Biostat 579 report

Introduction:

According to the World Cancer Research Fund, lung cancer is currently the second leading form of cancer worldwide[1]. While there are numerous well studied factors linked to risk of lung cancer, one in particular that is of question is the use of pesticides. A 2012 review of literature on pesticides and their links to various forms of cancer concludes that chemicals in every major class of pesticides were noted to have significant associations with an array of cancer sites[2]. Another study by Alavanja et al. concludes that seven individual pesticides, dicamba, metolaclor, pendimethalin, carbofuran, chlorpyrifos, diazinon, and dieldrin, were positively associated with lung cancer in particular[3]. Recently, a manuscript by Kangkhetkron and Juntarawijit was submitted for peer review, which studies that association between various pesticides and lung cancer in the country of Nahkon Sawan province of Thailand[4]. Out of the 17 pesticides that were studied, they found dieldrin, chlorpyrifos, and carbofuran to be significantly associated with higher risk of lung cancer. In this report, we offer an alternative analysis of the data provided in their study, which we believe to be necessary due to the nature of the data used.

Data and Sampling procedure:

The data was provided jointly with the manuscript of Kangkhetkron and Juntarawijit. The data collected was collected for the purpose of a case control study. Data on the cases was provided by the Cancer Based Program operated by the Thai National Cancer institute. Cases of lung cancer between the years of 2014 to 2017 were each individually contacted. Of the 299 total cases that were contacted, 229 responded and participated in the study. It is worth noting that these cases contacted were the cases who were currently living. The controls were neighbors of the cases who were of the same gender, and within 5 years of the case that they were associated with. For each case, 2 controls were randomly selected to be included in the study, for a total of 458 controls.

The data was collected in the form of a questionnaire. As such, most of the data is encoded as grouped data rather than continuous data. Important demographic data that was collected include data on gender, age, marital status, education, occupation, living duration in the community, distances between home and farmland, exposure to air pollution (i.e., cooking smoke, working in a factory with air pollution; asbestos, diesel engine exhaust, silica, wood dust, painting and welding exposure) and cigarette smoking. Data on pesticides was categorized into five groups: insecticides (organochlorine, organophosphate, carbamate, and pyrethoid), herbicides, fungicides, rodenticides, and molluscicides. This was encoded as either a binary exposure vs no

[1] https://www.wcrf.org/cancer-trends/lung-cancer-statistics/
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6276799/
[3] https://pubmed.ncbi.nlm.nih.gov/15496540/
[4] https://f1000research.com/articles/9-492/v1#ref-38

exposure, as well as number of days exposed. Summary statistics of each variable listed are provided in figure 1.

Methodology:

In the original manuscript by Kangkhetkron and Juntarawijit, data was analyzed by adjusted multiple logistic regression models, adjusting for gender (male, female), age, cigarette smoking (ever, never smoke), occupation (farmer, non-farmer), and exposure to air pollution. It is stated that these factors are fundamental confounding factors[5]. However, we believe that, due to the sampling scheme of cases and their paired controls, a matched analysis method is necessary. We provide three models to compare: an unconditional adjusted multiple logistic regression model, a conditional logistic model, and a mixed effects model. Due to privacy concerns, information on which cases were paired with which controls are not available. As such, we created a composite matching score based on the variables adjusted for in the logistic regression of the original authors, as well as the distance to nearest farmland variable, in hopes of recovering some information on the original pairing of neighbors to the cases.

