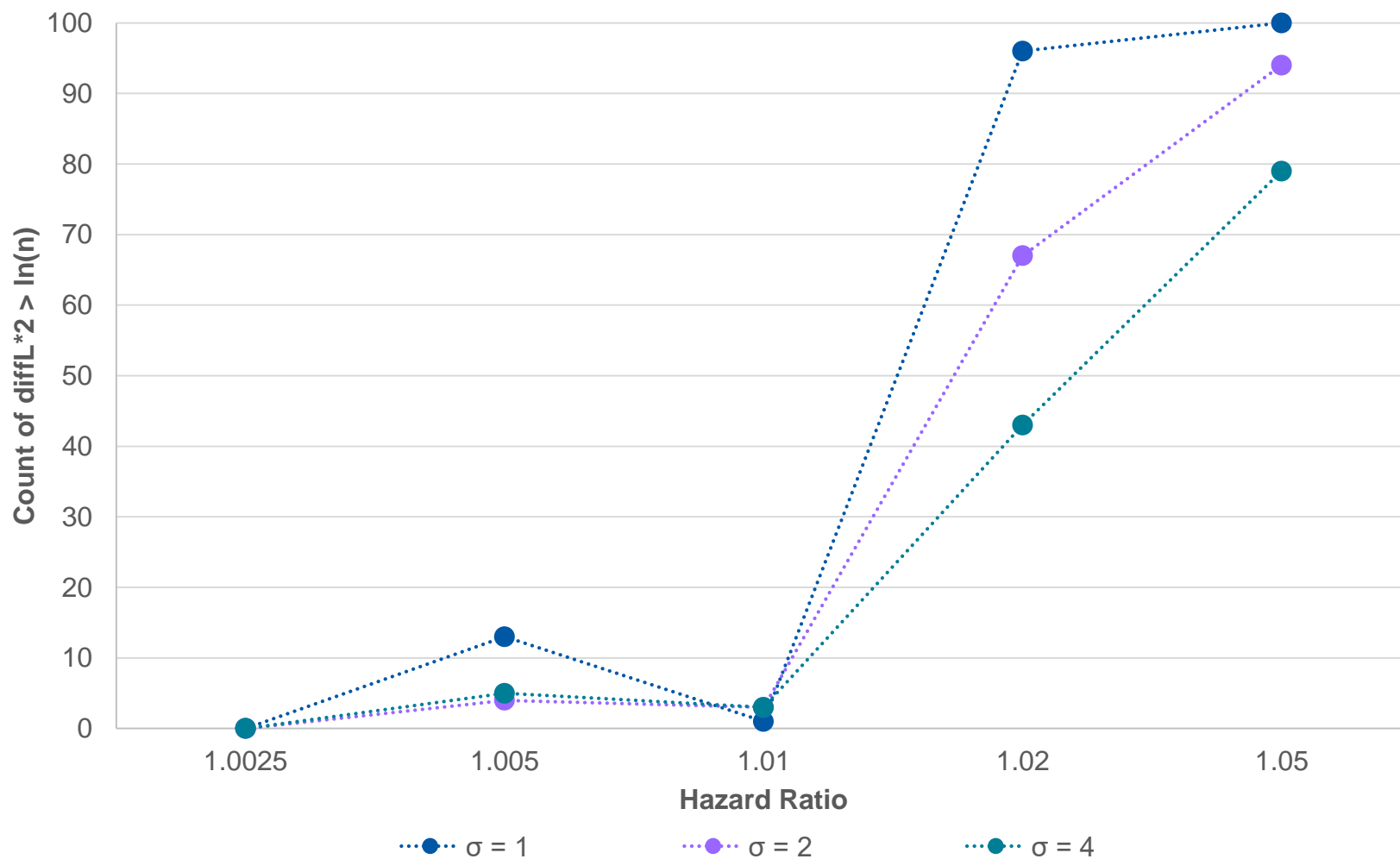
Threshold = 7, $2 \times \Delta LL > 2$



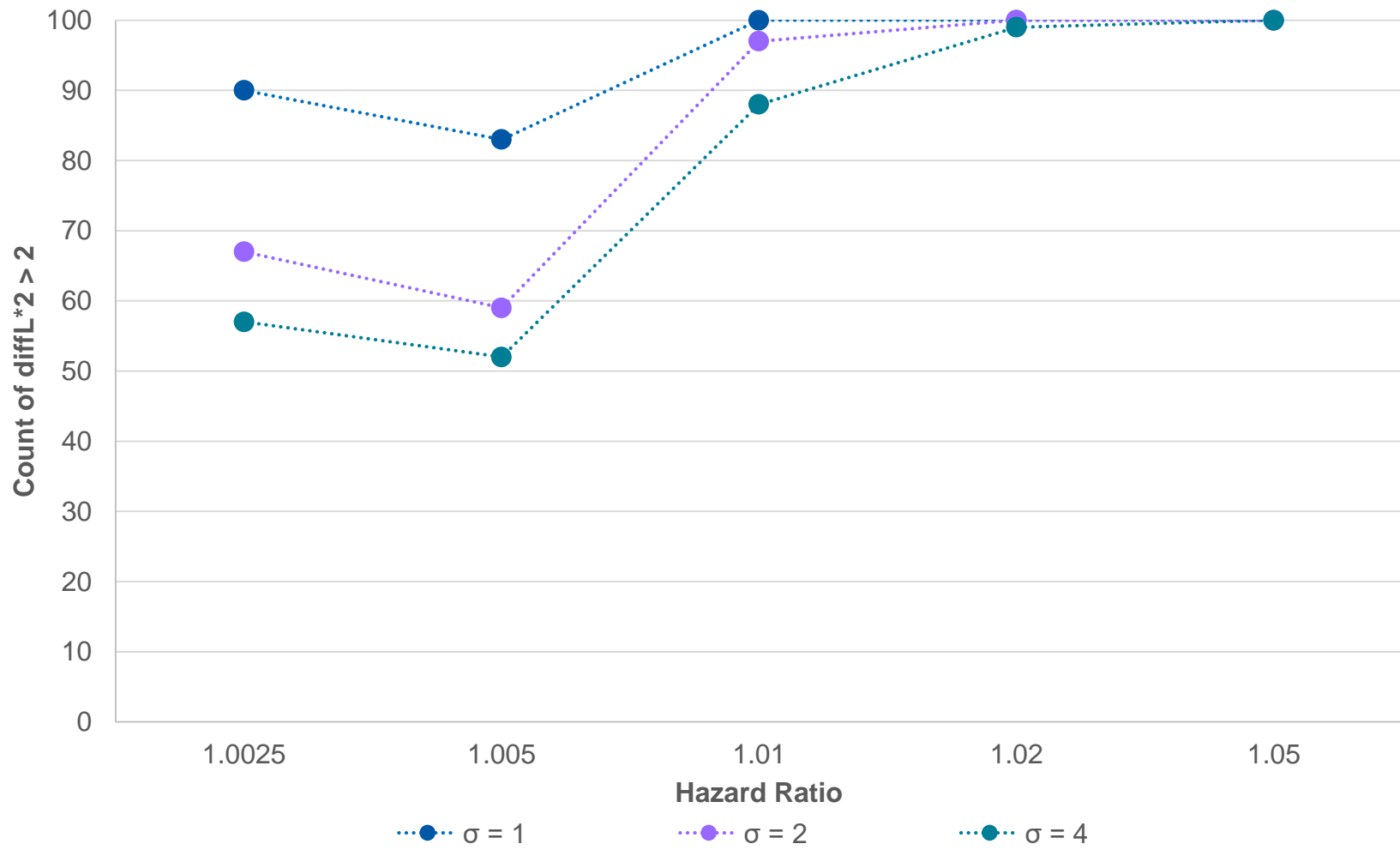
Threshold = 7, $2 \times \Delta LL > \ln(\text{nevents})$



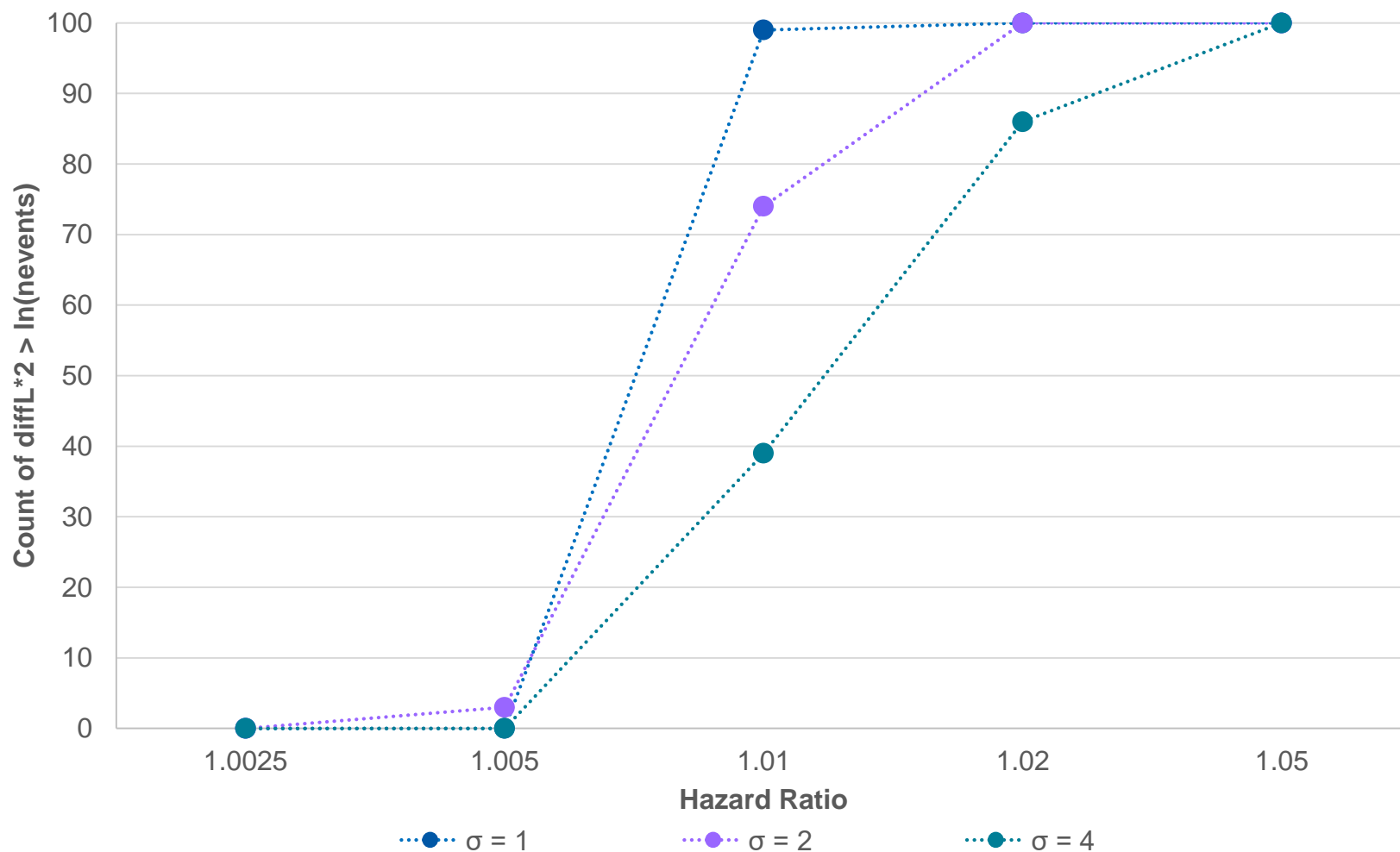
Threshold = 7, $2 \times \Delta LL > \ln(n)$



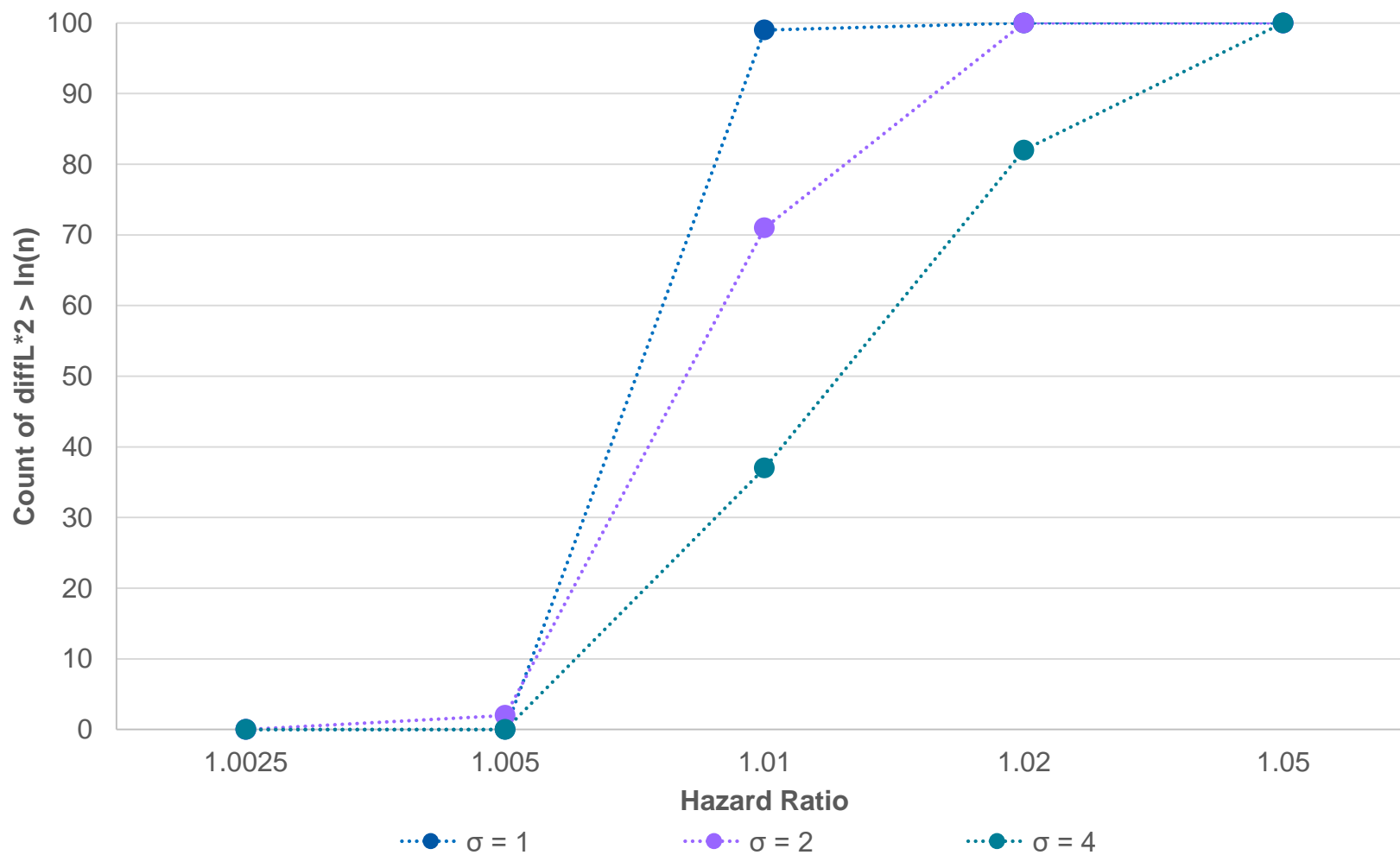
Threshold = 8.5, 2 × ΔLL > 2



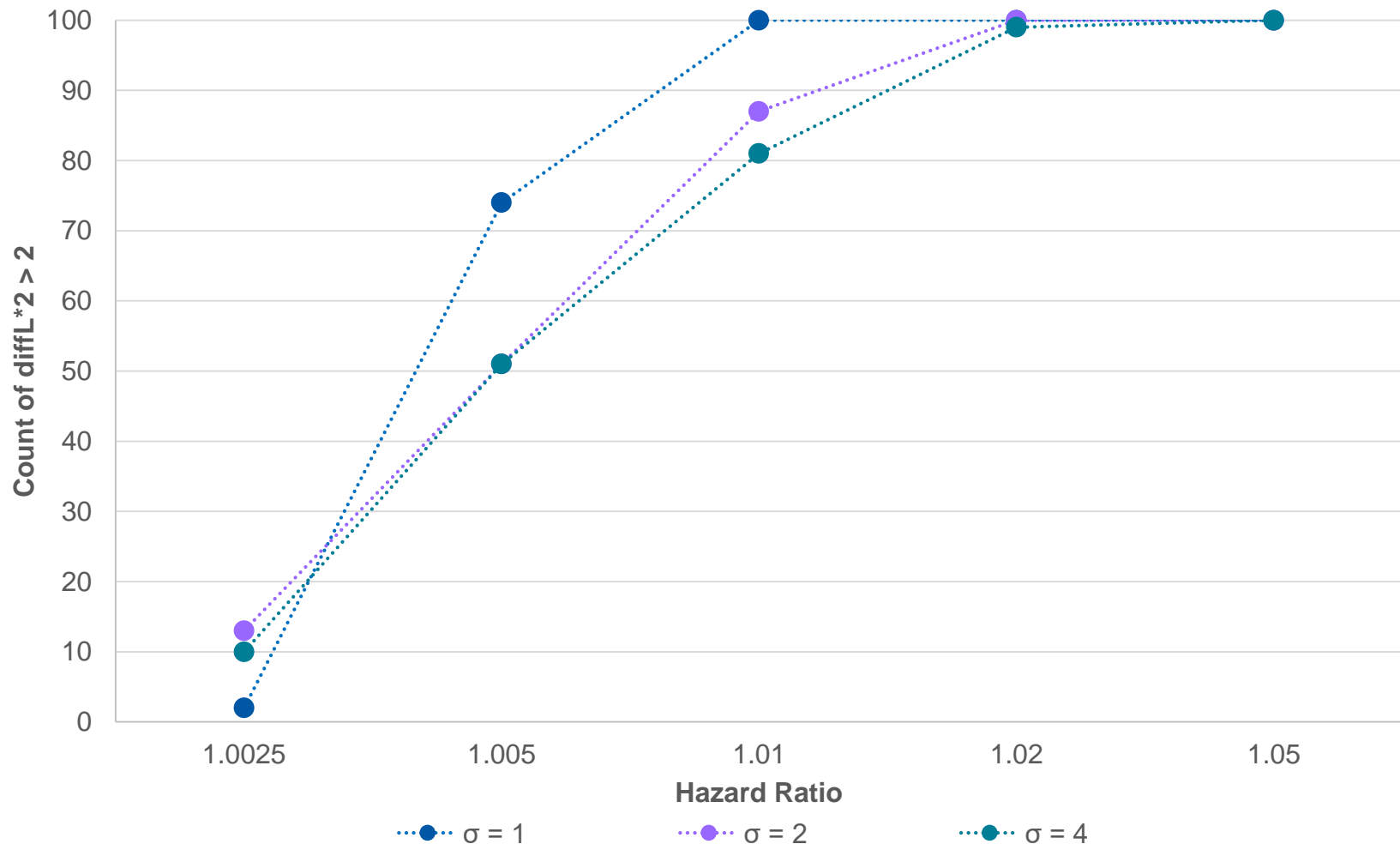
Threshold = 8.5, $2 \times \Delta LL > \ln(\text{nevents})$



