Data-Driven Characterization of Genetic Variability in Disease Pathways and Pesticide-Induced Nervous System Disease in the United States Population

Marissa B. Kosnik,1,2 Philipp Antczak,3,4,5 and Peter Fantke1

1Quantitative Sustainability Assessment, Department of Environmental and Resource Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark
2Department of Environmental Toxicology, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
3Faculty of Medicine and Cologne University Hospital, Center for Molecular Medicine Cologne, University of Cologne, Cologne, Germany
4Cluster of Excellence on Cellular Stress Responses in Aging-associated Diseases (CECAD), University of Cologne, Cologne, Germany
5Department II of Internal Medicine, University of Cologne, Cologne, Germany

BACKGROUND: Genetic susceptibility to chemicals is incompletely characterized. However, nervous system disease development following pesticide exposure can vary in a population, implying some individuals may have higher genetic susceptibility to pesticide-induced nervous system disease.

OBJECTIVES: We aimed to build a computational approach to characterize single-nucleotide polymorphisms (SNPs) implicated in chemically induced adverse outcomes and used this framework to assess the link between differential population susceptibility to pesticides and human nervous system disease.

METHODS: We integrated publicly available datasets of Chemical–Gene, Gene–Pathway, and SNP–Disease associations to build Chemical–Pathway–Gene–SNP–Disease linkages for humans. As a case study, we integrated these linkages with spatialized pesticide application data for the US from 1992 to 2018 and spatialized nervous system disease rates for 2018. Through this, we characterized SNPs that may be important in states with high disease occurrence based on the pesticides used there.

RESULTS: We found that the number of SNP hits per pesticide in US states positively correlated with disease incidence and prevalence for Alzheimer’s disease, Parkinson disease, and multiple sclerosis. We performed frequent itemset mining to differentiate pesticides used over time in states with high and low disease occurrence and found that only 19% of pesticide sets overlapped between 10 states with high disease occurrence and 10 states with low disease occurrence rates, and more SNPs were implicated in pathways in high disease occurrence states. Through a cross-validation of subsets of five high and low disease occurrence states, we characterized SNPs, genes, pathways, and pesticides more frequently implicated in high disease occurrence states.

DISCUSSION: Our findings support that pesticides contribute to nervous system disease, and we developed priority lists of SNPs, pesticides, and pathways for further study. This data-driven approach can be adapted to other chemicals, diseases, and locations to characterize differential population susceptibility to chemical exposures.

Introduction

Interactions between environmental factors and genetics underlie the majority of chronic human diseases,1 and genome-wide association studies (GWAS) have identified thousands of loci that may describe interindividual variability in disease occurrence. While some studies have described modifying effects of environmental factors like diet, alcohol, and smoking on disease outcomes, chemical exposures are rarely assessed in GWAS. This is in part due to the challenge of characterizing the spectrum of exposures an individual encounters (i.e., the exposome) and linking individual chemical exposures to particular genetic variants and disease outcomes.2 Because linking single-nucleotide polymorphisms (SNPs) to specific environmental exposures often requires highly exposed populations, pharmaceuticals are some of the only compounds this has been done for.3 Therefore, identifying SNPs that may play a role in differential disease development following exposure to a wider range of chemicals is challenging.

Of the many chemicals humans may be exposed to, pesticides have a unique set of characteristics. Pesticides are ubiquitously used, toxic by design, and intentionally applied to agricultural crops and released into the environment.4 Occupational exposure to pesticides among agricultural workers is well-recognized, and high levels of pesticide residues have been found in drinking water and on food products available for consumption.5 Further, different classes of pesticides have been implicated in nervous system diseases like Alzheimer’s disease, multiple sclerosis, and Parkinson disease through epidemiological studies of populations with chronic pesticide exposure.6–7 However, development of nervous system diseases following pesticide exposure is variable, suggesting some individuals may have more genetic susceptibility to pesticide-induced nervous system disease.8–9 Given that pesticide use is expected to increase with population growth and climate change,10 it is important to determine the potential role of pesticides in differential development of nervous system disease to protect genetically susceptible populations.

The need to consider differential population susceptibility to chemical exposures is recognized in chemical risk assessment.11,12 Conventional methods relying on arbitrary safety factors (e.g., dividing a toxic dose by 10) cannot capture the complex role of human genetics in response to chemical exposures,13 and new approach methodologies (NAMs) have been proposed as a potential way to address this gap.14 We previously developed a dataset of Chemical–SNP–Disease linkages by integrating disparate datasets together.15 Using these novel linkages, we characterized differences in SNPs implicated by consumer products and distinguished potential mechanisms of toxicity for different disease classes.16 While this dataset helps characterize potential gene–environment interactions, the linkages were limited to SNPs located in genes and did not consider toxicity pathways at the Chemical–Disease intersections in forming a linkage. Here, we developed a new framework to form Chemical–SNP–Disease linkages using a toxicity pathway-based approach to identify
SNPs implicated in differential population susceptibility to chemicals. Then, as an illustrative example of the value of this framework, we spatialized our newly developed Pesticide–SNP–Disease linkages using pesticide application data and nervous system disease data for the US. We focused on the US because it has publicly available spatialized pesticide application data, but this analysis can be repeated for other locations and chemical classes. Through our case study, we characterized the relationship between potential differential population susceptibility to pesticide exposure based on where pesticides are used and nervous system disease in the US. We then characterized the pesticides used and SNPs and genes implicated in different regions of the US and built priority lists of SNPs, genes, disease pathways, and pesticides for further study of differential population susceptibility to pesticide-induced nervous system disease. This data-driven approach to form Chemical–SNP–Disease linkages—and the resulting Pesticide–SNP–Nervous system disease dataset—can serve as a starting point to characterize the link between chemical exposures and varied health outcomes in a population to incorporate interindividual variability into chemical risk and impact assessment. Further, we demonstrate the applicability of this approach to explore the relationship between chemical exposure and disease in different geographic regions so that genetically susceptible populations can be better protected.

All associations generated here are based on computational linkages between distinct resources. The results provided here are meant to be seen as new data explorations and hypotheses that need to be tested and further validated as opposed to final associations or direct causes of observed disease rates in the US.

Methods

Chemical–Pathway–Gene–SNP–Disease Dataset Formation

A general overview of the data integration process to form Chemical–Pathway–Gene–SNP–Disease linkages is shown in Figure 1. Briefly, Chemical–Gene data were integrated from high-throughput screening data and a literature-curated database. Pathway–Gene–Disease linkages were formed using gene enrichment. Chemical–Pathway–Gene–Disease linkages were formed based on gene overlap in the Chemical–Gene and Pathway–Gene–Disease intersection using the Fisher’s exact test. SNPs were identified in Gene–Disease linkages. Additional SNPs located within 10,000 bases of a SNP identified in a Gene–Disease linkage were also included. Chemicals were reduced to pesticides applied in the US, and diseases were reduced to six nervous system disease with incidence and prevalence data for the US. Linkages were spatialized based on where pesticides were applied in the US.

Chemical–gene data integration. High-throughput screening (HTS) chemical–gene data from Tox21/ToxCast and literature-curated data from the Comparative Toxicogenomics Database (CTD) were merged to increase the chemical–gene associations used in the analysis, following the general protocol outlined in Kosnik et al. Briefly, HTS data from Tox21/ToxCast version invitrodb v4.1 summary files were downloaded from the United States Environmental Protection Agency (https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data). Chemical-assay hits were identified from the “mc3-6_winning_model_hits-flags” files by reducing chemical-assay associations to those with column “hitc” ≥ 0.9, meaning the chemical was active in that assay. To avoid associations with potential cytotoxicity, only chemical-assay hits with a half-maximal activity concentration (AC50) lower than the AC50 for that chemical in cytotoxicity assays were kept (i.e., the AC50 for a given chemical–assay test was below the AC50 for that same chemical in the “cytotox” file). Additionally, because cytotoxicity assays were not available for all chemicals, a general cytotoxicity value was determined from the median cytotoxicity AC50 across all cytotoxicity and viability assays (log AC50 = 1.47, which is conservatively below the 25th percentile of cytotoxicity values for chemicals with cytotoxicity data). If an assay had multiple gene targets, these were counted as separate chemical–gene associations.

