Applied methodologies demonstrated through three independent studies for assessing the effects of socioeconomic, environmental chemical exposure, and therapeutic factors on oncological diseases and COVID-19

Bayarmagnai Maggie Munkhjargal
University at Albany, State University of New York, bmunkhjargal@albany.edu

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Epidemiology Commons, and the Oncology Commons

Recommended Citation
https://scholarsarchive.library.albany.edu/legacy-etd/2758

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive. Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
Applied Methodologies Demonstrated through Three Independent Studies for Assessing the Effects of Socioeconomic, Environmental Chemical Exposure, and Therapeutic Factors on Oncological Diseases and COVID-19

by
Bayarmagnai Maggie Munkhjargal

A Dissertation

Submitted to the University at Albany, State University of New York

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

School of Public Health
Department of Environmental Health Sciences

2021
Acknowledgments

First, I would like to express my heartfelt appreciation to my advisor and Dissertation Committee Chair, Dr. David Carpenter. His personal and scholarly support was instrumental throughout the process. Without David’s help on multiple levels, none of this would have been possible.

I am deeply grateful to all the members of my Dissertation Committee. I appreciate Dr. Bell for career advice, methodological support, and bridging me with the other faculties for scientific collaborations. I am blessed to have Dr. Feingold encouraging me to gear my doctoral research towards things worthwhile with the potential of having a positive impact on public health, Dr. Welsh’s advice to consider biological plausibility in research beyond mere numbers, and Dr. Lessner for being a mobile text away to ready jump in for help with his insightful guidance.

I am grateful to Drs. Alan da Silva, Yu-Chen Su, and Bogdan Muresan for their expertise in spatial and biostatistics, Dr. Antonio Vaz de Macedo for his expertise in hematology and stem cell transplantations, Drs. Dimitrios Kouzoukas and Yoon JungLee for their cell-biology, clinical pharmacy, and professional writing support. I would like to thank my other collaborators, Drs. Tanima Bose, Gwang-Jin Kim, Paola A. Correa, Gwang-Jin Kim, and Santhi Pondugula on the COVID-19 research. Dear Ph.D. trainee Yeseul Ahn, and Drs. Gabriela Henriquez and Cristina Camargo, thank you for your subject matter expertise and for being ready for help, literally with anything you could help with. Thank you, all my research collaborators, for your trust and for allowing me to take the lead and being patient with my learning curves. I am proud of us and the research projects that we have accomplished together.

I appreciate the support of Drs. Felipe Fregni and Alma Sanchez, the Director and Coordinator of the Principles and Practice of Clinical Research Program at Harvard T.H Chan School of Public Health and fellow members of Cheeky Scientist Associates, the Ph.D. Professionals Organization.

I would like to congratulate my fellow friends Drs. Azad Mohammed, Delisha Bella, Sam Byrne, Stacy Pettigrew, Eva Tanner, Austin Roberts, Abigail Stamm, and Wayne Lawrence for accomplishing
your doctoral degrees at both the Departments of Environmental Health Science and Epidemiology and Biostatistics. Your success is an inspiration to me.

I thank the faculty, Drs. Shao Lin, Marya Schymura, Francis Boscoe, Melissa Tracy, and Recai Yucel for research opportunities and technical and funding support when I needed it.

Finally and foremost, I appreciate my husband, Seth, who has been a constant source of support and encouragement during the challenges of graduate school and life. I am thankful for having you in my life. I thank our son, Telmen Weinstein, who will celebrate his 5th year birthday soon, for being patient and abstaining from his hundreds of hours of spending time with mom throughout the process.

This thesis is dedicated to my extended family, who cheered me up from overseas in Mongolia. I could not have done this without your love and patience. It has been an honor to learn from you all and thank you for being part of this academic journey.
CHAPTER 1. INTRODUCTION

   1.1 Burden of Cancer
      1.1.1 What Causes Cancer?
      1.1.2 Cancer Treatment: Challenges, and Opportunities
   1.2 Burden of COVID-19 Pandemic
      1.2.1 Airborne transmission of viral agents and COVID-19
      1.2.2 Determinants of COVID-19 Incidence and Mortality
   1.4 Specific Aims
      1.4.1 EPC and Residential Exposure to Hazardous Organic Chemicals
      1.4.2 COVID-19 Vulnerability and Social & Epidemiological Factors
      1.4.3 Comparative Effectiveness of Emerging CAR T-Cells for the treatment of R-R DLBCL
   1.5 References

CHAPTER 2. Exocrine Pancreatic Cancer (EPC) and Environmental Exposure to Hazardous Organic Chemicals, New York State, USA.

2.1. Abstract
2.2. Background
   2.2.1 Environmental Chemical Exposures and EPC
   2.2.2 The use of SPARCS vs NYSCR in population-based cancer research
2.3 Materials and Methods
   2.3.1 Study Design and Population
   2.3.2 Exposure assessment
   2.3.3 Statistical analysis
2.4 Results
   2.4.1 Descriptive analysis
   2.4.2 Modelled results
2.5 Discussion
2.6 Conclusion
2.7 References

CHAPTER 3. A Methodological Blueprint to Identify COVID-19 Vulnerable Locales by Socioeconomic and Epidemiological Factors

3.1 Abstract
3.2 Background
3.3 Materials and Methods
   3.3.1 Study Design and Population
3.3.2 Conceptual Model
3.3.3 SES Measurement and Epidemiological Factors
3.3.4 Global Models
3.3.5 Local Spatial Models

3.4 Results
3.5 Discussion
3.6 Conclusion
3.7 Supplementary Materials
3.8 References

CHAPTER 4. Efficacy and Safety of Experimental Chimeric Antigen Receptor (CAR) T-Cells Versus Axicabtagene Ciloleucel (Yescarta) for the Treatment of Relapsed/Refractory Large B-Cell lymphoma (LBCL): Matching Adjusted Indirect Comparisons (MAICs).

4.1 ABSTRACT
4.2 BACKGROUND
4.3 Materials and Methods
   4.3.1 Summary of the evidence prior to this study
   4.3.2 Data sources
   4.3.4 Study outcomes
   4.3.5 Statistical analysis
4.4 Results and Discussion
4.5 Conclusion
4.6 References

CHAPTER 5. DISCUSSION

5.1 Ecological Study Design in Environmental Health and Cancer Research
   5.1.1 Summary
   5.1.2 Methodological Challenges and Opportunities
5.2 Mapping the COVID-19 Vulnerable Locales by Epidemiological and Socioeconomic Risk Factors
   5.2.1 Summary
   5.2.2 Methodological Highlights
5.3 Indirect Comparisons in Comparative Effectiveness Study
   5.3.1 Summary
   5.3.2 Challenges and Opportunities in Matching Adjusted Indirect Comparisons
5.4 References

APPENDIX

Appendix A: Chapter 3 Supplementary Materials
Appendix B: Chapter 4 Supplementary Materials
List of Tables

Chapter 2:
Table 2.1 Inpatient hospital discharge rate for pancreatic cancer during 1996-2013, NY State excluding NY city by demographics and poverty status.

Table 2.2 Adjusted rate ratios and 95% CIs of hospitalization for pancreatic cancer during 1996-2013, NY State excluding NY city. Negative Binomial Regression model results.

Table 2.3 Subgroup analysis. Relative risk and 95% CI by subgroups.

Chapter 3:
Table 3.1 Parameter estimates and 95% CI of the relative risk of COVID-19 associated with health and SES determinants during the entire study period (January 20–July 1, 2020).

Chapter 4:
Table 4.1 Summary of Clinical Trials, Pooled Populations by CAR T-cell structure, and Study Endpoints.

Table 4.2 Key Baseline Characteristics and MAICs of Experimental CAR T-cells versus Yescarta regarding Progression-Free Survival.

Table 4.3 MAICs of Experimental CAR T-cells versus Yescarta Regarding Grade ≥3 Cytokine Release Syndrome (CRS) and Neurotoxicity (NT).

List of Figures

Chapter 2:
Figure 2.1 Flowchart illustrating the Record Selection Process.

Figure 2.2 Deduplicated SPARCS data compared to The New York Cancer Registry by calendar year, age groups, and race and year.

Chapter 3:
Figure 3.1 Conceptual model of the causal relationship between the SARS-CoV-2 and area health/SE determinants.

Figure 3.2 The spatial distribution of COVID-19 cases across pandemic phases.

Figure 3.3 Relative Risk of COVID-19 associated with area health and SES determinants (GNBR models).

Figure 3.4 Spatial variation in the relative risk of COVID-19 associated with area-health and SES themes in the early, middle, and late phases of the pandemic (GWNBR models).
Appendix A: Chapter 3, Supplementary Materials

Table S.3.1 Confirmed COVID-19 cases diagnosed during January 20 - July 1, 2020, in South Korea.

Table S.3.2 Data sources reviewed and used for SES measurement.

Table S.3.3 PCA details showing factor scores with their weights, and selected area health and SES variables, and thematic composite indices.

Table S.3.4 GNBR model estimates with and without the estimated data for Daegu’s subparts throughout the study period (January 20 through July 1, 2020).

Table S.3.5 Parameter estimates and the Relative Risk of the COVID-19 incidence associated with health and SES determinants by three time periods corresponding with the early, middle, and late phases.

Table S.3.6 Matrix table of Pearson’s correlation coefficients (r).

Appendix B: Chapter 4, Supplementary Materials

Figure S.4.1. PRISMA Flow Diagram, Large B-Cell Lymphoma (LBCL)

Figure S.4.2. MAIC of experimental CAR T-cells and Yescarta regarding PFS among patients who received infusion. Kaplan Meier survival curves. Hazard Ratios and 95% Confidence Intervals computed through Cox Proportional Hazards Models.
Abstract

When trying to understand risk factors for a disease, even before we know the causal agents, it is necessary to create a surveillance data set that answers the questions of who, when, and where, and includes any potential covariates which may either promote or prevent the disease. There are a wide variety of surveillance data. For example, hospital discharge data, such as the New York Statewide Planning and Research Cooperative System (SPARCS) data, encompass all hospitalized cases in the state, while clinical datasets cover a specially constructed population in pursuit of research of a certain disease type. Analysis of surveillance data must begin with the selection of an appropriate study design and analytical approach. An ideal data source would include abundant information directly collected from individuals.

The goal of this dissertation is to demonstrate the limitations and strengths of ecological, cross-sectional, and systematic analysis with quantitative comparison in contrast to ideal methodological approaches given the practical constraints. Consequently, three independent studies have been conducted requiring different analytical approaches taking into consideration of the research questions and the types of data sources. The health outcome data sources used in the three different projects were intentionally vastly different from each other in terms of the granularity of data and exposure types. The granularity of data sources across the studies in this dissertation ranges from combined administrative units by the similarity of exposures (group of ZIP codes), aggregated geographical data by small administrative unit, and individual-patient level exposure data in the context of medical treatment. The scientific evidence obtained through these independent studies demonstrated in this dissertation is important to help harness the needed medical resources for the population with unmet preventative and treatment means to fight the disease.
CHAPTER 1. INTRODUCTION

The environmental exposure that humans experience encompasses a wide spectrum of health stressors including chemical, biological, or physical substances, and may harm people’s health. Thus, exposure types across the three projects also encompass distinctly different types of exposures: environmental, socioeconomic, and epidemiological factors, and medical interventions. Different data sources and availability of exposure data in each project presented unique challenges in choosing the appropriate study design, methods, and appropriate statistical techniques. Although a variety of strong epidemiological methods are available, each methodology has its own set of strengths and limitations that must be carefully evaluated before use.


1.1 Burden of Cancer

Cancer remains a leading cause of death, accounting for about 7.5 million deaths in 2008 which increased to 10 million deaths of which 606,520 deaths occurred in the U.S. in 2020 (Siegel et al., 2020; Sung et al., 2021). Despite the continued overall decline in cancer death rates in the U.S, death rates from some cancers are increasing, including the cancer of the pancreas, liver, kidney, oral cavity, and pharynx during 2013-2017 (Henley et al., 2020). Incidence rates for the cancers of the colorectum, esophagus, leukemia, myeloma, Non-Hodgkin Lymphoma (NHL), stomach, thyroid, cervix, and prostate remained the same compared to previous years (Siegel et al., 2020). Increased and stable rates of these cancers warrant further research in order to understand and reduce risk factors. Many people’s lives can be saved and live longer with their loved ones if cancers can be prevented, detected early, or treated effectively.
1.1.1 What Causes Cancer?

A better understanding of cancer causes is the basis of effective cancer prevention programs. Known cancer risk factors include (1) genetic mutations; (2) ionizing & non-ionizing, and UV radiation; (3) natural hormones (estrogen, androgen) and agonists; (4) carcinogenic chemical exposures, such as tobacco smoking, excessive alcohol, and red meat consumptions; (5) certain viruses including HPV and hepatitis C; (6) environmental chemical exposures in the ambient air, water, food, residential and workplace (Hussain et al., 2003). Mutagenic carcinogens directly damage DNA by inducing mutations in somatic and germline cells (Lawley, 1989). Ionizing radiation is the best-studied example of a mutagenic carcinogen (Ron, 1998). The non-mutagenic carcinogens act through indirect mechanisms, such as generating reactive oxygen species (ROS) and gene induction mechanisms (Furman et al., 2019). Some chemicals are endocrine disruptors, causing cancer via activation of steroid receptors (Soto & Sonnenschein, 2010). Many chemicals induce epigenetic changes that alter the risk of many diseases, including cancer (Pogribny & Rusyn, 2013).

Genetic factors: People may either inherit or acquire diseases such as altered genes in the body's cells, aberrant hormone levels in the circulation, or a compromised immune system, to name a few examples. Each of these variables has the potential to increase an individual's susceptibility to cancer (Bartsch & Hietanen, 1996). Genetics plays an essential role in carcinogenesis because it determines whether the cancer-initiating event will proceed to complete malignancy (Sugimura et al., 1992) or if cancer will be curable (Bild et al., 2006). Some individuals are more prone than others to acquire cancer as people vary in their capacity to remove cancer-causing chemicals from their bodies after being exposed or in their ability to repair DNA damage produced by such agents (Yard et al., 2016). Many genes have been discovered that significantly enhance a person's risk of getting various malignancies (such as colon, breast, and ovarian cancer) throughout their lifetime (Sugimura et al., 1992). Certain gene changes may also be inherited across generations, and this may explain why cancer
rates are greater in these families. Increased incidence of cancer in families may also be linked to common environmental exposures among family members (Borch-Johnsen et al., 1994). However, familial malignancies, which are cancers caused by these genes, account for just 1% (many sites) to 11% (prostate) percent of all cancers depending on the anatomical sites (Hemminki & Vaittinen, 1999). In addition, genetic variations may determine the sensitivity to environmental carcinogens among individuals through gene-environmental interactions (Friedenson, 2011).

**Lifestyle and Inflammation:** The fact that the most common cancer types vary between men and women suggests that hormones and health habits play a significant role in carcinogenesis (Öner et al., 2016). The five main behavioral and nutritional hazards are high body mass index, poor fruit and vegetable consumption, lack of physical exercise, cigarette use, and alcohol use, which account for around 30% of cancer fatalities (Ruiz-Núñez et al., 2013). Fruits and vegetables provide an enriched supply of vitamins, minerals, and antioxidants. Many epidemiological studies demonstrated the protective effect of fruits and vegetables from chronic diseases, including cardiovascular disease (CVD) and cancer (Block et al., 1992). The generation of reactive oxygen species (ROS) is a key mechanism of lifestyle-induced inflammation (Ruiz-Núñez et al., 2013). Also, ROS free radicals enhance the risk of cancer by causing mutations in DNA by indirect DNA damage, particularly in the presence of persistent free radicals as in chronic inflammatory conditions (Lonkar & Dedon, 2011). Chronic inflammation interacts synergistically with environmental and genetic factors to influence health outcomes, autoimmune diseases, certain cancers, neurodegenerative diseases, and CVD (Furman et al., 2019).

**Environmental pollutions:** Environmental and lifestyle exposures are strongly linked to the majority of the world's most prevalent cancers (IARC, 2015). The IARC press release in 2015 stated “The Panel was especially disturbed to learn that the actual burden of cancer caused by the environment has been severely underestimated. Exposure to possible environmental carcinogens is ubiquitous in the
United States, with over 80,000 chemicals on the market, many of which are used by millions of Americans in their everyday lives and are un- or understudied and generally uncontrolled”.

The environmental etiology of cancer is demonstrated by the longitudinal studies following up the people who immigrated from Asian countries to the USA and Australia over multiple generations (Parkin & Khlat, 1996). Asians immigrate to the USA and Australia from countries with low rates of prostate and breast cancer but high rates of stomach cancer, their prostate, and breast cancer rates gradually increase to equal or greater than the rates of these cancers in the United States. Likewise, their stomach cancer rates decline to as low as those in the United States. This phenomenon was also corroborated by the studies based on the Swedish Family-Cancer Database, one of the largest longitudinal data sources following up the immigrant over two generations (Hemminki et al., 2014).

Specific types of cancer have a more established etiology of environmental exposures than others. For example, the proven relationships between asbestos and lung cancer (Doll, 1955), benzidine and 4,4'-methylenebis, chemicals in some dyes, and bladder cancer (Letašiová et al., 2012). Smoking has been related to multiple cancers (lung, digestive organs, bladder, kidney, oral, cervical, and pancreatic cancer). While smoking is strongly associated with lung cancer, the proportion of lung cancer among never smokers have been explained by environmental risk factors including radon, second-hand tobacco smoke, and other indoor air pollutants (Samet et al., 2009). Many of the environmental hazards linked with cancer have been discovered via occupational exposure studies based on the workers that are exposed to the chemicals at greater levels than the general population. Therefore, general population-based studies are important as they provide more generalizable evidence in identifying the cancer risk factors in non-occupational settings.

The interplay of multiple risk factors: Although we know that certain genetic and environmental factors increase the risk of developing cancer, it is not possible to predict whether a specific environmental exposure will cause a particular person to develop cancer with adequate
certainty because of the complex interplay of many factors (Hankinson et al., 2004). This leads to a lack of consensus in terms of how much is the contribution of environmental pollution to cause cancer. According to Peto and Doll’s quantitative estimates of avoidable risks of cancer in the United States published in 1981, only 4% of cancer deaths in the U.S were attributable to occupational exposures (Doll & Peto, 1981). With the advent of cancer research methodologies and improved understanding of the involvement of environmental exposures into the pathways of carcinogenesis, The President’s Cancer Panel released a report in 2010 emphasizing that the majority of cancer is caused by environmental exposures, thus they consider the most effective method of preventing cancer is to limit one’s exposure to hazardous environmental pollutions (Christiani, 2011). The chance that an individual will develop cancer in response to a particular environmental agent depends on several interacting factors including the cumulative exposure dose (duration, frequency, and the amount) combined with genetic factors, diet, lifestyle, health, age, and gender (Hankinson et al., 2004). Furthermore, diet, alcohol consumption, and certain medications are known to affect the chemicals’ metabolic fates that break down cancer-causing substances (Guengerich, 2000). Therefore, research to determine the environmental causes of cancer requires various approaches to help disentangle the complex interplay of multiple potentially causal factors.

1.1.2 Cancer Treatment: Challenges, and Opportunities

**Conventional therapies:** Since the early 19th century, cancer management has evolved from just removing the entire cancerous tissues and organs and giving chemotherapy-based adjuvant therapies in the 1970s using cytostatic drugs and radiation (DeVita & Chu, 2008). The chemotherapies vary by the chemical structures and mechanisms of action to interfere with cell divisions. Some are alkylating agents (directly act on DNA, such as *cyclophosphamide*), antimitabolites (alter enzymes required for normal cell metabolism, such as *purine antagonists*), plant alkaloids (block cancer cell division, such as
doxorubicin), or antitumor antibiotics (prevent RNA synthesis, such as mitoxantrone) (Schirrmacher, 2019). Even though adjuvant chemotherapy had dramatically increased survival for cancer patients, it often causes severe systemic toxicity due to its lack of tumor specificity as it non-selectively destroys all cells undergoing cell divisions (Schirrmacher, 2019).