The composite matching score was created by standardizing each of the variables between 0 and 1 by dividing each individual value by the maximum over the dataset, and then taking the average of these standardized values over the individual. Due to computational concerns, the matching was done by greedy algorithm, starting from the smallest composite matching score of the cases. For the unconditional adjusted multiple logistic regression model, we aim to replicate the results of the original study, and therefore try to copy as much of their methodology as possible. However, in order to facilitate comparison across our other models, we will opt to test all 17 pesticides in one single model, rather than fit 17 different models. We will use Wald intervals to assess significance, using an alpha level of 0.05. For the conditional logistic regression model, each strata will be defined by the matching as done above. Wald intervals will be used to assess significance, using the same alpha level as above. For the mixed effects model, the random effect will be given to each strata, and the analysis will be done with Wald intervals as well. In both the conditional logistic regression model, as well as the mixed model, additional covariates will not be adjusted for, as they form the basis of the matching scheme. Due to the amount of missing data on the number of days exposed to each pesticide, the effect that will be tested will be on the binary exposure to each pesticide instead.

Results:

The full results of each model can be found in figure 3. From the unconditional adjusted multiple logistic regression model, we found a statistically significantly higher odds ratio of lung cancer for groups who were exposed to dieldrin (95% CI 1.33-5.11), chlorpyrifos (95% CI 1.97-6.69), and carbofuran (95% CI 1.34-3.95) after adjusting for gender, age, cigarette smoking, occupation, and exposure to air pollution. Due to the slightly different nature of our analysis over the original authors, these estimates do not

[5] https://f1000research.com/articles/9-492/v1#ref-21

exactly match. However, the 3 pesticides that were originally significant in the study do not change, and are still significant here. Of note is that none of the variables that were adjusted for were statistically significant.

For the conditional logistic regression model, we also found that the odds ratios for dieldrin (95% CI 1.19-4.93), chlorpylifos (95% CI 1.83-6.3), and carbofuran (95% CI 1.24-3.72) were all statistically significant, as in the case of the unconditional adjusted logistic regression model. Although alachlor appears to be statistically significant in the table, it is confirmed that this is due to rounding error, and the effect itself was not actually statistically significant at the 0.05 alpha level. Finally, for the mixed effects model, we also found that the odds ratios for dieldrin (95% CI 1.26-4.66), chlorpylifos (95% CI 1.85-6.15), and carbofuran (95% CI 1.33-3.83) were statistically significant as well. In the mixed effects model, the estimated variance of the random effect was 0.

Discussion and limitations:

In all 3 of the models, the same three pesticides that were labeled in the original study by Kangkhetkron and Juntarawijit were found to be significant. However, there is reason to be dubious of the results presented in this report. The original goal of this study was to introduce matched analysis onto the study, as it is believed to be necessary due to the matched nature of the sampling. However, due to how the data was provided, such information was not available. The matching done in this study was an attempt to recover that information from what was available. However, when looking at the results, it is possible that this attempt was not successful. We can see this from the fact that the estimated variance of the random effect in the mixed model was 0, which indicated that the strata had no effect. It is possible that this is due to the actual pairings having no effect. However, when considering the fact that the none of the variables adjusted for were significant, and the fact that these variables were what the composite matching score was based off of, it is also possible that the strata were simply capturing the effect of those variables and not much else, which would explain why the estimated variance of the random effect was 0. In addition to this, although not directly comparable, the effect size estimated in the conditional logistic regression model does not differ much at all from the estimated effect size in the other 2 models. Again, the same conclusions can be drawn in this case. Overall, it is impossible to come to any significant conclusions whether or not the matching done in this study was able to recover any information on the original matchings, and analysis would need to be done knowing those original matchings to be fully valid.

Other concerns include underlying bias in the sample. The questionnaire was sent to every recorded case, and only a portion of the lung cancer cases responded. If there is some underlying mechanism to why these cases did not respond to the survey, then this may induce some bias into the results. Ultimately, it is impossible to check how much this affects the results, as it is impossible to show randomness of response to the survey. The other major concern regarding bias is recall bias from the participants. According to the original study, with limited available information on the issue in Thailand, participants were not expected to be aware of pesticides as a causal factor for

lung cancer[6]. This is a problem as cases tend to memorize exposure better when they know that exposure to certain treatments can cause illness.