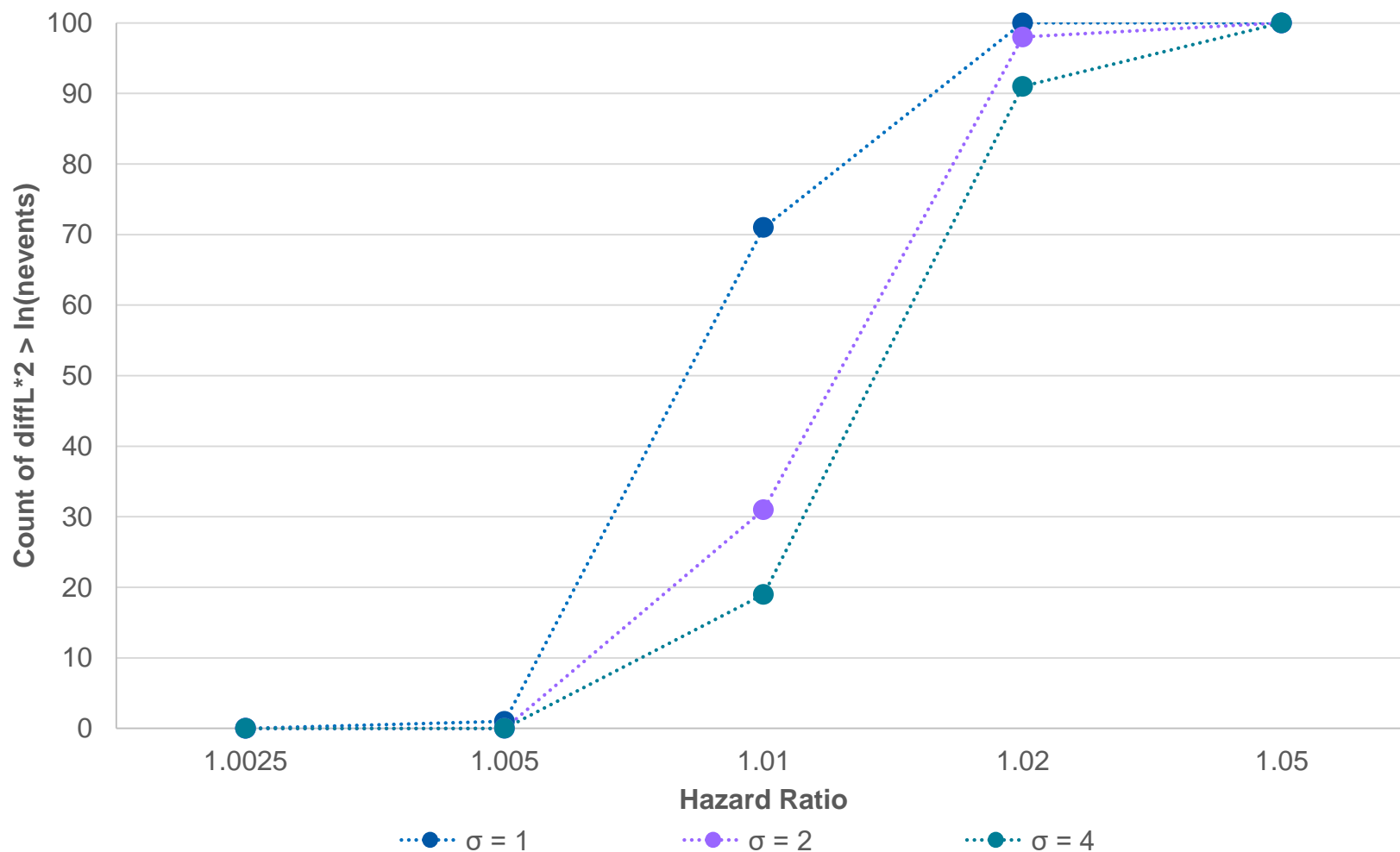
Threshold = 8.5, $2 \times \Delta LL > \ln(n)$



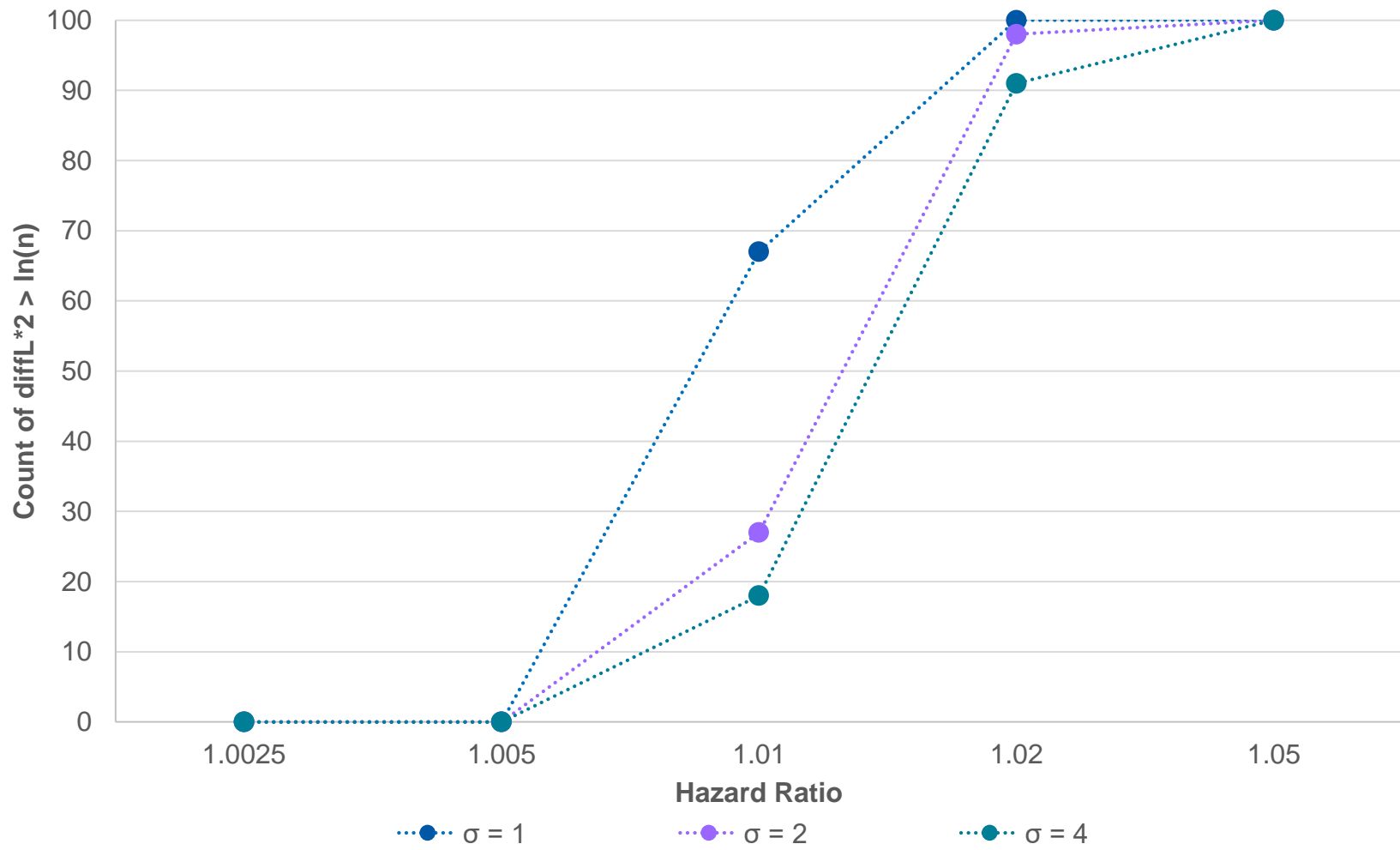
Threshold = 9.5, 2 × ΔLL > 2



Threshold = 9.5, $2 \times \Delta LL > \ln(\text{nevents})$



Threshold = 9.5, $2 \times \Delta LL > \ln(n)$





4. Estimated Threshold Locations With Cox Proportional Hazard Models Under Measurement Error

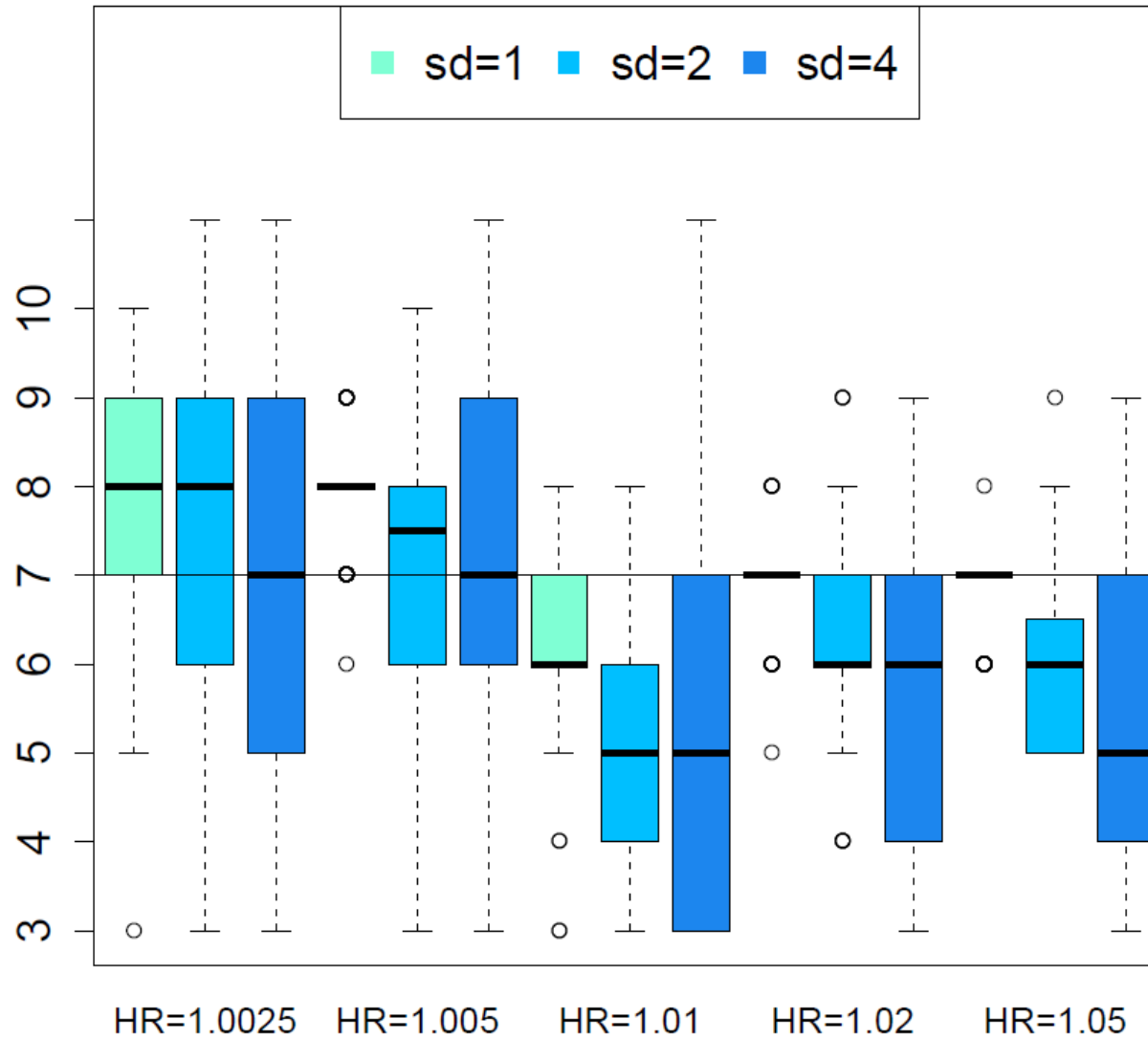
Estimating Threshold Location with the Cox Proportional Hazard Tests: Summary of Results

- With moderate amounts of measurement error ($sd = 2$), $HR=1.005$, and a threshold higher than the mean PM level (9.5), over 75% of the estimated threshold values were too low.
- Underestimates of the threshold increase in magnitude and frequency as measurement error increases.
- Underestimates of the threshold decrease in magnitude and frequency as the threshold increases.
- Underestimates of the threshold increase in magnitude and frequency as the hazard ratio increases.

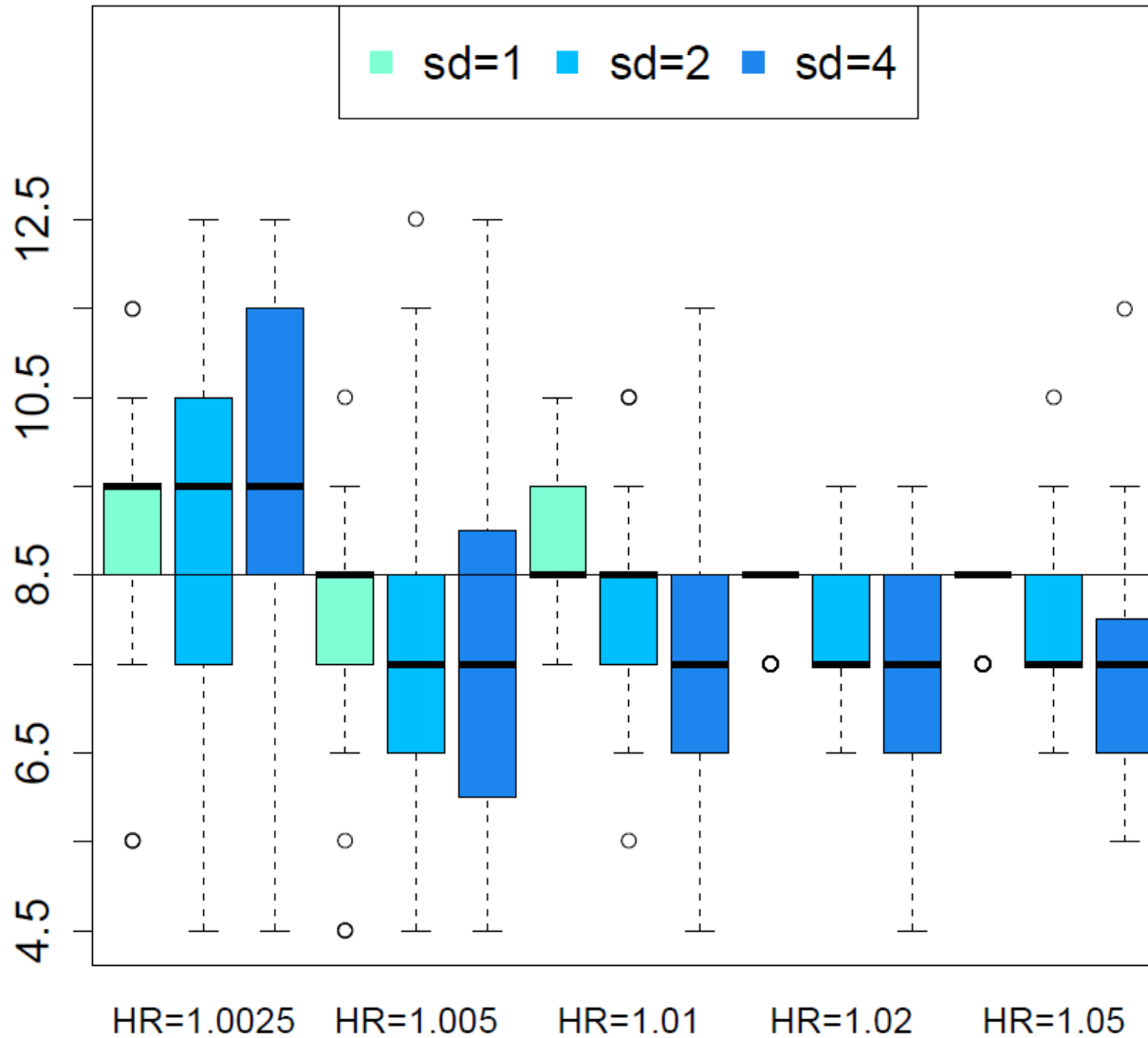
Threshold Estimation Results

- The plots below are boxplots.
 - The thick black line is the median estimate.
 - The colored box indicates the interquartile range (the middle 50% of the estimates).
 - The “whiskers” indicate the complete range of the data, excluding outliers.
 - The dots are outliers (defined as further than 1.5 times the interquartile range from the end of the interquartile range).

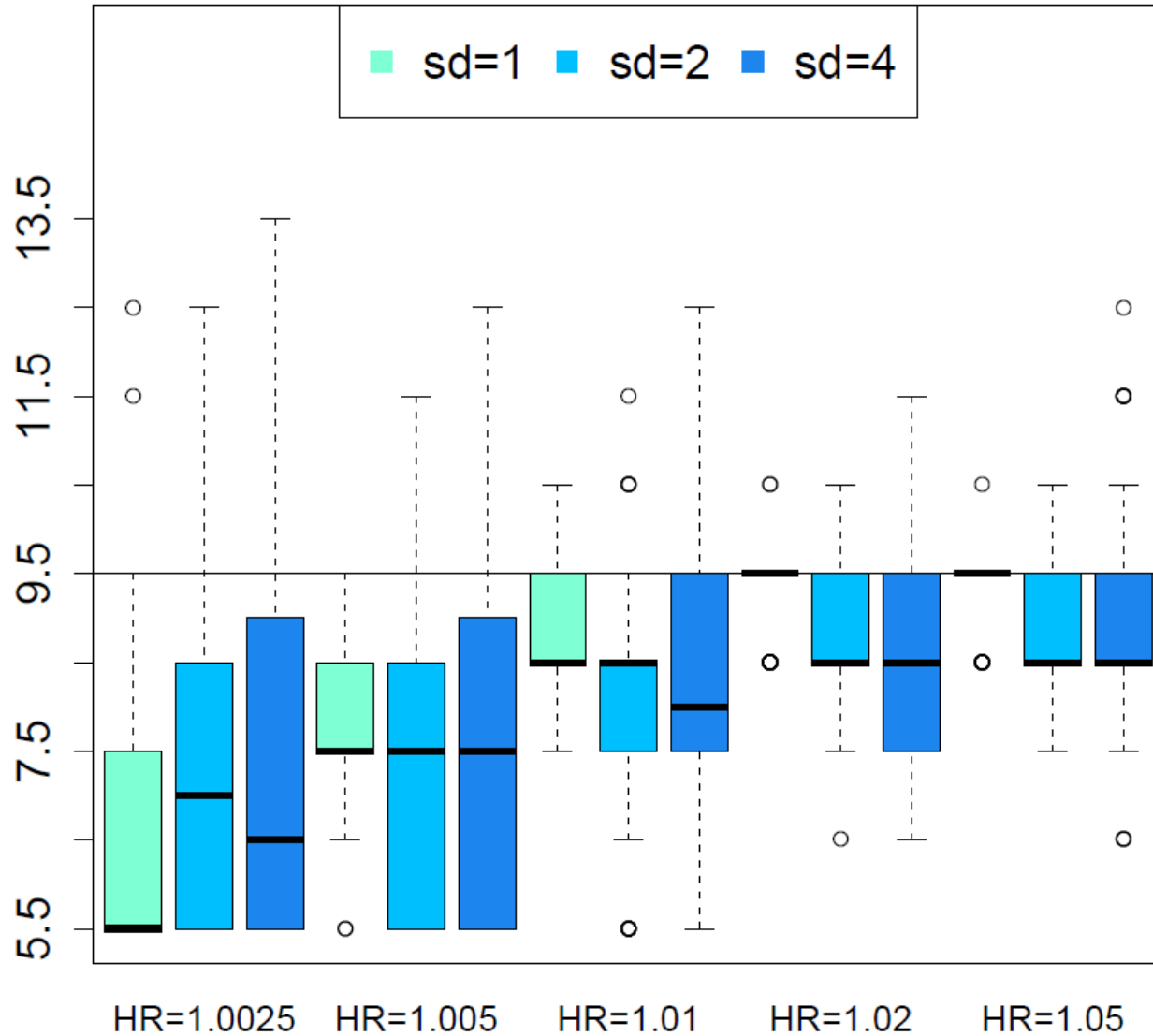
Estimated Thresholds When True Threshold = 7