Over the last half-century, advances in knowledge of cancer pathways and the role of immunology in carcinogenesis have led to the discovery of small-molecule inhibitors (SMIs) which target KIT oncogenic signal transduction pathways (i.e., sunitinib) and different types of monoclonal antibodies (MAbs) that target epidermal growth factor receptors (i.e., cetuximab), tumor-associated angiogenesis (i.e., bevacizumab) and cytotoxic T-lymphocyte associated protein 4 (CTLA-4) to support cytotoxic T-cell expansion (i.e., ipilimumab) (Zugazagoitia et al., 2016). Among diverse approaches, checkpoint inhibitory MAbs have yielded remarkable success in childhood malignancies, lung, breast, prostate, and colorectal cancers and chimeric antigen receptor (CAR) transfected T-cells shown to be more effective in hematological malignancies (Schirrmacher, 2019).

Immunotherapy: Immunotherapy has become one of four pillars in cancer therapy as a complemental treatment along with chemotherapy, radiotherapy, and surgery. This has improved the survival and quality of life for cancer patients. The underlying concept of immunotherapy is to modulate the tumor pathways to enhance the immune system’s ability to attack and eventually eliminate cancer. The most widely used form of anticancer immunotherapy, immune-checkpoint inhibitors (ICIs), does not involve attacking cancer cells directly but promotes immune responses to tumor cells (Zugazagoitia et al., 2016). The success of ICIs, demonstrated in melanoma, has propelled the research and introduced numerous different ICIs for the treatment of head and neck squamous cell carcinoma, renal cell carcinoma, non-small-cell lung cancer, and some other cancer types (Islami et al., 2020). Preclinical studies have disclosed diverse immunogenic pathways in tumorigenesis, specifically, ICIs, metabolic checkpoint inhibitors, oncolytic viruses, tumor microenvironment modulators, host’s microbiome
modulators, and adoptive cell therapies. These advances hold the promise for a longer and better quality of life for people living with cancers.

**Immune checkpoint inhibitors (ICIs):** Improved knowledge of the checkpoint proteins on T cells or cancer cells surfaces including PD-1/PD-L1 and CTLA-4/B7-1/B7-2 T led to the discovery of drugs that block these proteins, enabling T cells to more efficiently identify and destroy cancer cells (Schirrmacher, 2018). Specifically, the 2018 Nobel Prize was awarded for the discovery of monoclonal antibodies against the inhibitory immune checkpoints, CTLA-4 and PD-1, to trigger T-cells to more potently react to cancer cells (Zang, 2018). Unlike its success with the high objective response rate shown in melanoma, Hodgkin’s lymphoma, skin squamous cell carcinoma, and Merkel cell carcinoma, monotherapy of PD-1 inhibitors did not show appreciable benefit in some cancers, including pancreatic cancer and microsatellite-stable colonic adenocarcinoma (Esfahani et al., 2020).

**Enhancing the efficacy and reducing the toxicity of ICIs:** In pursuit of improved cancer outcomes and higher efficacy, various combination strategies have been evaluated including combining different ICIs or combining ICIs with new or conventional systemic chemotherapies (Kotecha et al., 2019; Le Tourneau et al., 2019). The combined therapy of CTLA-4 and PD-1 blockade showed a significantly higher response rate and improved survival benefit as opposed to PD-1 blockade alone in metastatic melanoma (Larkin et al., 2019) and metastatic renal cell carcinoma (Motzer et al., 2018, 2019). However, the effectiveness of combination strategies (Perez-Ruiz et al., 2019) is restricted by the amplified immune-related toxicity (IRAEs) (Hodi et al., 2018; Moslehi et al., 2018), excess fatality (Wang et al., 2018), and inflammation against the host’s healthy tissues resulting in de novo diabetes, pituitary dysfunction, and arthropathies. Hence, the focus of current research is to identify potent yet less toxic combinations, immunosuppressive agents to be used along with ICIs without hampering the efficacy, and the integration of tumor vaccines, tumor microenvironment modulators, and the host’s microbiome modulators into ICI therapies.
Another intensive research area is to understand the differential response rate of different tumors to the same therapeutic approaches and to transform the ‘cold’ tumors into ‘hot’ variants to boost the treatment outcomes. ‘Hot’ tumors show a good response to ICIs and are characterized by their abundance of tumor-infiltrating lymphocytes (TILs) in the tumor microenvironment, expression of anti–PD-L1 on tumor-associated immune cells, and genomic instability (Hegde et al., 2016) as opposed to the “cold” tumors. Radiotherapy and some chemotherapy had shown to be effective in converting the “cold” tumors into “hot” tumors. Whether tumor vaccines and tumor microenvironment modulators would enhance tumor responses to the ICIs is under active investigation (Golden & Apetoh, 2015; McGranahan et al., 2016).

**Precision medicine and personalized medicine:** In precision medicine, a patient receives a treatment that targets abnormal genes or molecular signaling pathways specific to that tumor. Some other cancers for which precision medicine is available to include breast, colorectal, stomach, and non-small-cell lung cancer, melanoma, and some other types of leukemia (Haslam & Prasad, 2019). Personalized neoantigen-based tumor-specific vaccines to sensitize the immune system against specific tumor antigens before the ICI and/or CTLA-4 treatment had shown promising results (Esfahani et al., 2020). The term *precision medicine* was first coined in 2015 and its goal was defined by the US National Research Council, as follows: “to identify a subset of patients, with a common biological basis of disease, who are most likely to benefit from a drug or other treatment and experience fewer side effects, thus reduced costs are likely to ensue. In such studies, compliance will likely be better, treatment duration longer and therapeutic benefits more obvious than is the case with traditional designs. Greater therapeutic differences could also result in more efficient regulatory approval and faster adoption by physicians and payers” (Ashley, 2016).

Identifying a subset of patients with a higher likelihood to benefit from certain drugs requires advanced biomarkers (Haslam & Prasad, 2019). In the context of ICIs, standard biomarker tests including
assay harmonization and standardization are used along with FDA-approved biomarkers: PD-L1 expression and Mismatch Repair Deficiency (MMR) (Emens et al., 2017). These markers are limited to tumor genetic characteristics but not indicative of other key tumor growth factors such as the tumor microenvironment, thus, emerging biomarkers of Tumor Mutation Burden (TMB) and Tumor-Infiltrating Lymphocytes (TILs) hold a high promise for a higher predictive power than existing biomarkers (Arora 2019).

Personalized medicine was defined by the President’s Council on Advisors on Science and Technology (PCAST) as “the tailoring of medical treatment to the individual characteristics of each patient. Preventative or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not” (PCAST, 2008). An example of personalized medicine is adoptive cell therapies such as autologous Chimeric Antigen Receptor CAR T-cell products. In this approach, peripheral T cells are collected from patients to genetically alter them to express CARs on their surface and infuse the resulting CAR-T cells back into patients to attack and destroy their tumor cells (June & Sadelain, 2018). Unlike autologous CAR T-cells, allogenic CAR T-cell treatments intend to use CAR T-cells derived from different individuals to a given patient and are currently under active investigation (Depil et al., 2020).

Tumor microenvironment (TME): The tumor microenvironment is the dynamic, physical, biological, and chemical barrier that shields the tumor cells from immune surveillance. On the other hand, interfering with the tumor microenvironment to allow immune cells arrive to at the tumor cells, enables tumor-infiltrating lymphocytes to enter the tumor bed and trigger the immune response. To achieve higher efficacy in ICI treatments, multiple tumor microenvironment modulators have been combined with ICI treatment oncolytic viruses, metabolic checkpoints, antitumor vaccines (Murciano-Goroff et al., 2020).
**Oncolytic viruses:** Oncolytic viruses selectively infect or replicate in tumor cells (Kaufman et al., 2015) to eventually destroy tumor cells by releasing cytokines and viral pathogen-associated molecular patterns (PAMPs) that enhance CD8+ T cell activation, and trigger NK cell-mediated innate immune responses (Kaufman et al., 2015). In addition, oncolytic viruses modulate TME by distorting the mesenchymal and collagen barriers to hinder tumor-infiltrating lymphocytes to enter the tumor bed as well as interfere with oncogenic kinase signaling to promote PD-L1’s expression (Ros & Vermeulen, 2018; Spranger et al., 2015). The first FDA-approved oncolytic virus, talimogene laherparepvec (T-VEC), a modified herpes simplex virus demonstrated robust T cell responses and multiple microenvironmental effects even beyond its injection site (Ott & Hodi, 2016). Clinical trials on T-VEC combinations with checkpoint inhibitors in melanoma indicated that this combination was more effective than ICI monotherapies (Chesney et al., 2018; Ribas et al., 2017; Sun et al., 2018).

**Metabolic checkpoints:** Tumor microenvironment attenuates T-cells antitumor capacity by supporting tumor metabolism and supplying the energy required for the accelerated divisions of tumor cells through increased glycolysis, uptake of L-arginine, tryptophan, and glutamine (Kouidhi et al., 2017; Martinez-Outschoorn et al., 2017). The state of increased energy demand exhausts the nutrients necessary for T-cell activation and TIL’s infiltration to tumor bed (Qorraj et al., 2017). Multiple therapeutic strategies to disturb tumor supportive microenvironment are currently being evaluated in early-stage clinical trials including attempts to interfere with glycolysis in cancer cells, thereby dropping lactate levels in the tumor microenvironment. For example, combining ICIs with rapamycin (Esfahani et al., 2019; Mineharu et al., 2014) or metformin (Ma et al., 2017; Rena et al., 2013) has been shown to improve objective response rate and survival for melanoma patients (Afzal et al., 2018).

**Microbiome:** The composition of the gut microbiome plays important role in immune responses and strongly influences the efficacy and toxicity of immunotherapies: anti PD1/PDL-1, and CTLA-4 antibodies (Yi et al., 2018). The immunomodulatory impact of the microbiome has been demonstrated
by key findings, as follows: (1) resistance to ICIs was consistently associated with antibiotic-induced dysbiosis in renal cancer, non-small-cell lung cancer, urothelial cancer, and melanoma (Routy et al., 2018); (2) specific subset of human microbiota were predictive of clinical response to ICI treatment (Matson et al., 2018), such as metastatic melanoma responders to anti–PD-1 therapy had enriched *Bifidobacterium longum, Collinsella aerofaciens,* and *Enterococcus faecium* (Matson et al., 2018), while, anti–PD-1 or PD-L1 therapy responders had enriched *A. muciniphila, Clostridiales, E. faecium,* *Eubacterium,* the *Firmicutes,* and *Ruminococcus* species (Derosa et al., 2018). The protective effect of a healthy microbiome from immune-related adverse events (IRAEs) has been evident in the reduced colitis after CTLA-4 blockade treatment in melanoma (Chaput et al., 2017). Current efforts to identify the optimal probiotic and immunotherapy combinations for enhanced efficacy and reduced toxicity of ICIs are underway (Gurbatri et al., 2020). Taken together, the future of cancer therapy lies in the optimal combination of ICIs with personalized medicine, including antitumor vaccines and other therapeutic approaches targeting the tumor microenvironment, tumor glycosylation, and microbiome (Emens et al., 2017; Esfahani et al., 2020; Islami et al., 2020).

**Cellular immunotherapy:** Besides immune checkpoint blockade, the cellular immunotherapy approach, also known as adoptive cell therapies, has yielded remarkable success, especially in hematological malignancies. This approach leverages killer T cells’ natural ability to bind to tumor surface antigens and eventually destroy them (June & Sadelain, 2018). While multiple approaches are under investigation, tumor-infiltrating lymphocyte (TIL) therapy, engineered T-cell receptor (TCR) therapy, chimeric antigen receptor (CAR) T-cell therapy, and natural killer (NK) cell therapy are better understood (Hayes, 2021). To date, three CAR T-cell products: axicabtagene ciloleucel (axi-cel; Yescarta), tisagenlecleucel (Kymriah), and lisocabtagene maraleucel (liso-cel; Breyanzi) have been approved by the regulatory agencies. As the CAR T-cell therapy is the most clinically advanced, all these three approved therapies have been recently incorporated into National Comprehensive Cancer Network (NCCN)
guidelines recommended for the third-line therapy for B-cell lymphoma, with the most recent addition being lisocabtagene maraleucel (JNCCN, 2020).

As reviewed in June and Sadelain 2018, CARs are transfected into T cells through viral vectors, plasmid transfection, or mRNA. Consequently, the peripheral T-cells acquire improved recognition of tumor cells with help of CARs expressed on their surfaces. Over time, CAR constructs evolved from the first-generation, consisting of only the CD3ζ signaling domain, to second-generation through additions of costimulatory endodomains, such as CD3ζ plus 41BB or CD28 signaling domains, and third-generation which consist of CD3ζ plus both 41BB and CD28 signaling domains. Advanced generation of T cells possesses better persistence and proliferation for sustained antitumor potency compared to their earlier prototypes (June & Sadelain, 2018).

The remarkable success that CAR T-cell therapies have shown in hematological malignancies has not translated into solid tumors (Marofi et al., 2021). The cause of this differential success has been described by the highly heterogeneous neoantigen expressions in solid tumors which are further complicated by hostile tumor microenvironment of oxidative stress, nutritional depletion, acidic pH, hypoxia, immunosuppressive factors including regulatory T cells, myeloid-derived suppressor cells, tumor-associated macrophages or neutrophils, and upregulation of cytoplasmic and surface inhibitory receptors (Newick et al., 2017). The underlying concept of CAR T-cell therapy to genetically manipulate the patients’ CAR T-cells provides a limitless opportunity to maneuver the killer T-cell’s function to accomplish the needed immune response.

In summary, although immunotherapy has demonstrated high efficacy and often manageable toxicity, response to these treatments is limited to a small fraction of patients, which is estimated to be only 8% of metastatic patients in the U.S (Haslam & Prasad, 2019; Marquart et al., 2018). In addition, precision medicine is costly as the treatment choice requires expensive tests for the characterization of
individual tumor profiles and genetic mutations in addition to their prohibitively high production costs (Cutler, 2020).

Regardless of the recent advances in immunotherapy, precision medicine, and biomarker research, unmet treatment needs persist for patients with late-stage cancers, and diseases that are non-responsive to these novel approaches such as pancreatic, liver, esophagus, stomach, and lung malignancies. The proportion of patients with no treatment options accounts for 21% of all cancer diagnoses and 41% of all cancer deaths in the USA (Jemal et al., 2017). In addition, relapsed and refractory hematological cancers at present have poor prognoses after multiple lines of the standard of care and even CAR T-cell therapy (Shah & Fry, 2019).

1.2 Burden of COVID-19 Pandemic

COVID-19, the disease caused by the SARS-CoV-2 coronavirus has spread worldwide since its first identification in December 2019 in Wuhan, China (Khan et al., 2021). WHO classified COVID-19 as a pandemic in March 2020 as the containment of the disease was hindered by the lack of antiviral treatment, lack of vaccines, and the existence of asymptomatic carriers (WHO, 2021). As of January 2021, the number of patients diagnosed with COVID-19 has exceeded 98 million, and more than 2.1 million patients have died (WHO 2021). As of April 2021, the world case fatality rate (CFR) was 2.15% (Hamed et al., 2021). Of all continents, the highest number of COVID-19 cases was reported from North America while most deaths were in Europe and South America had the highest CFR (Hamed et al., 2021). Unprecedented high health care demand of COVID-19 due to its incidence and mortality surges in a short time overwhelmed hospitals, strained resources, and stretched the workforce (Miller et al., 2020). The pandemic hit healthcare workers hard causing burnout and stress as the outbreaks emerged in waves covering from urban to rural areas of the countries (Miller et al., 2020). To reduce the healthcare
burden and economic damage through more efficient use of limited resources, epidemiologists and clinical researchers attempted to identify the people with the most urgent needs.

### 1.2.1 Airborne transmission of viral agents and COVID-19

Individuals are exposed to environmental contaminants and biological agents through three main routes: inhalation, ingestion, and skin contact. SARS-CoV-2 transmits from human to human through three principal ways: (1) Inhalation of air carrying droplets and aerosol particles that contain the infectious virus, (2) Deposition of virus carried in exhaled droplets and particles onto exposed mucous membranes, (3) Touching mucous membranes with hands soiled by exhaled respiratory fluids containing a virus or from touching inanimate surfaces contaminated with the virus. The risk of infection increases with proximity to an infectious source. WHO advises a physical distance of at least 1 meter between people, while the Centers for Disease Control and Prevention (CDC) recommends a physical distance of at least 1.8 meters between people (CDC, 2021).

### 1.2.2 Determinants of COVID-19 Incidence and Mortality

The COVID-19 incidence and mortality exhibited large inequities in terms of socioeconomic, demographic, clinical risk factors. COVID-19’s incidence, severity, and death risk increased with male gender, people aged over 70, low income, low education level, and being immigrants from low- or middle-income countries (Drefahl et al., 2020; Karmakar et al., 2021; Pijls et al., 2021). Studies in the U.S corroborated these findings with higher COVID-19 incidence and death rates in the counties with higher percentages of people identifying as racial and ethnic minorities in addition to specific factors, such as crowded housing and single-parent households (Karmakar et al., 2021). High COVID-19 incidence and fatality rates in lower socioeconomic areas were consistently associated with inadequate testing and lower adherence to social distancing policies compared with more affluent counterparts (Mena et al.,
Consequently, people experienced longer testing delays and a lack of access to health care to contain the spread of the epidemic in poor neighborhoods (Mena et al., 2021). To address inequities in the burden of the COVID-19 pandemic, more precise identification of COVID-19 vulnerable locales is needed, and their root causes must be addressed.

1.3 Data sources and Methodologies

A typical research question in epidemiology is whether human exposure such as airborne chemical pollutions or infectious disease agents has an impact on the incidence or mortality of diseases. Even though the randomized controlled trial is considered the gold standard that eliminates the undesirable effect of the confounding factors, randomizing people to different exposure groups is not feasible in epidemiologic research. Therefore, in epidemiological studies, common measures to prevent severe biases include the use of different study designs, such as cohort or case-control or statistical methods to account for the confounding factors using multiple regression, stratification, matching, or matching by propensity score models (Greenland & Morgenstern, 1989).

The choice of regression models depends on the measurement scale (categorical vs continuous) of the response variable (health outcome). Linear regression is compatible with model continuous outcomes, logistic regression for binary outcomes, Cox regression for time-to-event data, and Poisson or negative binomial regressions to model count data and rates. To model more complex distributions of outcome and response variables, we further extend basic regression models. Comparative Effectiveness Research (CER) is established on the epidemiological concept and lends epidemiological methodologies. As Sox and Greenfield defined “CER aims to compare the benefits and harms of different approaches to prevent, diagnose, treat and monitor diseases, or to improve the means of healthcare delivery” (Sox & Greenfield, 2009). Although various well-established epidemiological approaches are available, the viability of health research is largely influenced by data availability.
1.3.1 Data sources in epidemiology and comparative effectiveness studies

Secondary data, such as administrative medical databases, disease registries, claims data from Medicare and Medicaid systems, vital statistics records, and published clinical trial data are becoming increasingly utilized in epidemiological and comparative effectiveness studies. Although a wealth of secondary data is available, health researchers face significant gaps between their research needs and data availability. Particularly, individual patient data (IPD) is hard to obtain due to the protection of patient confidentiality. The data access gap is particularly prominent for rare diseases, such as pancreatic cancer, and novel treatment data such as CAR T-cell therapies. Furthermore, when disease occurrence is extremely rare, for example, as few as 5 counts per year given a small geographical area, the likelihood of identifiability of individuals increases. In new treatments, there is a significant lag time from data collection until data becomes available due to enough time needed for sufficient data to be accumulated. This delay in comparing novel medicines to existing treatments limits the utility of comparative effectiveness studies and the significance of epidemiological studies in helping policymakers, physicians, payers, and patients make better decisions.