Chemical–Gene interaction data were downloaded from the 29 November 2023 release of CTD (http://ctdbase.org/downloads), and data were reduced to species = Homo sapiens. HTS and CTD Chemical–Gene associations were merged based on overlapping CAS Registry Numbers (CASRN) and Entrez Gene IDs. Instances where there was a chemical or gene without a match in both datasets were also retained (i.e., a Chemical–Gene linkage did not have to be present in both datasets to be included). The integrated HTS/
Global Burden of Disease (GBD) were included for further analysis. The sensitivity analysis are shown in Figure S1. We found that requiring >10 genes per Pathway–Disease overlap was considered too restrictive, as it reduced the majority of pathway associations to Reactome pathways. Similarly, restricting the p-value to 0.001 removed many disease-relevant pathways (e.g., PPAR signaling pathway). Finally, requiring more than five genes per Chemical–Pathway–Disease linkage caused a large drop in the number of chemicals and Chemical–Disease associations. Therefore, the final set of Chemical–Pathway–Gene–Disease linkages (or Chemical–Disease linkages for simplicity) was formed by requiring a Pathway–Disease linkage to be enriched with an adjusted p-value of <0.001 and 10 genes overlapping and Chemical–Pathway–Disease linkages with a p-value of <0.01 and five genes overlapping.

Pathway–gene–disease integration. Pathway–Gene data from Reactome and WikiPathways were downloaded from the DAVID Knowledgebase on 6 December 2023. Pathway–Gene data from the Kyoto Encyclopedia of Genes and Genomes (KEGG) were collected using the get_gene_sets_list function. Disease associations and fewer SNPs were also kept to identify genes that are important in disease outcomes or may have unknown, disease-relevant SNPs in a population.

Chemical-SNP-disease linkages. SNPs were considered implicated in a Chemical–Disease linkage if they were implicated in the same disease and in or near a gene implicated by both the chemical and the disease. For each Chemical–Disease linkage, all genes implicated at the intersection were identified (i.e., genes present in pathways). Then, all SNPs located in the same genes and implicated in the same diseases were identified. This yielded Chemical–Pathway–Gene–SNP–Disease linkages (or Chemical–Disease/Chemical–SNP–Disease linkages depending on which data are being considered in the linkage intersection).

To identify intergenic SNPs, SNPs located within 10,000 bases of another SNP implicated in the same disease (the general maximum value accepted as a SNP association that is not far from a gene but likely not to be in protein coding genes) and implicated in a Chemical–Pathway–Gene–Disease linkage were identified. These SNPs were considered potentially implicated in the same Chemical–Disease association as the original (located in a gene) SNP. Variant consequences were available in DisGeNET, and consequence severity was determined from Ensembl release 111 (https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html). Chemical–Gene–Disease linkages without SNPs were also kept to identify genes that are important in disease outcomes or may have unknown, disease-relevant SNPs in a population.

Spatial Assessment of Pesticide–SNP–Disease Linkages

Pesticide–Pathway–Gene–SNP–Disease linkages were mapped to different US states based on where the pesticide was reported to be used in the USGS dataset. Disease data for the US was collected from GBD for 1992–2018 (https://www.healthdata.org/gbd/2019) the same years as pesticide application data from USGS. Disease incidence and prevalence rates were collected for “Alzheimer’s disease and other dementias,” “Parkinson’s disease,” “Idiopathic epilepsy,” “Multiple sclerosis,” “Brain and central nervous system cancer,” and “Migraine” and matched to the corresponding diseases in the Pesticide–SNP–Disease dataset: “Alzheimer’s Disease,” “Parkinson Disease,” “Epilepsy,” “Multiple Sclerosis,” “Brain Neoplasms,” and “Migraine Disorders.” Because disease data from GBD is only available per state, pesticide application data were aggregated across counties with the amount of pesticide applied per year per state as the metric for pesticide use. From the data on pesticide application between 1992 and 2018 curated from USGS as described above, if the kilograms of pesticide applied was reported with both a high- and low-value estimate rather than a single value (about 20% of county-level data entries), the median

---

Global Burden of Disease (GBD) were included for further analysis. The sensitivity analysis are shown in Figure S1. We found that requiring >10 genes per Pathway–Disease overlap was considered too restrictive, as it reduced the majority of pathway associations to Reactome pathways. Similarly, restricting the p-value to 0.001 removed many disease-relevant pathways (e.g., PPAR signaling pathway). Finally, requiring more than five genes per Chemical–Pathway–Disease linkage caused a large drop in the number of chemicals and Chemical–Disease associations. Therefore, the final set of Chemical–Pathway–Gene–Disease linkages (or Chemical–Disease linkages for simplicity) was formed by requiring a Pathway–Disease linkage to be enriched with an adjusted p-value of <0.001 and 10 genes overlapping and Chemical–Pathway–Disease linkages with a p-value of <0.01 and five genes overlapping.

Pathway–gene–disease integration. Pathway–Gene data from Reactome and WikiPathways were downloaded from the DAVID Knowledgebase on 6 December 2023. Pathway–Gene data from the Kyoto Encyclopedia of Genes and Genomes (KEGG) were collected using the get_gene_sets_list function. Disease associations and fewer SNPs were also kept to identify genes that are important in disease outcomes or may have unknown, disease-relevant SNPs in a population.

Chemical-SNP-disease linkages. SNPs were considered implicated in a Chemical–Disease linkage if they were implicated in the same disease and in or near a gene implicated by both the chemical and the disease. SNP–Disease data were collected using the disgenet2r package version 0.99.3. For each Chemical–Disease linkage, all genes implicated at the intersection were identified (i.e., genes present in pathways). Then, all SNPs located in the same genes and implicated in the same diseases were identified. This yielded Chemical–Pathway–Gene–SNP–Disease linkages (or Chemical–Disease/Chemical–SNP–Disease linkages depending on which data are being considered in the linkage intersection).

To identify intergenic SNPs, SNPs located within 10,000 bases of another SNP implicated in the same disease (the general maximum value accepted as a SNP association that is not far from a gene but likely not to be in protein coding genes) and implicated in a Chemical–Pathway–Gene–Disease linkage were identified. These SNPs were considered potentially implicated in the same Chemical–Disease association as the original (located in a gene) SNP. Variant consequences were available in DisGeNET, and consequence severity was determined from Ensembl release 111 (https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html). Chemical–Gene–Disease linkages without SNPs were also kept to identify genes that are important in disease outcomes or may have unknown, disease-relevant SNPs in a population.

Spatial Assessment of Pesticide–SNP–Disease Linkages

Pesticide–Pathway–Gene–SNP–Disease linkages were mapped to different US states based on where the pesticide was reported to be used in the USGS dataset. Disease data for the US was collected from GBD for 1992–2018 (https://www.healthdata.org/gbd/2019) the same years as pesticide application data from USGS. Disease incidence and prevalence rates were collected for “Alzheimer’s disease and other dementias,” “Parkinson’s disease,” “Idiopathic epilepsy,” “Multiple sclerosis,” “Brain and central nervous system cancer,” and “Migraine” and matched to the corresponding diseases in the Pesticide–SNP–Disease dataset: “Alzheimer’s Disease,” “Parkinson Disease,” “Epilepsy,” “Multiple Sclerosis,” “Brain Neoplasms,” and “Migraine Disorders.” Because disease data from GBD is only available per state, pesticide application data were aggregated across counties with the amount of pesticide applied per year per state as the metric for pesticide use. From the data on pesticide application between 1992 and 2018 curated from USGS as described above, if the kilograms of pesticide applied was reported with both a high- and low-value estimate rather than a single value (about 20% of county-level data entries), the median
was taken. This yielded a final dataset with the amount of pesticide applied per state per year for 1992–2018. Many pesticides were applied in all states in all years including 2,4-dichlorophenoxyacetic acid (2,4-D), dicamba, paraquat, chlorpyrifos, and glyphosate, of which the latter was also used at a greater mass than other pesticides (4.5 times greater than the next most-applied pesticide, 2,4-D). All six nervous system diseases occurred in all states.