Estimated Thresholds When True Threshold = 8.5



Estimated Thresholds When True Threshold = 9.5



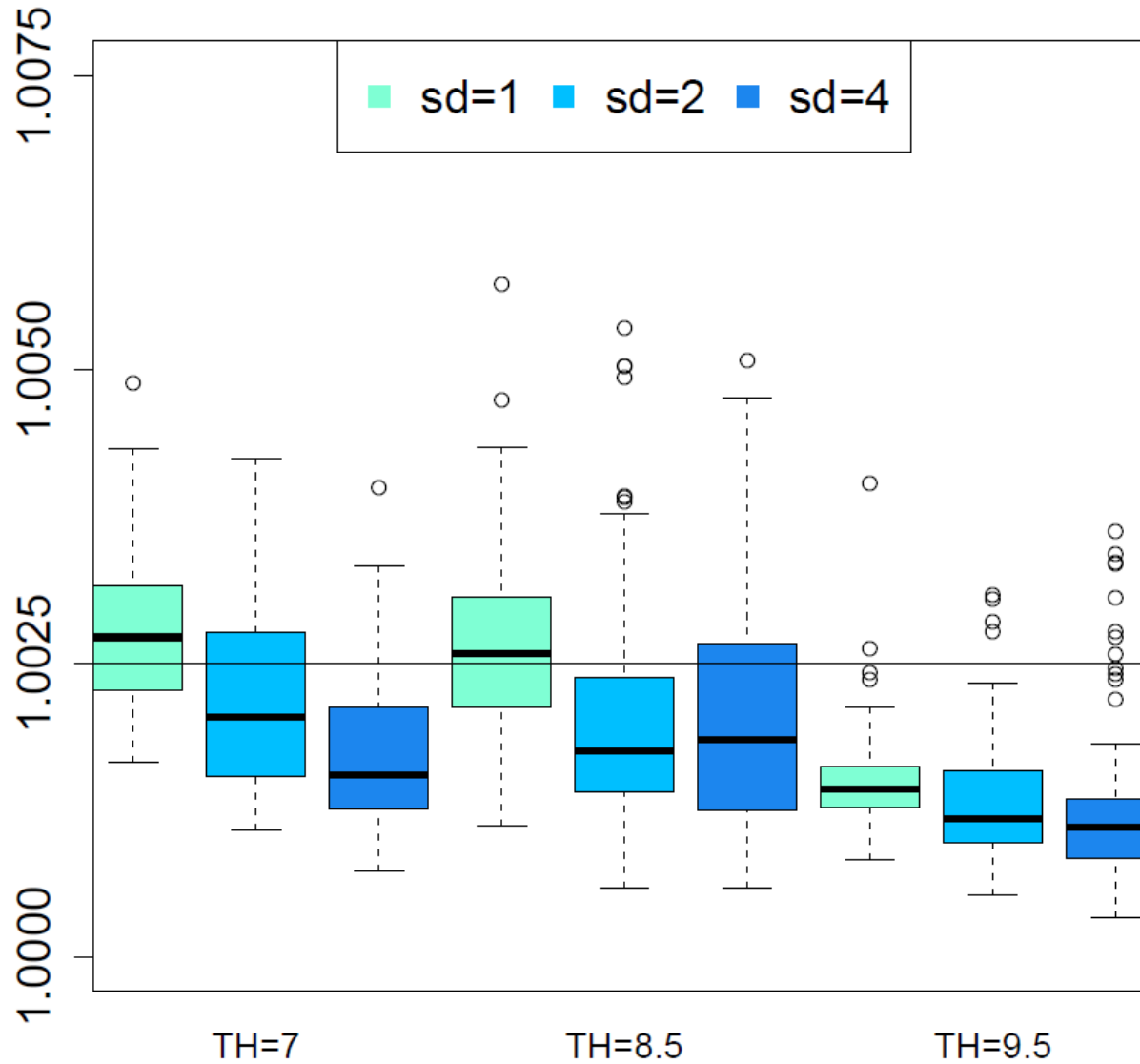


5. Estimated Hazard Ratios With Cox Proportional Hazard Models Under Measurement Error

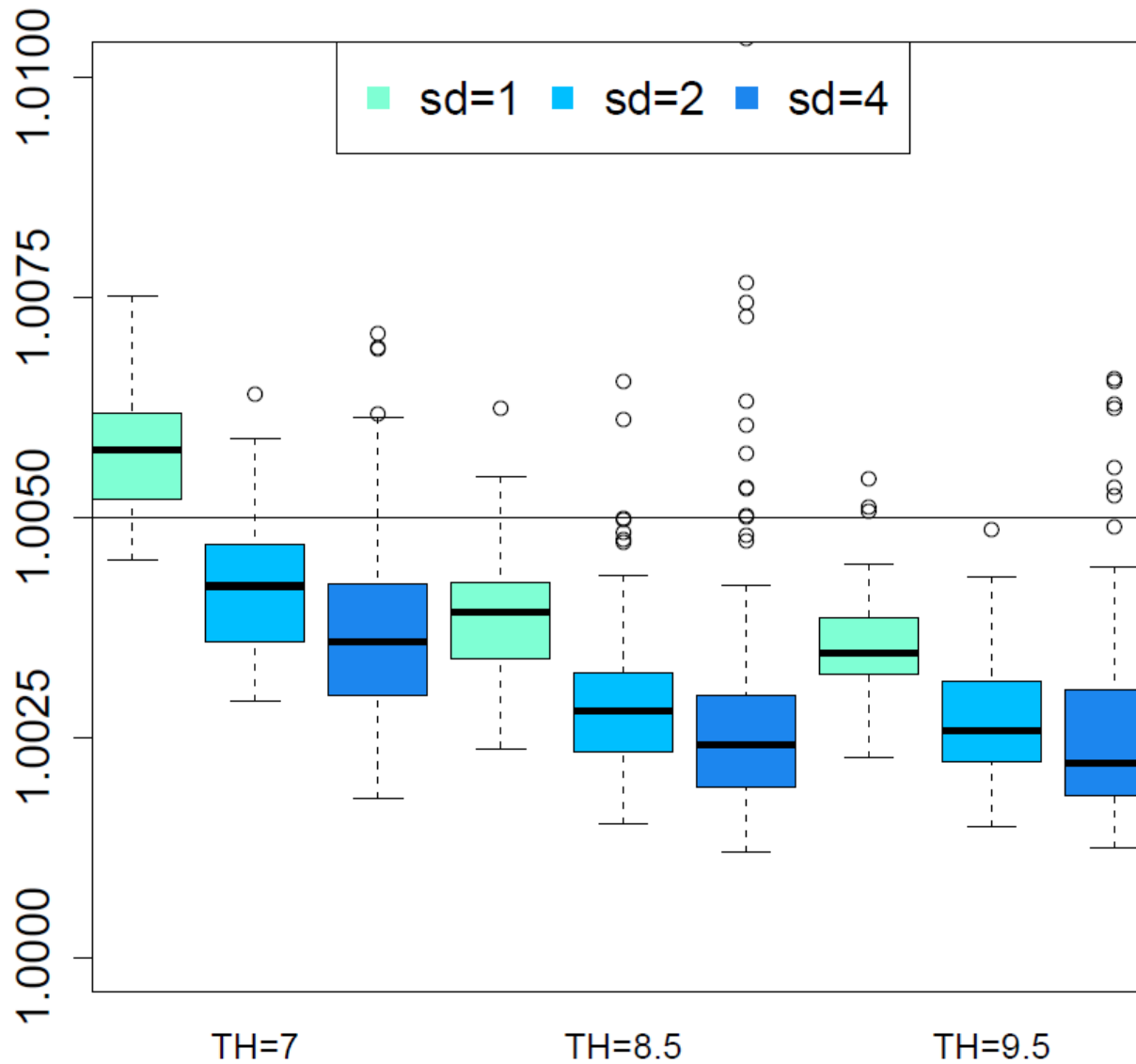
Estimating Hazard Ratios with the Cox Proportional Hazard Tests: Summary of Results

- With moderate amounts of measurement error ($sd = 2$), and a threshold higher than the mean PM level (9.5), all of the estimated hazard ratios were too low when the true $HR=1.005$, with the median estimate approximately half of the true HR.
- The hazard ratios become more attenuated as measurement error increases.
- The hazard ratios become more attenuated as the threshold increases.

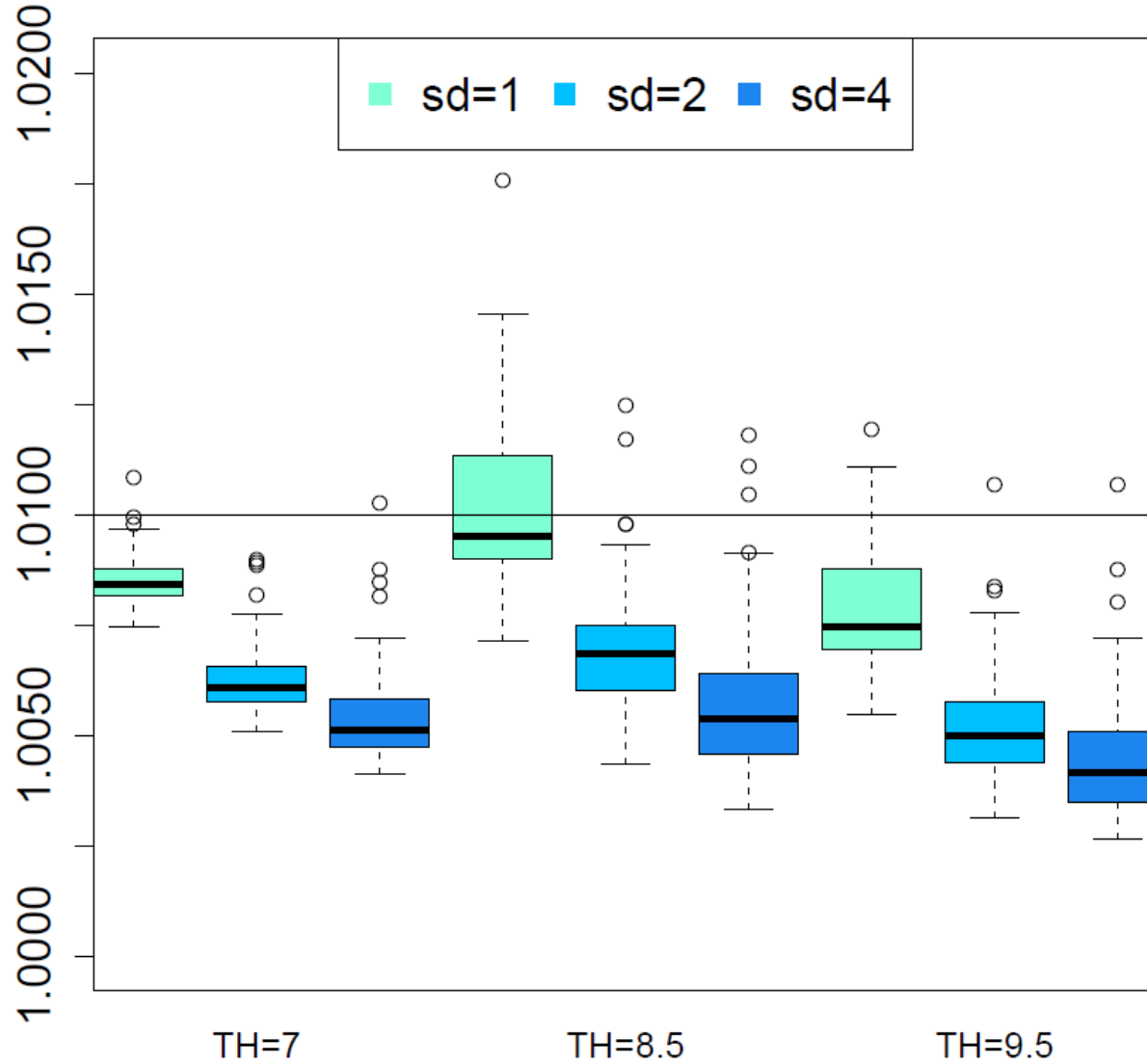
Estimated Hazard Ratios When the True Hazard Ratio is 1.0025



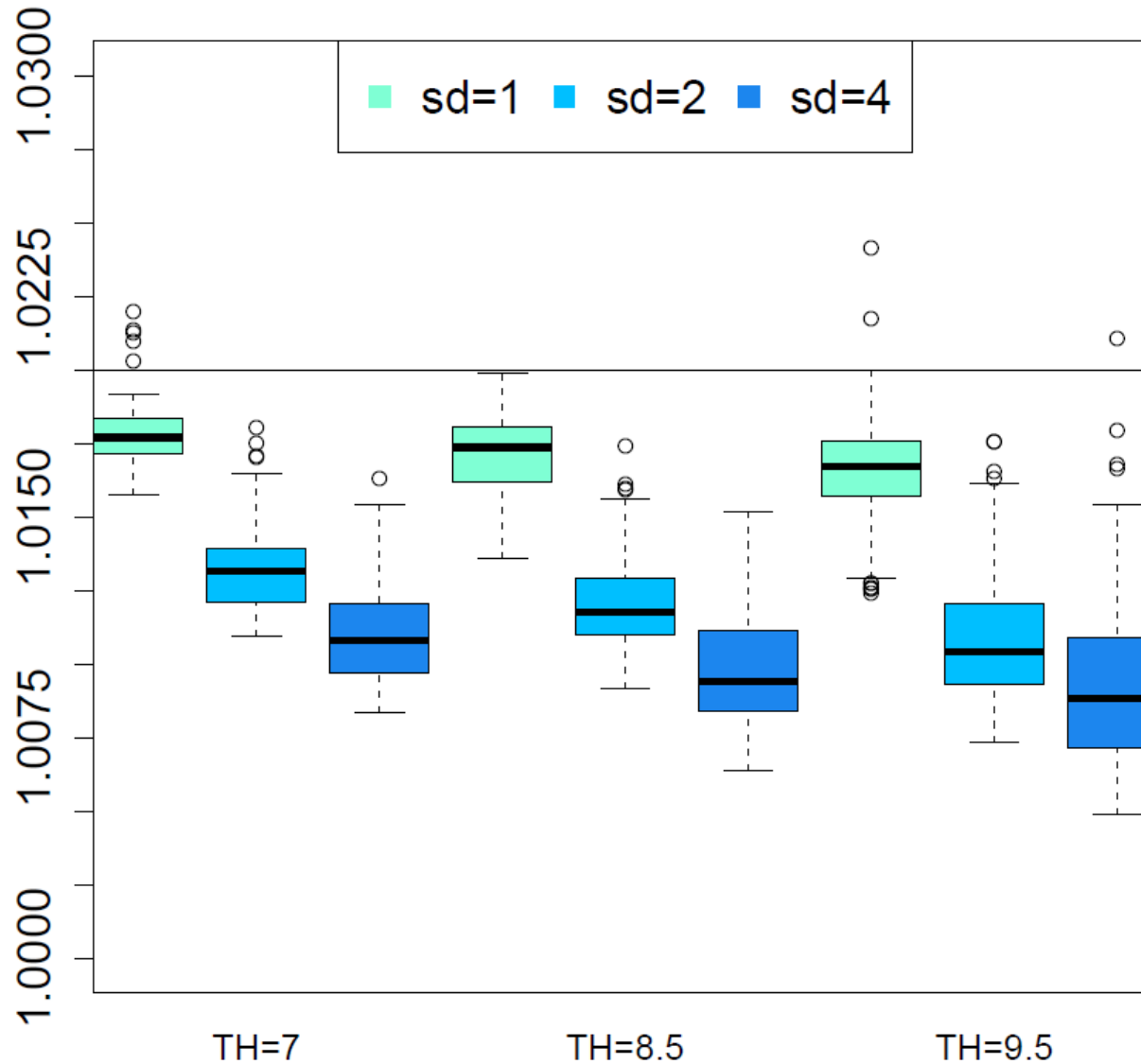
Estimated Hazard Ratios When the True Hazard Ratio is 1.005



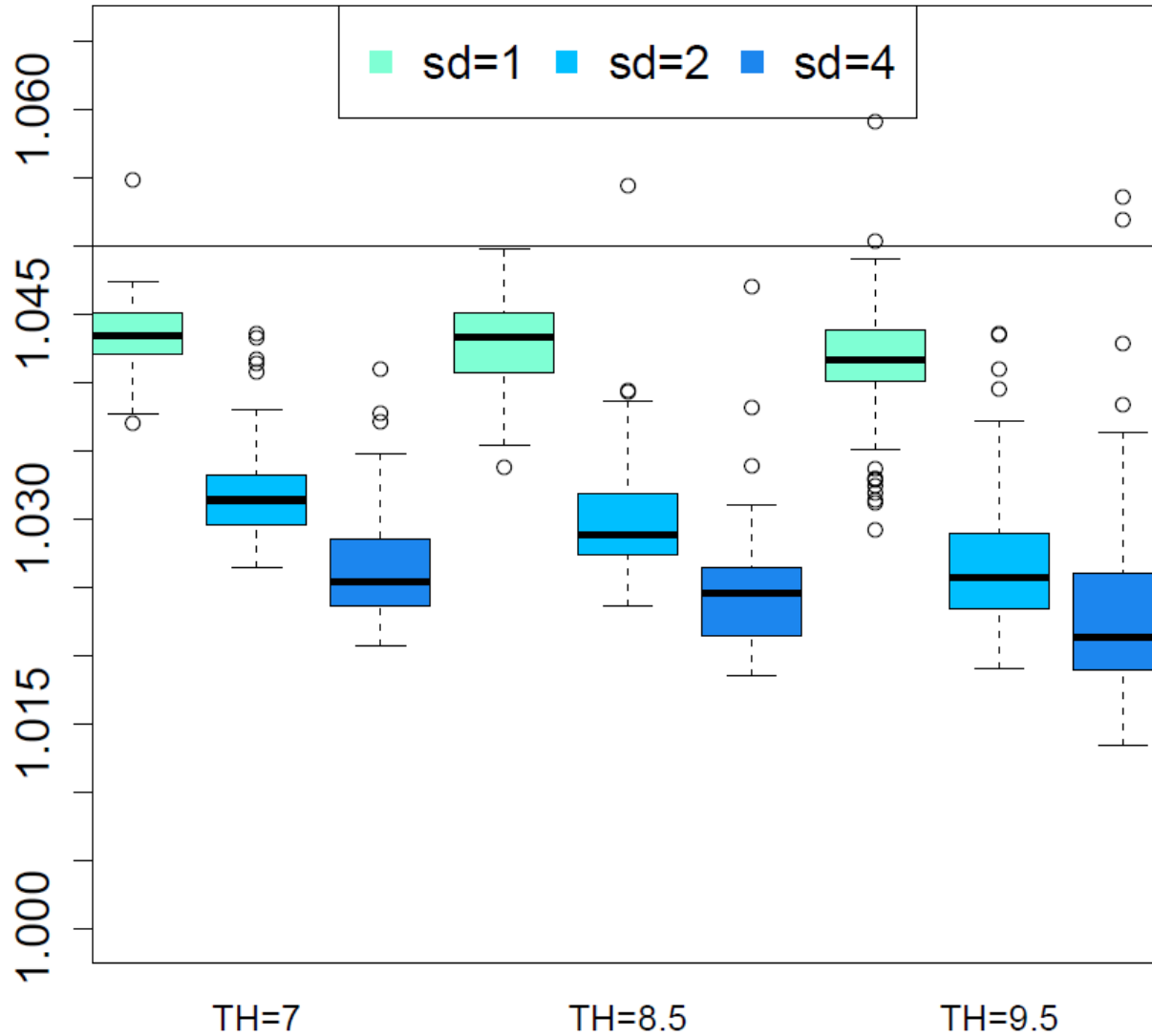
Estimated Hazard Ratios When the True Hazard Ratio is 1.01