To address the data gaps, researchers use a variety of strategies. For example, deduplication techniques reduce the risk to count the same individual multiple times and the Matching Adjusted Indirect Comparison technique allows to use of aggregate published data for the comparison arm when individual patient data is available only for the intervention arm.

The quality of CER depends on the appropriateness of the study design and the availability of individual patient data. Ideally, randomized clinical trials (RCTs) are preferred as the gold standard for examining causal relationships, despite their high cost and long time commitment. Likewise, individual patient data (IPD) from both trials that had originally evaluated the therapies are desired as that would allow estimating the full joint distribution of both comparator studies (Phillippo et al., 2018). But frequently in practice, particularly in oncology which has significant unmet medical demand for cancer
patients, single-arm studies are used to fulfill the requirement for quicker regulatory review and decision making. As for data granularity, many researchers have IPD for trials of one treatment, for example, as the principal investigators of a new drug, but only aggregate data for trials of comparator treatments (Signorovitch et al., 2012). Recently developed, Matching Adjusted Indirect Comparison (MAIC) bridges this data gap by modifying average patient characteristics in trials with IPD to match those reported in studies without IPD.

1.3.2 Methodologies

Regression modeling is a universal tool for data analysis that can be applied in all study designs. The overarching goal of regression modeling is to estimate the effects of explanatory variables (i.e., environmental exposures, socioeconomic risk factors, or medical interventions) on a response variable such as disease incidence or survival outcomes regardless of the type of disease. However, our choice for specific regression modeling techniques depends on the measurement scale of the response/dependent variable and the study design. Commonly, we use linear regression for continuous outcomes, logistic regression for binary outcomes, Cox regression for time-to-event data, and Poisson regression for frequencies and rates. In this section, I discuss the basics of most common regression models and briefly how these models are extended to suit to test specific research questions and three different study designs: ecological and spatial study designs as well as the indirect comparison of medical interventions.

1.3.2.1 Global regression models

Linear Regressions: Accuracy of linear regression increase if the true association between response and explanatory variables: Y and X is linear, or the residual term is normally distributed with mean 0 and variance $\sigma^2$. However, the requirement for a normally distributed residual term loosens
when the sample size is large enough. In practice, pure linear regression is not commonly used in environmental health research as the underlying linearity assumption doesn’t adequately reflect the often non-normal distribution of environmental exposures. Multiple linear regressions allow accounting for the effects of confounders. The linear regressions compute the parameter estimates (effect of an explanatory variable on the response variable) through ordinary least squares (OLS) or linear least squares. This method minimizes the sum of squared vertical distances between the observed and predicted observations in the linear approximation (Draper & Smith, 1998).

**Logistic regression:** Multiple logistic regression is extensively used in cancer epidemiology as response variables such as the development of cancer (yes/no) or cancer mortality (0,1) is often expressed as binary variables and it applies to both cohort and case-control studies (Anderson, 1972; Mantel, 1973). Logistic regression estimates the strength of the assessed association by maximum likelihood and the Hosmer-Lemeshow test tests its predictive performance (Hosmer et al., 1997).

**Odds Ratios (ORs):** In logistic regression output, the OR indicates the probability of an event happening in the presence of a specific exposure, as opposed to the probability of the result occurring in the absence of the exposure. Although odds ratios are often employed in case-control studies, they can also be used in cross-sectional and cohort research with modified interpretations (Szumilas, 2010). If the assessed exposure-outcome relationship depends on the other explanatory variables, it is impossible to describe the effect of a variable by a single value of OR. In this case, we use stratification and compute strata-specific ORs. When a continuous explanatory variable is concerned OR describes the factor by which the odds of an event change for each one-unit increase of explanatory variable. In studies concerned with rare events, ORs can be interpreted as risk ratios because the corresponding numbers are approximately identical (Bender, 2009).

**COX regression:** COX regression model is used to assess the effect of several continuous explanatory variables $X_1, \ldots, X_k$ on the response variable $T$, in relation to the survival time. Unlike linear
and logistic regressions, the Cox model estimates the hazard function rather than the mean parameter estimates depending on the explanatory variables (Cox, 1972). In a Cox proportional hazards (PH) regression model, the measure of effect is the hazard rate which represents the likelihood of the event, considering the person has survived up to a specific time point (or time until an event occurs). The distribution of the time-to-event as a function of survival distribution function (S) can be expressed as $S(t) = P(T \geq t)$ (Cox, 1972). The impact of a categorical explanatory variable on the response variable can be assessed by Kaplan-Meier curves for each group (Kaplan & Meier, 1958). Cox PH model assumes that the ratio of hazard of two persons is constant over time. However, the issue of nonproportionality can be reasonably resolved by stratified analysis or the inclusion of time-dependent covariates in the model. Also, alternative models developed to address the violation of proportional hazard assumption are available including Restricted Mean Survival Time (Irwin, 1949) and Accelerated Failure Time Model (Park & Wei, 2003).

**Poisson and Negative binomial regressions:** Poisson regression is commonly used in health research as it fits the discrete response variable Y using the common measurement scales of counts, frequencies of disease, or rates. As reviewed by Bender in 2009, Discrete data is considered to have a Poisson distribution if the average of the response variable is the same as its variance. However, the stringent assumption that the variance is equal to the mean made by the Poisson model is frequently violated due to heterogeneity or overdispersion of data. The negative binomial is a generalization of Poisson regression as it has the same mean structure as Poisson regression. Negative binomial regression (NBR) uses an extra parameter to model the data over-dispersion. Hence, NBR estimates are more accurate for overdispersed data than Poisson regression, thereby resulting in tighter confidence intervals as compared to those from a Poisson regression model (Bender, 2009).
1.3.2.2 Model extensions

**Geographically Weighted Regression (GWR):** Global models assume that geographic location does not influence the relationship between response and independent variables. This assumption is also referred to as spatial stationarity. In perfect spatial stationarity, data has a constant mean and variance in each location. Thus, the models with local spatial effects were developed for a better understanding of the spatial process. Global models compute mean values over space and resulting averages obscure underlying geographic processes if existed. The technique, known as Geographically Weighted Regression (GWR) extends global linear regression to consider spatial variations of relationships by considering all other observations as a function of the distance to that point. GWR accomplishes this task by dividing the space into stationary subregions and models the spatial tendency in a continuous way, with parameters varying in space (Brunsdon et al., 1996). For the computation of local average, GWR uses two types of kernel density estimation. The first kernel function assumes that the bandwidth at each regression point is a fixed constant across the study area while the second kernel function assumes bandwidth varies at each regression point. The bandwidth is the radius of that buffer around each regression point. Mapping out resulting the regression estimates based on both constant and varying kernel densities present the local variations in the parameter estimate (Brunsdon et al., 1996).

**Geographically Weighted Negative Binomial Regression (GWNBR):** The use of classic GWR is limited to Gaussian data distribution and is not appropriate when the response variable represents a count data. In addition, GWR is prone to spatial error autocorrelation (Leung et al., 2000), multicollinearity (Wheeler & Tiefelsdorf, 2005), and extreme coefficients including sign reversal (Farber & Páez, 2007). To address these limitations of GWR, Geographically Weighted Poisson Regression (GWPR) and Geographically Weighted Negative Binomial Regression (GWNBR) were developed as Poisson and negative binomial are the most adequate distributions for the modeling of count data.
(Nakaya et al., 2005). However, GWNBR is proven to be more robust than GWPR given that the assumption of equality between mean and variance in Poisson distribution is relaxed in GWNBR by using an additional parameter to account for data overdispersion. GWNBR is becoming increasingly popular in health research because of its ability to model count data, particularly when the data exhibit non-stationarity and overdispersion, while also incorporating the capabilities of other models, such as Poisson regression, Generalized Weighted Poisson Regression, and Negative Binomial Regression.

**Matching Adjusted Indirect Comparison:** MAIC extends conventional crude direct standardization to cope with the absence of IPD on one study and only aggregate data available on the other to allow indirect comparisons of the treatment effects across the trials. At its core, the MAIC method utilizes the individual patient data (IPD) from the trial of intervention (usually a manufacturer's own product) and re-weights the patients with IPD such that their characteristics are balanced with those of the patients from the trial of a comparator intervention (Signorovitch et al., 2012). The MAIC is available in both anchored and unanchored variants, with the anchored method requiring a common comparator, such as the placebo group, in both comparator trials. The presence of imbalanced prognostic factors should not affect anchored MAIC estimations because of the common comparator (Phillippo et al., 2018). When a common comparator is unavailable and the results from the re-weighted IPD and the published aggregate data must be compared directly, the unanchored MAIC method is employed. Unanchored MAIC is appreciated for its utility to improve “unadjusted” or naive indirect comparisons by accounting for the differences in prognostic factor and effect modifier distributions between the two studies. Unanchored MAIC, on the other hand, is vulnerable to residual bias due to unobserved prognostic variables and effect modifiers because of the lack of a common comparator across trials. However, no research has been done on the degree of residual bias that might exist. Future studies will be required to determine the degree of error in unanchored estimates (Phillippo et al., 2018).
1.4 Specific Aims

My goal in this dissertation was to learn diverse analytical approaches through three distinctly different research projects. I aimed my learning endeavor would also contribute to alleviating public health burdens from COVID-19 and rapidly fatal cancers with vast unmet treatment needs. Research question specific data and methodologies were used to accomplish the following aims:

**Specific Aim 1: Examine the relationship between the hospitalization rate of Exocrine Pancreatic Cancer (EPC) and Exposure to Hazardous Organic Chemicals in New York State.**

We hypothesized that environmental exposure to persistent and volatile organic chemicals is positively associated with the incidence of exocrine pancreatic cancer. Also, we demonstrated the use of SPARCS data without personal identifying ID with help of a data deduplication technique that reduced the likelihood of counting the same person multiple times.

**Specific Aim 2: Identify COVID-19 vulnerable locales and examine the effects of epidemiological and socioeconomic risk factors on the spatial distribution of COVID-19 incidence.**

We hypothesized the area level epidemiological and socioeconomic risk factors such as area morbidity, risky health behaviors, crowding, population mobility, social distancing, healthcare access, and education would have a significant effect on the COVID-19 incidence rate, therefore, the spatial distribution of COVID-19 will be significantly heterogeneous in association with these epidemiological and socioeconomic factors. We also hypothesized the strength of the observed relationship will vary across the early, middle, and late phases of the pandemic during the study period.

**Specific Aim 3: Compare the efficacy and safety of novel experimental CAR T-Cell products for the treatment of relapsed/refractory (R-R), Diffuse Large B-Cell Lymphoma (DLBCL) to the FDA approved CAR T-cell, Yescarta.** We hypothesized that novel and experimental CAR T-Cell products are significantly better or worse than the currently approved CAR T-cell products, namely Yescarta, in terms
of efficacy and safety. We chose the first approved CAR T-cell product, Yescarta thereby harboring the longest follow-up data available to date.

This dissertation encompassed three different research projects that provide methodological ideas and epidemiological evidence to help fight the disease through public health and biomedical interventions.

1.5 References


2.1. Abstract

The etiology of exocrine pancreatic cancer (EPC) remains unknown except for family history and smoking. Despite recent medical advances, rates of pancreatic cancer incidence and mortality are increasing. Although existing evidence suggests a potentially causal relationship between environmental chemical exposures and pancreatic cancer, whether residential exposure impacts pancreatic cancer rates remains unknown.

We identified 28,941 patients diagnosed with exocrine pancreatic cancer in New York State exclusive of New York City for the years 1996–2013. Descriptive statistics and negative binomial regression were used in this ecological study to compare pancreatic cancer hospitalization rates among patients who lived in ZIP-codes with hazardous waste sites (HWSs) containing persistent organic pollutants (POPs) and volatile organic pollutants (VOCs) compared with Clean ZIP codes with no identified hazardous waste sites. We assessed the effect of selected known and suspected human carcinogens on the EPC hospitalization rates by subgroup analyses.

Compared with the Clean sites, the pancreatic cancer hospital discharge rate in the 'VOCs without POPs' and 'VOCs and POPs' sites, after adjustment for potential confounders were 1.06 [95% confidence interval (CI), 1.03–1.09] and 1.05 (95% CI, 1.01–1.08), respectively. In the analysis by specific chemicals, rate ratios for the benzene (1.12) and ethylbenzene (1.34) in the non-chlorinated VOCs group, trichloroethylene (1.07) and tetrachloroethylene (1.11) in the chlorinated VOCs group, chlorinated pesticides (1.11) and PCBs (1.05) in the POPs groups were statistically significant (p-values<0.05) compared with Clean sites.

Residential exposure to both volatile and semi-volatile organic compounds coming from identified HWSs is associated with an elevated risk of being hospitalized for exocrine pancreatic cancer. We attribute the exposure to inhalation. These results are important because while the exposures are
much lower than seen in occupational settings, residential exposure is continuous, and we have identified several specific chemicals showing significant associations.

2.2. Background

Exocrine pancreatic cancer (EPC) is a rapidly fatal malignancy with a 5-year survival rate of 10% (SEER, 2020) and is the third leading cause of cancer mortality in the USA (Gordon-Dseagu et al., 2018). According to the latest cancer statistics, overall cancer incidence in the U.S has declined during 2012-2016 except for the increased rate for the five cancers including pancreatic cancer (Henley et al., 2020a). The reason for the increasing trend is not understood. The same increasing trend for pancreatic cancer was seen in New York State (NYS), in which the 5-year average incidence rate (14.4%) was even higher than the national rate (12.7%) during the same period (NYSCR, 2020a). The etiology of EPC remains largely unknown. Less than 10% of all pancreatic cancers are estimated to be associated with genetic factors (Landi, 2009; Raimondi et al., 2010) and about 25% of new cases have been attributed to smoking (Schottenfeld & Jr, 2006; Silverman et al., 1994; Weiderpass et al., 1998). Specific medical conditions and lifestyle-related risk factors have been suggested (diabetes mellitus, chronic pancreatitis, obesity, poor diet, excess alcohol consumption, poor oral hygiene, infections of H. pylori, hepatitis B, and C) (Barone et al., 2016), but there is a lack of consistency across studies. Pancreatic cancer incidence has been reported higher among African Americans, males, and the elderly (Michaud, 2004; NYSCR, 2020b). Evidence for an effect of poverty on pancreatic cancer incidence is inconsistent, ranging from having a little effect (Boscoe et al., 2014) to being an important determinant dependent on race (Brotherton et al., 2016; Cervantes et al., 2019; Noel & Fiscella, 2019; Silverman et al., 2003).

2.2.1 Environmental Chemical Exposures and EPC

Several occupational, hospital-based case-control and agricultural studies have reported some evidence of an association between environmental chemical exposures and EPC. A large meta-
analysis (Ojajärvi et al., 2001) of 92 occupational studies and 23 occupational chemicals reported excess pancreatic cancer risk from occupational exposure to chlorinated hydrocarbons, nickel and chromium compounds.

However, occupational studies usually lack statistical power due to the small sample size, and are prone to the ‘healthy workers effect’, and exposure misclassifications (Blair et al., 2007). A few hospital-based case-control (Antwi et al., 2015; Fryzek et al., 1997) and self-reported exposure studies (Lo et al., 2007; Silverman, 2001) reported varying findings, but these types of studies are prone to information bias and misclassification in the control and case ascertainment processes (Althubaiti, 2016). The Agricultural Health Study Cohorts reported significantly elevated rate ratios for pancreatic cancer associated with exposure to the pesticides, S-ethyl-N,N-dipropylthiocarbamate and pendimethalin, but the results remain inconclusive given the potential exposure misclassification (Weichenthal Scott et al., 2010). Reasonable consistency has been seen in a few studies reporting a positive association between pancreatic cancer and exposures to chlorinated solvents, organochlorine pesticides, polychlorinated biphenyls (PCBs), and polyaromatic hydrocarbons (PAHs) (Alguacil et al., 2000; Beard John et al., 2003; Carpenter, 2015; de Basea et al., 2011; Hoppin et al., 2000; Porta et al., 1999).

The relationship between long-term residential exposure to organic chemicals and pancreatic cancer remains unknown. Moreover, there is an increasing need for scientific inquiry given the increasing incidence of this cancer, and the ubiquity of exposure to these chemicals. The situation is further complicated by the long latency period between exposure and development of cancer which necessitates an extended follow-up time and individual-level data.

We examined the association between residential exposure to different groups of organic chemicals and pancreatic cancer hospitalization rates among NYS residents using state-wide population-based hospital discharge data recorded over an 18-year period. We have used several datasets
maintained by NYS to explore associations, if any, between residence near hazardous wastes containing known contaminants and pancreatic cancer. The New York State Department of Environmental Conservation (NYSDEC) maintains a listing of hazardous wastes sites (HWS) that pose “a threat to public health”. The listing include the geographical location of the HWS along with the major hazardous chemicals contained in each site (NYSDEC, 2021).

2.2.2 The use of SPARCS vs NYSCR in population-based cancer research

For the statewide pancreatic cancer data, the New York State Cancer Registry (NYSCR) is the principal source of information on cancer incidence and mortality and provides residential address, sociodemographic, and tumor-specific data for all NYS residents diagnosed with cancer. However, accessibility to the registry is limited due to patient confidentiality concerns, and the publicly available format of the registry data for pancreatic cancer is only at the county level. As an alternative data source, the Statewide Planning and Research Cooperative System (SPARCS) can be used (SPARCS, 2020). SPARCS contains hospital discharge data that is mandated to be reported by all state-regulated hospitals to the New York State Department of Health (NYSDOH), so does not include Veterans Administration or Indian Health Services facilities. SPARCS data provides all diagnoses (up to 15), procedures, demographic information (age, sex, race, and ethnicity), and patient’s residential address upon hospital discharge. The publicly available SPARCS data used in this study included the 5-digit ZIP codes of the patients’ residential addresses, but not the patient’s name or street address. We have matched the hospitalization rate for pancreatic cancer by ZIP codes to the ZIP codes containing HWS with different contaminants.

The validity of using SPARCS data as hypothesis-generating information that is subsequently confirmed in studies with individual-level exposure assessment has been demonstrated for two different diseases. Kouznetsova et al. (2007) (Kouznetsova et al., 2007) and Huang et al. (2006) (X. Huang et al., 2006) reported significant elevations in hospitalization rates for diabetes mellitus and hypertension,
respectively, among individuals living in ZIP codes containing POPs waste sites. Most POPs sites contained primarily PCBs, but some also had chlorinated pesticides and dioxins or furans. Subsequent studies by Aminov et al. (2016) (Aminov Zafar et al., 2016) and Goncharov et al, 2008) (Goncharov et al., 2008) confirmed these associations with direct measurement of diabetes markers and blood pressure, respectively, in relation to serum PCB concentrations. A further test of the hypothesis that living near waste sites containing specific chemicals will result in an elevation in hospitalization for a disease known to be caused by that chemical comes from the study of Boberg et al. (2014) (Boberg et al., 2011).

Benzene is known to cause leukemia(Collins et al., 2003; Hayes et al., 2001; IARC, 2018). They found statistically significant increases in hospital discharge rates for chronic lymphocytic leukemia and lymphoma among people who lived in ZIP codes containing benzene waste sites (Boberg et al., 2011).

2.3 Materials and Methods

2.3.1 Study Design and Population

We conducted a population-based cross-sectional study. In the state-wide SPARCS data, we identified a total of 107,572 hospital discharge records of patients diagnosed with EPC between 1996 and 2013. EPC cases were identified using the international classification of diseases ICD-9-CM codes initialized with ‘157’ excluding the records of ‘1574’ for the malignancy of Langerhans islets (endocrine tumors).