Based on where pesticides were applied, SNP–Disease associations were mapped to that state (e.g., Copper–rs669–Alzheimer’s disease was mapped to all states where copper was applied). This spatialization was not determined based on where SNPs were known to occur in the population in a given state due to a lack of such data. What this spatialization intends to describe is which SNPs may be most relevant in a geographic region based on the pesticides applied there, and the SNPs we identified as hits by these same pesticides. The number of unique pesticides and kilograms/square mile of pesticide applied are shown in Figure S3, along with the total number of SNP and gene hits.

**Pesticide–SNP–Disease linkages and disease occurrence by state.** As multiple pesticides can implicate the same SNP/gene and the same pesticide can be applied in multiple years, the same SNP/gene can have multiple “hits” in a single year and across years. Therefore, to consider the cumulative potential for a SNP/gene to be affected by pesticide application, we assessed SNP/gene frequency over the years based on pesticide-induced hits. An example figure of SNP hit calculation is shown in Figure S4. To assess the relationship between SNP/gene hits and disease occurrence, the SNP/gene hits and disease incidence/prevalence rates were compared using a linear model. We compared the SNP/gene hits cumulatively and the disease rates for each year in 1992–2018. The Pearson’s correlation coefficient and robust linear model root mean square error (RMSE) were used to assess the correlations between SNP/gene hits and disease outcomes over time. For each year of disease data between 1992 and 2018, we compared the cumulative SNP hits from all pesticides applied prior to and including that year to the disease incidence and prevalence rate (e.g., for Alzheimer’s disease incidence/prevalence in 2008, the SNPs and genes implicated based on pesticide use from 1992 to 2008 were compared). We did this for overall disease rates, age-standardized disease rates, and disease rates in age 55+ group for each year (1992–2018). Ultimately, to include as many previous years of potential pesticide exposure as possible (i.e., SNP/gene hits from 1992 to 2018), we used 2018 as the year of analysis and general population disease rates for the remainder of the methods. We assumed long-term pesticide exposure and did not consider acute exposure for this analysis.

**State-to-state migration flows.** To test if the population in US states changed over time, we used census data on state-to-state migration flows for the US.29 These data provided information on how many people in a state lived in the same state in the previous year and were available for 2010–2018.

**Bootstrapping of pesticide–SNP linkages per state.** To assess if associations between SNP hits per pesticide per year and disease occurrence were random, a bootstrapping approach was used. Pesticides applied in each state in each year in 1992–2018 were randomly shuffled (thus changing the SNPs implicated in each state), keeping the same number of pesticides applied overall in each state. For 1,000 randomly generated datasets, we compared the disease incidence/prevalence for 2018 with the number of SNP hits per pesticide per year. From this correlation, we checked the Pearson’s correlation coefficient and developed a robust linear model, calculating the mean absolute error (MAE) and RMSE for each random correlation. The distribution of the random Pearson’s correlation coefficients, MAEs, and RMSEs were compared to the same values for the true dataset. This approach was also used to assess if the relationship between kilograms of pesticide applied and disease occurrence was random (where the pesticides used/kilogram applied were shuffled).

**High/low-probability state selection.** To identify the high-probability states (those with higher nervous system disease occurrence and more SNP hits per pesticide per year) and low-probability states (those with lower nervous system disease occurrence and fewer SNP hits per pesticide per year) for each disease, the values for the SNP hits per pesticide per year for each disease were scaled between 0 and 1. The disease incidence and prevalence were also scaled between 0 and 1, and the scaled incidence and prevalence were summed for each state to represent the disease occurrence. The disease occurrence value was then multiplied by the scaled SNP hits value to identify the high- and low-probability states for each disease. Per disease, the 10 highest and lowest probability states were identified.

**Statistical Analysis**

**LASSO.** To assess the relationship between the different Pesticide–SNP hits and disease occurrence in each state, multilinear regression models were built with the number of SNP hits implicated by each pesticide in each state as the predictor variable (the number of predictors was the number of pesticides). For all diseases, too many pesticides were implicated to build a normal multilinear regression model owing to the limits on the degrees of freedom (48 states included in the model vs. 265–280 predictors per disease). Least absolute shrinkage and selection operator (LASSO) regression was conducted using the c060 R package, version 0.2-930 with the stabpath function. The size of the training set was 0.66, and 10,000 steps were used. The stabset function was used to select the stable variable set based on a per-comparison error rate of <0.05. A separate model was built with the dependent variable as incidence or prevalence for Alzheimer’s disease, multiple sclerosis, and Parkinson disease.

**Frequent itemset mining.** Pesticide use patterns over time were assessed using frequent itemset mining. Association rules for pesticides were developed per state based on pesticides used in that state. For each state, each year of USGS data from 1992 to 2018 was used as an independent transaction with all unique pesticides used in that year as the input items (a visual showing this approach is in Figure S5). Rules between two and three pesticides were identified with support set to 0.18 (meaning the combination had to appear in at least 5 years) and confidence set to 0.8. Analysis was done using the arules R package version 1.7-531 with the a priori function. Because we were more interested in the sets of pesticides identified than the directionality of the rule, only the resulting combinations of pesticides were kept, not the rules themselves (e.g., rules A,B => C and B,C => A were considered the same rule: A,B,C). The output of the frequent itemset mining analysis was a set of association rules per state describing recurring pesticide use patterns in 1992–2018. Association rules were mapped to diseases based on pesticides implicated in that disease (e.g., copper and fipronil were linked to Alzheimer’s disease in our dataset, so the association rule copper–fipronil for Connecticut was linked to Alzheimer’s disease). If an association rule for a state did not have any pesticides linked to that disease, then the association rule was excluded. If an association rule had at least one pesticide implicated in a disease, then the association rule was kept because it is possible the currently unlinked pesticide could still be contributing to the disease, even if not directly linked in our dataset.

**Cross-validation.** Ranks between different lists were analyzed using the Mann-Whitney test (p < 0.05 = ranks significantly differ between lists, p ≥ 0.05 = ranks are not significantly different between lists). Overlap between different lists was determined using the hypergeometric test (p < 0.05 = there is significant
overlap between lists, $p \geq 0.05$ = items in lists do not significantly overlap).

**Priority List Development**

To develop priority lists of SNPs/genes, pathways, and pesticides, five states were randomly selected from the 10 high-probability states and 10 low-probability states. Then, the SNPs, genes, pathways, and pesticides implicated in these five high-probability states compared to those in the five low-probability states were compared to develop priority lists as described in the following sections. This process was repeated 10 times to identify SNPs/genes, pathways, and pesticides consistently implicated in different subsets of high-priority states compared to low-priority states. Then, the overall priority list was developed based on the sum of ranks across each of the 10 repetitions and the number of repetitions that the SNP/gene, pathway, or pesticide was prioritized in. The differences in the SNP ranks between repetitions may come from variability in the bottom 10 states, as we conducted an additional set of 10 repetitions where the five low-probability states were kept constant between repetitions while the five high-probability states were randomly selected from the 10 high-probability states. This resulted in more consistency in the priority lists between the 10 repetitions. For the full analysis, we chose to compare 10 high-probability states to 10 low-probability states for consistency.

**SNP and gene list development.** For each disease, a priority list of SNPs and genes was developed by determining the SNP frequency and rule value. The SNP frequency was the number of SNP/gene hits in 1992–2018 across the five randomly selected high-probability states vs. the five randomly selected low-probability states. If the hit frequency for a SNP/gene in high-probability states was higher than the hit frequency in low-probability states, that SNP/gene was retained. For each disease, the priority SNPs and genes were ranked based on the number of states, implicating that SNP/gene and the number of pesticides present in association rules in high-probability states that implicate that SNP/gene (representing different patterns in SNP/gene hits based on pesticide use). Rules were matched to SNPs/genes if at least one pesticide in that rule was linked to the SNP/gene in the Pesticide–Gene–SNP–Disease linkage. For each SNP/gene–rule combination, the rule value was assigned as follows:

1. If all pesticides in a rule implicated that SNP/gene in that disease in at least three out of five randomly selected high-probability states, that SNP/gene–rule combination was assigned a rule value of “5,” representing higher overlap.
2. If fewer than all pesticides in a rule implicated that SNP/gene in that disease in at least three out of five randomly selected high-probability states, that SNP/gene–rule combination was assigned a rule value of “1,” representing lower overlap.