Estimated Hazard Ratios When the True Hazard Ratio is 1.02



Estimated Hazard Ratios When the True Hazard Ratio is 1.05



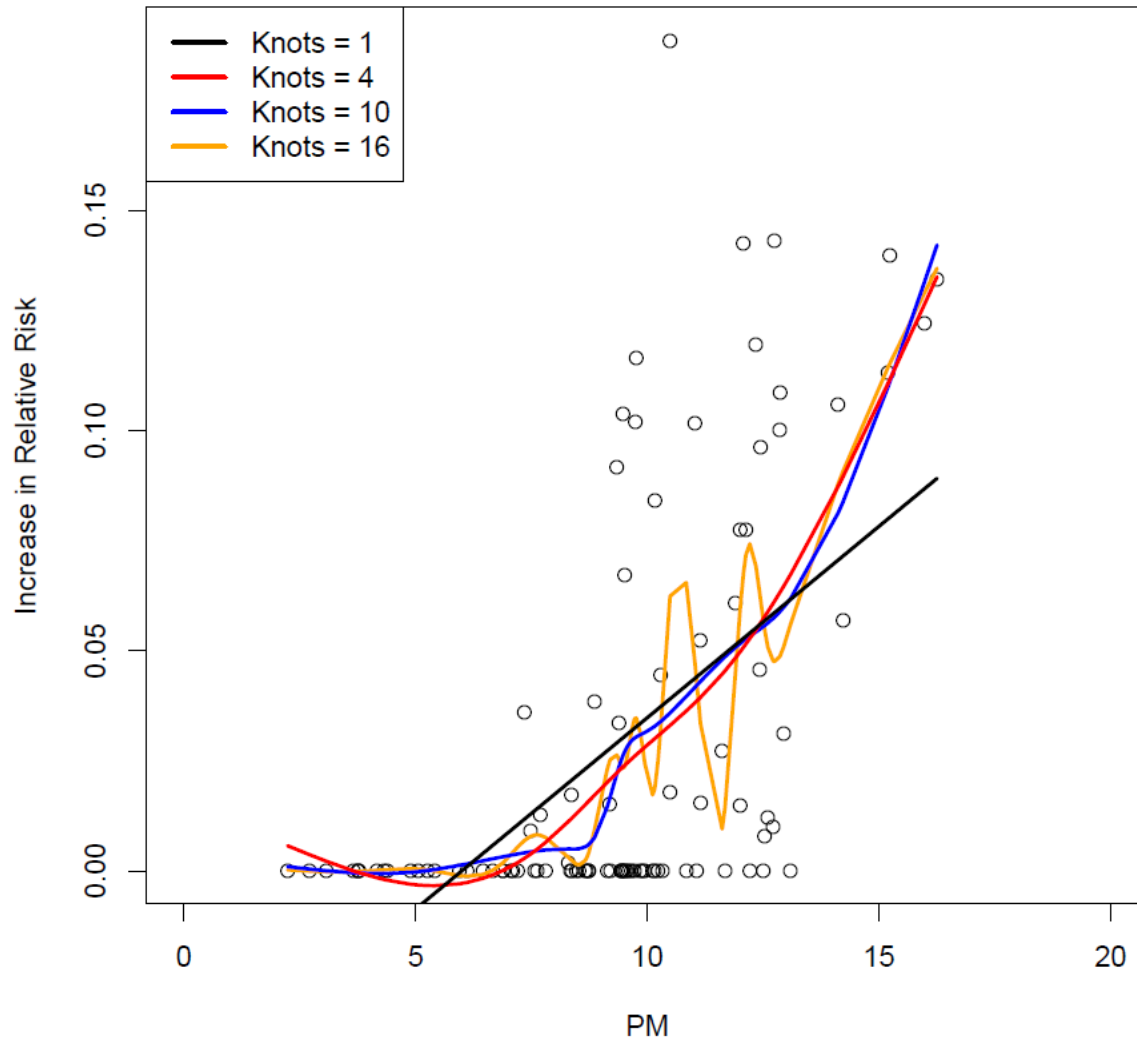


6. Using Nonparametric Regressions to Examine Mortality Data for Thresholds

Nonparametric Regression Techniques

- Splines fit piecewise polynomial functions between a set of “knots.”
 - Knots are usually set at the quantiles of the data (e.g., 3 knots would be at the 25th, 50th, and 75th quantiles of PM).
 - More knots = more “wiggly” lines.
- Loess runs a series of regressions on a “span” of data (e.g., 20%) around each data point. The loess line connects the predicted points.
 - Data further from the point being predicted gets a lower weight.
 - Smaller spans = more “wiggly” lines.
- Splines and loess can be made to resemble each other arbitrarily closely. We examine splines below.

Examples of Splines



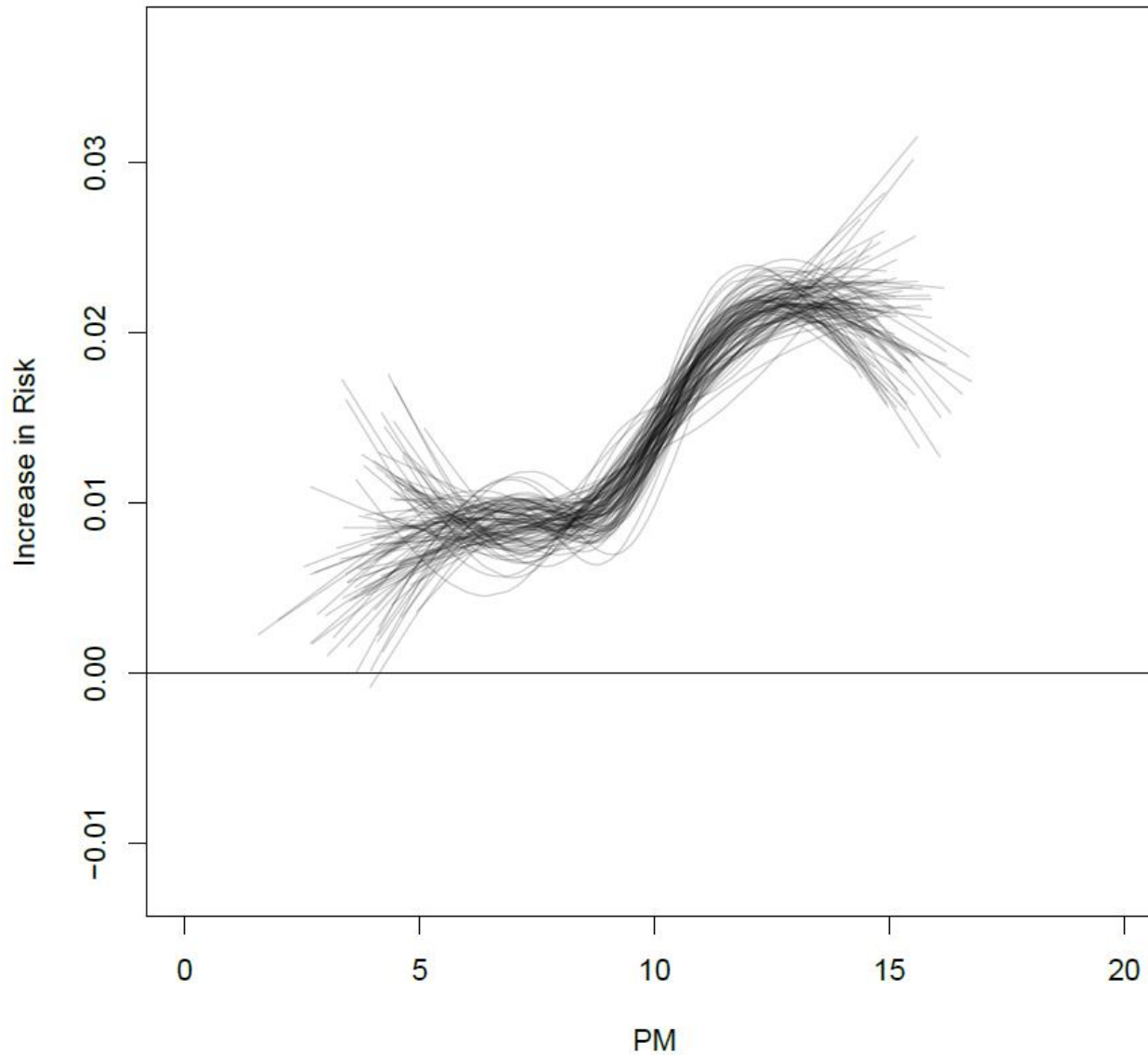
Fitting Splines to the Simulated Data

- We fit splines to the PM mortality data.
- The dependent variable is relative risk, with relative risk defined as 1 for the lowest level of PM. We estimate the spline on the increase in risk (we subtracted 1 from RR – this makes no different to model fit).
- Our splines had 4 knots. There is no agreed upon standard in the literature – fit is as much “art” as science.
- Some literature has tested for nonlinearity or thresholds by testing a spline against a linear regression. The properties of these tests aren’t clear, especially since the fit of the spline depends in part of the judgement of the researcher.

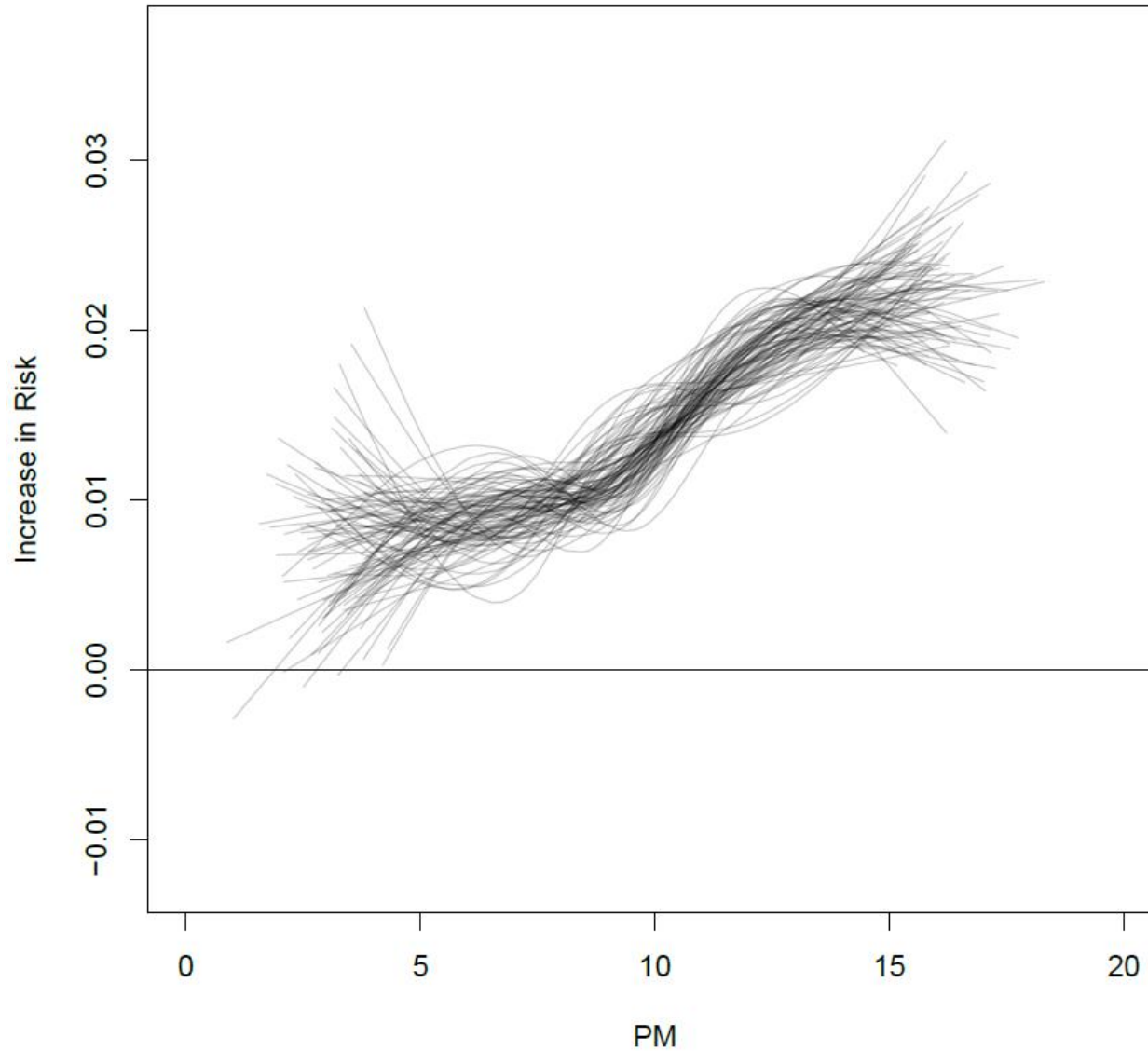
Examples of Splines in the Simulated Data

- Below we plot all splines estimated across 100 simulations for a threshold of 9.5 and a hazard ratio of 1.005, for varying levels of measurement error.
- As measurement error increases, the increasing attenuation of the hazard ratio and the decreasing ability to detect the threshold are clear.
- Note the increases in risk are positive for most cities because a few low PM cities had relatively low mortality. Nevertheless, the threshold shape is still apparent at $sd=1$.

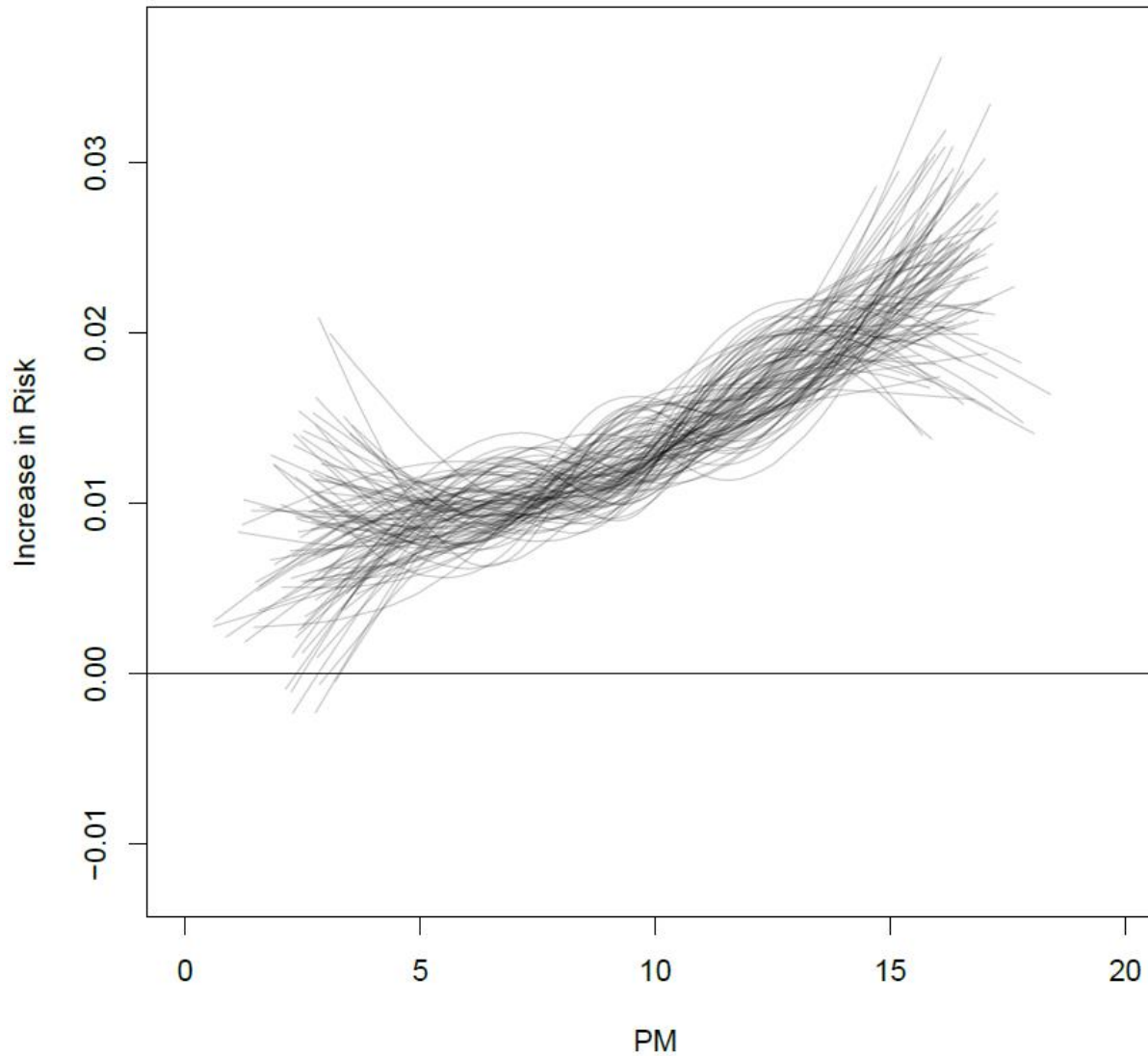
Splines Across 100 Simulations, Threshold = 9.5, HR = 1.005, SD=1



Splines Across 100 Simulations, Threshold = 9.5, HR = 1.005, SD=2



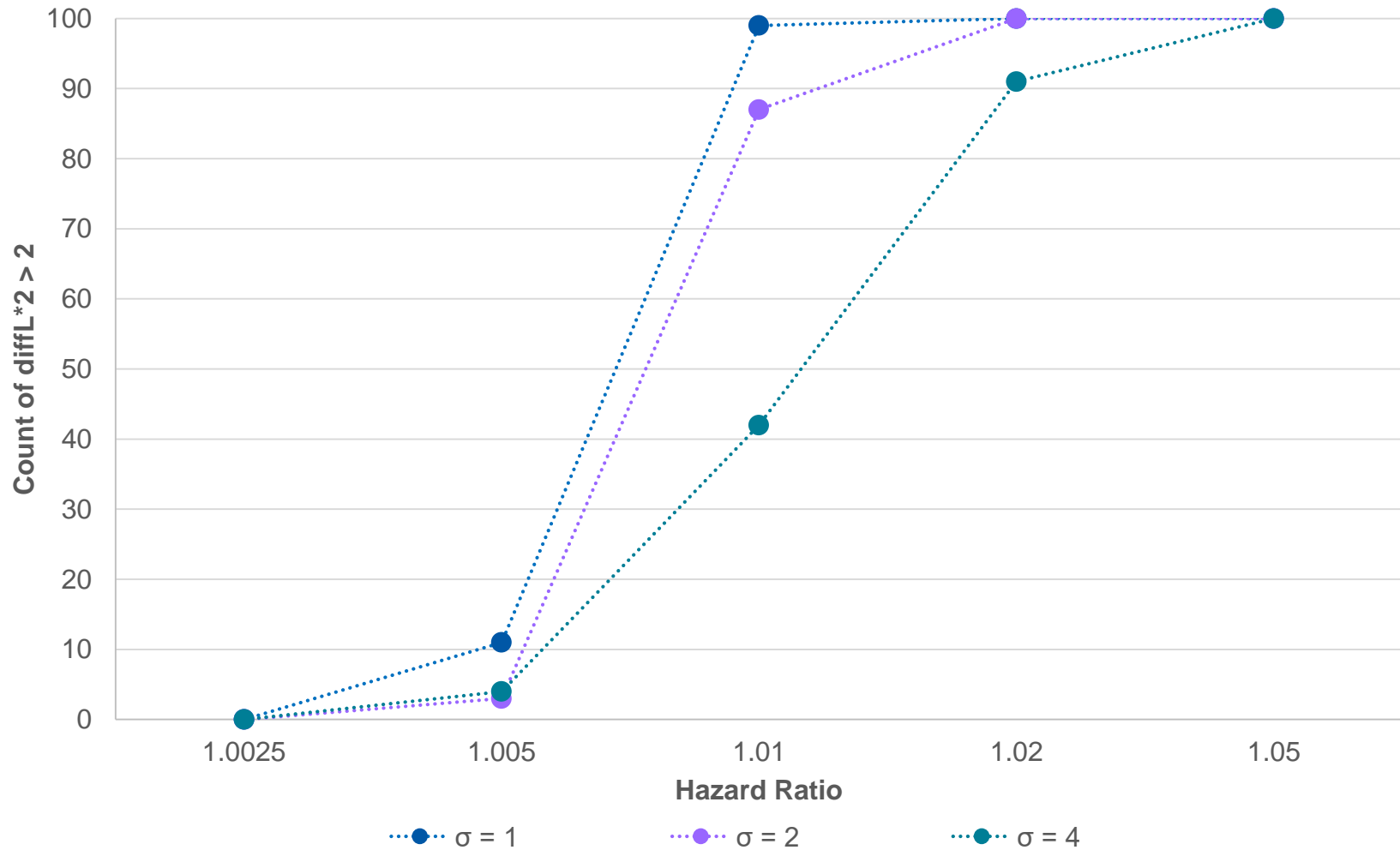
Splines Across 100 Simulations, Threshold = 9.5, HR = 1.005, SD=4