The flowchart, Figure 1, shows the record selection process. Since New York City (NYC) operates an independent hospital discharge data system, we excluded 47,201 records of the NYC. There were 51,867 records after restricting data to the non-Hispanic white population (whites) and African Americans (blacks). The resulting data was deduplicated to 29,527 to achieve a record per patient. Details for the deduplication method and algorithms are described below. The records-data was linked to the U.S Decennial Census 2000 and Census 2010 datasets for the proxy of person-years corresponding with the hospital discharges in 1996-2004 and 2005-2013, respectively, in the SPARCS
data based on the 5-digit ZIP codes. A total of 571 records without matching ZIP codes to the Census ZIP codes were eliminated as these were post office box ZIP codes that did not reflect the site of residence. ZIP-code area-poverty level was assigned to each record based on the Summary Files of Census 2000 and 2010 to determine the percentage of people in the ZIP-codes below the federally defined poverty line. The area-poverty categories used were <5%, 5%-< 10%, 10%-< 20%, and>20% to create a categorical area-poverty indicator. There were 15 records with no poverty data and they were excluded, which led to 28,941 records in the final analytical dataset.

The SPARCS data version used in this study does not identify patients with multiple hospitalizations. To distinguish multiple hospitalizations of the same patient and select the earliest single hospitalization record for each patient, we deduplicated the multiple hospitalizations per patient data (Figure 1) by the deterministic linkage method based on the key variables, such as 5-digit ZIP-code of residential address, patient’s birth year, birth month, race, sex, and calendar year of hospital discharges. We used the NYSCR’s publicly available incidence case data for pancreatic cancer as a gold standard (reference) to assess the accuracy of the deduplication process. The combination of variables used for deduplication that approached closest to the NYSCR incidence data by calendar years, 5-year age groups, sex, and race was used to produce the final record per patient hospital discharge data.
2.3.2 Exposure assessment

The exposure variable was created based on the information whether the given ZIP code was listed in the NYSDEC database that contains a listing of all state HWS with known chemical pollutants. We used the list of active hazardous waste sites from the NYSDEC as of 1995 which did not include planned or complete sites. For the study of cancer, this older list is appropriate because of the latency for the development of the disease. We assume that the major route of exposure is the inhalation of volatile or semi-volatile chemicals. Each ZIP code area was categorized as containing Volatile Organic Compounds (VOCs), Persistent Organic Pollutants (POPs), Other, or Clean The NYSDEC data contains a listing of all of the major chemicals at each site. This was the case for every site on the NYSDEC list. Almost all of the sites had multiple toxic chemicals, which is why we dichotomously categorized the ZIP codes as containing or not
containing POPs, containing or not containing VOCs, or containing neither POPs nor VOCs. Then we identified individual POPs and categories of VOCs, such as chlorinated or not chlorinated in each ZIP code using the same dichotomous rule.

We further refined the ZIP-codes with “VOCs without POPs” (228 ZIP-codes), “POPs without VOCs” (85 ZIP-codes), “VOCs and POPs” (110 ZIP-codes), “Other” (32 ZIP codes), and “Clean” (1,169 ZIP-codes) to reduce the cross contaminations. “Clean” is not meant to indicate that that zip code had no sources of contamination but only that it did not contain an identified hazardous waste site.

2.3.3 Statistical analysis

We calculated EPC hospitalization rates per 100,000 population as the number of hospital discharge diagnoses of EPC divided by the total population residing in the ZIP codes at each category of age, race, sex, and area-poverty status. We modeled the rates of pancreatic cancer hospitalization in the exposure categories of VOCs without POPs, VOCs and POPs, POPs without VOCs, and Other in comparison to the ZIP codes with no identified HWSs (Clean) using negative binomial process adjusting to age, race, sex, and area-poverty status. We calculated 95% confidence intervals (CI) of rate ratios (RRs) for each exposure category accounting for the effects of the potential confounders including age, race, sex, and area-poverty status. Given that epidemiological studies and meta-analyses have reported that both race and poverty impact pancreatic cancer rates (Silverman et al., 2003), we assessed pancreatic cancer hospitalization rates by race and poverty status by including the interaction term in the models.

The negative binomial model was set as:

\[
\text{Number of pancreatic cancer discharges} = \exp (\beta_0 + \beta_1 \text{exposure} + \beta_2 \text{age} + \beta_3 \text{race} + \beta_4 \text{sex} + \beta_5 \text{area-poverty status} + \beta_6 \text{race} \times \text{area-poverty (interaction term)}) + \epsilon
\]

where, \(\beta_0, \ldots, \beta_6\) was the intercept and regression coefficients, whereas \(\epsilon\) was the model random error.
Variable levels and reference groups were defined as:

Exposure levels: Clean, VOCs without POPs, VOCs and POPs, POPs without VOCs, and Other using Clean group for a reference.

Age: <54, 54-74, and over 75-years using <54 group for a reference.

Race: Whites and Blacks using Whites for a reference.

Sex: Males and Females using Females for a reference

Area-poverty status: <5%, 5%--<10%, 10%--<20%, and>20% using <5% group for a reference.

2.4 Results

The goal of deduplication was to identify duplicate data rows potentially associated with a single patient based on a subset of variables, as follows: 5-digit ZIP-code of residential address, patient’s birth year, birth month, race, sex, and calendar year of hospital discharges. The data deduplication was repeated for all possible combinations of the subset of variables and the performance of each deduplication result was compared to the publicly available NYSCR’s pancreatic cancer incidence data as a gold standard. The minimum amount of disparity between the deduplicated SPARCS and the NYSCR’s incidence data was achieved by the subset of the above-mentioned variables except for the calendar year. Figure 2.2 presents the resulting deduplicated hospital discharge records of pancreatic cancers in the SPARCS and the incident cases of pancreatic cancers in the NYSCR during 1996-2013.

SPARCS contains hospital discharge data that is mandated to be reported by all state-regulated hospitals to the New York State Department of Health (NYSDOH), so does not include Veterans Administration and Indian Health Services (IHS) facilities which covers veterans and American Indian/Alaska Native [AI/AN] population. Whereas all cancer cases are mandated to be reported to the Cancer Registry except for the veterans’ hospitals and military hospitals are exempt from the reporting requirements. However, in practice, some VA hospitals voluntarily report to the Cancer Registry (health.ny.gov/statistics/cancer/registry/about.htm). We didn’t have data regarding volunteering VA
hospitals. However, the disparity between the SPARCS and NYSCR due to veterans and AI/AN population may not significantly distort the study results given a relatively small amount of cases from two sources such as, about an annual average of 19 new cases reported among veterans per U.S state (Zullig et al., 2017) and an annual average of 16 new cases of pancreatic cancer reported among AI/AN population in NY State (wonder.cdc.gov). The overall difference between the two datasets was only 199 for the entire study period after the data deduplication. Unlike chronic diseases, pancreatic cancer has a high hospitalization rate in general due to the necessity of intensive medical care, therefore, it's unlikely to miss the impactful number of cases from non-hospitalized cases.

These results indicate that with the deduplication procedures applied, the use of the SPARCS data provides comparable information to that of the NYSCR, at least for rapidly fatal cancers such as pancreatic cancer, and that this can be done without having access to unique identifying information.

Figure 2.2. Deduplicated SPARCS data compared to The New York Cancer Registry by calendar year, age groups, and race and year, excluding New York City.

The orange line represents the number of hospital discharge cases for exocrine pancreatic cancer (International Classification of Diseases [ICD] codes: 157) after data deduplication. The blue line represents the number of pancreatic cancer incidence reported by NYSCR. The panels compare the data reported from the two data sources from 1996 to 2013 excluding New York City by calendar years, 5-year interval age groups, and White and Black races clockwise.
2.4.1 Descriptive analysis

Table 2.1 compares crude and age-adjusted rates of pancreatic hospital discharge rates by exposure groups and key socio-demographics. Both crude and age-adjusted rates were lowest in the *Clean* exposure group, whereas the highest rates were observed in the 'VOCs without POPs' group followed by 'VOCs and POPs' and 'POPs without VOCs' groups. The ‘Other’ group (HWSs with only metals and some nuisance materials) had only 733 cases over the 18 years, thus these rates lacked meaningful interpretation. As expected for pancreatic cancer, the majority of patients were older than 54-years. In record per patient data, age ranged from 3 to 107 with an average of 70 years and median of 71 years. Among the cases used in the study, 93% were non-Hispanic white, 7% were African Americans (*blacks*).

The age-adjusted rate is higher among *blacks* whereas the crude rate is higher among *whites*. The sex-specific crude rate ratio is almost one whereas the age-adjusted rate is higher among males after adjusting to age, probably because of higher rates of smoking among men. The area-poverty level-specific rates stratified by race shows that both crude and age-adjusted rates are highest in the wealthiest group (percentage below poverty lower than 5%) among *whites*. In contrast, the rates were highest in the poorest group (percentage below poverty greater than 20%) among the *blacks*. The observed trends were not linear and were slightly more prominent in the crude rates. The disparity between the crude and age-adjusted rates within the same groups shows that the age distribution of *whites* and *blacks* is substantially different.
Table 2.1. Inpatient hospital discharge rate for pancreatic cancer during 1996-2013, NY State excluding NY city by demographics and poverty status.

<table>
<thead>
<tr>
<th></th>
<th>Hospital discharge N(%)</th>
<th>Person-years N(%)</th>
<th>Hospital discharge rate per 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total N (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>14,124 (49%)</td>
<td>85,012,839 (49%)</td>
<td>16.6</td>
</tr>
<tr>
<td>Female</td>
<td>14,817 (51%)</td>
<td>89,607,546 (51%)</td>
<td>16.5</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whites</td>
<td>26,919 (93%)</td>
<td>157,835,898 (90%)</td>
<td>17.1</td>
</tr>
<tr>
<td>Blacks</td>
<td>2,022 (7%)</td>
<td>16,784,487 (10%)</td>
<td>12.0</td>
</tr>
<tr>
<td><strong>Age (in years)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under 54</td>
<td>3,475 (12%)</td>
<td>127,542,060 (73%)</td>
<td>2.7</td>
</tr>
<tr>
<td>55-74</td>
<td>13,596 (47%)</td>
<td>33,824,727 (19%)</td>
<td>40.2</td>
</tr>
<tr>
<td>Over 75</td>
<td>11,870 (41%)</td>
<td>13,253,598 (8%)</td>
<td>89.6</td>
</tr>
<tr>
<td><strong>Whites by poverty level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under 5%</td>
<td>11,520 (43%)</td>
<td>61,235,694 (39%)</td>
<td>18.8</td>
</tr>
<tr>
<td>5%-10%</td>
<td>8,085 (30%)</td>
<td>49,803,606 (32%)</td>
<td>16.2</td>
</tr>
<tr>
<td>10%-20%</td>
<td>5,954 (22%)</td>
<td>38,616,948 (24%)</td>
<td>15.4</td>
</tr>
<tr>
<td>Over 20%</td>
<td>1,360 (5%)</td>
<td>8,179,650 (5%)</td>
<td>16.6</td>
</tr>
<tr>
<td><strong>Blacks by poverty level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under 5%</td>
<td>276 (14%)</td>
<td>2,293,173 (14%)</td>
<td>12.0</td>
</tr>
<tr>
<td>5%-10%</td>
<td>467 (23%)</td>
<td>3,985,857 (24%)</td>
<td>11.7</td>
</tr>
<tr>
<td>10%-20%</td>
<td>645 (32%)</td>
<td>5,616,369 (33%)</td>
<td>11.5</td>
</tr>
<tr>
<td>Over 20%</td>
<td>634 (31%)</td>
<td>4,889,088 (29%)</td>
<td>13.0</td>
</tr>
</tbody>
</table>

The distribution of the cases across the poverty categories is strikingly different in the direction that more **Blacks** live in a poor neighborhood (31% living in the area of a poverty line >20%) compared to 5% among **Whites**. Taking everything into account, the final statistical analyses included categorical **age** variable (under 54 as a reference group), **race** (Black vs Whites as a reference), **gender** (male as a reference), and **poverty status** (poverty <5% as a reference) as covariates for the adjustment in addition to the main exposure variable using **Clean** sites as a reference.

### 2.4.2 Modelled results

Table 2.2 presents the results of a multiple negative binomial regression model. We found a 6% and 5% increase in pancreatic cancer hospitalization rates in the population in the **VOCs without POPs** and **VOCs and POPs** groups, respectively, compared with **Clean** sites after adjusting for the potential
confounders (age, race, sex, and area-poverty). The association for the \textit{POPs without VOCs} group was marginally significant (p-value 0.110). As expected, the adjusted RRs significantly increased among the elderly. Males had a 24% excess rate compared with females. We observed a monotonic increase in rate ratios across increasing area-poverty levels among \textit{blacks} peaking at 1.29 (95% CI: 1.11-1.50) in the poorest group (percentage of people living below poverty >20%). However, RRs across area-poverty levels among \textit{whites} were not linear in trend but were consistently lower compared with the reference group (the wealthiest area).

Table 2.2 presents the sub-group analyses to assess the relationship between the EPC hospitalization rates and selected human carcinogens as identified by the International Agency for Research on Cancer (IARC). Data shown in Table 2.2 for the \textit{VOCs without POPs} and \textit{POPs without VOCs} groups are also included in Table 2.3 for comparison purposes. \textit{Any VOCs} or \textit{Any POPs} groups represent the ZIP codes that contain HWSs with VOCs and POPs, respectively, without excluding the cross-contaminations of POPs or VOCs. EPC hospitalization rate for \textit{Any VOCs} group was significant (1.06, 95%CI: 1.03-1.08), similar to \textit{VOCs without POPs} group (1.06, 95% CI: 1.03-1.09). The association for \textit{Any
POPs was statistically significant (1.05, 95%CI: 1.02-1.08) and slightly higher than the association for
POPs without VOCs group.

Table 2.3. Subgroup analysis. Relative risk and 95% CI by subgroups.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Rate Ratio</th>
<th>95% CI</th>
<th>N of ZIP codes</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VOCs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any VOCs</td>
<td>1.06</td>
<td>1.03 - 1.08</td>
<td>338</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VOCs without POPs</td>
<td>1.06</td>
<td>1.03 - 1.09</td>
<td>228</td>
<td>0.001</td>
</tr>
<tr>
<td>ncVOCs</td>
<td>1.09</td>
<td>1.05 - 1.13</td>
<td>111</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Benzene</td>
<td>1.12</td>
<td>1.07 - 1.16</td>
<td>62</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ethylbenzene</td>
<td>1.34</td>
<td>1.26 - 1.42</td>
<td>22</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>cVOCs</td>
<td>1.06</td>
<td>1.03 - 1.09</td>
<td>261</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Trichloroethylene</td>
<td>1.07</td>
<td>1.04 - 1.11</td>
<td>101</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Tetrachloroethylene</td>
<td>1.11</td>
<td>1.07 - 1.15</td>
<td>83</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Vinyl chloride</td>
<td>1.05</td>
<td>1.00 - 1.11</td>
<td>22</td>
<td>0.07</td>
</tr>
<tr>
<td>*Chlorinated benzenes</td>
<td>0.99</td>
<td>0.92 - 1.06</td>
<td>16</td>
<td>0.73</td>
</tr>
<tr>
<td><strong>POPs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any POPs</td>
<td>1.05</td>
<td>1.02 - 1.08</td>
<td>195</td>
<td>0.002</td>
</tr>
<tr>
<td>POPs without VOCs</td>
<td>1.04</td>
<td>0.99 - 1.09</td>
<td>85</td>
<td>0.110</td>
</tr>
<tr>
<td>†Chlorinated pesticides</td>
<td>1.11</td>
<td>1.05 - 1.18</td>
<td>25</td>
<td>0.004</td>
</tr>
<tr>
<td>PCBs only</td>
<td>1.05</td>
<td>1.01 - 1.09</td>
<td>186</td>
<td>0.013</td>
</tr>
<tr>
<td>PAHs (non-chlorinated)</td>
<td>1.00</td>
<td>0.90 - 1.10</td>
<td>19</td>
<td>0.93</td>
</tr>
</tbody>
</table>

*Chlorinated benzenes included chlorobenzene, dichlorobenzene and monochlorobenzene.
†Chlorinated pesticides included chlordane, DDE, DDD, DDT, dieldrin, endosulfan, endrin, and heptachlor.

For chlorinated and non-chlorinated VOCs, non-chlorinated VOCs had a slightly higher RR (1.09, 95% CI: 1.05-1.13) than chlorinated VOCs (1.06, 95%CI: 1.03-1.09). From non-chlorinated VOCs, we selected benzene, a known carcinogen, and ethylbenzene, a possible carcinogen to humans for further analysis. We found a statistically significant association for benzene 1.12 (95%CI: 1.07-1.16) and an even stronger association for ethylbenzene at 1.34 (95%CI: 1.26-1.42). For the chemicals in the chlorinated VOCs group, we found statistically significant associations for trichloroethylene 1.07 (95%CI: 1.04-1.11), tetrachloroethylene 1.11 (95%CI: 1.07-1.15), and marginally significant association for the vinyl chloride. Chlorinated benzenes were not associated with an increased EPC rate. Chlorinated pesticides (chlordane, DDE, DDD, DDT, dieldrin, endosulfan, endrin, and heptachlor) and PCBs showed increased RR, respectively, 1.11 (95%CI: 1.05-1.18) and 1.05 (95%CI: 1.01-1.09). There was no significant association for PAHs. The quality of fit of the negative binomial model was satisfactory given the value of
the Pearson $\chi^2$ and the deviance divided by the number of degrees of freedom was close to 1. All statistical analyses were conducted as two-sided with an alpha level of 0.05 using the Proc Genmod procedure in SAS software version 9.4 (SAS, 2020).

2.5 Discussion

We assessed whether living near a hazardous waste site containing volatile and persistent organic pollutants is associated with rates of pancreatic cancer in the general population, as this has not been studied before. We have evidence that residence within a ZIP-code containing HWSs with specific VOCs and POPs is associated with an elevated risk of EPC after adjustment for age, sex, race, and area-poverty status. Overall, the strength of association for VOCs was stronger than POPs, although this must be considered in light of the fact that VOCs are much more volatile than semi-volatile POPs.

As this is an ecological study we didn’t aim to establish causality but tested the hypotheses of whether living in ZIP codes containing Superfund hazardous waste sites that abut specific organic chemicals are associated with pancreatic cancer incidence. We have the evidence that the ecologic studies showing elevated rates of hospitalization for hypertension and diabetes in PCB sites (Kouznetsova et al., 2007; Huang et al., 2006) were then followed by studies in populations where PCB concentrations in blood were correlated with hospitalizations rates for hypertension and diabetes (Aminov Zafar et al., 2016; Goncharov et al., 2008).

In this study, noticeably the strongest associations were found for ethylbenzene and benzene. IARC classified ethylbenzene as possibly carcinogenic to humans. Ethylbenzene is found at the highest concentration in styrene and plastic production industries (Ott et al., 1980) and it is one of the most commonly found substances at HWSs (ATSDR, 2010). A dose-response relationship between the occupational exposure to ethylbenzene and excess pancreatic cancer risk was reported in a large cohort study among the workers in Denmark, Finland, Italy, Norway, Sweden, and the UK (Kogevinas et al., 1994). This finding was also corroborated by the Danish sub-cohort of the study with the significantly
increased incidence rate ratio (Kolstad et al., 1995) and North American occupational study with increased standardized mortality ratio in styrene chemical plants (Bond et al., 1992). Other occupational studies based on styrene factories also reported suggestive evidence of excess death from pancreatic cancer in the U.S (Ruder et al., 2004; Wong et al., 1994). Benzene is a known human carcinogen (IARC, 2018) especially for hematological cancers (Boberg et al., 2011; Collins et al., 2003; Hayes et al., 2001). Our result for benzene was consistent with a major clinic-based case-control study in the U.S (Antwi et al., 2015) that reported a statistically significant (OR=1.7, 95%CI: 1.23-2.35) association between pancreatic cancer and benzene exposure. However, an occupational exposure study in Spain (Alguacil et al., 2000) did not find conclusive evidence (OR = 0.93, CI 95%: 0.47-1.83).