Rules present in fewer than three out of five randomly selected states were not included in these calculations, as they were not considered to represent patterns in high- or low-probability states. Then, for each SNP/gene, the rank value was calculated as the sum of rule values multiplied by the hit frequency. Each individual repetition’s SNP/gene priority list was then ordered by this rank value. This process was repeated 10 times with 10 different subsets of five high-probability and five low-probability states (or the five bottom states fixed). Then, the overall SNP/gene priority list was determined by summing the ranks for all SNPs/genes across the 10 repetitions and dividing the summed rank by the number of repetitions the SNP/gene appeared in [e.g., if a SNP was ranked one in five reps and two in five reps, its overall rank would be as follows: $(1 \times 5 + 2 \times 5)/10$ repetitions.] Only SNPs/genes present in high-priority lists for $\geq 6/10$ repetitions were included.

**Pathway list development.** Pathways with more SNP/gene hits in high-probability states than low-probability states were identified for each repetition of five randomly selected high- and low-probability states. In each repetition, a threshold requirement was set that at least three out of five of the randomly selected high-probability states had to have more hits in a pathway compared to low-probability states so as to only include pathways characteristic of several high-probability states. For each individual repetition, pathways were then ranked based on the fold greater hits in high-probability states compared to low-probability states. This process was repeated 10 times with 10 different subsets of five high-probability and five low-probability states (or the five bottom states fixed). Then, the overall pathway priority list was determined by summing the ranks for all pathways across the 10 repetitions and dividing the summed rank by the number of repetitions the pathway appeared in. Only pathways present in high priority lists for $\geq 6/10$ repetitions were included.

**Pesticide list development.** For each repetition of five randomly selected high-probability states and five randomly selected low-probability states, priority pesticides were identified from pesticides acting on the high-priority SNPs and genes identified (as described in “SNP and gene list development”). Of these pesticides, the list was reduced to those pesticides used in at least three out of five randomly selected high-probability states, occurring in association rules, and applied with greater average mass in high-probability states than low-probability states. Then, for each individual repetition, pesticides were ranked ordered with priority given to pesticides only in association rules in high-probability states but not low-probability states, the number of SNPs and genes implicated, and difference in kilograms applied in high-probability vs. low-probability states. This process was repeated 10 times with 10 different subsets of five high-probability and five low-probability states. Then, the overall pesticide priority list was determined by summing the ranks for all pesticides across the 10 repetitions and dividing the summed rank by the number of repetitions the pesticide appeared in. Only pesticides present in high priority lists for $\geq 6/10$ repetitions were included.

Specific pesticide modes of action (MoAs) were determined from insecticide, fungicide, and herbicide resistance action committee classifications [Insecticide Resistance Action Committee (IRAC),32 Fungicide Resistance Action Committee (FRAC),33 and Herbicide Resistance Action Committee (HRAC)34; accessed 1 November 2021]. Not all pesticides in our dataset had a corresponding MoA. Other MoAs (e.g., rodenticides) did not have a comparable number of pesticides in our dataset and were therefore excluded.

**Data Analysis and Visualization**

Analyses were conducted using R (version 4.2.2; R Development Core Team). Figures were generated using ggplot2 (version 3.4.1).35 All code used in the analysis and for figure generation is available on GitHub (https://github.com/Uttox-Kosnik/Genetic_variability_disease_pathways_and_pesticide-induced_nervous_system_disease_in_US).

**Results**

We generated Pesticide–Pathway–Gene–SNP–Disease linkages by integrating multiple publicly available sources. For simplicity, these are termed Pesticide–SNP–Disease linkages or Pesticide–Gene–Disease linkages if no SNP was implicated. The final data set contained 285 pesticides implicating 2,372 total SNPs and 3,395 total genes across 1,285 toxicity pathways in six nervous
system diseases of interest: Alzheimer’s disease, Parkinson disease, multiple sclerosis, brain neoplasms, epilepsy, and migraine disorders (Figure 1). These data formed 1,506 Pesticide–Disease associations with at least one SNP implicated and a median of 19 SNPs per Pesticide–Disease association (range, 1–489) for a total of 49,542 unique Pesticide–SNP–Disease linkages. An additional 42,825 Pesticide–Gene–Disease linkages had no SNPs implicated. Pesticides with the most SNPs implicated include copper-based compounds (copper had 1,021 SNPs implicated), biostenticides like hydrogen peroxide and rotenone (1,069 and 1,023 SNPs, respectively), and synthetic pesticides like atrazine and paraquat (738 and 677 SNPs, respectively). Alzheimer SNPs, respectively), and synthetic pesticides like hydrogen peroxide and rotenone (1,069 and 1,023 based compounds (copper had 1,021 SNPs implicated), biopesticidal agents in the literature. For example, multiple studies identified SNPs in HLA–DRA, ABCB1, APEX, PON1, and PPARGC1A as influencing pesticide-induced Parkinson disease in a population, and our approach identified these same SNPs in the Pesticide–Disease intersection.36–43 The only SNPs we identified in the literature but not in our dataset were SNPs not incorporated in DisGeNET, the SNP database we used to form the linkages (e.g., rs671 in ALDH244; ALDH2 was implicated in our dataset with different SNPs). Additionally, Pesticide–Disease disease associations curated from epidemiology studies in CTD, and all of these CTD Pesticide–Disease linkages were also identified through our data integration process. This increased confidence in the linkages formed between pesticides, SNPs, and diseases in the present analysis. The full set of Pesticide–Pathway–Gene–SNP–Disease linkages is available in Excel Table S1.

Genetic Variants, Pesticides, and Nervous System Disease Rates

To characterize the potential role of genetic susceptibility in pesticide-induced nervous system diseases, we used spatially refined pesticide application data for the contiguous US from 1992 to 2018 to map the Pesticide–SNP–Disease linkages in our dataset to different states based on where each pesticide was applied. This mapping was not based on where SNPs occur in a population, as these data are not available. Instead, this mapping of Pesticide–SNP–Disease linkages shows which SNPs might be relevant in a geographic region based on the pesticides applied there. To assess the cumulative effect of pesticide exposure on SNPs/gene in a disease outcome, each instance of a pesticide potentially acting on a SNP/gene was considered a “hit.” Because multiple pesticides can implicate the same SNP/gene and pesticides can be applied in multiple years, this means the same SNP/gene may be implicated multiple times in the same region (see Figure S4 for a visual representation of this calculation). This enabled us to identify US states where populations with specific SNPs may be more vulnerable to pesticide exposures (based on the pesticides applied in that region, not based on SNPs known to occur in that region). Using these spatialized Pesticide–SNP linkages, we looked at the relationship between pesticide use, implicated SNPs/genes, and disease outcomes in the US. The rate of disease incidence and prevalence by state for the six diseases under study were collected from GBD for the years 1992–2018. To determine the most informative year for analysis, we built a linear model for each year in 1992–2018 to compare the SNP hits per pesticide per year vs. disease incidence and prevalence rate. The Pearson’s correlation coefficient and robust linear model RMSE from each model is in Figure S6. Generally, the relationship between the SNP hits per pesticide and disease occurrence got stronger over time. While the RMSE also increased over time for most diseases, this is likely due to the increased range of disease cases between 1992 and 2018. For example, in 1992 the Parkinson disease incidence rate per 100,000 people was between 11.2 and 21.2 compared to 31.6 and 68.5 for 2018. A similar trend was observed for gene hits per pesticide per year and disease occurrence (Figure S7). We also compared each year of age–standardized disease incidence/prevalence and disease incidence/prevalence in age 55+ to the number of SNP hits per year (Figure S8 and S9). Multiple sclerosis and Parkinson disease occurrence was correlated with SNP hits per pesticide in all years, but Alzheimer’s disease was not significantly correlated. We also analyzed state-to-state migration flows for 2010–2018 (the years data were available for) to test the chronic exposure assumption that individuals in a state do not change over time. We found that no more than 2.4% of the total US population had lived in a different state in the previous year (maximum 6.2% analyzing state-by-state) with most living in the same house as the previous year (84.5%). Therefore, while there is likely some migration into and out of states, we expect it to account for a very small portion of the total population in a region and, thus, have minimal influence on this analysis.