For the assumptions of the logistic regression models, it is possible that independence of observations is violated. However, when checking the residual plots for all three models, as shown in figures 5, 6, and 7, it seems that there is random scatter with respect to the index of each individual, so we will claim that independence is satisfied. In the unconditional logistic regression model, we must assume a linear relationship between the log odds of lung cancer and age. Visually, this can be checked for in figure 4. It does not seem that there is a very linear relationship between age and the log odds of lung cancer. Finally, we can check the assumption of low multicollinearity through a correlation heatmap, which is shown in figure 2. Visually, it seems that there is no strong correlation between the covariates used in the models.

Conclusions:

Our study found that there was a significant association between dieldrin, chlorpylifos, and carbofuran, and the incidence of lung cancer of those living in the Nakhon Sawan province in Thailand. Our results in this study were consistent with the results produced by Kangkhetkron and Juntarawijit, despite using different analysis methods, as well as other previous literature on the topic. However, despite these findings, we find that there are various reasons that these results may still be dubious, the most prominent of which would be the possibility of failure to recover the original pairings present in the data, which is necessary for the matched analysis used in this study. We suggest that further analysis be done using the original pairings of case to controls in order to have the most valid results.

[6] https://f1000research.com/articles/9-492/v1#ref-21

Figures:

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
LungCA	680	0.34	0.47	0	0	1	1
Gender	680	0.59	0.49	0	0	1	1
age	680	66	11	31	58	74	98
age_group	680	2.6	1	1	2	3	4
status	680	2.1	0.44	1	2	2	3
education	680	1.1	0.34	1	1	1	3
Occupation	680	1.4	0.5	1	1	2	2
Residency	680	2.6	0.66	1	3	3	3
Distances	680	2	0.94	1	1	3	3
Cooking_fumes	680	1.7	0.47	1	1	2	2
Air_Pollution_Exposure	680	0.49	0.5	0	0	1	1
CigSmoke1	680	0.65	0.48	0	0	1	1
CigSmoke2	680	0.53	0.78	0	0	1	2
Cigarette_total	678	47537	103755	0	0	54750	876000
Cigarette_year	680	10	17	0	0	15	70
Cigarette_number	679	4.1	7.1	0	0	6	45
CigSmoke_status	680	1.4	0.67	1	1	2	3
Glyphosate_use	680	0.41	0.49	0	0	1	1
Glyphosate_days	680	300	654	0	0	338	5760
Paraquat_use	680	0.35	0.48	0	0	1	1
Paraquat_days	680	229	582	0	0	160	5760
two_four_Dichlorophenoxy_use	680	0.17	0.38	0	0	0	1
two_four_Dichlorophenoxy_days	680	73	302	0	0	0	3600
Butachlor_use	680	0.056	0.23	0	0	0	1
Butachlor_Days	680	15	127	0	0	0	1920
Propanil_use	680	0.047	0.21	0	0	0	1
Propanil_days	680	8.9	60	0	0	0	660
Alachlor_use	680	0.063	0.24	0	0	0	1
Alachlor_days	680	40	249	0	0	0	3600
Endosalfan_use	680	0.12	0.32	0	0	0	1
Endosalfan_days	680	72	334	0	0	0	5460
Dieldrin_use	680	0.065	0.25	0	0	0	1
Dieldrin_days	44	668	512	30	315	960	2450
DDT_use	680	0.074	0.26	0	0	0	1
DDT_days	50	331	373	2	82	320	1280
Chlorpylifos_use	680	0.1	0.3	0	0	0	1
Chlorpylifos_days	70	629	366	80	360	960	1440
Folidol_use	680	0.15	0.36	0	0	0	1
Folidol_days	104	530	564	12	130	858	2400
Mevinphos_use	680	0.056	0.23	0	0	0	1
Mevinphos_days	38	668	550	7	210	1080	2160
Carbaryl_Savins_use	680	0.071	0.26	0	0	0	1
Carbaryl_Savins_days	48	819	726	60	270	1200	3000
Carbofuran_use	680	0.12	0.33	0	0	0	1
Carbofuran_days	85	733	1043	1	100	1200	7200
Abamectin_use	680	0.2	0.4	0	0	0	1
Abamectin_days	134	478	711	20	123	540	5460
Armure_Propiconazole_use	680	0.12	0.32	0	0	0	1
Armure_Propiconazole_days	80	357	414	2	80	480	1400
Metal_aldehyde_use	680	0.063	0.24	0	0	0	1
Metal_aldehyde_days	43	593	399	2	345	790	1680
Morphology_Group	680	0.9	1.6	0	0	1	6