Pesticides, organochlorine insecticides, and PCBs: We found a significantly elevated risk of EPC hospitalization (RR=1.11, CI95%: 1.05-1.18) associated with chlorinated pesticides (chlordane, DDE, DDD, DDT, dieldrin, endosulfan, endrin, and heptachlor). An occupational study has reported an increased standardized mortality ratio among the workers exposed to DDT (Beard John et al., 2003). Two studies based on individual-level serum concentrations of specific organochlorines (DDE, DDT, and PCBs) in Spain (Porta et al., 1999) and the U.S (DDE, PCB, and trans-nonachlor) (Hoppin et al., 2000) reported significantly elevated odds ratios of EPC among the patients with higher serum concentrations of these specific organochlorines compared to controls. However, confidence in our result is limited by the small number of ZIP codes that contained these chlorinated pesticides. Nevertheless, our observation corroborates existing evidence and is biologically plausible. Questions remain as to which of this group of chlorinated pesticides are responsible for the associations seen.

A significant association for PCBs in this study is consistent with the observations of Porta et al. who reported excess pancreatic cancer incidence among individuals with higher serum PCBs (Porta et al., 1999). Hoppin et al. (Hoppin et al., 2000) also reported an elevated risk of pancreatic cancer in relation to serum PCB levels in a case-control study. PCBs as a group consists of several different congeners
depending on the number of chlorines and their positions around the biphenyl rings. We assume the route of exposure in our study is inhalation, and it is the PCB congeners with fewer chlorines that are more volatile, whereas those with more chlorines are more persistent and more likely found in food. Our results suggest that lower chlorinated, more volatile PCB congeners contribute to the risk of EPC, not just the higher chlorinated congeners that dominate in serum samples. We have previously demonstrated elevated cancers due to inhalation of PCBs (Carpenter, 2015).

**Chlorinated hydrocarbon solvents**: Residential exposure to trichloroethylene, tetrachlorethylene, and vinyl chloride may constitute a non-negligible risk of EPC. Trichloroethylene and vinyl chloride are *human carcinogens* and tetrachloroethylene is a *probable human carcinogen* according to IARC. Meta-analyses have reported elevated EPC risk associated with chlorinated hydrocarbons (CHC) and weak associations for individual CHCs, trichloroethylene, vinyl chloride, and tetrachloroethylene (Andreotti & Silverman, 2012; Ojajärvi et al., 2001, 2007). Positive associations related to CHCs were also corroborated by more recent hospital-based case-control studies in Spain (PANKRAS-II) (Alguacil et al., 2002) and the USA (Antwi et al., 2015). The potentially causal effect of VOCs was suggested because KRAS mutated cases were more likely to have been occupationally exposed to chlorinated hydrocarbon solvents compared to non-mutated cases (Alguacil et al., 2002).

For comparison, we applied the same study design, statistical methods, and data deduplication algorithm to kidney cancer using the same SPARCS dataset. Hospital discharge cases for kidney cancer were identified by ICD-9-CM code (189) and resulting records were deduplicated by the residential ZIP, birth-year, birth-month, sex, race, and ethnicity to reduce the multiple hospital records per person. We conducted multiple negative binomial regressions adjusting to the same demographic variables used for pancreatic cancer. We did not find any statistically significant increase in the kidney hospitalization rates associated with living in ZIP-codes that contained HWS containing volatile and persistent organic...
pollutants in comparison to *Clean* sites. While this study has not been published it assures that the findings with pancreatic are not non-specific.

The key limitation of this study is the ecological fallacy stemming from reported inferences was based on the aggregate measures of the response variable (number of hospital discharges for EPC per ZIP code of NY) and group-level non-biological exposure assignment (living in ZIP with hazardous waste sites). Therefore, study results cannot be assumed to be true for individuals.

Our exposure assessment is limited to residence in a ZIP code containing an HWS at the time of diagnosis. This is certainly poor exposure assessment, and its validity is not supported by any chemical measurements in the individuals. Moreover, it does not account for residential movement. Exposure assignment to ZIP codes is not adequately specific given that VOCs, PAHs, and PCBs are ubiquitous in the ambient air as they are regularly produced from motor vehicle emissions, cigarette smoke, and polluted food and consumer production processes (US EPA, 2015). Also, these chemicals occur in a mixture with other pollutants that make measuring and tracking these chemicals in a residential setting very difficult (Carpenter et al., 1998). In addition, potential confounders such as the area of ZIP codes and the residents' (centroid of ZIP code areas) proximity to the waste sites have not been accounted for in the exposure assessment.

The 24% excess rate from the multi-variable model in the current analysis compared with females may partially reflect smoking as one of the PDAC (pancreatic ductal adenocarcinoma) risk factors tend to be higher among men. Even though, smoking doesn’t account for the difference in the observed disparity given that most cancer incidence and death rates (except women’s cancers) are higher in males than females not just due to smoking (Sung et al., 2021).

Also, we don’t have individual-level smoking and food consumption data which would account for the significant amount of person-level exposure to benzene, ethylbenzene (Tang et al., 2000), PAHs, and PCBs (Carpenter, 2006). Even though direct personal exposure to cigarette smoke would achieve a higher
PDAC risk, the attributable fraction of risk from the exposure to hazardous waste sites is necessary. Residential exposure is known to be many times lower than occupational or point-source exposures such as direct handling of chemicals (i.e., pesticide applications). Our exposure assessment is not inclusive of the baseline concentrations of these chemicals in the ambient air.

Despite these limitations, primarily due to poor exposure assessment, the strengths of the study out-way the limitations. Using the SPARCS dataset we have a large number of cases of EPC with demographic information that can be matched to the HWS exposure information. Our conclusions are in general consistent but expand upon previous occupational studies. Our results have important implications because we are studying low-level residential exposures that affect a large number of people. The results reported in this study are likely to be the underestimations of actual relationships between the exposure and pancreatic cancer hospitalizations. At a minimum, our results demonstrate the need for additional study of the contribution of exposure to organic chemicals to EPC. Our results also show the merit of using hospitalization data when cancer registry data is not available for discerning patterns of cancer. Furthermore, the result of this study is generalizable to other locals since there are HWSs everywhere.

2.6 Conclusion

Living near to an HWS containing hazardous organic chemicals is associated with a statistically significant elevation in rates of hospitalization for EPC after adjustment for some of the potential confounders. This finding is consistent with the hypothesis that inhalation of these volatile and semi-volatile chemicals poses an increased risk of the development of EPC.

2.7 References

Alguacil, J., Group, for the P. I. S., Porta, M., Group, for the P. I. S., Malats, N., Group, for the P. I. S., Kauppinen, T., Group, for the P. I. S., Kogevinas, M., Group, for the P. I. S., Benavides, F. G., Group, for the P. I. S., Partanen, T., Group, for the P. I. S., Carrato, A., & Group, for the P. I. S.


CHAPTER 3. A METHODOLOGICAL BLUEPRINT TO IDENTIFY COVID-19 VULNERABLE LOCALES BY SOCIOECONOMIC AND EPIDEMIOLOGICAL FACTORS

3.1 Abstract

COVID-19 has severely impacted socioeconomically disadvantaged populations. To support pandemic control strategies, geographically weighted negative binomial regression (GWNBR) mapped COVID-19 risk related to epidemiological and socioeconomic risk factors using South Korean incidence data (January 20, 2020, to July 1, 2020). We constructed COVID-19-specific socioeconomic and epidemiological themes using established social theoretical frameworks and created composite indexes through principal component analysis. The risk of COVID-19 increased with higher area morbidity, risky health behaviors, crowding, and population mobility, and with lower social distancing, healthcare access, and education. Falling COVID-19 risks and spatial shifts over three consecutive time periods reflected effective public health interventions. This study provides a globally replicable methodological framework and precision mapping for COVID-19 and future pandemics.

3.2 Background

The reported gap between the COVID-19 rates of the most and least advantaged populations (Jung et al., 2020) presents a potential for reducing the outbreak through targeted interventions. Unlike non-modifiable factors (Glanz & Bishop, 2010; Zulman et al., 2008), such as populations with genetic predispositions, population vulnerability to infectious disease outbreaks can be remediated through targeted interventions. Population vulnerability to respiratory infectious diseases is characterized by multiple interrelated factors, such as family income, education, employment status, health behavior, healthcare access, and other area-health indicators (Blumenshine et al., 2008; Oakes & Rossi, 2003). Hence, identifying key socio-economic determinants for COVID-19 and mapping the vulnerable locales will enable policymakers to target specific modifiable factors in high-risk areas.
Among various approaches to disentangle how socioeconomic status impacts health, Coleman’s social theory has been regarded as exceedingly useful because of its treatment of SES (socioeconomic status) beyond access to material resources but also a function of social and human capital that ‘uniquely locate the individual’s status in the social structure’. Blumenshine furthered the understanding by illustrating the mechanistic pathways between the socio-economic position and health disparity, as the underlying socioeconomic determinants of individuals can determine their likelihood of being exposed to the pandemic virus, contracting the disease, and timely and effective treatment after the disease developed (Blumenshine et al., 2008).

Although prior studies provided COVID-19 risk factors, none identified COVID-19-vulnerable locales associated with SES and COVID-19-specific epidemiological factors with acceptable generalizability and methodological capacity. Current COVID-19 studies relying on health disparity measures use arbitrary SES variables based on researchers’ preferences, irrespective of their COVID-19 relevance. Consequently, the SES measures across these studies are incomparable, limiting their usefulness. To address this, we integrated Coleman’s Social Theory and Blumenshine’s mechanistic framework (Figure 1), which formulates a universal SES definition and SES indicator selection mechanistically/causally relevant to the COVID-19 health outcome. This approach can inform public health interventions to alleviate SES factor-related COVID-19 risk.

Since SES and epidemiological data cover an expanse of highly intercorrelated variables, the composite SES index, derived from multiple unique SES variables helps garner the most explanatory information from the contributing indicators. A single universal composite measure, like the commonly-used area deprivation index (Jarman et al., 1991), is limited to controlling SES effects as confounders, but not when the study’s goal is assessing the effect of multiple SES determinants on a health outcome (e.g., COVID-19 in this study). Therefore, a multiple composite SES index approach helps quantify each composite SES index’s effect on COVID-19 (CDC, Surgo Foundation, 2020).
Thus far, one study used multi-scale Geographically Weighted Regression (GWR) to map the US COVID-19 incidence rate, while accounting for selected SES variables (median household income, income inequality, percentage of nurse practitioners, and black female population) (Mollalo et al., 2020). Since multi-scale GWR does not fit a beta distribution typical for infectious disease rates (Zhu & Chen, 2021), we recommend Geographically Weighted Negative Binomial Regression (GWNBR) to improve methodological accuracy. GWNBR directly uses discrete count data without further transformation and is robust in overdispersion, spatial/temporal clustering, and false positives (da Silva & Rodrigues, 2013; Ma, 2020). Globally, the COVID-19 pandemic emerged in waves with country-specific mitigation strategies producing sharp declines. To help improve public health interventions by precision targeting of high-risk locales, this study identified key SES and epidemiological risk determinants and their geographic distribution. Our study’s goals were to; (1) provide a methodological framework for identifying COVID-19-vulnerable locales associated with SES and epidemiological determinants, and (2) operationalize the framework using South Korean data to demonstrate its value.

3.3 Materials and Methods

3.3.1 Study Design and Population

We used COVID-19 incidence data from January 20 through July 1, 2020, released from the Korea Centers for Disease Control and Prevention (KCDC) (KCDC, 2020) and prepared by the DS4C project (Kim, 2020). The dataset is based on the report materials of KCDC and local governments from this time period. These data are available under the KOCL (Korea Open Government License). Analytical data consisted of 11,811 COVID-19 cases aggregated by 250 districts (Table S.3.1) aligned to SAS’s South Korean geographic matrix. Since the data were unavailable for Daegu’s subparts, we estimated the incidence from KCDC’s press release cluster reports.
3.3.2. Conceptual Model

Figure 3.1 shows the Coleman-Blumenshine Framework (CBF) refined approach, based on Coleman’s Social Theory and Blumenshine’s mechanistic framework (Blumenshine et al., 2008; Coleman, 1990). The model defines SES as a function of material, social, and human capitals (Coleman, 1990; Oakes & Rossi, 2003) and emphasizes pathways by how SES indicators differentially increase SARS-CoV-2 exposure and susceptibility to developing COVID-19 (Blumenshine et al., 2008). Based on the CBF model and COVID-19 risk factors literature (de Lusignan et al., 2020; Koh, 2020; Oh et al., 2020; Richardson et al., 2020), we identified seven area-level health and SES factors that determined the SARS-CoV-2 exposure level and the likelihood of developing COVID-19 after exposure.

Figure 3.1. Conceptual model of the causal relationship between the SARS-CoV-2 and area health/SE determinants. Abbreviations: material capital (MC), human capital (HC), social capital (SC), socioeconomic status (SES). Subscripts (i, j, k) indicate the number of variables used from the data sources. *Material, Human and Social Capital refers to latent structural components of the SES and COVID-19-specific determinants. Health and SES connected by arrows indicate the inter-relatedness of area health and SES, hereinafter, denoted as a health/SE. ↓Area-health/SE themes identified relevant to COVID-19 based on the current person and population-level literature. As per Coleman’s social theory and contributing data underlying each health/SE theme, crowding, healthcare access and social distancing relates to material capital, health behavior and area morbidity relates to human capital, whereas, crowding, education and population mobility to social capital. Modified from source (Blumenshine et al., 2008; Quinn & Kumar, 2014). Data sources: Korean Community Health Survey by the KCDC, Health Insurance Statistics by the National Health Insurance Services, Disability Status by the Ministry of Health and Welfare, Death Cause Statistics by the National Statistics Agency, Korean Census Bureau, Internal Migration Statistics by the Statistics Korea, and the State of Urban Planning Report by the Ministry of Land, Infrastructure, Transport, and Tourism.
3.3.3 SES Measurement and Epidemiological Factors

All SES and epidemiological-related data were retrieved from the Korean Statistical Information Service’s (KOSIS) online data archive (kosis.kr/index/index.do). KOSIS offers a convenient one-stop service for South Korea’s major domestic statistics. Table S.3.2 presents the data sources used for SES measurement. Table S.3.3 shows 24 data items out of 124 candidates relevant to the seven health/SE areas. We used an independent variable proxy for education, and by Principal Component Analysis (PCA) created six thematic composite indices: healthcare access, health behavior, crowding, area morbidity, difficulty to social distancing, and population mobility. Factors were computed as linear combinations of the original variables selected for each health/SE theme. We used the first component scores (Johnson & Wichern, 2008) in calculating the composite scores since they explained the largest data variation.

Then we computed each variable’s weight by dividing each factor score by the sum of all variable factor scores as,

\[
Weight_i = \frac{Score_i}{\sum_{i=1}^{p} Score_i}
\]

where \(i\) relate to each theme’s variable and \(p\) is the number of each theme’s variables. Each thematic composite index was computed as the weighted average for all 250 district values. For example, the composite index for health behavior was calculated as:

\[
Health\ behavior_k = 0.438 \times \text{obesity by measurement}_k + 0.429 \times \text{alcohol drinking}_k + 0.100 \times \text{current smoking}_k + 0.033 \times \text{self-reported obesity}_k
\]

where \(k\) is the original variable’s value for district \(k\). Note that weights sum to 1 (0.438 + 0.429 + 0.100 + 0.033 = 1). Six thematic composite indices and an individual proxy for education (percentage of high school educated people) were used in the final models as independent variables. Our model outcome was the confirmed case counts of COVID-19 aggregated by 250 districts. Global negative binomial regression (GNBR) and GWNBR (da Silva & Rodrigues, 2013) computed relative risk of COVID-19 associated seven area health/SE themes.
3.3.4. Global Models

GNBR models calculated relative COVID-19 risk for the entire study period and each pandemic phase. The global model was set as,

\[
COVID-19 = \exp(\beta_0 + \beta_1 \text{healthcare access} + \beta_2 \text{health behavior} + \beta_3 \text{area morbidity} + \beta_4 \text{education} + \beta_5 \text{difficulty to social distancing} + \beta_6 \text{population mobility} + \varepsilon).
\]

where, \(\beta_0, ..., \beta_n\) were the intercept and regression coefficients, whereas \(\varepsilon\) was the model random error.

3.3.5. Local Spatial Models

We used Gaussian GWNBR to model discrete count data and handle overdispersion issues. GWNBR computed parameter estimates for all districts following,

\[
y_j \sim NB\left(t_j \exp(\sum_k \beta_k(u_j, v_j)x_{jk}), \alpha(u_j, v_j)\right)
\]

where \((u_j, v_j)\) are the locations (coordinates) of the data points \(j\), for \(j = 1, \ldots, n\). The models empirically computed bandwidth, via the cross-validation criterion, and achieved minimal Akaike's information criterion (AIC) as,

\[
CV = \sum_{j=1}^{n} \left(y_j - \hat{y}_{j(b)}\right)^2
\]

where \(\hat{y}_{j(b)}\) is the estimated value for point \(j\), omitting the observation \(j\), and \(b\) is the bandwidth. The likelihood of false positives was corrected by the method of da Silva and Fotheringham (da Silva & Fotheringham, 2016). All statistical analyses including specific macro programs for spatial weight matrices and GWNBR models were implemented using SAS (version 9.4). Missing data (2%) were excluded from the analyses.

3.4 Results

Figure 3.2 compared the spatial COVID-19 distribution across pandemic phases. The initial outbreak wave occurred in Daegu which then spread to Gyeongsangbuk-do and surrounding provinces
in the early phase (Oh et al., 2020). The second wave occurred in Seoul and its surrounding metropolises, Ulsan and Busan, and Gyeonggi-do province in the late phase of the pandemic.

Early Phase

Middle Phase

Late Phase

Figure 3.2. The spatial distribution of COVID-19 cases across pandemic phases. Early phase: from January 20 to March 20, 2020. Middle phase: March 21 to April 15, 2020. Late phase: April 16 to July 1, 2020. The shade intensity and bar heights both indicate the number of COVID-19 cases during each in each district.

Global and Local Spatial Models

Throughout the entire study period model, GNBR suggested that the COVID-19 risk is associated with increased risky health behavior, area morbidity, and difficulty to social distancing (Table 3.1).

Inverse associations indicate an increased COVID-19 risk with reduced healthcare access, lower education, and increased efflux in population mobility. No substantial risk was associated with crowding.

Table 3.1. Parameter estimates and 95% CI of the relative risk of COVID-19 associated with health and SES determinants during the entire study period (January 20–July 1, 2020).

<table>
<thead>
<tr>
<th>Health/SE themes</th>
<th>Estimates</th>
<th>Relative Risk (95% CI)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthcare access</td>
<td>−0.13</td>
<td>0.88 (0.84–0.92)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Health behavior</td>
<td>0.04</td>
<td>1.04 (1.01–1.07)</td>
<td>0.019</td>
</tr>
<tr>
<td>Crowding</td>
<td>0.05</td>
<td>1.05 (0.89–1.25)</td>
<td>0.545</td>
</tr>
<tr>
<td>Area morbidity</td>
<td>0.04</td>
<td>1.04 (1.03–1.06)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Education</td>
<td>−0.09</td>
<td>0.91 (0.86–0.97)</td>
<td>0.002</td>
</tr>
<tr>
<td>Difficulty to social distancing</td>
<td>0.06</td>
<td>1.06 (1.01–1.12)</td>
<td>0.017</td>
</tr>
<tr>
<td>Population mobility</td>
<td>−0.22</td>
<td>0.80 (0.69–0.93)</td>
<td>0.003</td>
</tr>
<tr>
<td>Dispersion^a</td>
<td>2.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>1850</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

^aThe variance of a negative binomial distribution; Abbreviations: Akaike’s information criterion (AIC), a measure of goodness of model fit; confidence interval (CI); socioeconomic (SE).