To better consider the potential that Pesticide–SNP interactions increased the likelihood of developing nervous system disease in a general population [e.g., at younger ages as was found with multiple sclerosis SNPs (Table S1)], we did not use age–standardized disease rates in our analysis. Additionally, we relate pesticide use from 1992 to 2018 to disease rates per 100,000 people for the year 2018 for the remainder of the analysis, to include as many years of pesticide application as possible representing chronic pesticide exposure preceding a disease. To assess if pesticide use in a state may influence disease occurrence, we compared the kilograms of pesticide applied per state to disease incidence and prevalence rates for 2018 and found no clear association (Figure S10). While the kilograms of pesticide applied was significantly correlated with multiple sclerosis incidence and prevalence (Pearson’s p < 0.05), this association was found to be random based on a bootstrapping analysis showing the relationship was significant regardless of the pesticides applied in a given state (p < 0.05 in ≥80% of random runs across diseases). This is not surprising since the total kilograms of pesticide applied may not necessarily equate to a potential biological effect. To better account for the action pesticides may have on people who encounter them, we assessed the SNP hits per pesticide per year. Interestingly, by relating the number of SNP hits implicated per pesticide for 1992–2018 to disease occurrence in 2018, we found a significant, positive correlation for Alzheimer’s disease, multiple sclerosis, and Parkinson disease [Pearson’s correlation coefficient, 0.3; p < 0.05 (Figure 2); see Figure S4 for how SNP hits per pesticide per year are calculated and Figure S11 for all disease correlations]. We also see a similar relationship between disease incidence/prevalence and the number of gene hits per pesticide applied per year and the number of unique SNPs (not hits) per pesticide [Pearson’s correlation coefficient, 0.3; p < 0.05 (Figure S12 and S13)]. When evaluating whether just the type of pesticide, number of pesticide applications, or the total number of SNP hits implicated in a state showed a similar correlation with disease
incidence/prevalence, we found that these resulted in negative correlations (Figures S14–S16). This showed that, it is not the raw number of times pesticides are applied or just the number of SNPs hit that correlated with increased disease incidence/prevalence. To ensure that the number of pesticides alone was not driving the relationship between SNP hits per pesticide per year and disease occurrence, we reduced the total pesticide set to only those pesticides present in the five states with the fewest pesticides, which reduced the total number of pesticides in our analysis from 285 to 183. We found that even by restricting all states to the same subset of pesticides, there is still a positive correlation between SNP hits per pesticide per year and disease occurrence for Alzheimer’s disease, multiple sclerosis, and Parkinson disease (significant for multiple sclerosis and Parkinson disease; Pearson’s $p < 0.05$) (Figure S17).

Based on the positive correlations we identified, pesticides applied in states with higher nervous system disease incidence/prevalence implicated more unique SNPs per pesticide on average, and pesticides may act on these SNPs more frequently. To account for SNPs potentially affected by pesticide application over time, we focused on SNP hits per pesticide per year for the remainder of the analysis (Figure 2A).

Interestingly, a bootstrapping approach highlighted that Alzheimer’s disease, multiple sclerosis, and Parkinson disease correlation coefficients were higher than expected by chance. Less than 10% of 1,000 random runs had significant correlations between SNPs per pesticide per year and disease occurrence (Pearson’s $p < 0.05$). Further, when the distribution of the Pearson’s correlation coefficients, MAEs, and RMSEs from the random runs were compared to the same values for the true dataset (Figures S18 and S19), we found that the true Pearson’s correlation coefficient was higher and the MAE and RMSE were lower than at least 90% of the randomly generated datasets for Alzheimer’s disease, multiple sclerosis, and Parkinson disease. For migraine disorder disease prevalence, brain neoplasms incidence, and epilepsy incidence and prevalence, the bootstrapping approach suggested that the relationships between SNP hits per pesticide per year and disease occurrence could not be sufficiently explained by the Pesticide–SNP–Disease linkages on their own. We therefore excluded these diseases for the remainder of the analysis.

Identifying Pesticides That May Contribute to Disease Occurrence

The observation that SNP hits per pesticide per year for Alzheimer’s disease, Parkinson disease, and multiple sclerosis resulted in a positive correlation led us to hypothesize that the relationship between the SNPs and pesticides is more important for these diseases. To identify pesticides that may contribute to disease occurrence, we used a statistical model that included both the number of pesticides and their application rates. The model predicted that the highest probability states for each disease were those with the highest SNP hits per pesticide per year, while the lowest probability states had the lowest SNP hits per pesticide per year. The model equations are as follows:

- Alzheimer’s disease: $y = 2.99x + 32.42$ (incidence), $y = 10.55x - 208.93$ (prevalence)
- Multiple sclerosis: $y = 0.09x - 2.73$ (incidence), $y = 7.52x - 62.79$ (prevalence)
- Parkinson disease: $y = 1.02x + 5.01$ (incidence), $y = 21.74x + 253.35$ (prevalence)

Data underlying Figure 2A appear in Excel Table S9. Note: $r$, Pearson’s correlation coefficient; RMSE, root mean square error; SNP, single-nucleotide polymorphism.

Figure 2. SNP hits per pesticide per year vs. disease incidence and prevalence. (A) Scatterplots of SNP hits per pesticide per year for pesticides applied in 1992–2018 vs. disease incidence and prevalence in 2018. Each point is a US state. Colors/shapes indicate geographic region of each state. Trend lines indicate robust linear model. All correlations are significant, $p < 0.05$. (B) The five highest and lowest probability states for each corresponding disease in panel A. High-probability state indicates high disease occurrence and more SNP hits per pesticide per year (closer to top right corner of panel A), and low probability state indicates low disease occurrence and fewer SNP hits per pesticide per year (closer to bottom left corner of panel A). RLM equations are as follows:

- Alzheimer’s disease = 2.99x + 32.42 (incidence), 10.55x − 208.93 (prevalence), multiple sclerosis = 0.09x − 2.73 (incidence), 7.52x − 62.79 (prevalence), Parkinson disease = 1.02x + 5.01 (incidence), 21.74x + 253.35 (prevalence). Data underlying Figure 2A appear in Excel Table S9. Note: $r$, Pearson’s correlation coefficient; RMSE, root mean square error; SNP, single-nucleotide polymorphism.
in disease occurrence than just the types of pesticides or SNP hits per state. To assess whether individual pesticides were driving the associations between disease occurrence and SNP hits per pesticide, we applied a regularized regression approach (LASSO) to identify which predictors may contribute more significantly to the association. For Alzheimer’s disease (267/280), multiple sclerosis (254/265), and Parkinson disease (266/278), most pesticides were selected in the stable set of variables (full set of coefficients in Excel Table S2). For all three diseases, the pesticides that did not receive a coefficient were used with the same application pattern in each state, meaning that there were no differences in the number of SNP hits per state (e.g., chlorpyrifos, glyphosate, paraquat, etc.), or were pesticides only applied in one state (resmethrin, carbopenthion, dithiopyr). By reducing the input for the robust linear model to just those compounds that received a coefficient in LASSO regression, we saw a slight decrease in model performance, with Alzheimer’s disease not significantly correlated, but the models still had better performance than most randomly generated datasets (Figure S20).