Figure 1: Summary statistics

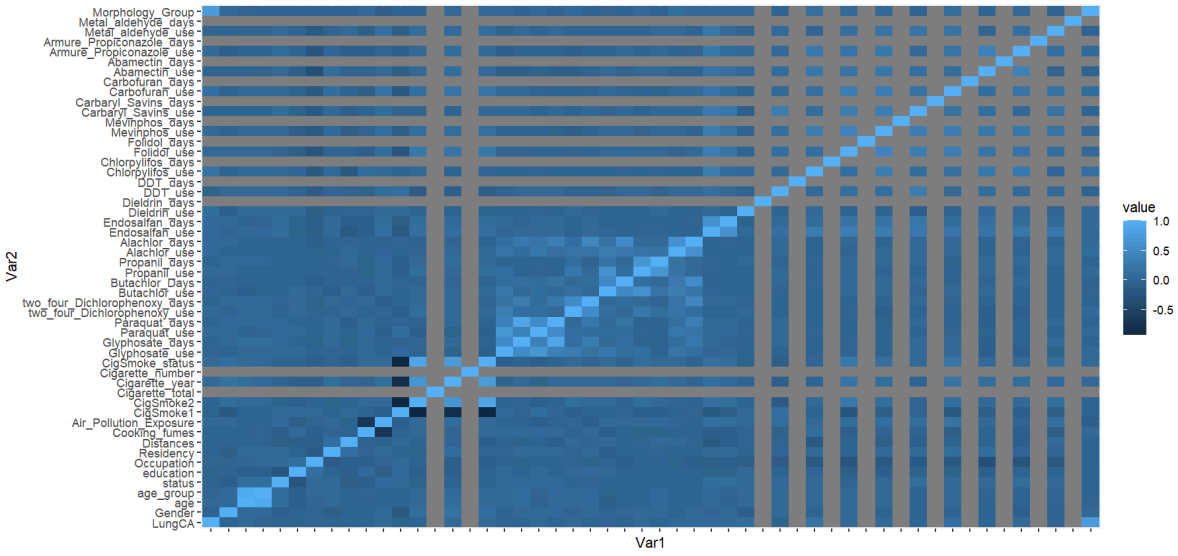

Figure 2: Correlation heatmap

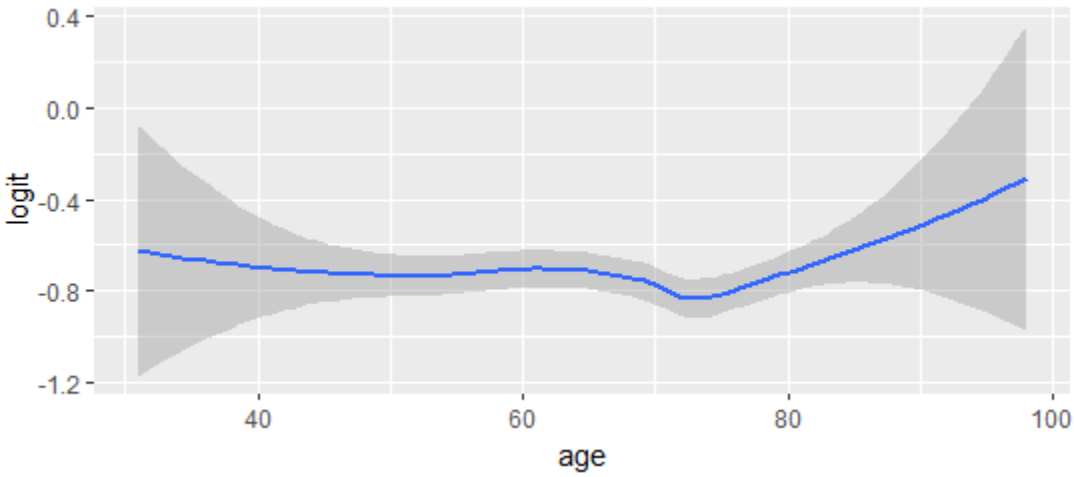

Figure 4: Relationship between age and log odds of lung cancer

Comparison of the 3 models

	Dependent variable:		
	LungCA *conditional logistic* (1)	LungCA *generalized linear mixed-effects* (2)	*logistic* (3)
Glyphosate_use	-0.080	-0.147	-0.130
	(0.251)	(0.245)	(0.247)
Paraquat_use	-0.088	0.007	0.014
	(0.251)	(0.240)	(0.242)
two_four_Dichlorophenoxy_use	0.066	0.115	0.097
	(0.256)	(0.248)	(0.250)
Butachlor_use	-0.037	-0.007	-0.023
	(0.521)	(0.483)	(0.487)
Propanil_use	0.032	0.070	0.060
	(0.554)	(0.515)	(0.520)
Alachlor_use	0.688*	0.547	0.551
	(0.392)	(0.374)	(0.374)
Endosalfan_use	0.432	0.365	0.323
	(0.319)	(0.304)	(0.308)
Dieldrin_use	0.884**	0.883***	0.956***
	(0.364)	(0.334)	(0.341)
DDT_use	-0.562	-0.478	-0.374
	(0.388)	(0.370)	(0.379)
Chlorpylifos_use	1.223***	1.216***	1.280***
	(0.315)	(0.306)	(0.310)
Folidol_use	0.020	0.095	0.094