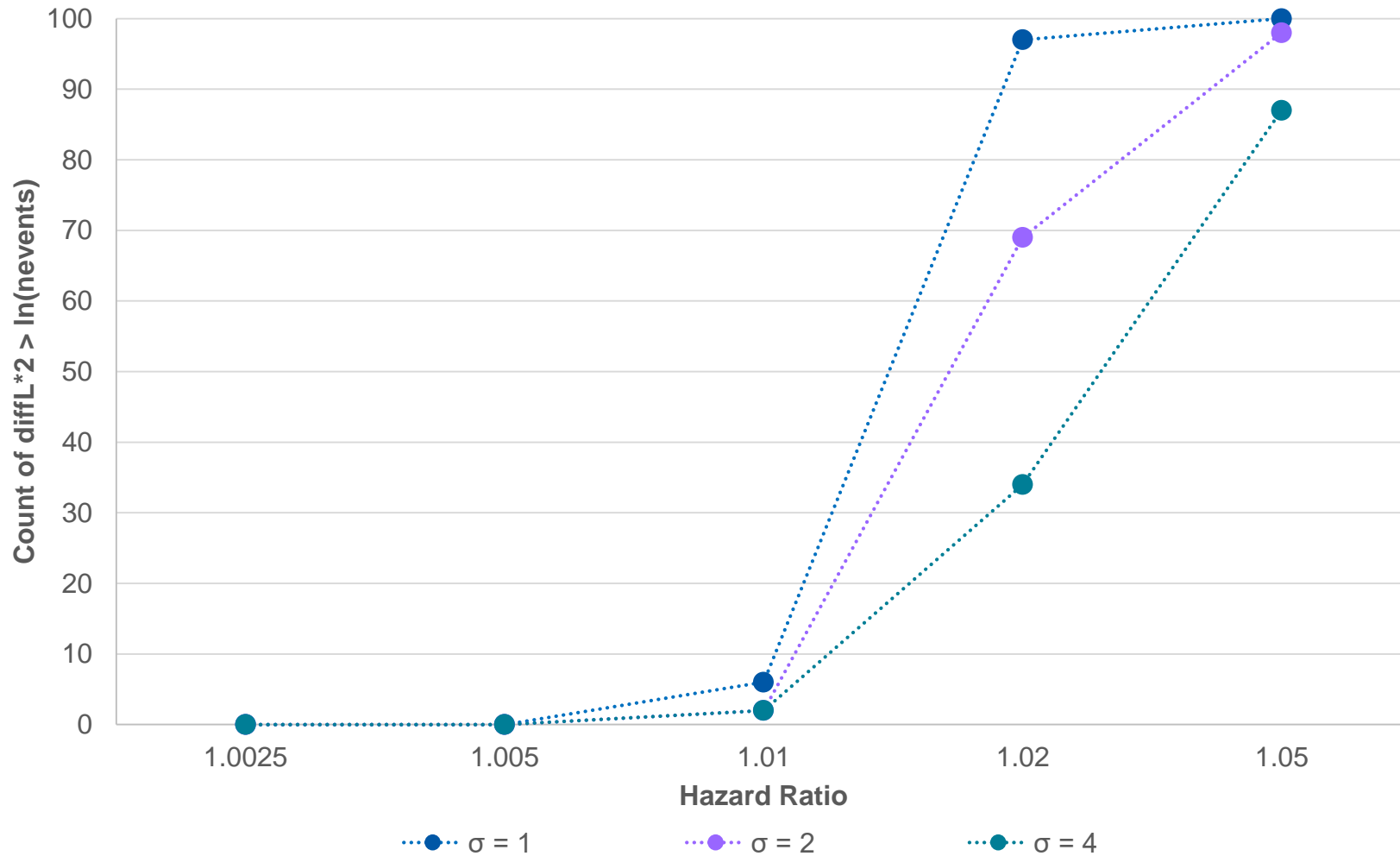


7. Tests for Thresholds With Cox Proportional Hazard Models Under Measurement Error, No Random Variation Across Cities

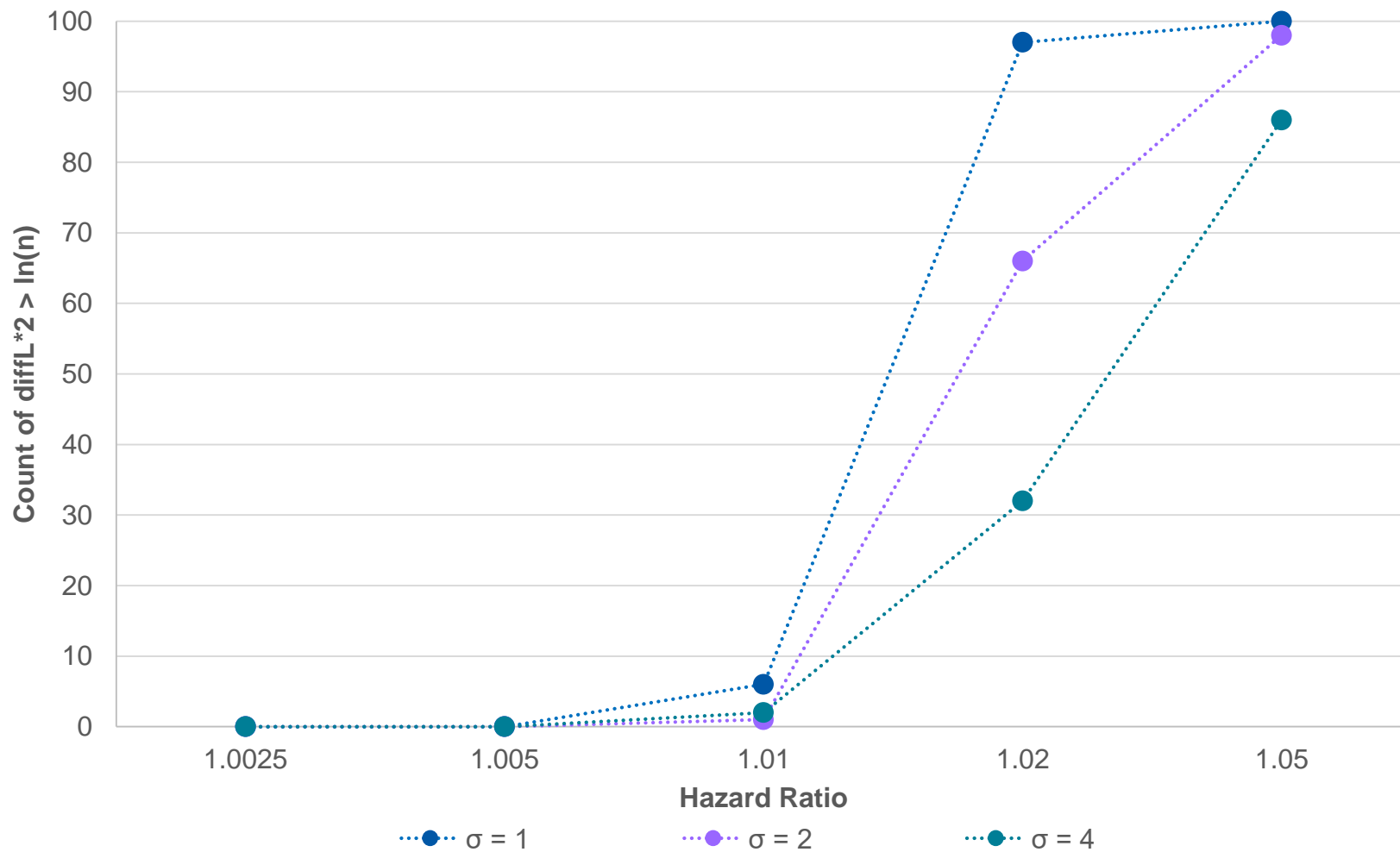
Threshold = 7, $2 \times \Delta LL > 2$



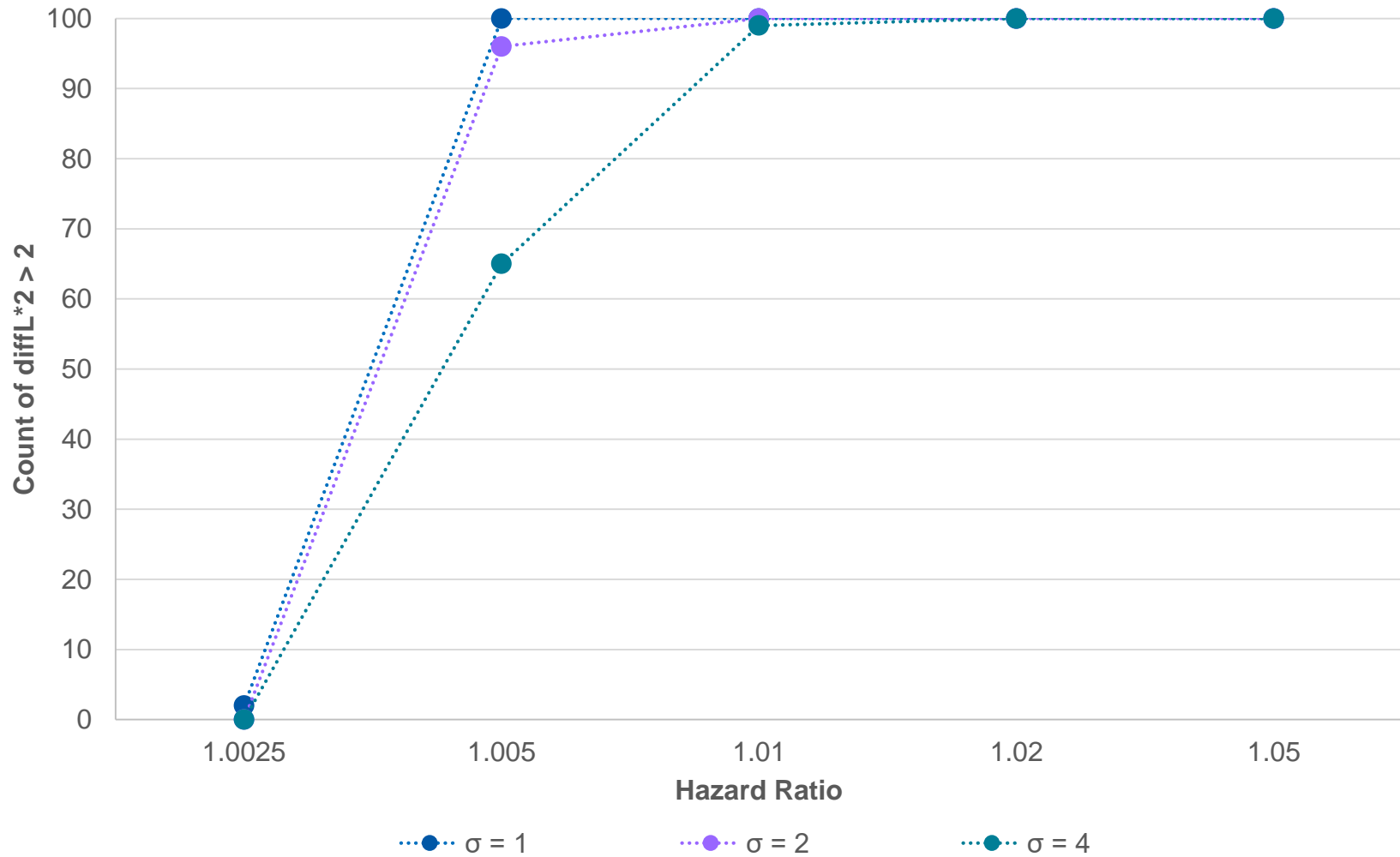
Threshold = 7, $2 \times \Delta LL > \ln(\text{nevents})$



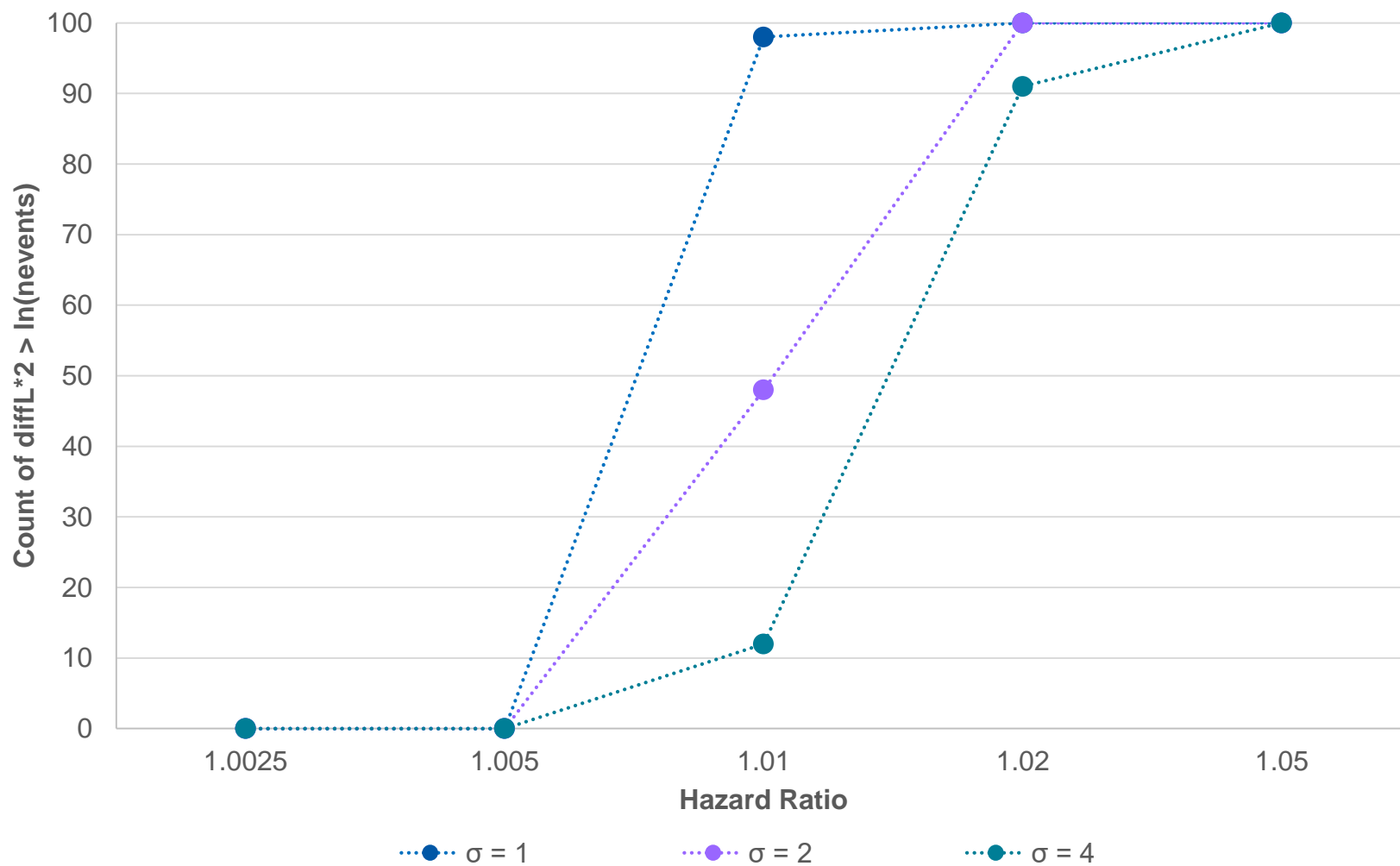
Threshold = 7, $2 \times \Delta LL > \ln(n)$



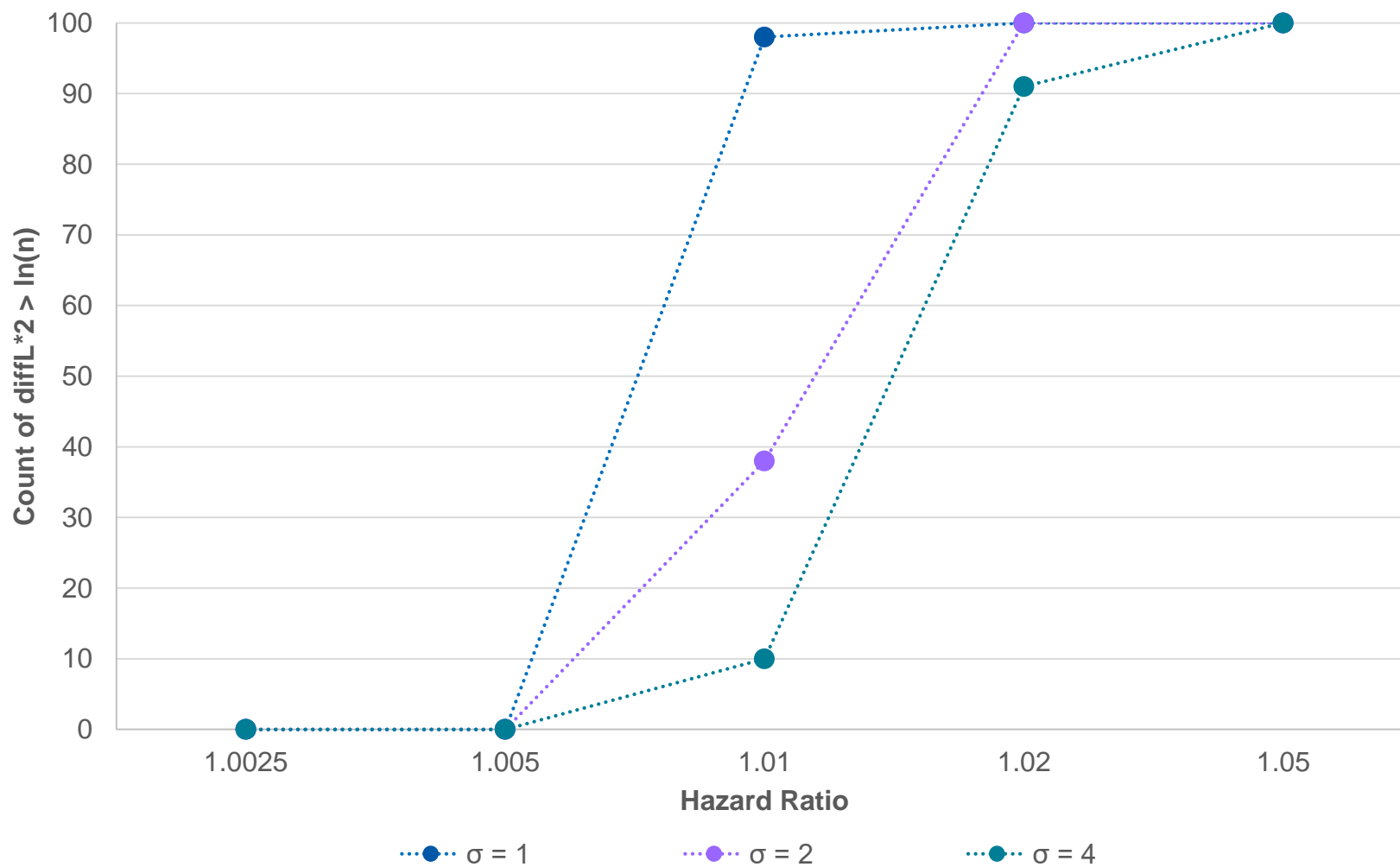
Threshold = 8.5, $2 \times \Delta LL > 2$