We implemented global and local spatial models for the early, middle, and late pandemic phases. Figure 3 presents the relative risk of COVID-19 with its 95% CI from GNBR models, and Figure 4,
the relative risk spatial distribution from GWNBR associated with seven thematic areas by pandemic phases. Table S.3.4 provides more details on the stratified GNBR models. GWNBR fit data better than the global model given smaller AIC for the middle and late phases, respectively, ($\text{AIC}_{\text{gwnbr}} \sim 1034$ vs $\text{AIC}_{\text{gnbr}} \sim 1044$, $\text{AIC}_{\text{gwnbr}} \sim 1038$ versus $\text{AIC}_{\text{gnbr}} \sim 1074$) except for the early phase of the pandemic ($\text{AIC}_{\text{gwnbr}} \sim 3533$ vs $\text{AIC}_{\text{gnbr}} \sim 1527$). This reflects the large spatial cluster emerging from Daegu church (Her, 2020; Oh et al., 2020) activities during the early phase that subsequently spread to its neighboring districts.

The GNBR and GWNBR model results agreed across all pandemic phases. In the early phase, lower healthcare access and education, and increased risky health behavior, area morbidity, difficulty to social distancing, and population mobility are associated with higher COVID-19 risk. The crowding-associated risk was not significant in GNBR. In the middle phase, healthcare access, area morbidity, education, and difficulty to social distancing remained important risk determinants. In the late phase, only healthcare access, health behavior, and increased crowding significantly determined the COVID-19 risk.

![Figure 3.3. Relative Risk of COVID-19 associated with area health and SES determinants (GNBR models). Each panel shows the relative risk and its 95% confidence intervals associated with each thematic area. Colours represent the pandemic’s early (January 20 to March 20, 2020), middle (March 21 to April 15, 2020), and late phases (April 16 to July 1, 2020). The dashed line shows the reference level (1). Values over or below the reference line indicate statistically significant results at $\alpha = 0.05$. Corresponding P-values can be found in Table S.3.5.](image)
Healthcare
Access
Educ ation
Health
Beh aviour
Area
Morbidity
Difficulty to
Social
Distancing
Population
Mobility

Figure 3.4. Spatial variation in the relative risk of COVID-19 associated with area-health and SES themes in the early, middle, and late phases of the pandemic (GWNBR models). In the maps, the color gradient corresponds with larger (darker) to lower (lighter) relative risk. Areas in white indicate the relative risks are statistically not significant (α = 0.05). The figure consists of A, B, and C columns referring to Early, Middle, and Late phases, referring to the pandemic’s early (January 20 to March 20, 2020), middle (March 21 to April 15, 2020), and late phases (April 16 to July 1, 2020).

GWNBR created early phase maps showing higher risk in non-contiguous districts (Figure 4A).

This higher risk reflected virus transmission in the initially affected districts before spreading over larger areas. During the early phase, we found protective effects of improved healthcare access, higher
education, and outbound population mobility, whereas the disease risk was increased in the districts with higher risky health behavior, area morbidity rate, and difficulty to social distancing.

In the middle phase (Figure 4B), only healthcare access, area morbidity rate, education, and difficulty to social distancing remained as the key risk-determinants with the same directions but reduced strengths. Spatial shifts from the early phase were from the northwest toward the capital and southwest regions. In the late phase (Figure 4C), healthcare access, risky healthy behavior, and area crowding were primary risk determinants. We observed the protective effect of improved healthcare access, while risky health behavior was the significant risk factor. In contrast to the early phase, we found higher disease risk in more crowded districts. In the late phase, risk determinants concentrated around the capital and middle regions. Taking the type of risk-determinants and their spatial distributions across the three phases together, our results showed that the pandemic had evolved from lower to higher density areas. This led to the second wave that emerged in Seoul and its surrounding areas.

We observed noticeable spatial shifts in the risk determinants over the study period (Figure 4). Difficulty to social distancing increased COVID-19 risk in the capital and middle regions in the early phase which then shifted to the country’s southeast part in the middle phase. Area morbidity-associated risk was concentrated in the western part which then gradually shifted north in the middle phase. The education-associated risk was higher in the west in the early phase until it shifted southwest in the middle phase. Population mobility elevated COVID-19 risk only in the early phase for South Korea’s northern, eastern, and western parts.

We investigated the correlations between all pairs of composite indices (Table S.3.6). The largest Pearson’s r was 0.603 between healthcare access and area morbidity. We verified no multicollinearity given that the model’s standard error of both healthcare access (0.025) and crowding (0.09) was small.

For a direct comparison between non-spatial and spatial models, GNBR and GWNBR were carried out
with the same variables and stratified by the same periods (NBR: Figure 3.3, and GWNBR: Figure 3.4). AIC and dispersion coefficients were used to compare the models’ goodness of fit.

3.5 Discussion

GWNBR created a continuous surface of relative COVID-19 risk for all 250 districts associated with area-health and socioeconomic determinants by the pandemic phases (Figure 3.4). Our findings are consistent with individual and population-level studies that reported elevated COVID-19 risk associated with less healthcare access (Ji et al., 2020), and education (Bavel et al., 2020; Bruine de Bruin & Bennett, 2020), and more risky health behavior, crowding, specific comorbidities (de Lusignan et al., 2020; Richardson et al., 2020), difficulty to social distancing (Koh, 2020; Quinn et al., 2011) and population mobility (Jia et al., 2020). Our study’s high internal validity was shown since the GNBR and GWNBR results agreed except for crowding in the early phase.

Our approach captured statistically and noticeably high spatial variation by pandemic phases for all themes, consistent with the reported pattern of COVID-19 distribution in the country (Her, 2020; Oh et al., 2020). Since its first confirmed case on January 20th, 2020, South Korea experienced two major outbreak waves in Daegu and Seoul, and the surrounding Gyeonggi-do province, respectively, in February (early phase) and May 2020 (late phase).

After the first confirmed case, the KCDC invoked a four-level alert for the public’s emergency awareness (blue-attention, yellow-caution, orange-alert, red-serious) commensurate with the number of new confirmed cases [http://www.koreabiomed.com]. The epidemic’s initial wave in Daegu, caused by the local church activities, triggered the country-wide directives of hospital-based isolation/quarantine, contact tracing followed with free testing and treatment, strengthening medical centers for rapid diagnostics, emergency medical responses, and treatment aids (Oh et al., 2020). These specific measures along with high public adherence to the school and business closures, personal hygiene, and social distancing significantly dropped the case counts by mid-March. The second wave
erupted in May when non-essential businesses reopened [28] in Seoul, which spread to its surrounding metropolises, Ulsan and Busan, and Gyeonggi-do.

*The types of risk determinants changed over the pandemic phases.* Analysis stratified by periodic phases found that the initially high risk in the early period gradually decreased except healthcare access, health behavior, and crowding-associated risk, which increased in strength and concentrated in the capital and its surrounding provinces in the late phases (Figure 3 & Figure 4). Risk reductions could be explained by the impact of effective control measures that lowered the risk associated with these determinants and drop in an effective reproduction number (Re), as the number of infection-susceptible people decreased over time (Anderson et al., 2020).

In the early phase, all health/SE themes were statistically significantly associated with COVID-19 incidence. As anticipated, there was no excess risk at Daegu and its surrounding areas since the abnormally high spike of COVID-19 cases was caused by local church’s activities (Her, 2020; Oh et al., 2020) without relevance to the local area’s social status.

The increase in health behavior-associated risk is consistent with reports showing greater risk with poor emotional health (Bavel et al., 2020), smoking (Vardavas & Nikitara, 2020), and obesity (de Lusignan et al., 2020; Lighter et al., 2020). Individual patient-level COVID-19 risk-factors analysis Lusignan et al.,2020 (de Lusignan et al., 2020) reported smoking was a protective factor. However, the authors warned that the low proportion of current smokers in their study sample (11.4%), resulted in a wide confidence interval of the reported odds ratio, 0.59 (0.42–0.83). This increases the uncertainty of their result. Greater COVID-19 risk among the people with lower education has been explained as a reduced awareness of disease risk and low-income to obtain an education. This relationship was seen in our results as higher education is associated with lower COVID-19 incidence. The difficulty to social distancing in this study directly reflected the inability to afford unemployment possibly resulting in the reduced exertion of protective measures (Bavel et al., 2020; Bruine de Bruin & Bennett, 2020). The risks
associated with social distancing and area-morbidity peaked in the study’s early phase appeared to reduce in the middle phase, and completely remediated in the study’s late phase.

In the middle phase, all the previous risk factors except for risky health behaviors, population mobility, and crowding were high. The middle phase’s lessened risk associated with risky health behaviors, population mobility, and crowding may reflect the impact of the Prime Minister’s declaration. This implemented active interventions for social distancing, community health education, testing with a local contact, tracing, and hospital-based or self-isolation during March’s first weeks.

Notably, the late phase findings are consistent with the risk factors reported associated with the second wave in early May. During our study’s late phase, South Korea scaled up free testing and treatment through its existing health care centers (Her, 2020; Tanne et al., 2020), which may have improved healthcare access a key measure for combating COVID-19. Our finding that healthcare access exerts a stronger protective effect in the late phase compared with the earlier phases supports this. Our findings of increased risk associated with risky health behaviors may have captured behavioral fatigue at a population scale in response to the country’s multiple quarantine period extensions (Brooks et al., 2020) that likely were exacerbated by entertainment business re-openings (i.e., night clubs, karaoke) in early May. Elevated risks associated with increased crowding in the study’s late phase reflect the outbreak’s second wave, which occurred in South Korea’s most crowded region: Seoul and its surroundings (Figure 3).

*Spatial variation in the SES-related risk factors across the pandemic phases potentially reflect the geography-specific control measures and/or the differential public response to the measures.* GWNBR models revealed the pandemic phase-specific spatial variation for all health/SE themes except for population mobility which was not significant beyond the early phase. This may indicate that the effectiveness of the control measures varied over time potentially due to differential interventions or public response across the municipal districts. Our findings may also indicate a dynamic change in
population vulnerability throughout the pandemic “a person not considered vulnerable at the outset of a pandemic can become vulnerable depending on the policy response” as a Lancet editorial stated (The Lancet, 2020).

The factors increasing our recommended framework’s robustness include: 1) SES measurement and relationship conceptualization of the exposure (health/SE themes) and outcome (COVID-19 incidence) based on the refined conceptual framework; 2) joint use of conceptual and statistical modeling; 3) complementary use of global and local spatial statistics; and 4) stratified analysis by pandemic phases that enable us to capture the spatial variation over pandemic phases. However, this methodological framework relies on carefully collected country-specific data.

Our study is subject to ecological fallacy inherent to the study design. However, our empty hierarchical mixed model accounting for the individual and district-level data shows that 61% of the COVID-19 incidence distribution variation was explained by the district-level factors, leaving 39% of the variability for an explanation by individual factors.

We verified that the data estimation for Daegu city subparts did not affect the study results. The comparison of the intercept, standard error, relative risk, and P-value between the models with and without the estimated data showed that the intercept and standard error were diminished by 2.2% and by 9.2%, respectively in the models, including estimated data [each calculated by $100 \times (−12.29 - (−12.56))/−12.29$ and $100 \times (1.88−2.95)/1.88$, respectively]. A significance level change was observed for none of the model estimates, except for crowding. The P-value changed from ~0.06 to ~0.04 when the estimated data were excluded. However, the crowding-associated risk remains significant at $P = 0.1$.

Model details are provided in Table S.3.4. To assess the periodic trend in the relative COVID-19 risk associated with SES factors, we conducted stratified analyses by the early, middle, and late phases corresponding with 20 January-20 March, 21 March-15 April, and 16 April-1 July 2020.
Population-based COVID-19 studies are prone to response bias, which would not exist if everybody was tested. However, multiple factors determine testing coverage, therefore, the number of confirmed cases, such as easy access for testing and its accuracy (Omori et al., 2020), contact tracing strategies (Kretzschmar et al., 2020), under-testing of asymptomatic patients (Kinoshita et al., 2020). Also, psychological factors, fear of COVID-19 (Harper et al., 2020), risk perception (Dryhurst et al., 2020; Irigoyen-Camacho et al., 2020; Wise et al., 2020), and stigma-related testing avoidance (Baldassarre et al., 2020) impact the testing rate. Potential bias in this study is expected to be low given South Korea’s anti-pandemic strategies. Importantly, the country’s COVID-19 relief programs supported with 15 billion Korea won, dedicated to supporting vulnerable populations may have reduced the potential testing disparity by socioeconomic status. All Koreans and foreigners were entitled to free testing and treatment while testing access were more convenient through an extended number of rapid diagnostic centers and testing prompts through mobile phones.

Contact tracing-based testing increases the likelihood of capturing asymptomatic cases. Korean tracing system has been reported as the global best practice lending to its advanced information technology system and data extensions through large consumer and healthcare databases (global positioning system, credit card transactions, closed-circuit television, and medical facility use records) (Lee et al., 2020; Park et al., 2020). Testing avoidance from fear of stigma (Baldassarre et al., 2020; Logie & Turan, 2020) would likely have affected the early period of the analysis, which strongly reflected the abnormally high spike of cases in Daegu city traced to the local church. The city has not disclosed the data with the necessary granularity for a further investigation of this matter, as of writing. However, given these factors would likely result in the under-estimation of confirmed cases, any bias in our results should be toward the null.
3.6. Conclusions

The demonstrated methodology guides to design of multiple-determinant targeted interventions and pinpoint high-risk locales to remediate the excess COVID-19 risk attributable to socioeconomic disadvantages. Overall, our work has demonstrated that the anti-pandemic measures taken by the South Korean government were effective.

1. The completely remediated risk associated with area-morbidity and difficulty to social distancing is likely to be explained by the country’s emergency relief programs that targeted vulnerable individuals with socioeconomic disadvantages: Foreign workers, homeless, poor urban residents, disabled people, and the elderly. The assistance programs provided free testing, financial support, food assistance, health check-up visits, as they acknowledged excess hardship in adhering to social distancing rules because of the inability to afford unemployment.

2. The observed overall protective effect of improved healthcare access and higher education in our study support the rationale behind the country’s primary anti-pandemic agenda to strengthen healthcare facilities for rapid diagnostic and therapeutic services, combined with actionable health promotion rules which reportedly gained high public compliance.

3. However, we found risky health behavior was a persistent risk factor during both major outbreaks in Daegu and Seoul. Elevated crowding associated risk coincided with the Seoul outbreak, as anticipated.

4. Persistently high risks associated with health behavior and crowding, combined with the reduced protective effect of healthcare access and education in the study’s late phase may corroborate the finding that a prolonged pandemic induces adherence fatigue and lessened risk perception (Brooks et al., 2020; Bruine de Bruin & Bennett, 2020).

South Korean public health interventions have been discussed in detail elsewhere (Her, 2020) and we endorse the country’s anti-pandemic interventions as guidance to international policymakers.
The main highlights were: (1) Targeting vulnerable locales to COVID-19 and aiming to address multiple risk factors considering emergency relief programs to provide financial support, food assistance, health check-up visits; (2) implementing social distancing while assisting individuals with difficulty to social distancing. Social distancing measures may include school/business closures, hospital-based or self-isolated quarantine; (3) improving healthcare access for expanded testing and treatment with priority health services made available to the individual with extenuating medical conditions; (4) strengthening existing healthcare facilities and extending rapid diagnostic centers to enable easily accessible and free testing; (5) enhancing case identification capacity through information technology, such as contact tracing, mobile phone-based testing prompts and general risk alerts, rapid case-isolation by automated test result delivery to the testee’s mobile phones.

We emphasize the importance to anticipate adherence, behavioral, and mental fatigue over the course of a prolonged epidemic. In the latter phase of the epidemic, we recommend paying intensified attention to the urban and highly crowded areas to prevent a potential outbreak as well as promoting creative social networking solutions (drive-through services, virtual social events, telehealth, etc.) and ensuring emerging vaccine accessibility for the socially disadvantaged population (Persad et al., 2020). We intend that this framework can be replicable to both, international researchers, and policymakers, in order to enable rapid pandemic responses. As socioeconomic disparity is a global problem, nationwide programs with an intensified focus on the vulnerable populations at excess risk to pandemic ensure the efficacy and efficiency of pandemic alleviation efforts.

Future research should assess the mortality, and mortality and incidence ratio as a crude surrogate for survival using the same study design and methodology. Understanding the impact of socioeconomic and epidemiological risk factors on mortality compared with incidence would clarify the extent of the potentially preventable deaths through modifications of the assessed risk factors. The
overall and spatial disparity between mortality, incidence, and mortality/incidence ratio would inform where intensified public health and intensified healthcare services are needed.

### 3.7 Supplementary Materials

The following are available online at www.mdpi.com/xxx/s1, Table S.3.1 Confirmed cases of COVID-19 diagnosed during 20 January –1 July 2020 in South Korea by pandemic phases, Table S.3.2: Data sources reviewed and used for SES measurement, Table S.3.3: PCA details showing factor scores with their weights, and selected area health and SES variables and thematic composite indices, Table S.3.4: GNBR model estimates with and without the estimated data for Daegu’s subparts throughout the study period (January 20 through July 1, 2020), Table S.3.5: Parameter estimates and the Relative Risk of the COVID-19 incidence associated with health and SES determinants by three-time periods corresponding with the early, middle and late phases, Table S.3.6: Matrix table of Pearson’s correlation coefficients (r).

### 3.8 References


Quinn, S. C., Kumar, S., Freimuth, V. S., Musa, D., Casteneda-Angarita, N., & Kidwell, K. (2011). Racial disparities in exposure, susceptibility, and access to health care in the US H1N1 influenza


CHAPTER 4. EFFICACY AND SAFETY OF INNOVATIVE EXPERIMENTAL CHIMERIC ANTIGEN RECEPTOR (CAR) T-CELLS VS. AXICABTAGENE CILOLEUCEL (Yescarta) FOR THE TREATMENT OF RELAPSED/REFRACTORY LARGE B-CELL LYMPHOMA (LBCL): MATCHING ADJUSTED INDIRECT COMPARISONS (MAICs).