	(0.282)	(0.280)	(0.287)
Mevinphos_use	-0.290	-0.336	-0.346
	(0.461)	(0.460)	(0.463)
Carbaryl_Savins_use	-0.301	-0.309	-0.271
	(0.408)	(0.403)	(0.408)
Carbofuran_use	0.764***	0.814***	0.830***
	(0.281)	(0.270)	(0.275)
Abamectin_use	-0.459*	-0.414	-0.378
	(0.271)	(0.255)	(0.261)
Armure_Propiconazole_use	-0.253	-0.321	-0.307
	(0.338)	(0.319)	(0.321)
Metal_aldehyde_use	-0.090	-0.050	-0.009
	(0.377)	(0.368)	(0.375)

Gender			-0.061
			(0.178)
age_group			-0.023
			(0.087)
CigSmoke1			-0.194
			(0.194)
Occupation			0.220
			(0.191)
Air_Pollution_Exposure			-0.057
			(0.178)
Constant		-0.874***	-0.980**
		(0.134)	(0.440)
Observations	669	669	669
R^2	0.068		
Max. Possible R^2	0.519		
Log Likelihood	-221.478	-402.117	-400.896
Akaike Inf. Crit.		842.234	847.791
Bayesian Inf. Crit.		927.844	
Wald Test	39.560*** (df = 17)		
LR Test	47.025*** (df = 17)		
Score (Logrank) Test	47.310*** (df = 17)		
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 3: Results and model diagnostics