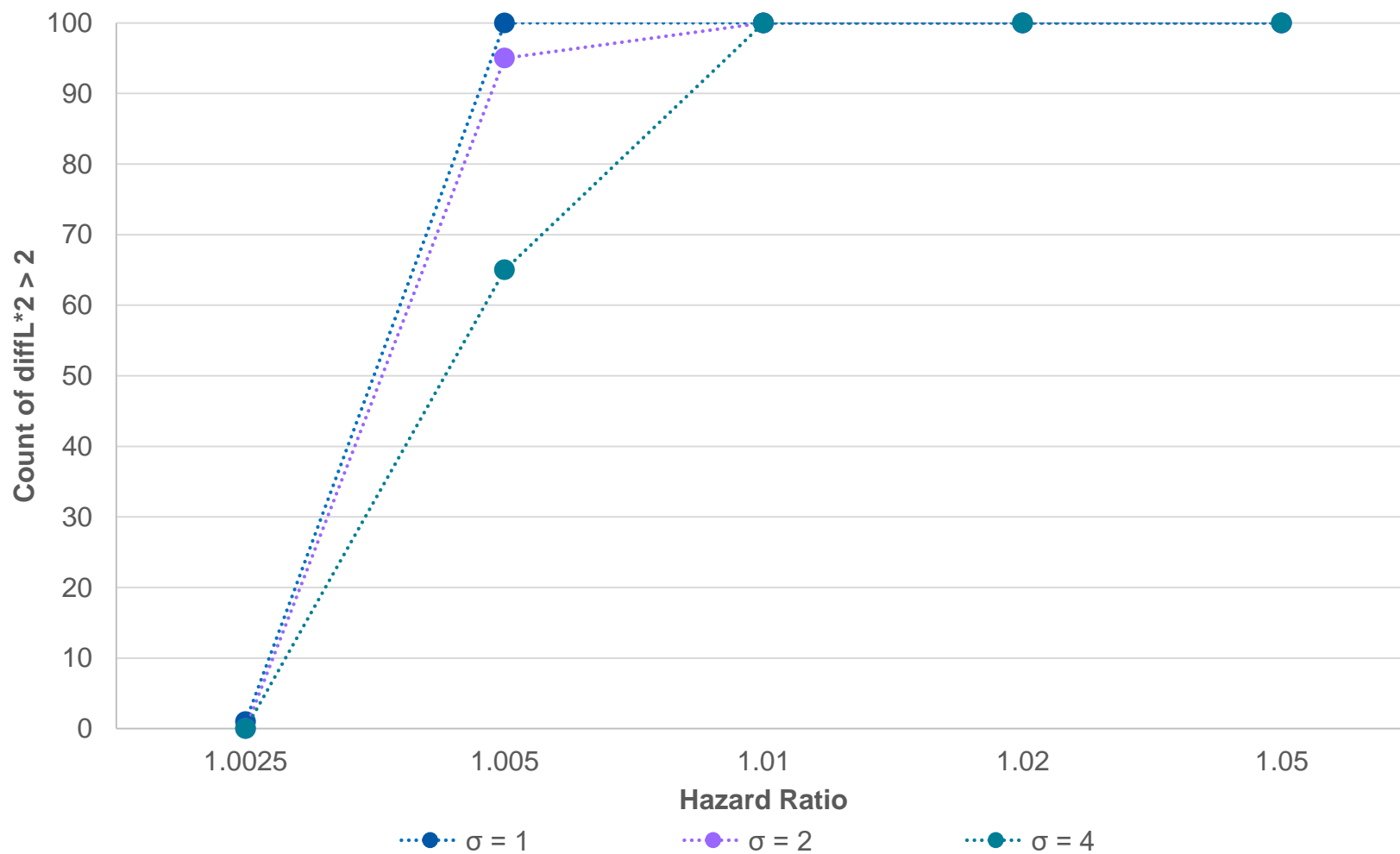
Threshold = 8.5, $2 \times \Delta LL > \ln(\text{nevents})$



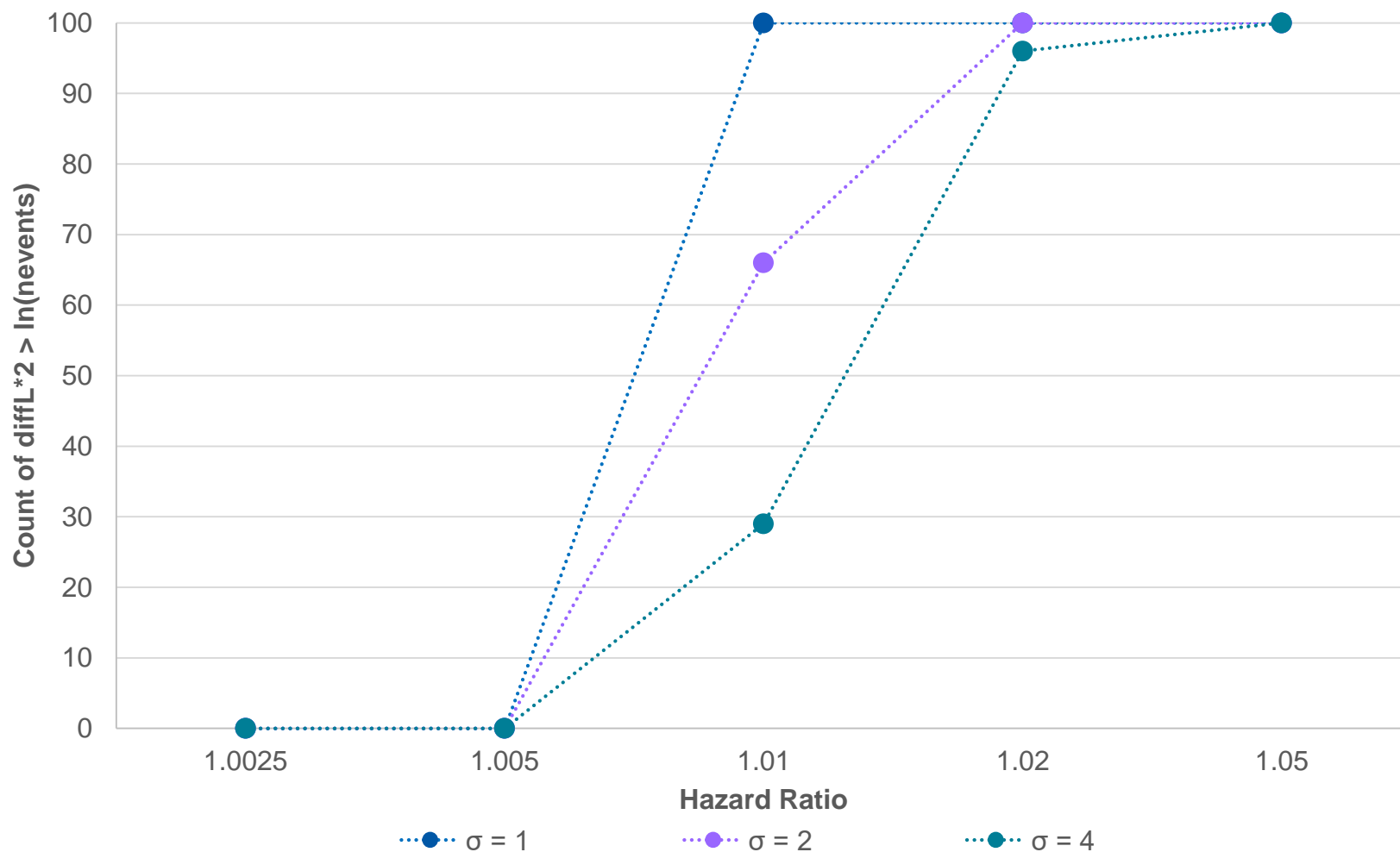
Threshold = 8.5, $2 \times \Delta LL > \ln(n)$



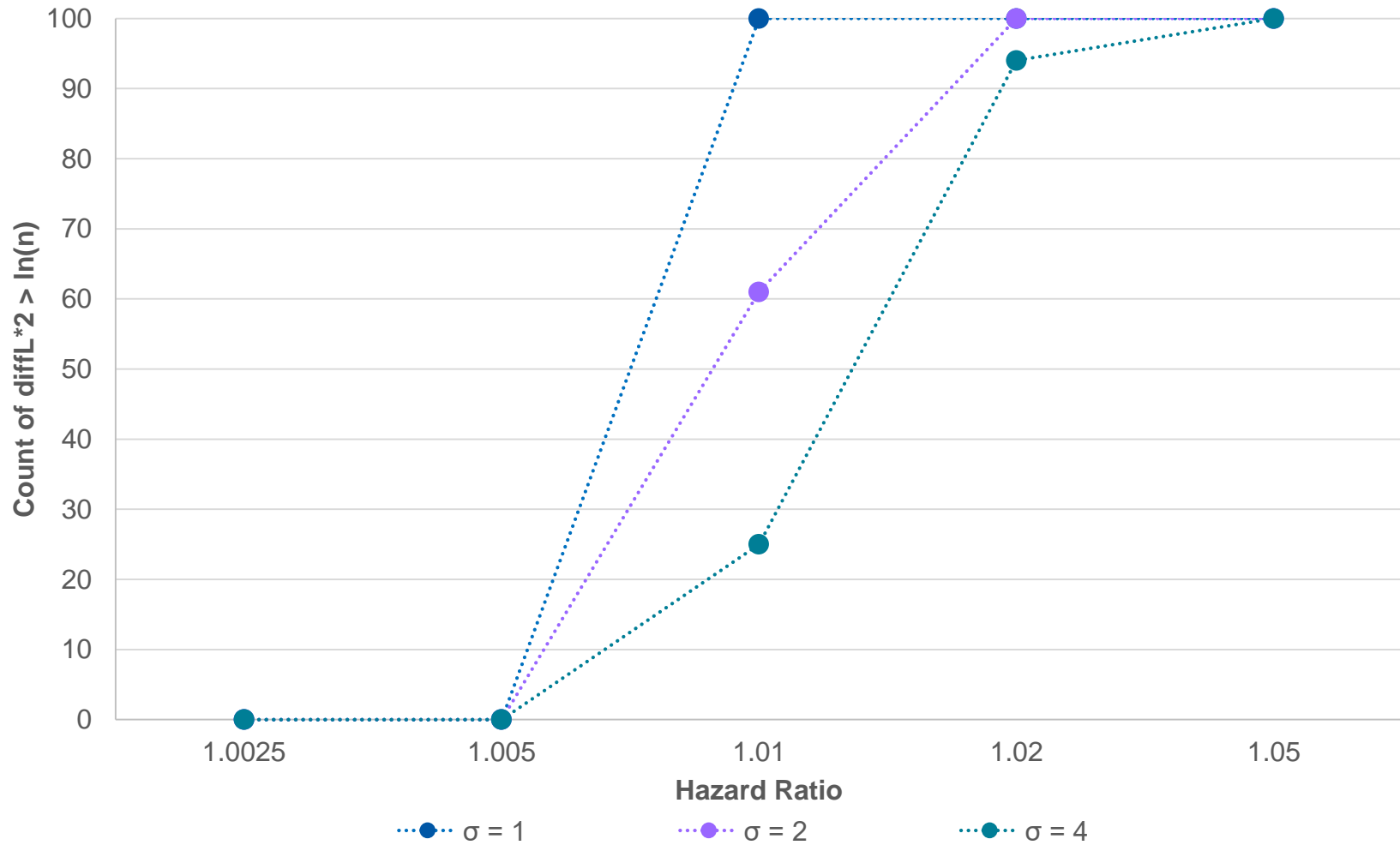
Threshold = 9.5, $2 \times \Delta LL > 2$



Threshold = 9.5, $2 \times \Delta LL > \ln(\text{nevents})$



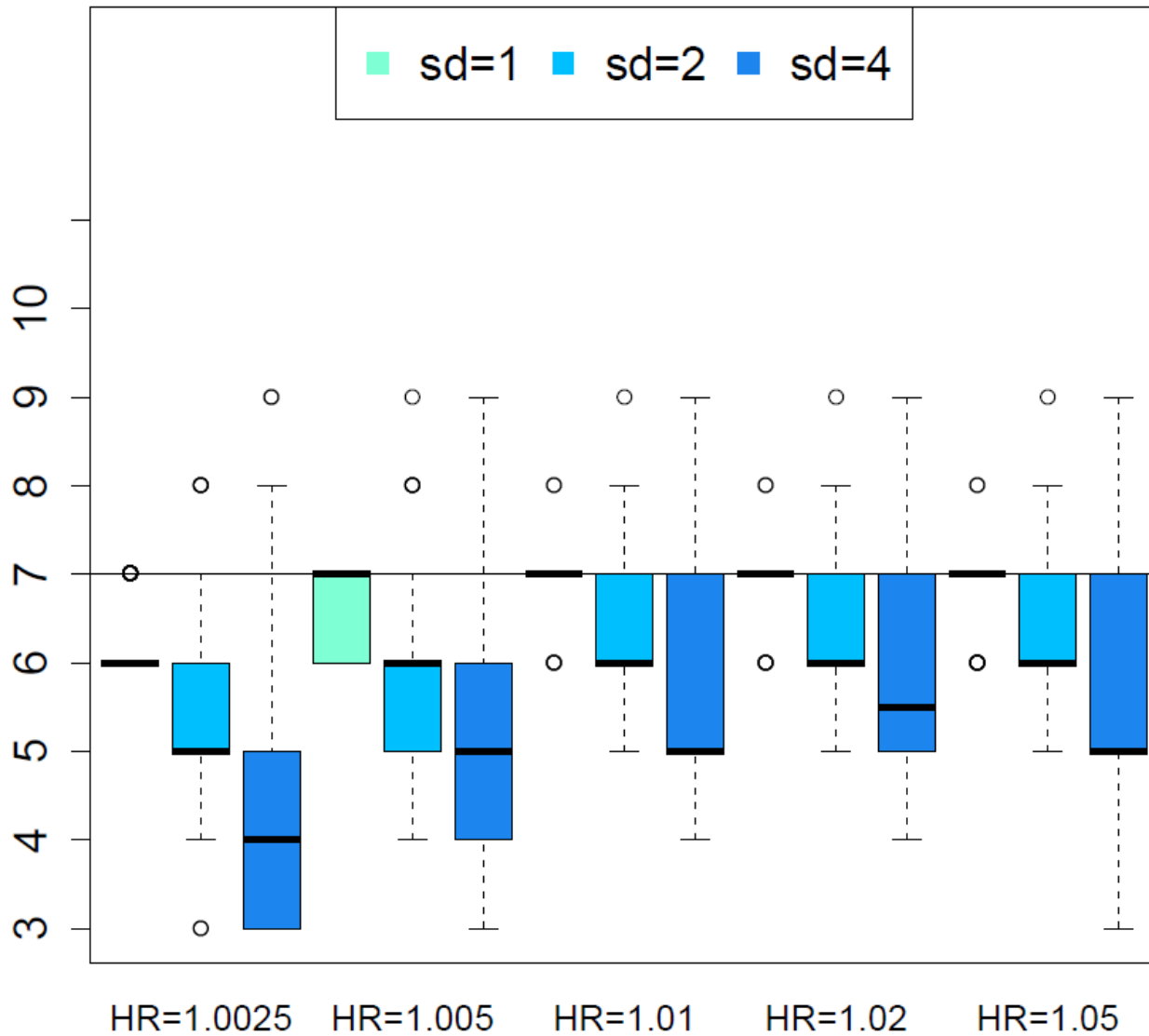
Threshold = 9.5, $2 \times \Delta LL > \ln(n)$



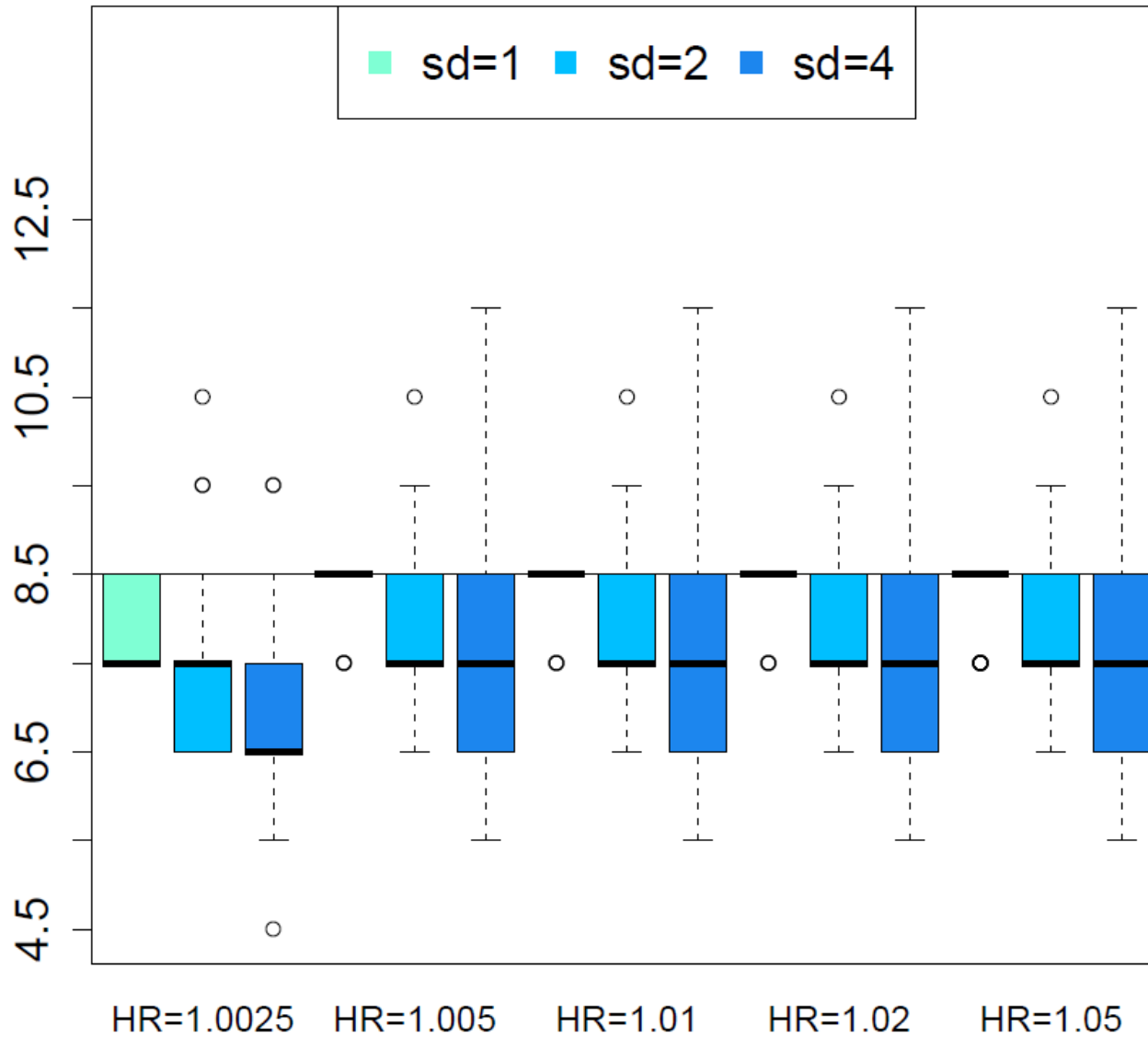


8. Estimated Threshold Locations With Cox Proportional Hazard Models Under Measurement Error , No Random Variation Across Cities