For Alzheimer’s disease, the herbicide alachlor, insecticides clothianidin and omite, fungicides thiophanate and azoxytrobin, and cloquintocet-mexyl received the largest coefficients (0.63–0.81). The top five pesticides for multiple sclerosis were insecticides ethoprop, deltamethrin, and tebufenozide and fungicides tolclofos-methyl and zoxamide (0.68–0.88), and Parkinson disease had herbicides alachlor and trifluralin, insecticide isofenphos, fungicide thiophanate, and isazophos as the top five pesticides (0.57–0.69). While the ranks of the top 100 pesticide coefficients differed between the three diseases (Mann-Whitney test p < 0.05), the ranks of all pesticide coefficients did not significantly differ. This suggests some differences in pesticides driving the correlation between disease occurrence and SNP hits per pesticide per year in US states. However, this analysis cannot disentangle the exact relationship between pesticide use and high disease occurrence.

Characterizing Linkages Implicated in Different Geographic Regions
To determine what may underlie the relationship between disease occurrence and Pesticide–SNP linkages implicated in differential population response, we analyzed the top 10 and bottom 10 states with the highest and lowest values for disease incidence/prevalence and SNP hits per pesticide per year, i.e., those closest to the top right and bottom left corners of the disease correlation plots (Figure 2, referred to as high and low probability states, respectively) for the three correlated diseases: Alzheimer’s disease, multiple sclerosis, and Parkinson disease.

Across the three diseases, Rhode Island, New York, New Hampshire, Maine, Vermont, Connecticut, and Massachusetts were in the top 10 high-probability states, and Oklahoma, Mississippi, Arkansas, Texas, Georgia, and Louisiana were in the bottom 10 low-probability states (Figure S21). Seven pesticides were applied in high-probability states in 1992–2018 but were never used in low probability states, including chlorpropham, chlorobenzilate, and tolclofos-methyl. However, despite the unique pesticides used in high-probability states, no SNPs were implicated in high-probability states that were not also implicated by pesticides used in low-probability states. Therefore, we hypothesize that, even if the same SNPs are implicated by pesticides in high- and low-probability states, the frequency and efficacy at which pesticides act on SNPs differs. For example, the insecticide clofentezine was applied in 9/10 high-probability states for Parkinson disease in an average of 13.8 years (median of 14) but was only applied in 5/10 low-probability states in an average of 5.4 years (median of 3). Similarly, the fungicide zoxamide was applied in all 10 high-probability states for multiple sclerosis in an average of 10.7 years (median of 10.5) but was only applied in 5/10 low-probability states in an average of 4.4 years (median of 6). This means the SNPs and genes implicated by these pesticides would be more frequent hits in high-probability states than low-probability states. We therefore predict that pesticide use over time (and, thus, SNPs/gene hits over time) may underlie the differences observed in disease occurrence between high- and low-probability states.

Identifying Pesticides Used Over Time in Different Regions
To differentiate pesticide use patterns over time between the high- and low-probability states, we performed frequent itemset mining on the annual USGS data per state. For this process, each individual year of pesticide application data (1992–2018) was treated as a unique transaction with all the pesticides applied in a state in that year considered input items (see Figure S5 for a visual of how this was done). Association rule directionality was not considered. This analysis identified sets of two to three pesticides that were applied together in at least 5 years in the same state (i.e., pesticide usage patterns).

Across all 48 states included in the analysis, 1,682,960 unique rules (i.e., sets of two to three pesticides applied together) were identified with frequently applied pesticides like glyphosate, atrazine, and 2,4-D implicated in multiple rules across states. However, other pesticides were used differently between states (e.g., clofentezine and zoxamide) leading to different association rules. Across diseases, the 10 high-probability states had 581,195 total rules while the 10 low-probability states had 883,275 total rules, suggesting a more variable pesticide usage (Figure 3A). Comparing association rules in the high- and low-probability states across diseases, only 19% of association rules were the same, meaning the pesticides most frequently used together in 1992–2018 differed between the states (hypergeometric test p = 1, meaning no significant overlap between rules in high- and low-probability states). For the full set of high- and low-probability states, herbicides were most frequently implicated (Figure 3A).

To characterize differences in pesticide use in high- and low-probability states for each disease, we mapped the identified association rules for each state to Alzheimer’s disease, multiple sclerosis, and Parkinson disease based on the pesticides implicated in each disease in our dataset (association rules for the five highest probability states are in Excel Table S3). As with the overall set of association rules across diseases, there was no significant overlap in association rules between high- and low-probability states for each disease (hypergeometric test p = 1). Then, we reduced the association rules to those present in at least 5/10 states to ensure the rule is characteristic of high- or low-probability states rather than just representing a pesticide use pattern in one state (Figure 3B). Interestingly, by requiring at least five states to have an association rule in common, the number of association rules in low-probability states dropped below the number of association rules in high-probability states for Alzheimer’s disease and Parkinson disease. Further, insecticides were implicated most in high-probability states while herbicides were implicated most in low-probability states.

We repeated this assessment after removing common pesticides, here defined as pesticides used across 45+ states in 25+ years (14 pesticides total, including chlorpyrifos, glyphosate, and atrazine). Through this analysis of 1,658,309 association rules, the findings were the same as when all pesticides were included in rules (Figure S22). Briefly, there was no significant overlap in rules in high-probability and low-probability states, and by requiring at least five states to have an association rule, more insecticides were implicated in high-probability states and more herbicides were implicated in low-probability states.
SNPs and Genes Frequently Implicated in Disease Pathways

To prioritize SNPs, genes, and pathways that may drive differences in the relationship between the SNP hits and disease occurrence in high- and low-probability states, we generated 10 sets of 5/10 randomly selected high-probability states and 5/10 randomly selected low-probability states. Then, we identified the SNPs and genes that were more frequent hits by pesticides in the most association rules in the five randomly selected high-probability states compared to the five randomly selected low-probability states (i.e., those SNPs and genes that pesticides implicate more in high-probability states) and the corresponding disease pathways. We then compared these priority lists across the 10 repetitions to identify frequently implicated SNPs, genes, and pathways in high-probability states compared to low-probability states (i.e., those SNPs and genes that pesticides implicate more in high-probability states) and the corresponding disease pathways. Across the 10 repetitions, the SNPs present in ranked lists overlapped with the overall priority list (hypergeometric test $p < 0.05$) (Table S2). The rank order between the individual and overall priority list was not significantly different in individual repetitions for six Alzheimer’s and Parkinson disease repetitions and one multiple sclerosis repetition (Mann-Whitney $p \geq 0.05$), meaning there were some differences in the SNP ranks between runs, but the SNPs overlapped. With the five bottom states fixed, this increased to the rank order not being significantly different for eight Alzheimer’s disease and Parkinson disease repetitions and nine multiple sclerosis repetitions (Table S3), meaning the ranks were consistent across repetitions.

We found 1,187 SNPs and 2,018 genes that were more frequent hits in high-probability states than low-probability states compared to 737 SNPs and 901 genes that were more frequent hits in low-probability states than high-probability states. This supports our hypothesis that variable pesticide use in low-probability states may result in variable SNP/gene hits compared to high-probability states where the same SNPs and genes are frequently implicated. We ranked the high-priority SNPs and genes based on the hit frequency, the number of association rules implicating each SNP and gene, and how many of the 10 repetitions the SNPs and genes were implicated in (see “Methods” for details). The full set of high-priority SNPs and genes is in Excel Table S4.

While some SNPs overlapped between the three diseases, the majority of the priority SNPs were unique to each disease (Figure 4A). Of the 1,187 priority SNPs, 65 were found in two overlapping diseases (mostly between Alzheimer’s disease and Parkinson disease), and only six were in all three diseases. The 10 highest priority SNPs for Alzheimer’s disease included several SNPs in *GSK3B*, *EGFR*, *DIO2*, and one in *MAPK1* and *SGK1*. Multiple sclerosis also has a high-priority SNP in *MAPK1*, *GSSTM1*, *CXCR4*, *AGT*; two in *RPS6KB1*; and several in *CYP24A1*, while Parkinson disease had two in *ICAM1* and *PTEN* and six in *PARK7*. While most of the high-priority SNPs for Alzheimer’s disease and multiple sclerosis are modifiers according to Ensembl classifications, the top SNP for Alzheimer’s disease is moderate severity, and all top 10 Parkinson disease SNPs are moderate severity.
To determine the pathways that may underlie the increased disease occurrence in high-probability states, pathways were ranked by the fold difference in SNP/gene hits between high-probability states vs. low-probability states in each of the 10 repetitions of five randomly selected high-probability states compared to low-probability states, and the number of repetitions the pathways were implicated in. Across the 10 repetitions, the pathways present in ranked lists overlapped with the overall priority list (hypergeometric test \( p < 0.05 \)) (Table S2). The rank order between the individual and overall priority list was not significantly different in individual repetitions for all 10 Alzheimer’s disease and Parkinson disease repetitions and nine multiple sclerosis repetitions (Mann-Whitney \( p \geq 0.05 \)). With the five bottom states fixed, this increased to the rank order not being significantly different for any repetitions (Table S3), meaning the ranks were consistent across repetitions.