4.1 ABSTRACT

Despite favorable results of CAR T-cell therapy for relapsed/refractory large B-cell lymphoma (R/R LBCL), several challenges remain, including incomplete response, immune-mediated toxicity, and antigen-loss relapse. We delineated the relative clinical benefit of the novel approaches compared to the currently approved CAR T-cell therapies. In the absence of head-to-head comparisons and randomized controlled trials, we performed Matching Adjusted Indirect Comparisons to quantify the relative efficacy and safety of experimental CARs against Axicabtagene ciloleucel (Yescarta), the first FDA-approved CAR. A total of 182 R/R LBCL patients from 15 clinical trials with individual patient data (IPD) were pooled into eight populations by their CAR T-cell constructs and +/- ASCT status. The study endpoints were Progression-Free Survival (PFS), grade ≥ 3 cytokine release syndrome (CRS), and grade ≥ 3 neurotoxicity (NT). Tandem CD19.CD20.4-1BBζ CARs indicated favorable efficacy and safety, whereas the co-infusion of CD19 & CD20 with 4-1BBζ showed no clinical benefit compared to Yescarta. Third generation CD19. CD28. 4-1BBζ, and sequential administration of autologous stem cell transplantation (ASCT) and CD19. CARs presented statistically insignificant yet improved PFS and safety except for ASCT combined intervention which had suggestively higher NT risk than Yescarta. CARs with modified co-stimulatory domains to reduce toxicity (Hu19. CD8.28Zη and CD19. BBz.86ζ) presented remarkable safety with no severe adverse events; however, both presented worse PFS than Yescarta. Third-generation CARs demonstrated statistically significantly lower NT than Yescarta. CD20. 4-1BBζ data suggested targeting CD20 antigen alone lacks clinical, or safety benefit compared to Yescarta. Further comparisons with other FDA-approved CARs are needed.
4.2 BACKGROUND

Large B-cell lymphomas (LBCL) comprise diverse types of B-cell Non-Hodgkin Lymphoma (NHL), of which diffuse large B-cell lymphoma (DLBCL) is the most common histologic subtype, accounting for approximately a quarter of NHL cases worldwide (Morton et al., 2006). Survival rates have greatly improved over the past decades, particularly in the immunochemotherapy era, with a 5-year relative survival rate reported between 55.4% and 62.0% in developed countries (Howlader N, Noon AM, 2015). However, despite the advances achieved with rituximab-based regimens, up to 50% of patients with advanced-stage de novo DLBCL, for instance, will eventually relapse, even after achieving a complete response (CR) (Friedberg, 2011). If progression occurs during the initial treatment phase or soon after a brief CR, only 30% to 40% will respond to salvage chemotherapy and will be able to undergo consolidation with autologous stem cell transplantation (ASCT) (Neste et al., 2016). Even so, among these patients, roughly half will ultimately relapse after transplantation (Gisselbrecht et al., 2012). The prognosis, in such cases, is poor, especially for those who have high-risk factors or relapse within 12 months post-ASCT (Gisselbrecht et al., 2012; Neste et al., 2016). Thus, effective treatment for R/R LBCL remains a highly unmet need.

To date, only three CAR T-cell products (axicabtagene ciloleucel (Axi-cel, Yescarta), tisagenlecleucel (Tisa-cel, Kymriah), lisocabtagene maraleucel (Liso-cel, Breyanzi) are approved by the FDA for R/R LBCL (Abramson et al., 2020; Neelapu et al., 2017; Schuster et al., 2019). Despite the unprecedentedly high efficacy of these CAR T-cell therapies compared to historical outcomes for patients with R/R LBCL, current challenges, such as incomplete response, immune-mediated toxicity, and post-treatment relapse, remain. For example, in the ZUMA-1 trial for R/R LBCL, only 39% of patients maintained a CR to the therapy at the median of 27-month follow-up despite the initially high (82%) objective response rate (ORR) achieved (Locke et al., 2019; Neelapu et al., 2017). In an attempt to optimize CAR T-cell characteristics to address these inadequacies, pre-clinical researches identified
tumor antigen escape and CD19 antigen downregulation as potential causal factors for the suboptimal response and relapse observed after CAR T-cell therapy (Ruella et al., 2016). Tumor antigen escape leads to low antigen density via transfer of target antigens from the tumor cells to the CAR T-cells. This process, known as trogocytosis, has been observed with CD19, CD22, mesothelin, and B-cell maturation antigen (BCMA) (Hamieh et al., 2019).

Existing evidence prior to this study encompasses diverse strategies focused on advancing CAR T-cell performance. Specific approaches already notable for both their feasibility and clinical and safety benefit include (i) Multi-antigen targeting CAR T-cells obtained through co-infusion or sequential administration of single-targeted CAR T-cells against different antigens. Alternatively, tandem and bicistronic constructs expressing two different CARs on a single or a separate chimeric protein(s), respectively (Shah et al., 2019); (ii) Third and advanced generation CAR T-cells using integrated co-stimulatory domains (Enblad et al., 2018; C. Huang et al., 2020; Ramos et al., 2018); (iii) Enhanced co-stimulatory domains intended at reducing toxicity and preserving potency (Brudno et al., 2020; Ying et al., 2019); (iv) Combination therapy of CAR T-cells and immune checkpoint inhibitors (Cao et al., 2019); (v) Co-administration of ASCT and CAR T-cells (Kebriaei et al., 2016; Sauter et al., 2019; X. Wang et al., 2016); (vi) Alternative antigen targetings (other than CD19), such as CD20, CD22, CD27, ICOS, and OX40 (Schneider et al., 2021).

Under the backdrop of such key pre-clinical and clinical discoveries, with several studies evaluating innovative approaches toward CAR T-cell therapy for LBCL patients to extending the durability of response beyond that achieved by the currently approved CAR T-cell products would seem highly desirable, thus setting the basis for the novel approaches currently under investigation. Yet, to date, no data exist regarding the comparative efficacy and safety of experimental CAR T-cell products versus currently approved CAR T-cell therapies.
Hence, we aimed to compare the efficacy and safety of the currently available experimental CAR T-cell products to Yescarta, the first FDA-approved CAR T-cell therapy, thereby harboring the longest follow-up data available to date. Also, to overcome limitations of all currently available CAR T-cell trials being single-arm trials, and individual patient-level data only available for experimental CAR T-cell products and not for ZUMA-1 trial, the comparator (Phillippo et al., 2018), we used unanchored matching-adjusted indirect comparison (MAIC) as a primary method. The MAIC techniques attenuate bias in comparing multiple treatments assessed in different studies by matching patient-level data from the clinical trials of one treatment to aggregate data by comparator trials. Additionally, MAIC provides a more robust adjustment for cross-trial differences in patient characteristics than traditional meta-regressions due to its higher accuracy obtained from individual patient data than from aggregate data (Signorovitch et al., 2012). We believe this systematic review-based quantitative comparison may provide guiding insights into the ongoing efforts to advance CAR T-cell therapy for the treatment of R/R LBCL.

4.3 Materials and Methods

4.3.1 Summary of the evidence prior to this study

Dual target CAR T-cells: Preclinical studies demonstrated high anti-tumor potency with tandem CD19/CD20 CAR T-cells (Schneider et al., 2017; Zah et al., 2016), sequential infusion of CD19 and CD79b CARs (Ormhøj et al., 2019), co-infusion of CD19 and CD38 CAR T-cells (Mihara et al., 2010), and CD19/CD37 constructs (Scarfò et al., 2018). The clinical benefit of tandem CD19/CD20 CARs (Shah et al., 2020; Tong et al., 2020; Y. Zhang, 2020), co-infusion of CD19 and CD20 CARs (Sang et al., 2020), and mixed infusions of CD22 and CD19 CAR T-cells (N. Wang et al., 2020; Zeng et al., 2020) has been evaluated in small early phase clinical trials with demonstrated feasibility and varying levels of efficacy and safety.
Among the next-generation CAR T-cells, more mature data exist for the third-generation CAR T-cells incorporating both CD28ζ and 4-1BBζ co-stimulatory signaling domains. In mice models, third-generation CAR T-cells demonstrated improved T-cell persistence and stronger antitumor potency compared to second-generation constructs (Zhong et al., 2010). In addition, the clinical benefits of third-generation CARs in LBCL patients were evaluated in early phase trials (Enblad et al., 2018; C. Huang et al., 2020; Ramos et al., 2018). However, whether the addition of 4-1BBζ co-stimulatory domains to a common CD28ζ domain enhances such clinical benefits compared to second-generation CAR T-cells in this population is still unclear.

Variations of CAR T-cells with modified co-stimulatory domains aimed at reducing treatment-related toxicity include (1) Hu19. CD8.28Z, containing a fully human single-chain variable fragment (scFv) and CD8α-based hinge and transmembrane domains (Brudno et al., 2020); (2) CD19. BBz.86, with an 86-amino-acid fragment from human CD8α (Ying et al., 2019); and. Both CD19. BBz.86 and Hu19. CD8.28Z CAR T-cells demonstrated exceptional safety, yet attenuated efficacy, based on the CR rates of 29% and 39% observed, respectively, compared to the 54% CR rate noted among the LBCL patients receiving Axicel (Yescarta).

ASCT and CAR T-cell therapy: Multi-center randomized clinical trials are underway to determine the comparative efficacy and safety of CAR T-cell therapy alone vs. ASCT combined with systemic therapies for the treatment of R/R LBCL (refer to Discussion section for additional details). The study compared locally manufactured CD19. CD28ζ CAR T-cells in China to ASCT alone (NCT03196830) demonstrated superior efficacy and safety of the CAR T-cell product compared to ASCT in R/R NHL patients (Li et al., 2019). Whether the sequential administration of ASCT and CAR T-cells hold higher clinical benefits than CAR T-cells alone remains to be elucidated. Of note, this has already been shown to be feasible and safe in three clinical trials (Kebriaei et al., 2016; Sauter et al., 2019; X. Wang et al., 2016).
CD20. 4-1BBζ CAR-T cells demonstrated high antitumor activity against LBCL in pre-clinical studies (Y. Wang et al., 2014), and few clinical trials tested second-generation CD20 CAR T-cells in this disease (Y. Wang et al., 2014; W. Y. Zhang et al., 2016, 2017). Clinical trials evaluating third-generation CD20 CARs are currently underway in China (NCT02710149), in the USA (NCT03277729) - evaluating MB-106, a fully human third-generation CD20.4-1BBζ,CD28 ζ CAR T-cell constructs - and in Germany (NCT03664635), with MB-CART20.1 CARs (Borchmann, 2020; Shadman et al., 2019). Targeting CD20 was shown to be exceptionally more efficacious in follicular lymphoma, as demonstrated with the success of rituximab, an anti-CD20 monoclonal antibody, which led to the current rituximab-based first-line combination treatment for most NHL types (Salles et al., 2017). The fact that 30–40% of LBCL patients relapse after rituximab suggests that targeting CD20 alone is not enough (Salles et al., 2017). This has set the basis for comparative insights between CD20 and CD19-targeted CAR T-cells, thereby shedding light on the development of the dual targeting approaches mentioned above.

4.3.2 Data sources

Individual Patient Data (IPD): A Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline-based systematic review identified 15 single-center clinical trials for experimental CAR T-cell products (Table 4.1) with individual patient data (IPD), as presented in the PRISMA flow diagram in Figure S1, Supplementary Materials. Although LBCL is the most common histologic subtype of NHL, accounting for around 25% of cases, the number of patients with R/R LBCL who fulfill the eligibility criteria to receive CAR T-cell therapy is less common. This explains the small size of experimental CAR T-cells for this disease. To increase statistical power and test the hypothesis by distinct types of CAR T-cell interventions, we pooled IPD of 182 LBCL patients identified across 15 trials and subdivided them into eight distinct groups, as follows:

(1) dual targeting strategies, such as tandem CD19/CD20 CAR infusion;
(2) co-infusion of CD19 and CD20 CARs;
(3) third-generation CARs;
(4) CD19 CARs with modified constructs for reduced toxicity, including Hu19.CD8.28Z; and
(5) CD19.BBz.86-based CARs;
(6) sequential administration of ASCT and CD19.CD28ζ CARs; and
(7) CD20. 4-1BBζ CARs;
(8) CD19. 4-1BBζ CARs manufactured in China (Table 4.2).

We ensured that the trials had at least similar CAR T-cell constructs to be pooled together while excluding a few trials for unique CAR T-cell constructs yet with less than 10 patients (Figure S.4.1).
Table 4.1. Summary of Clinical Trials, Pooled Populations by CAR T-cell structure, and Study Endpoints.

<table>
<thead>
<tr>
<th>Intervention strategies</th>
<th>Pooled populations</th>
<th>Target Antigens</th>
<th>Signaling domains</th>
<th>Clinical Trial Registry Numbers</th>
<th>Disease Histology</th>
<th>N infusion</th>
<th>Endpoints of trials</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual targeting</td>
<td>Tandem CD19, CD20 with 4-1BB</td>
<td>Tandem CD19, CD20</td>
<td>4-1BB</td>
<td>NCT03019055, NCT03097770</td>
<td>DLBCL, trDLBCL, RS</td>
<td>14, 19</td>
<td>ORR, PFS, OS, CRS, NT</td>
<td>Shah2020 [32], Tong2020 [31]</td>
</tr>
<tr>
<td></td>
<td>Co-infusion CD19 &amp; CD20 with 4-1BB</td>
<td>Co-infusion of CD19 &amp; CD20</td>
<td>4-1BB</td>
<td>NCT03207178</td>
<td>DLBCL, trDLBCL</td>
<td>21</td>
<td>ORR, PFS, OS, CRS, NT</td>
<td>Sang2020 [33]</td>
</tr>
<tr>
<td>Third generation</td>
<td>CD19 with CD28 &amp; 4-1BB</td>
<td>CD19</td>
<td>CD28-41BB</td>
<td>NCT01853631, NCT02132624, NCT03121625</td>
<td>DLBCL, trDLBCL, RS</td>
<td>13, 4, 9</td>
<td>ORR, PFS, OS, CRS, NT</td>
<td>Ramos2018 [15], Enblad2018 [13], Huang2020 [14]</td>
</tr>
<tr>
<td>Modified constructs for reduced toxicity</td>
<td>Hu19.CD8.28Z</td>
<td>CD19</td>
<td>Human-CD28</td>
<td>NCT02659943</td>
<td>DLBCL, trDLBCL</td>
<td>19</td>
<td>ORR, EFS, CRS, NT</td>
<td>Brudno2020 [16]</td>
</tr>
<tr>
<td></td>
<td>CD19, 8Bz.86</td>
<td>CD19</td>
<td>4-1BBz.86</td>
<td>NCT02842138</td>
<td>DLBCL, trDLBCL</td>
<td>21</td>
<td>ORR, DOR, CRS, NT</td>
<td>Ying2019 [17]</td>
</tr>
<tr>
<td>ASCT+ CAR T-cell</td>
<td>Sequential ASCT and CD19, CD28</td>
<td>CD19</td>
<td>CD28</td>
<td>NCT01497184, NCT01840566, NCT01318317</td>
<td>DLBCL, trDLBCL, RS</td>
<td>7, 13, 4</td>
<td>ORR, PFS, CRS, NT</td>
<td>Kebriaei2016 [19], Sauter2019 [20], WangX2016 [21]</td>
</tr>
<tr>
<td>Alternative target antigen</td>
<td>CD20. 4-1BB</td>
<td>CD20</td>
<td>4-1BB</td>
<td>NCT01735604, NCT01735604</td>
<td>DLBCL, trDLBCL</td>
<td>6, 8</td>
<td>ORR, PFS, CRS, NT</td>
<td>WangY2014 [38], Zhang2016 [40]</td>
</tr>
<tr>
<td>Alternative co-stimulatory domain</td>
<td>CD19. 4-1BB</td>
<td>CD19</td>
<td>4-1BB</td>
<td>NCT03156101, ChiCTR1500076</td>
<td>DLBCL, HGBCL, trDLBCL</td>
<td>10, 14</td>
<td>ORR, PFS, CRS, NT</td>
<td>Chen2020 [59], WangT2016 [60]</td>
</tr>
</tbody>
</table>

ASCT - autologous stem cell transplantation; CAR - chimeric antigen receptor; CD - cluster of differentiation; ChiCTR - Chinese clinical trial registry; CRR - complete response rate; CRS - cytokine release syndrome; DLBCL - diffuse large B-cell lymphoma; DOR – duration of response; HGBCL – high-grade B-cell lymphoma; Hu - human; NCT - national clinical trial; NT - neurotoxicity; ORR - objective response rate; OS – Overall survival; PFS - progression-free survival; EFS – event-free survival; RS - Richter's transformation to DLBCL; trDLBCL - transformed DLBCL.
The trials for CAR T-cells that eventually evolved into Yescarta, and any early phase trials of the CAR T-cells developed into Kymriah and Breyanzi, currently approved products, were also excluded from this study. We have previously published the study protocol describing the study selection, data extraction, and risk of bias assessment process for the present study (Weinstein et al., 2021.).
Reconstructed patient-level progression-free survival (PFS) data for ZUMA-1 trial: for the calculation of the hazard ratio (HR) and its 95% confidence interval (CI) associated with the PFS of each pooled CAR T-cell population versus Yescarta, we reconstructed individual patient PFS data from the ZUMA-1 trial through a validated algorithm developed by Guyot and colleagues (2012) (Guyot et al., 2012). This was achieved by obtaining the number of patients at risk and the total number of events along with the geometric coordinates of the published PFS Kaplan-Meier (KM) curve associated with Yescarta over a 24-month follow-up time using Origin digitizing software.

4.3.4 Outcomes assessed

We chose PFS for efficacy outcome given that the purpose of the study was to determine the relative benefit of experimental CAR T-cell products compared to Yescarta in terms of response durability. The reasons why we chose PFS as the primary endpoint in this study are as follows. Firstly, one of the key therapeutic challenges in CAR T-cell therapy is that almost half of the patients experience the disease relapse after the initial infusion (Shah & Fry, 2019). In addition, 12 out of eligible 16 clinical trials in the current analysis have not reached the median follow-up time of overall survival time. Given lower grade toxicities of CAR T-cell therapies are more rapidly reversible and relatively well managed, we chose grade ≥ 3 cytokine release syndrome (CRS) and neurotoxicity (NT) events for safety outcomes.

Table 4.1 presents various endpoints reported by the contributing studies. Overall survival (OS) was not used since multiple studies had not reached the median follow-up time at the time of this analysis. Likewise, we did not focus on the objective response rate or the initial response types, as these
measures do not directly reflect the durability of response over time, and there is notable cross-trial variation in the timing of response measurements.

4.3.5 Statistical analysis

Given the existing evidence limited to single-arm trials, we conducted unanchored MAICs to adjust for cross-trial heterogeneity in baseline characteristics. In MAIC, patients in experimental CAR T-cell trials with IPD were re-weighted to match the mean baseline characteristics in ZUMA-1 with only aggregate data. The weights were estimated by the method of moments, applied to the IPD, so the summary statistics of the baseline characteristics of the IPD becomes similar to those of the aggregate data (Signorovitch et al., 2012).