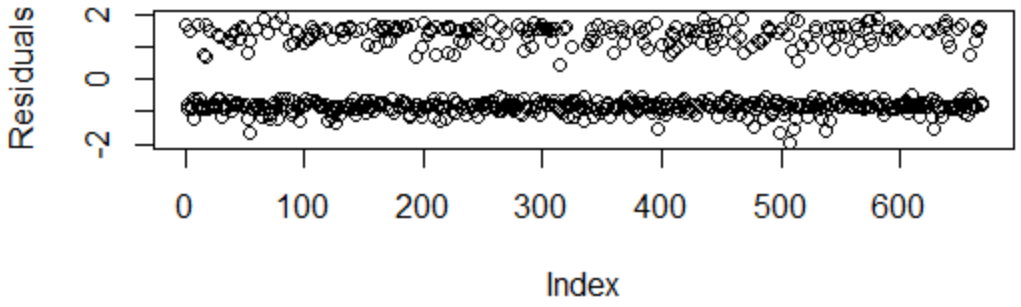

Figure 5: Residual plot of unconditional logistic regression model

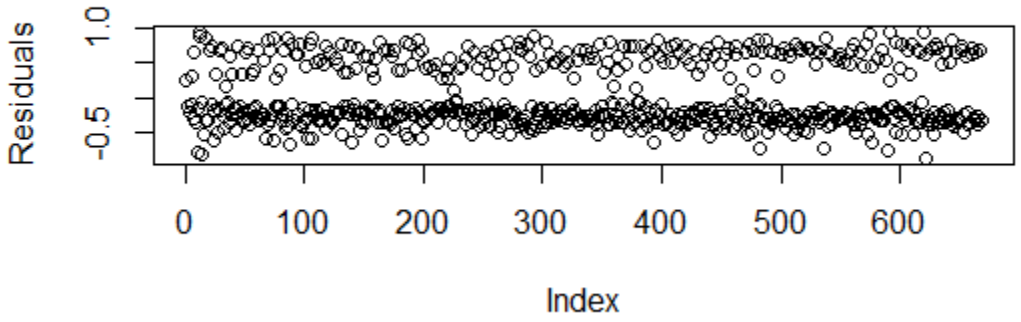

Figure 6: Residual plot of conditional logistic regression model

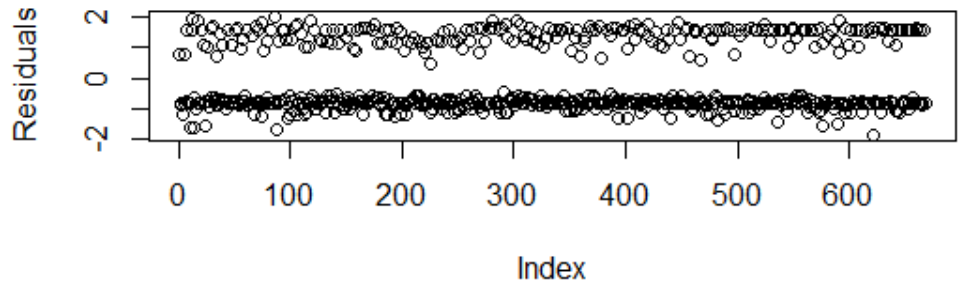

Figure 7: Residual plot of mixed effects model

References:

1. Alavanja, Michael. “Occupational Pesticide Exposures and Cancer Risk: A Review - PubMed.” *PubMed*, https://pubmed.ncbi.nlm.nih.gov/22571220/. Accessed 25 May 2023.
2. ---. “Pesticides and Lung Cancer Risk in the Agricultural Health Study Cohort - PubMed.” *PubMed*, https://pubmed.ncbi.nlm.nih.gov/15496540/. Accessed 25 May 2023.
3. Kangkhetkron, Teera, and Chudchawal Juntarawijit. “Pesticide Exposure and Lung Cancer Risk: A... | F1000Research.” *F1000Research* | *Open Access Publishing Platform* | *Beyond a Research Journal*, https://f1000research.com/articles/9-492/v1#ref-21. Accessed 25 May 2023.
4. “Lung Cancer Statistics | World Cancer Research Fund International.” *WCRF International*, https://www.wcrf.org/cancer-trends/lung-cancer-statistics/. Accessed 25 May 2023.