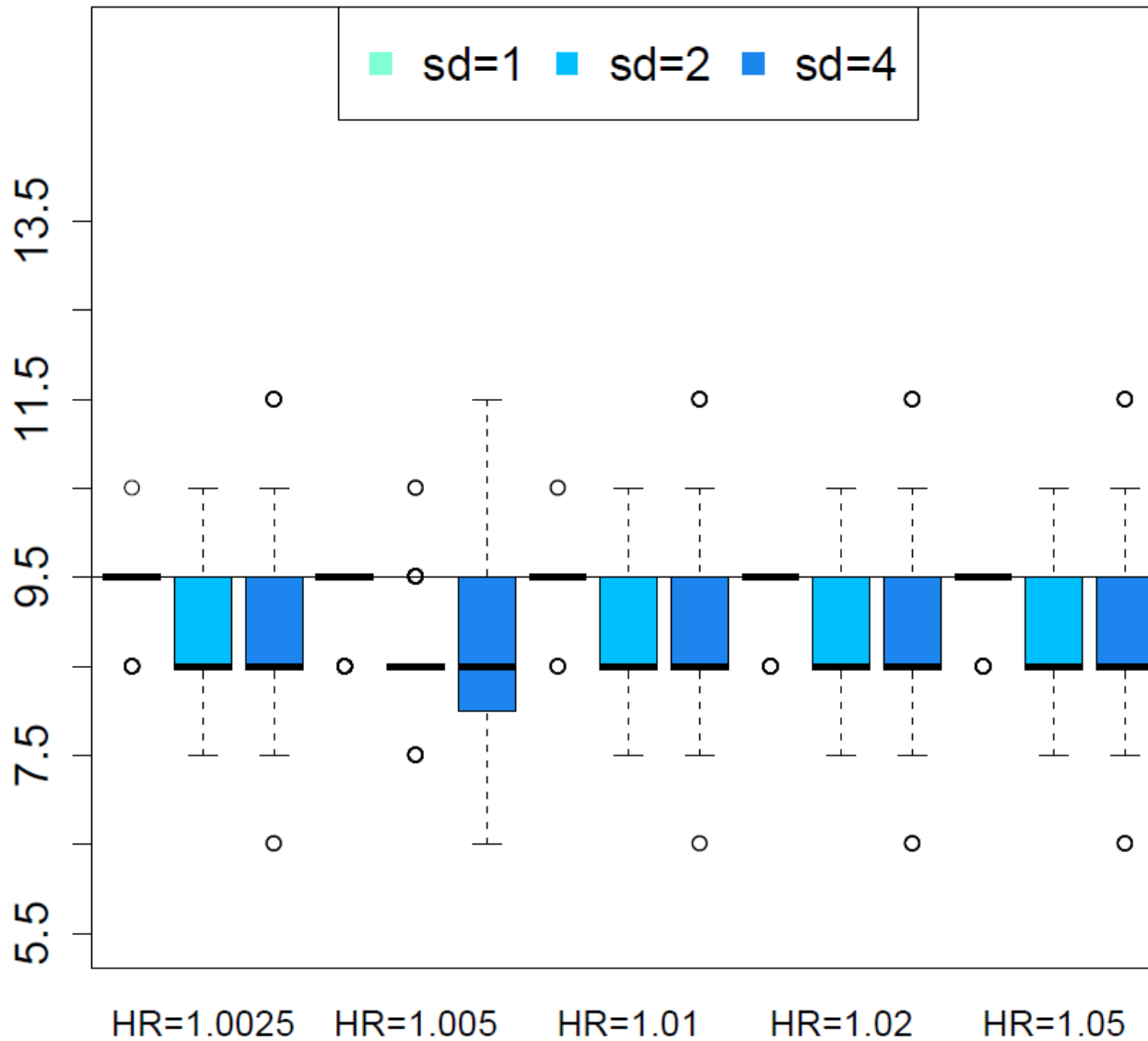
Estimated Thresholds When True Threshold = 7



Estimated Thresholds When True Threshold = 8.5



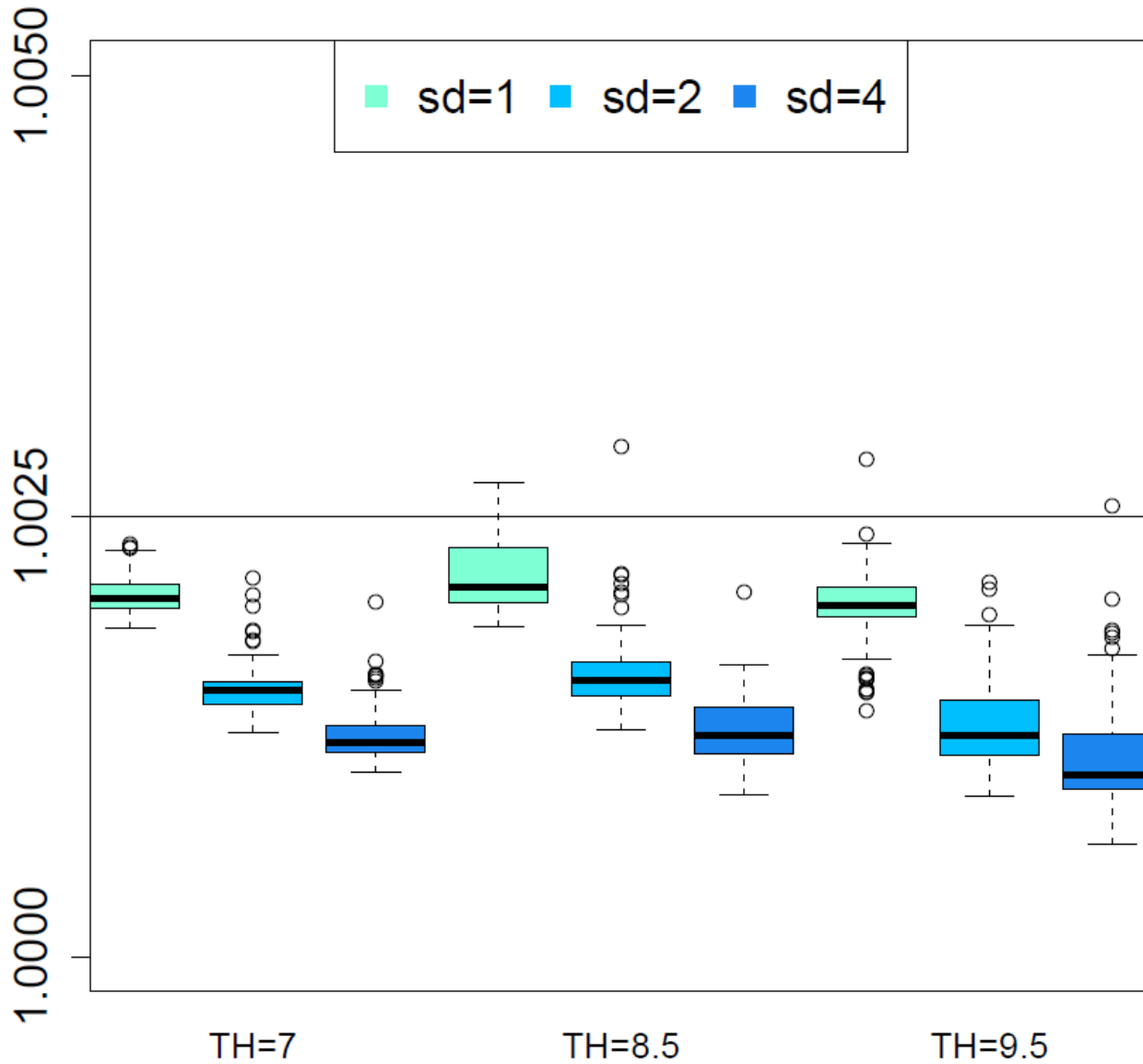
Estimated Thresholds When True Threshold = 9.5



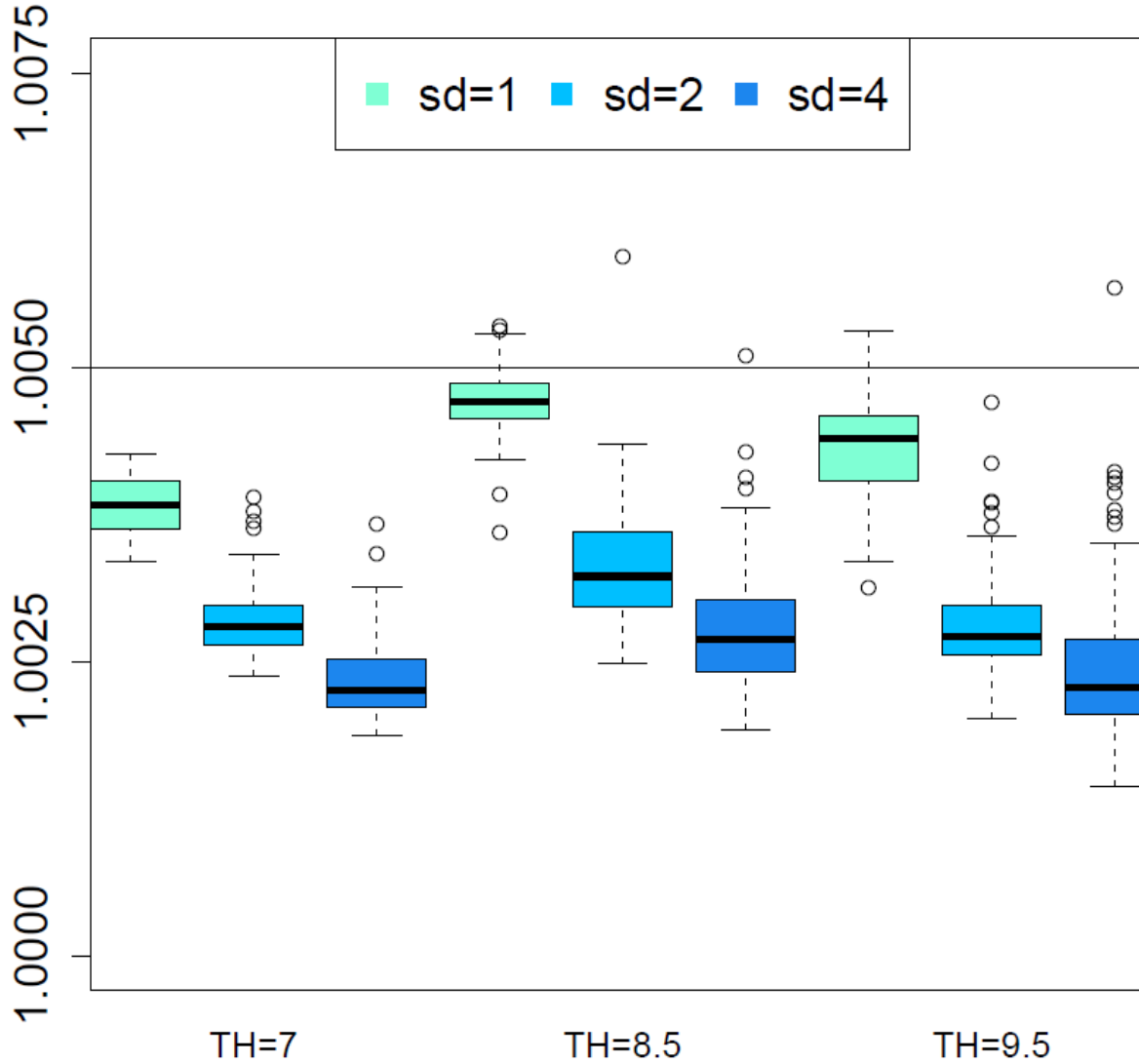


9. Estimated Hazard Ratios With Cox Proportional Hazard Models Under Measurement Error , No Random Variation Across Cities

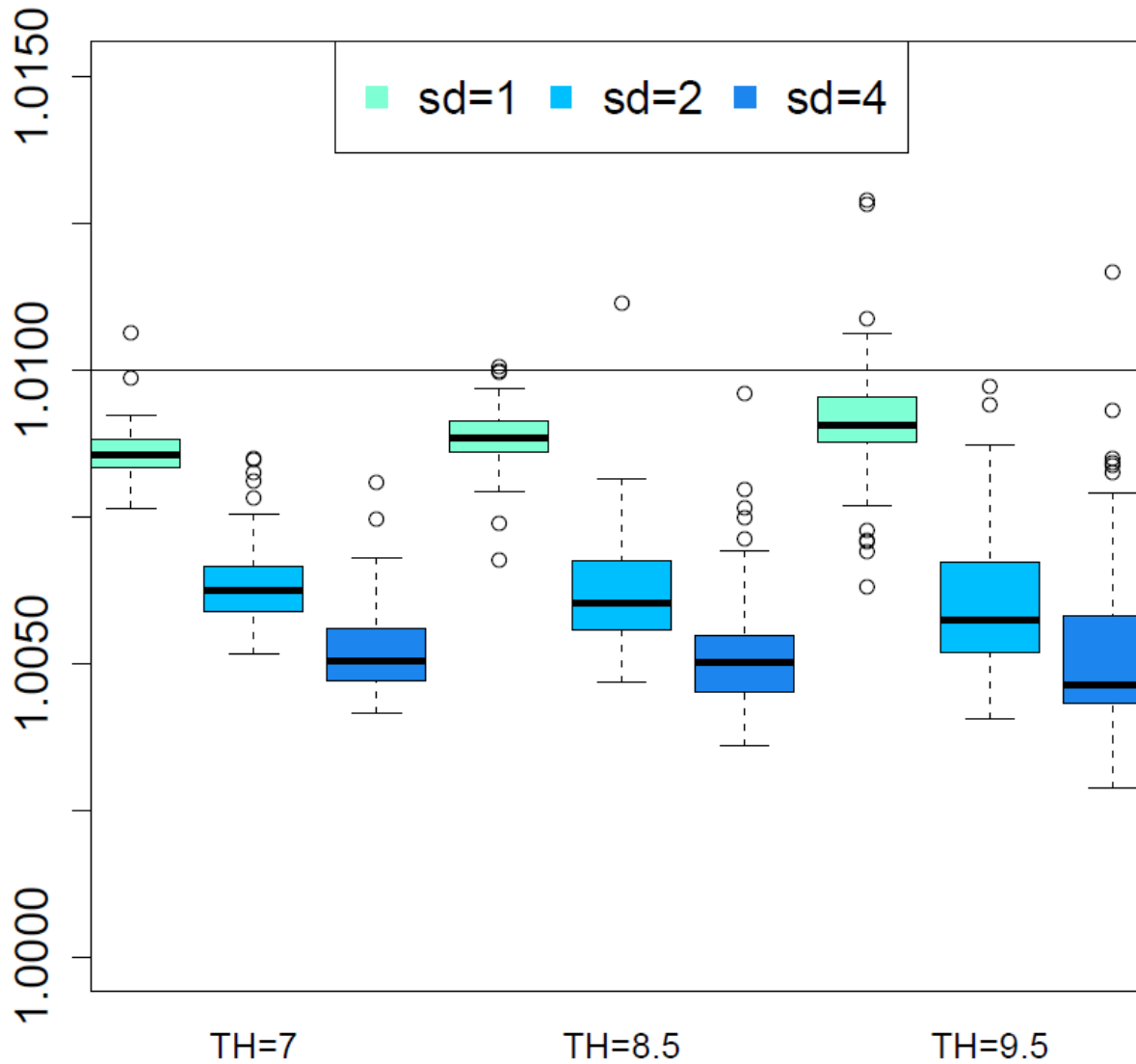
Estimated Hazard Ratios When the True Hazard Ratio is 1.0025



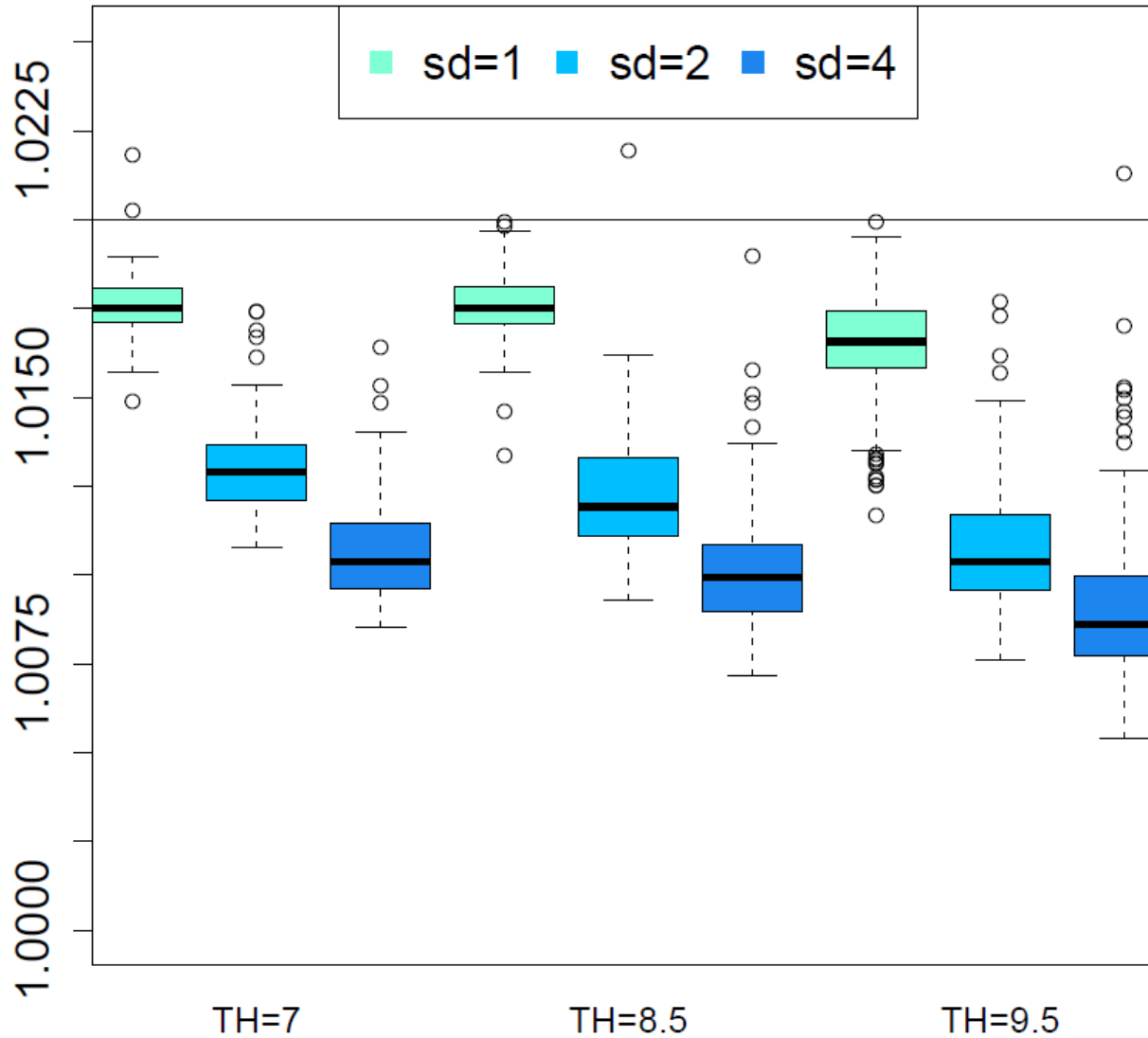
Estimated Hazard Ratios When the True Hazard Ratio is 1.005