For Alzheimer’s disease, the top-ranked pathways for the high-probability states included Alzheimer’s disease and pathways of neurodegeneration, while multiple sclerosis had immune system-related pathways like signaling by interleukins and Parkinson disease pathways included Parkinson disease and pathways of neurodegeneration (Figure 4B). The full set of ranked pathways implicated in high-probability states is in Excel Table S5. Of the 1,065 pathways implicated in high-probability states, all pathways were also implicated in low-probability states. Therefore, we suspect that the SNPs/gene implicated in the high- and low-probability states lead to different activity of the same pathways following exposure to a pesticide.

Figure 4C shows the fold greater or lower Gene–Pathway hits from pesticides used in high-probability or low-probability states (i.e., how many times more or less a gene was implicated by pesticides in a disease pathway in high-probability vs. low-probability states). For Alzheimer’s disease, 92% of pathways had more SNP/gene hits in high-probability states than low-probability states, while 84% of multiple sclerosis pathways and 89% of Parkinson disease pathways had more SNP/gene hits in high-probability states than low-probability states. This may mean there is more potential for gene/SNP activity in disease pathways from pesticides used in high-probability states compared to low-probability states.

**Pesticides Implicating SNPs in High-Probability States**

To prioritize pesticides that may drive the differences between high- and low-probability states, we generated 10 sets of 5/10 random high-probability states vs. 5/10 random low-probability states, just as in identification of priority SNPs, genes, and pathways. Then, we identified pesticides applied in greater quantities in the five randomly selected high-probability states than the five randomly selected low-probability states and ranked these by the mass applied, number of implicated high-priority SNPs (SNPs in Figure 4A), and presence in association rules, meaning they were applied at least 5 years in 1992–2018 (see “Methods” for details). The overall set of pesticides was then determined based on the ranks across the 10 repetitions of five random high- and low-probability states and how many repetitions the pesticide was implicated in. This resulted in 47 pesticides of interest (Figure 5, Excel Table S6). Across the 10 repetitions, the pesticides present in ranked lists overlapped with the overall priority list (hypergeometric test \( p < 0.05 \)) (Table S2). The rank order between the individual and overall priority list was not significantly different in individual
Figure 5. Priority pesticides based on use in high-probability states compared to low-probability states and application patterns over time. (A) The number of priority genes and SNPs (Excel Table S4) in priority Pesticide–Disease linkages. (B) The number of association rules with priority pesticides. Number repetitions indicate number of individual cross-validation repetitions with the pesticide prioritized. (C) Sample priority pesticide application in kilograms/square mile for 1992–2018. Red trend lines indicate top five high-probability states, and blue trend lines indicate bottom five low-probability states. High-probability state indicates high disease occurrence and more SNP hits per pesticide per year, and low-probability state indicates low disease occurrence and fewer SNP hits per pesticide per year. A year missing from a trend line means the pesticide was not applied in the state in that year. Data underlying Figure 5A,B are in Excel Table S13 and data underlying Figure 5C are in Excel Table S14. Note: SNP, single-nucleotide polymorphism.
replications for eight Alzheimer’s disease and Parkinson disease replications and seven multiple sclerosis replications (Mann-Whitney \( p \geq 0.05 \)). With the five bottom states fixed, this increased to the rank order not being significantly different for eight Alzheimer’s disease replications, nine Parkinson disease replications, or any multiple sclerosis replications (Table S3), meaning the ranks were consistent across replications. Further, more pesticides were identified with priority with the five bottom states fixed (Excel Table S7).

Copper was in the top three prioritized pesticide for all three diseases and uniquely implicated in association rules in all high-probability states. Additionally, the ethylenebisdithiocarbamates (EBDC) fungicides maneb, mancozeb, and polymarcine (metiram) were in the top 10 pesticides for Alzheimer’s disease with maneb and mancozeb also in the top 10 for Parkinson disease and mancozeb and polymarcine in the top five for multiple sclerosis. Thiram, hydrogen peroxide, cyprodinil, dimethomorph, triflumizole, pyrimethanil, captan, and zoxamide were also high-priority pesticides for all three diseases. Based on potential to cause harm in humans from common targets for a given pesticide MoA, we would expect more insecticides and/or fungicides to be prioritized in high-probability states than low-probability states. For both Alzheimer’s disease and Parkinson disease, a third of the priority pesticides with MoA data were insecticides. Further, compared to the full set of pesticides used across all states, all three diseases were enriched for fungicides targeting amino acid protein synthesis, mitosis and cell division, and multisite contact activity (hypergeometric test \( p < 0.05 \)). In contrast, only five herbicides were identified with priority across all three diseases.

**Discussion**

We integrated diverse datasets to form Chemical–Pathway–Gene–SNP–Disease linkages to better account for individual human genetic susceptibility to chemical exposures in risk assessment. These linkages did not depend on existing evidence that a chemical leads to an adverse outcome in genetically susceptible individuals, which is important for predicting chemical effects before an adverse outcome occurs to protect potentially exposed populations. While we do not know the frequency of SNPs in different states, our data-driven analysis identified SNPs that may be most relevant in different geographic regions based on the pesticides applied there. With the huge number of marketed chemicals and the concern of how chemical exposures contribute to adverse outcomes, NAMS are necessary to address the growing number of possible chemical-adverse outcome combinations and account for interindividual variability. One proposed approach that differs from ours is to identify SNPs in Adverse Outcome Pathway (AOP) genes using data from the AOP Wiki. Romano et al. used this approach to identify SNPs of interest in liver cancer by linking UK Biobank genetic data to liver cancer AOPs and identifying SNPs implicated at the intersection. Future analyses could combine this AOP-based approach with our approach to increase confidence in the validity of associations formed using each respective method or even identify new AOPs from our set of Pesticide–Pathway–Gene–SNP–Disease linkages.

We found support that pesticides contribute to Alzheimer’s disease, multiple sclerosis, and Parkinson disease and that the SNPs and genes implicated at the Pesticide–Disease intersection explain the extent to which this occurs. To our knowledge, our approach using spatialized pesticide application data to identify SNPs that may be relevant to disease occurrence in different geographic regions (in the absence of data on which SNPs actually occur in different geographic regions) has not been done before. Eccles et al. used spatialized exposure data to predict the potential for different geographic regions to be affected by chemicals. While the focus of their methods was to characterize exposure to chemical mixtures based on potential molecular target perturbations (without consideration of specific SNPs, specific chemicals, or specific adverse outcomes), integrating some of their methods with our spatialized Pesticide–SNP–Disease linkages represents one approach to quantify the potential for the pesticides we identified to trigger the implicated pathways in disease progression.

We developed priority lists of SNPs and genes more frequently implicated in high-probability states than low-probability states, pesticides that may be most likely to act on these SNPs/gene, and biological pathways that may be most important in driving disease occurrence. These lists can serve as a starting point to explore the role of pesticides in nervous system disease and to account for interindividual variability in chemical risk assessment. For example, future studies investigating disease occurrence in high-probability states may evaluate if the SNPs we identified with priority are more frequent compared to states with low disease occurrence. Of the 1,187 priority SNPs identified, very few SNPs overlapped between the diseases, suggesting that the points of differential population susceptibility to chemical exposures vary between diseases. This is in line with the different toxicity mechanisms of importance that we identified between the diseases in our list of priority pathways. Many of the top SNPs and genes that we identified have extensive external support for their importance in the identified diseases. For example, the highest-ranked gene for Alzheimer’s disease is MAPK1, which is important in Alzheimer’s disease progression and other nervous system diseases.