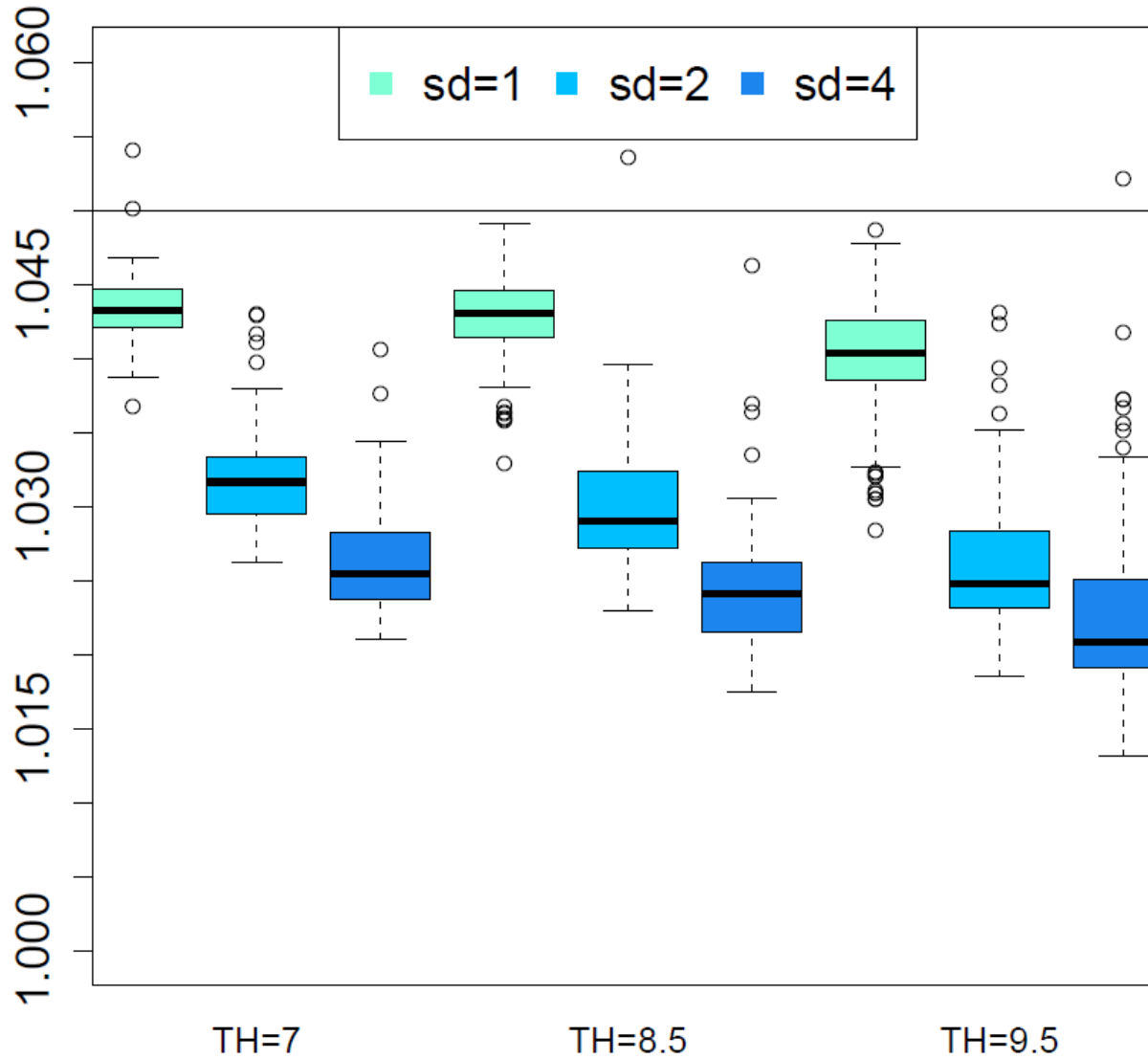
Estimated Hazard Ratios When the True Hazard Ratio is 1.01



Estimated Hazard Ratios When the True Hazard Ratio is 1.02



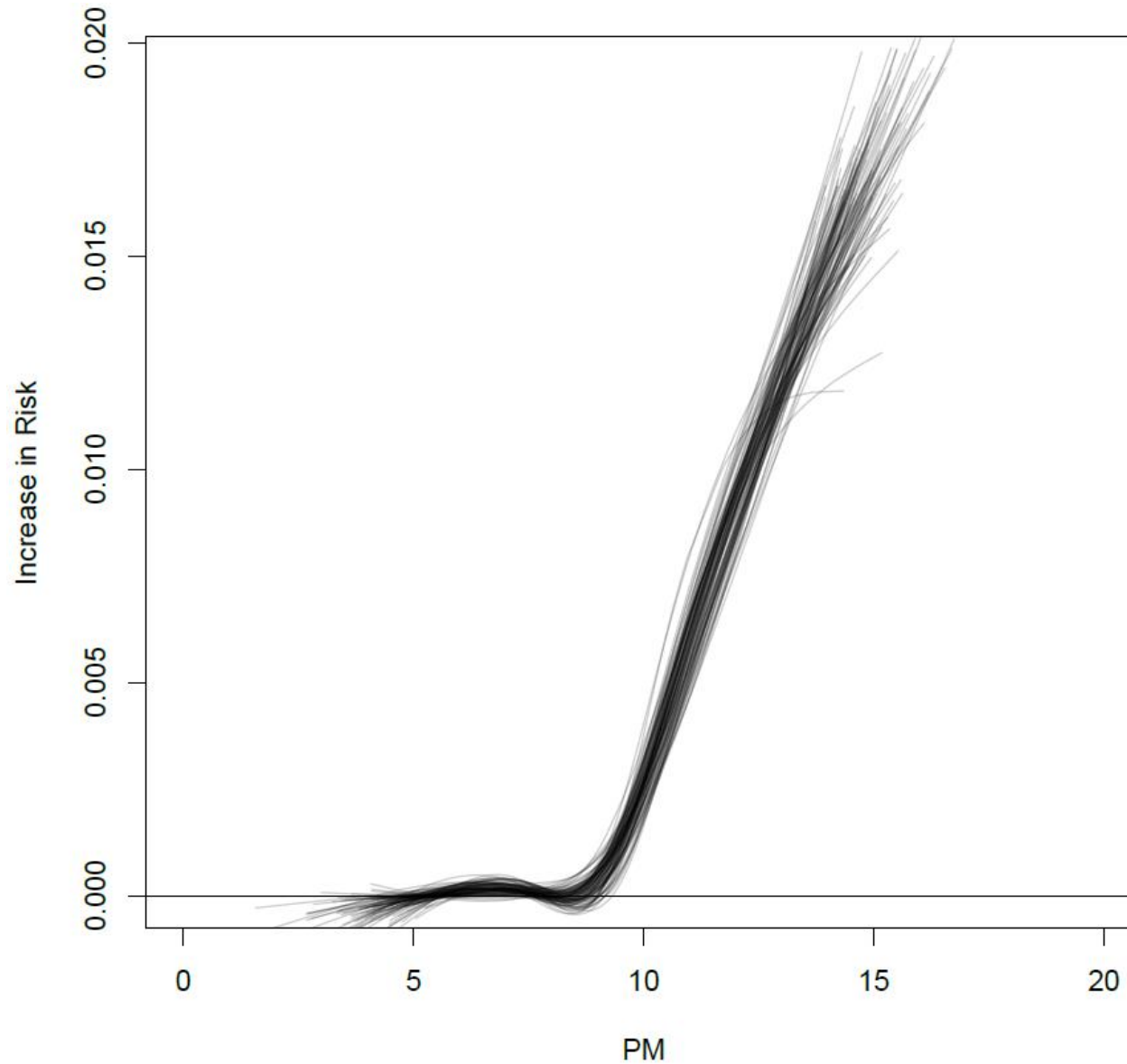
Estimated Hazard Ratios When the True Hazard Ratio is 1.05



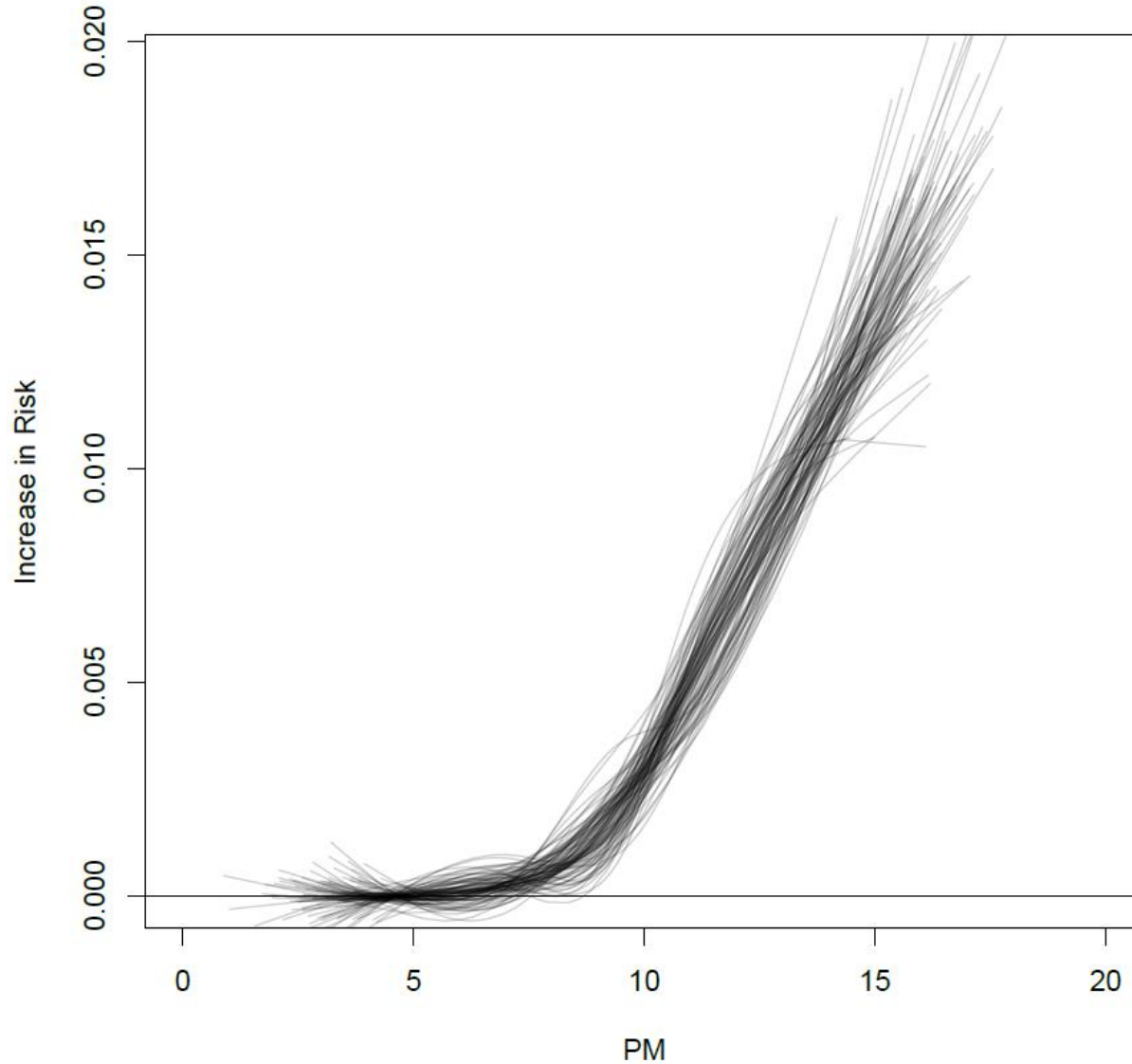


10. Using Nonparametric Regressions to Examine Mortality Data for Thresholds, No Random Variation Across Cities

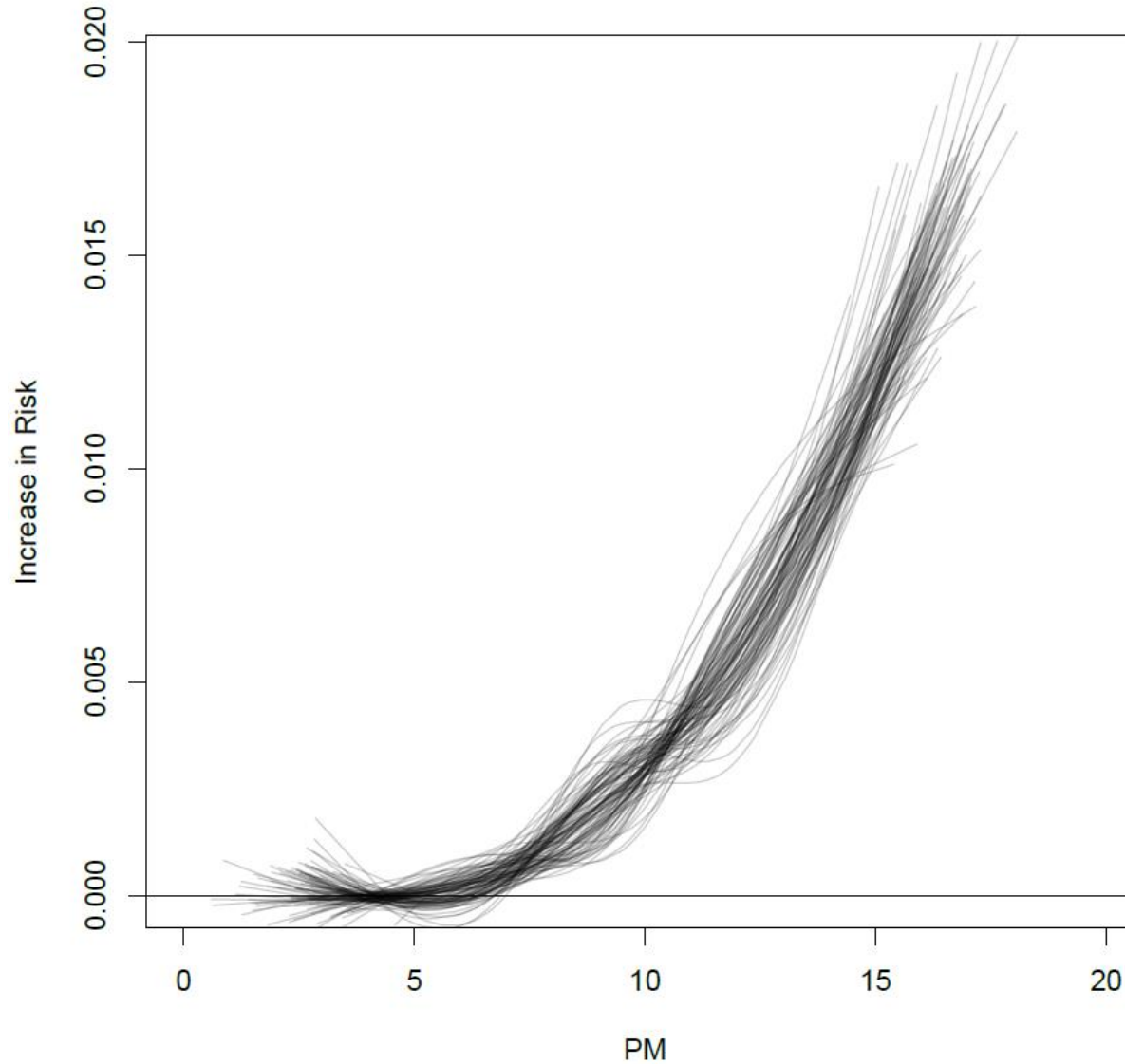
Splines Across 100 Simulations, Threshold = 9.5, HR = 1.005, SD=1



Splines Across 100 Simulations, Threshold = 9.5, HR = 1.005, SD=2



Splines Across 100 Simulations, Threshold = 9.5, HR = 1.005, SD=4

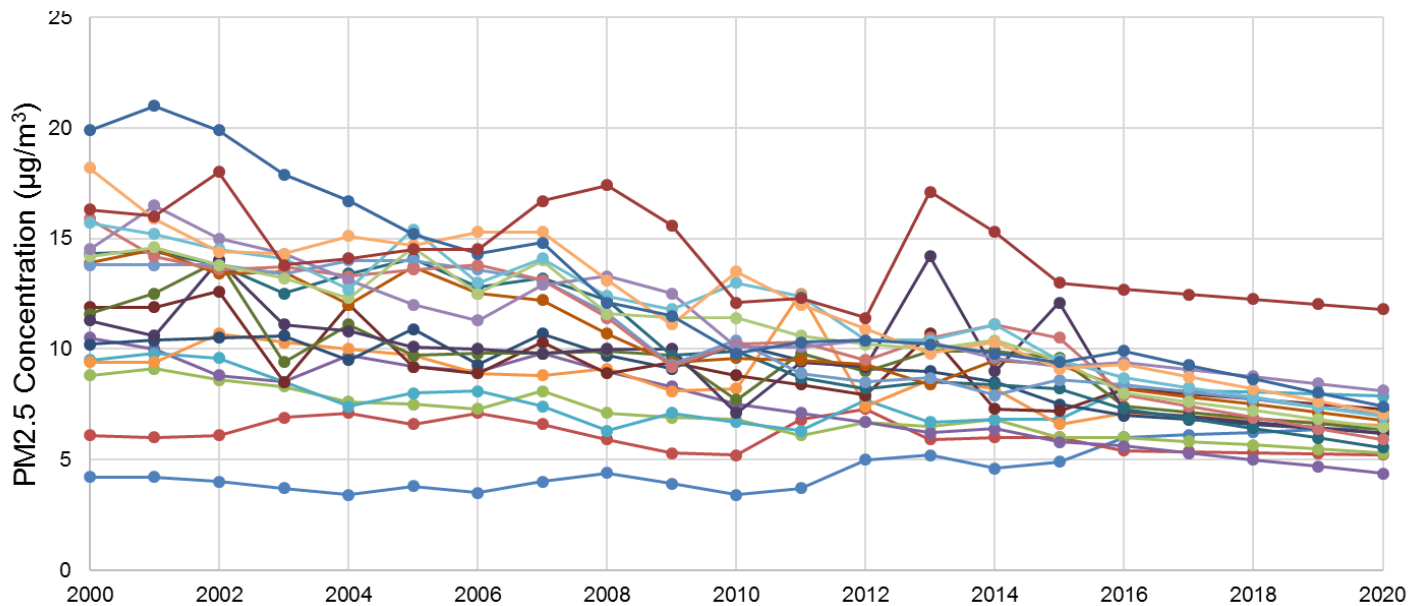




11. Estimated Hazard Ratio Using a “Snapshot” of PM From a Single Year When True PM Trends Downward Over Time

Variation in PM Over Time in 20 US Cities

- “True” PM in each city was assumed to trend downward over 20 years.
- Given mortality in year 20, the estimated relationship between mortality and PM will depend on which year of PM data is used in the Cox proportional hazards model.



Effect of Using “Snapshot” of PM Concentration in Given Year

- We use a simulated cohort of 60 year old men for 20 cities, each with 100,000 identical people (2 million total observations).
- “True” PM in each city for each year was based on the previous slide
 - Cohort mortality outcomes were simulated assuming a linear/no threshold true C-R relationship
 - The true HR was assumed to be 1.005
- We estimated two Cox proportional hazards models, one using the PM measures from year 1, and one using the PM measures from year 20
 - Using year 1: HR = 1.0023
 - Using year 20: HR = 1.0047



12. References

References for Other Simulation Studies of Long-Term Air Pollution Risk Estimation Methods

- Abrahamowicz et al. (2004). Bias due to Aggregation of Individual Covariates in the Cox Regression Model. *American Journal of Epidemiology*, 160 (7), 696-706.
- Gassama et al. (2017). Comparison of methods for estimating the attributable risk in the context of survival analysis. *BMC Medical Research Methodology*, 17, 1-11.
- Gryparis et al. (2009). Measurement error caused by spatial misalignment in environmental epidemiology, *Biostatistics*, 10 (2), 258-274.
- Kim, S.Y., Sheppard, L., and Kim, H. (2009). Health Effects of Long-term Air Pollution: Influence of Exposure Prediction Methods, *Epidemiology*, 20 (3), 442-450.
- Lee, A., Szpiro, A., Kim, S.Y., Sheppard, L. (2015). Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics*, 26 (4), 255-267.
- Moolgavkar et al. (2018). An Assessment of the Cox Proportional Hazards Regression Model for Epidemiologic Studies. *Risk Analysis*, 38 (4), 777-794.
- Shinozaki, T., Mansournia, M., Matsuyama, Y. (2017). On Hazard Ratio Estimators by Proportional Hazards Models in Matched-pair Cohort Studies. *Emerging Themes in Epidemiology*, 14 (6), 1-14.
- Szpiro et al. (2011). Efficient measurement error correction with spatially misaligned data. *Biostatistics*, 12 (4), 610-623.
- Wang, W. & Albert, J. (2017). Causal Mediation Analysis for the Cox Proportional Hazards Model with a Smooth Baseline Hazard Estimator. *Center of Biostatistics and Bioinformatics*, 66 (4), 741-757.
- White, A. Yu, J., Jerrett M., Coogan P. (2016) Temporal aspects of air pollutant measures in epidemiologic analysis: a simulation study. *Scientific Reports*, 6, 19691; doi: 10.1038/srep19691.
- Xue et al. (2013). Testing the proportional hazards assumption in case-cohort analysis. *Medical Research Methodology*, 13 (88), 1-10.