Further, six of the top 10 SNPs prioritized for Parkinson disease are located in the \( PARK7 \) gene, which is implicated in early onset Parkinson disease when defects are present in the gene. Evidence also suggests \( PARK7 \) protective activity can be inhibited by thiocarbamates like thiram, and thiram was a priority pesticide for Parkinson disease in our analysis. Additionally, many of the SNPs/genes we found with literature supporting their role in pesticide-induced nervous system disease (Table S1, e.g., rs1130409 in \( APEX \)) were identified as high-priority SNPs/gene in our analysis.

We identified priority pesticides associated with nervous system diseases based on their potential to affect genetically susceptible individuals. Rotenone was identified with priority in Alzheimer’s disease and Parkinson disease and is a well-recognized contributor to neurotoxicity, with a “rotenone model for Parkinson disease” being a common model to mimic environmental toxicant-induced Parkinson disease. We also identified cyprodinil as a high-priority pesticide in all three diseases, and cyprodinil applied in mixtures has been found to cause adverse nervous system effects in mice, which supports our hypothesis that the combination of pesticides applied in high-priority states may be more important in the relationship between disease occurrence and SNP hits per pesticide per year than any individual pesticide. Polymarcine (metiram), mancozeb, and maneb were also prioritized in different combinations across all three diseases. Polymarcine, mancozeb, and maneb are EBDC fungicides, and while polymarcine toxicity is under-studied compared to other EBDCs, both rodent models and epidemiological studies have implicated maneb in Parkinson disease, especially with co-exposure to paraquat (which was applied in all years in all states in our study). We also identified polymarcine as a high-priority chemical for multiple sclerosis, and while polymarcine has not been linked to multiple sclerosis in the literature, it has been linked to amyotrophic lateral sclerosis (ALS), so it may also be worth investigating further. Lastly, we identified copper among the top three priority pesticides in all three diseases. Copper is an essential metal with a pivotal role in the human nervous system and is used heavily in organic farming. However, there is some evidence that exposure to copper...
may increase the risk for Parkinson disease, and multiple studies have found environmental copper exposure increases the risk for Alzheimer’s disease in susceptible individuals. This literature support combined with our findings highlights copper as an important compound for further investigation in differential development of nervous system disease.

While we found that the high-probability states implicated more SNP/gene hits in disease pathways than low-probability states and identified pesticides that, on average, were applied more in high-probability states than low-probability states, there were some differences in priority lists when subsets of five high-probability states were compared to five low-probability states. These differences were reduced when subsets of five high-probability states were compared to only a constant set of five low-probability states, suggesting there is more variability in low-probability states than high-probability states. Further, more pesticides were identified with priority when the five bottom states were fixed. One possible reason for this is that the low-probability state California was removed from the analysis when the five bottom states were fixed, and California is different from other states. California was a low-probability state for both Alzheimer’s disease and multiple sclerosis, but California used the highest number of pesticides and the highest amount of pesticides of all US states (1.8 x 10^9 kg applied between 1992 and 2018, over 1.8 times as much pesticide as the state with the next most pesticides applied). Because of this, some priority pesticides had higher application rates in California than in high-probability states (e.g., cyprodinil). One reason that California differs from other low-probability states is that it is the top agricultural exporting state in the US and has been since data were first collected in 2000 (https://www.ers.usda.gov/data-products/state-agricultural-trade-data/annual-state-agricultural-exports/). California is also one of the largest economies in the world, meaning it may not be as representative of other low-probability US states. However, we still identified priority SNPs, genes, pathways, and pesticides even with California as a low-probability state, and the total SNP hits per pesticide per year implicated by California were lower than for high-probability states. This shows that, even with California as a low-probability state, the relationship we found between SNP hits, pesticides, and disease occurrence was consistent.

Our analysis complements epidemiological studies using a broad, data-driven approach. Because epidemiology studies are time-consuming and expensive, there are not many studies covering the same pesticides and geographic regions in our study, and there are even fewer that include a wider range of pesticides and whole-genome sequencing. Data-driven approaches like ours can be useful for pinpointing hotspots where more targeted methods should be used (e.g., when assessing disease occurrence in high-probability states, our priority lists identified SNPs that can be searched for, pesticides that may be more relevant to ask residents about when determining use scenarios, etc.). Further, all code used in this analysis are available in GitHub so that these datasets/linkages can be further explored or so these methods can be repeated for other chemical classes, disease end points, or geographic regions (e.g., this analysis could be repeated at the county level rather than the state level if county-level disease occurrence data become available). For future mechanistic studies, our study can also serve as a complement to analyses using genetically diverse in vitro cell systems such as organoids as we identified SNPs/genes and pathways to assess following pesticide exposure.

The primary limitation in our analysis is the assumption that pesticides used in a geographic region affect individuals living in the same region. Specific routes of exposure are not considered (e.g., oral vs. inhalation), nor the likelihood that an individual would come in contact with a particular pesticide based on its physicochemical properties (e.g., persistence). Temporal differences in exposure patterns are also not considered in our analyses (e.g., when a pesticide is applied to a crop or the life stage of an individual at the time of exposure). Future analyses could focus on likely highly exposed groups (e.g., farmers, bystanders, people who eat more locally grown produce), incorporate trade data on crops, or incorporate likely pesticide minimum residue levels to better predict actual pesticide concentrations that individuals are exposed to. Additionally, it is unlikely that the diseases we are analyzing only occur due to pesticide exposure, so it is possible that some of the associations we found have other factors underlying them in our or other more epidemiological analyses (e.g., diet, lifestyle, other chemical exposures). Further, when evaluating pesticide association rules as pesticides used together in a state, we are not accounting for mixture effects or the possibility that pesticides used in the same year were not applied in the same area or at the same time. While these are important considerations to determine how individuals in a region are exposed to pesticides, our preliminary approach was still capturing a signal between spatialized pesticide usage and disease occurrence, which may indicate local exposure pathways were still relevant. The pesticide association rules we identified also represent an area for further research where pesticide mixtures may be particularly relevant in adverse outcomes since we found different pesticide use patterns between high- and low-probability states. Finally, because our analysis depends on existing datasets, there is a chance that data in our analysis are biased toward pesticides and SNPs/genes that have been better studied. While we still see the same positive correlation between SNPs per pesticide per year and disease occurrence even when the pesticide set is reduced to pesticides present across states, it is possible we are missing relevant gene targets for many of the pesticides in our analysis simply because those associations have not been tested. Future analyses should continually incorporate new chemical–gene relationships as they are identified and search for confirmatory exposure–SNP–adverse outcome relationships.

Conclusions
We formed a dataset of Pesticide–Pathway–Gene–SNP–Disease linkages to characterize the mechanisms by which pesticides can lead to nervous system disease in individuals with different genetic susceptibility. Our analysis serves as a broad, complementary approach to experimental and/or mechanistic studies by harnessing data from well-established public sources and finding new associations. We found that pesticides may contribute to Alzheimer’s disease, multiple sclerosis, and Parkinson disease in the United States, and the priority lists of SNPs, genes, pathways, and pesticides we developed identify potentially important contributors to this relationship. This work represents a novel approach to consider interindividual variability in disease outcomes where existing Chemical–SNP–Disease associations are limited. The approach we demonstrated for data integration and subsequent analysis of Chemical–Pathway–Gene–SNP–Disease linkages can be used to assess additional chemical classes, diseases, and/or geographic regions to characterize the relationship between chemicals, genetic variants, and disease outcomes in a population in order to protect vulnerable individuals and advance chemical risk and impact assessment in support of improving public health.

Acknowledgments
This work was financially supported by the SPRINT project (grant agreement no. 862568) funded by the European Commission through Horizon 2020, and by the PARC project (grant no. 101057014) funded under the European Union’s Horizon Europe Research and Innovation program.