Based on the calculated weights, individual patient-level PFS and percentage of grade ≥3 CRS and NT were re-weighted for further survival and logistic regression analyses. For the comparator arm, reconstructed individual patient level PFS for Yescarta (see Materials section) was used. Given these data, Cox proportional hazards (PH) model estimated the HR and its 95% CI for PFS associated with each pair of eight pooled CAR T-cell populations versus Yescarta. Corresponding weighted KM curves were created assuming exact ties according to the Kalbfleisch-Prentice method. Finally, logistic regression models were used to compute the odds ratio (OR) and its 95% CI based on the re-weighted data for both safety outcomes: grade ≥ 3 CRS and NT. Table 4.2 presents the effective sample size (ESS) and the weighted versus unweighted values of matching baseline covariates across each pooled CAR T-cell population versus ZUMA-1.
### Table 4.2. Key Baseline Characteristics and MAICs of Experimental CAR T-cells versus Yescarta regarding Progression-Free Survival †

<table>
<thead>
<tr>
<th>Intervention strategies</th>
<th>Pooled populations</th>
<th>Mutually reported variables</th>
<th>ZUMA-1 (% median age)</th>
<th>Post / Pre - weighting (% median age)</th>
<th>N patients, pooled population</th>
<th>ESS</th>
<th>PFS, HR (95CI%)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dual targeting</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1). Tandem CD19, CD20 with 4-1BB</td>
<td></td>
<td>DBCL, Prior chemo ≥3, CD19 status, CD4 &amp; CD8 ratio, Relapse after ASCT</td>
<td>0.76 / 0.70, 0.90 / 0.48, 0.21</td>
<td>0.76 / 0.86, 0.90 / 0.86, 0.48 / 0.46, 0.21 / 0.16</td>
<td>33 / 25, 0.58 (0.33-1.01)</td>
<td>Shah2020 [32], Tong2020 [31]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2). Co-infusion CD19 &amp; CD20 with 4-1BB</td>
<td></td>
<td>Age, median, Prior chemo ≥3, Disease Stage I or II</td>
<td>58 / 0.70, 0.15</td>
<td>58 / 0.70, 0.57, 0.15 / 0.14</td>
<td>21 / 15, 1.33 (0.70-2.54)</td>
<td>Sang2020 [33]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Third generation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3). CD19 with CD28 &amp; 4-1BB</td>
<td></td>
<td>DBCL, Prior chemo ≥3, Refractory</td>
<td>58 / 0.76, 0.78, 0.70</td>
<td>58 / 0.70, 0.69, 0.78 / 0.77, 0.70 / 0.69</td>
<td>26 / 23, 0.85 (0.43-1.66)</td>
<td>Ramos2018 [15], Enblad2018 [13], Huang2020 [14]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Modified constructs for reduced toxicity</strong></td>
<td></td>
<td>DBCL, Prior chemo ≥3, Relapse after ASCT</td>
<td>0.76 / 0.70, 0.70, 0.21</td>
<td>0.76 / 0.74, 0.78 / 0.63, 0.70 / 0.63, 0.21 / 0.26</td>
<td>19 / 17, *2.00 (1.01-3.96)</td>
<td>Brudno2020 [16]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4). Hu19.CD8.28Z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5). CD19, BBr.86</td>
<td></td>
<td>DBCL, Prior chemo ≥3, Relapse after ASCT</td>
<td>58 / 0.76, 0.78, 0.70, 0.76</td>
<td>58 / 48, 0.78 / 0.71, 0.70 / 0.71, 0.76 / 0.76</td>
<td>21 / 20, 1.67 (0.90-3.09)</td>
<td>Ying2019 [17]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ASCT+ CAR T-cell</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6). Sequential ASCT and CD19. CD28</td>
<td></td>
<td>DBCL, Prior chemo ≥3, Refractory</td>
<td>58 / 0.76, 0.78, 0.21</td>
<td>58 / 58, 0.76 / 0.77, 0.78 / 0.38, 0.21 / 0.42</td>
<td>24 / 13, 0.73 (0.30-1.74)</td>
<td>Kebriaei2016 [19], Sauter2019 [20], WangX2016 [21]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Alternative target antigen</strong></td>
<td></td>
<td>DBCL, Prior chemo ≥3, Disease Stage I or II</td>
<td>58 / 0.70, 0.15</td>
<td>58 / 0.61, 0.643, 0.286</td>
<td>14 / 13, 1.04 (0.52-2.06)</td>
<td>WangY2014 [38], Zhang2016 [40]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7). CD20. 4-1BB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8). CD19. 4-1BB</td>
<td></td>
<td>Male, Extramal disease</td>
<td>58 / 0.68, 0.70</td>
<td>58 / 0.43, 0.68 / 0.63, 0.70 / 0.50</td>
<td>24 / 11, 0.47 (0.18-1.28)</td>
<td>Chen2020 [59], WangT2016 [60]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

† Hazard ratio (HR) and 95% CI (confidence interval) based on Cox proportional hazards models. ASCT - autologous stem cell transplantation; CAR - chimeric antigen receptor; CI - confidence interval; CD - cluster of differentiation; ESS - effective sample size; HR - hazard ratio; Hu - human; MAIC - matching adjusted indirect comparison; PFS - progression-free survival; *statistical significance at α=0.05; ZUMA-1 – name of Yescarta clinical trial.
A recent MAIC study of Yescarta vs. Kymriah identified the LBCL-specific key prognostic covariates and demonstrated refractory status and number of prior therapies as the most influential variables on the CAR T-cell treatment outcomes (Oluwole et al., 2020). We identified the mutually reported key baseline covariates and used categorizations as follows: age (<58 years), disease stage (<3), histology (diffuse LBCL/other types), refractory status, number of prior lines of therapy (>4), and extranodal disease status. The pack of mutually reported covariates varied for each pair of distinct pool CAR T-cell population and Yescarta, as this was dictated by the size of the IPD pooled population and the mutual availability of the data in both the IPD and ZUMA-1 trials, as shown in Table 4.2. The degree of overlap between pairwise comparisons reflects in ESS (Table 4.2). All analyses were performed using R version 4.1.0 (2021). The survival package, along with the necessary supporting functions, was used to estimate alternative survival functions by trial.

4.4 Results and Discussion

Dual targeting strategies versus Yescarta: The MAIC weighted dual-targeting approach using tandem CD19.CD20.4-1BBζ CAR T-cells presented suggestive evidence of increased PFS (HR = 0.58; 95% CI, 0.33-1.01), reduced grade ≥ 3 CRS (OR=0.70; 95% CI, 0.18-2.76) and statistically significantly lower odds of grade ≥ 3 NT (OR=0.14; 95% CI, 0.02-0.78) compared to Yescarta (Table 4.2 & 4.3).
Table 4.3. MAICs of Experimental CAR T-cells versus Yescarta Regarding Grade ≥3 CRS and NT

<table>
<thead>
<tr>
<th>Intervention strategy</th>
<th>Pooled populations</th>
<th>N infusion</th>
<th>Sum of weights</th>
<th>Number of CRS, grade ≥ 3</th>
<th>OR (95%CI) CRS, grade ≥ 3</th>
<th>Number of NT, grade ≥ 3</th>
<th>OR (95%CI) NT, grade ≥ 3</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual targeting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1). Tandem CD19. CD20. 4-1BB</td>
<td>33</td>
<td>28</td>
<td>3</td>
<td>0.70 (0.18-2.76)</td>
<td>2</td>
<td>*0.14 (0.02-0.78)</td>
<td>Shah2020 [32] Tong2020 [31]</td>
<td></td>
</tr>
<tr>
<td>2). Co-infusion CD19 &amp; CD20 with 4-1BB</td>
<td>21</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Sang2020 [33]</td>
</tr>
<tr>
<td>Third generation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3). CD19 with CD28 &amp; 4-1BB</td>
<td>26</td>
<td>26</td>
<td>1</td>
<td>0.20 (0.02-2.12)</td>
<td>1</td>
<td>*0.20 (0.04-0.94)</td>
<td>Ramos2018 [15] Enblad2018 [13] Huang2020 [14]</td>
<td></td>
</tr>
<tr>
<td>Modified constructs for reduced toxicity</td>
<td>4). Hu19. CD8.28Z</td>
<td>14</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Brudno2020 [16]</td>
</tr>
<tr>
<td>5). CD19. BBz.86</td>
<td>21</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Ying2019 [17]</td>
</tr>
<tr>
<td>ASCT+ CAR T-cell</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternative target antigen</td>
<td>7). CD20. 4-1BB</td>
<td>14</td>
<td>13</td>
<td>2</td>
<td>1.04 (0.20-5.38)</td>
<td>NA</td>
<td>NA</td>
<td>WangY2014 [38] Zhang2016 [40]</td>
</tr>
<tr>
<td>Alternative co-stimulatory domain</td>
<td>8). CD19. 4-1BB</td>
<td>24</td>
<td>12</td>
<td>4</td>
<td>0.95 (0.15-5.94)</td>
<td>NA</td>
<td>NA</td>
<td>Chen2020 [59] WangT2016 [60]</td>
</tr>
</tbody>
</table>

ASCT - autologous stem cell transplantation; CAR - chimeric antigen receptor; CI - confidence interval; CD - cluster of differentiation; CRS - cytokine release syndrome; OR - odds ratio; Hu - human; MAIC - matching adjusted indirect comparison; *statistical significance at α=0.05; NT - neurotoxicity; ZUMA-1 – name of Yescarta clinical trial.
In contrast, co-infusion of CD19 and CD20 CAR-T cells had statistically insignificant but worse PFS (HR=1.33, 95% CI: 0.70-2.54) than Yescarta. IPD for safety outcomes was not available from this study. However, naive direct comparison shows this co-infusion approach presented higher grade ≥ 3 CRS (28.5% vs. 13% in ZUMA-1) and lower neurotoxicity (9.5% vs. 28% in ZUMA-1) than Yescarta. The improved PFS associated with tandem CAR T-cells whereas reduced PFS associated with co-infusion of CD19 and CD20 CAR T-cells versus Yescarta corroborates pre-clinical studies that demonstrated the higher efficacy and safety of tandem CAR T-cells than that of co-infusions (Ruella et al., 2016). The reduced survival benefit and increased CRS associated with the co-infusion of different CAR T-cell targets may be associated with (1) additive toxicity from stronger cytokine storm through the amplified number of targetable antigens; (2) competitive targeting limit the expansion of other CAR T-cells; (3) compromised engraftment due to the interference of multiple antigens (Guo et al., 2021; Shah et al., 2019; Shah & Fry, 2019).

Among the alternative strategies to target more than one antigen receptor at a time, more comprehensive data is available for CD19 and CD22 covering various administration approaches (sequential and co-infusion) and constructs (tandem and bicistronic). Sequential administration of CD19 and CD22 CAR T-cells in 12 DLBCL patients (Zeng et al., 2020) and co-infusion in 36 NHL patients (N. Wang et al., 2020) resulted in objective response rates of 77% and 83% and grade ≥3 CRS rates of 14% and 21%, respectively. Tandem CD19 and CD22 CAR T-cells appeared to be feasible and potentially efficacious in R/R B-cell acute lymphoblastic leukemia (B-ALL) (Hossain et al., 2018). Although these preliminary results are somewhat comparable to the 82% ORR and 13% grade ≥3 CRS observed in ZUMA-1, longer follow-up data are required to assess whether sequential and mixed infusion approaches reduce post-CAR T-cell therapy relapse. Bicistronic CD19. CD22 trials in pediatric and adult R/R B-ALL (Amrolia et al., 2019; Yang et al., 2018) demonstrated unprecedentedly high CR rates (100%) and a notable safety profile with a single occurrence of grade ≥3 CAR T-cell related encephalopathy.
syndrome (CRES) in the pediatric trial. As for B-cell lymphoma, bicistronic CD19.CD22 trials led by Shah and colleagues (NCT03448393), Miklos and colleagues (NCT03233854), and Pulsipher and colleagues (NCT03330691) are currently in progress.

**Third-generation CAR T-cells versus Yescarta:** We found suggestive evidence of improved PFS with HR=0.85 (95% CI: 0.43-1.66) and safety in terms of grade ≥3 CRS with OR=0.20 (95% CI: 0.02-2.12) and NT with OR=0.20 (95% CI: 0.04-0.94) associated with third-generation CAR T-cells versus Yescarta. Although this was not statistically significant, the slight protective effects observed may be due to the multifunctional cytokine secretion and improved persistence of T-cells from the concurrent expression of CD28ζ and 4-1BBζ co-stimulatory domains versus that of the CD28ζ co-domain alone (Carpenito et al., 2009). In addition, in-vivo studies demonstrated that adding 4-1BBζ to the second-generation construct protects CD28ζ tumor-specific cells from activation-induced cell death while supporting central memory cells and mitochondrial functions (Gomes-Silva et al., 2017). In alignment with this existing evidence, three contributing IPD trials of the pooled population in our study reported high expansion and improved persistence of T-cells in common. Moreover, all three trials of third-generation CAR T-cells analyzed in this study have highlighted that the patients with less tumor burden and who were prior responders to chemotherapy had higher tumor clearance benefits than patients with more tumor burden and non-responders to chemotherapy (Enblad et al., 2018; C. Huang et al., 2020; Ramos et al., 2018).

**Anti-CD19 CARs with modified co-stimulatory domains for reduced toxicity:** We separately compared Hu19.CD8.28Z and CD19. BBz.86 CAR T-cells designed to minimize toxicity while preserving antitumor potency to Yescarta. Both products presented excellent safety profiles with the absence of grade ≥3 CRS and NT. However, both CAR T-cells presented reduced PFS: Hu19.CD8.28Z with HR=2.00; 95% CI, 1.01-3.96 and CD19. BBz.86 with HR=1.67; 95% CI, 0.90-3.09 compared to Yescarta, though without statistical significance in CD19. BBz.86. Reduced cytokine-mediated toxicity is often
accomplished by attenuating CAR signal strength and enhancing T-cell persistence (Brudno et al., 2020; Feucht et al., 2019; Ying et al., 2019). Consequently, this enables tumor immune escape and hampers the antitumor potency of CAR T-cells, particularly for low antigen density tumors (Majzner et al., 2020). Multiple pre-clinical studies are underway toward determining the CAR structure that might achieve minimum toxicity and maximum efficacy for low antigen density tumors. A recent in-vivo leukemia model demonstrated the high potency of a new CD19. CD28ζ/T-4-1BBζ construct, despite the low antigen density of the leukemic cells, while accomplishing a similar efficacy to Yescarta (Majzner et al., 2020). Further studies are required to assess how these pre-clinical findings translate into clinical benefits for lymphoma patients.

**ASCT + CD19. CD28ζ versus Yescarta:** We conducted a MAIC of three small clinical trials that evaluated the safety and efficacy of the sequential administration of ASCT and CD19. CD28ζ CAR T-cell therapy in LBCL patients (Kebriaei et al., 2016; Sauter et al., 2019; X. Wang et al., 2016). The authors hypothesized that the sequential administration of ASCT and CAR T-cell reduces cytokine production and enhances antitumor potency by increasing the expansion and persistence of CAR T-cells. ASCT prior to CAR T-cell administration is believed to reduce tumor burden, diminish immuno-suppressive microenvironment, and boost lymphodepletion, thereby reducing the number of regulatory T-cells and myeloid cells. Our findings demonstrated suggestive favorable PFS (HR=0.73; 95% CI, 0.30-1.74) and reduced grade ≥ 3 CRS (OR=0.25; 95% CI, 0.02-3.68) but increased NT (OR=1.78; 95% CI, 0.60-5.28) for the sequential administration of CD19. CD28ζ CAR T-cells within 2 to 6 days after ASCT compared to Yescarta. None of these findings reached statistical significance. The direction of the results remained unchanged in separate analyses for the U.S. and Chinese trials.

Nevertheless, the feasibility and clinical benefit of the concurrent administration of ASCT with CAR T-cell therapy may not be justifiable since this combined regimen would be limited to only about half of the patients who are transplant-ineligible due to chemo-refractory disease and half of those who
received ASCT yet still at risk for disease relapse post-autografting (Philip et al., 1995). Of note, combining ASCT with CAR T-cell therapy may not be necessary if CAR T-cell therapy alone is superior to ASCT, as previously shown (Li et al., 2019). Intensive efforts are underway to understand whether CAR T-cell therapy is efficacious and safe to replace ASCT in earlier lines of treatment of LBCL. A few randomized multi-center clinical trials are underway comparing the FDA-approved CAR T-cells - Yescarta in the ZUMA-7 (NCT03391466), Kymriah in the BELINDA (NCT03570892), and Breyanzi in the TRANSFORM (NCT03575351) trials versus the standard of care comprised of systemic therapies followed by ASCT.

**CD20. 4-1BBζ versus Yescarta:** We did not find any notable clinical benefit or harm in this pooled population versus Yescarta in terms of PFS (HR= 1.04; 95% CI, 0.52-2.06) and CRS (OR=1.04; 0.20-5.38). IPD for neurotoxicity was not available from this trial. CD19 has been a primary target in CAR T-cell therapy for LBCL due to its pan B-cell expression and increased expression in B-cell leukemias and lymphomas (June & Sadelain, 2018). In contrast, CD20 and CD22 have limited expression in mature B cells. Nevertheless, targeting both CD19 and CD20 has an additive effect, given CD20 antigen’s higher average density of surface molecules per tumor cell, combined with CD19’s pan B-cell lineage cell expression, with extended-expression in certain CD20-negative tumor subsets (Horna et al., 2019). Since all patients in this trial were treated with rituximab before CD20.4-1BBζ CAR T-cell administration, the question of whether CD20. CAR T-cells would be more efficacious among rituximab-näive patients remain to be clarified.

**CD19. 4-1BBζ in Chinese patients versus Yescarta:** The MAIC of CD19. 4-1BBζ CARs based on the pooled population of two small trials conducted in China, Shanghai (Chen et al., 2020; T. Wang et al., 2020) versus Yescarta showed no significant difference but suggestively better PFS than that of Yescarta, with HR=0.47 (95% CI, 0.18-1.28) and slightly reduced grade ≥ 3 CRS (OR=0.95; 95% CI, 0.15-5.94) than Yescarta. These two trials did not provide neurotoxicity data. Despite the small sample size, the slight
improvement in safety was associated with the CD19. 4-1BBζ CAR T-cells versus Yescarta in these studies is consistent with the recent MAIC of Kymriah and Breyanzi to Yescarta (Oluwole et al., 2020). 4-1BBζ is one of the well-established co-stimulatory domains incorporated into the CD19 CARs in Kymriah and Breyanzi, while Yescarta contains a CD28ζ co-domain. The impact of the CD28ζ versus 4-1BBζ co-stimulatory domains on CAR T-cell behavior has been studied in in-vivo studies and multiple clinical studies in B-ALL. CD19. CD28ζ CAR T-cells show a faster and higher peak expansion, yet reduced T-cell persistence compared to 4-1BBζ-containing CARs (Majzner & Mackall, 2019). Nonetheless, it is not fully clear whether CAR T-cell persistence is a strong determinant of response durability in LBCL as it is for B-ALL.

A recently published MAIC of Yescarta versus Kymriah demonstrated superior efficacy of Yescarta, with higher CR rates (RR=1.62, 95% CI: 1.16-2.27) and improved OS (HR=0.51, 95% CI: 0.31-0.83), yet increased toxicity, with grade 1-2 CRS with OR = 6.20 (95% CI: 2.76-13.93) and grade ≥ 3 NT with OR=2.20 (95% CI: 0.98-3.60) in R/R LBCL (Oluwole et al., 2020). Another recently published MAIC of Yescarta versus Breyanzi demonstrated similar efficacy, slightly favoring Yescarta (PFS with HR=1.30; 95%, 0.96-1.77). However, Breyanzi presented a significantly safer profile than Yescarta (grade ≥3 CRS and NT with OR= 0.16; 95% CI, 0.06-0.47, and OR=0.31, 95% CI, 0.18-0.54, respectively (Maloney et al., 2020). Among the currently approved CAR T-cells, based on the existing MAICs, Yescarta appears to present higher efficacy than Kymriah and comparable efficacy to Breyanzi. In contrast, the latter two incorporating 4-1BBζ co-domains demonstrate a safer profile regarding CRS and NT than Yescarta. The lower toxicity and similar efficacy observed with Breyanzi versus Yescarta relates to its ability to induce a low variability in cytokine production (IL-2, IFN-γ, TNF-α, etc.). This was accomplished through controlled manufacturing to maintain the ratio of CD4+ and CD8+ to 1:1 under optimized culture conditions. In-vivo studies are underway to clarify the exact underlying mechanisms in this regard (Ramsborg et al., 2017).
**Strengths, Limitations, and Future study**

To our knowledge, this study is the first to report the indirect comparison of experimental CAR T-cells to Yescarta, an FDA-approved CAR T-cell product. This study incorporated a systematic literature review with MAICs, the only statistical tool to address studies in the absence of direct head-to-head comparisons and the presence of single-arm trials only. Consequently, we were able to systematically retrieve comprehensive data from an IPD bank of experimental CAR T-cells. Although a larger case series are needed to confirm these results, our findings are biologically plausible and clinically meaningful while corroborating with existing pre-clinical and clinical data. This study was limited to the eligible trials' follow-up times, including those that have not yet reached their median follow-up times. However, this limitation is less of an issue in our study, given that our primary outcome was PFS and not OS. However, we could not draw any definitive conclusions from our analysis due to the lack of statistical significance resulting from the small sample sizes of the contributing trials in the statistical analysis.

Even though we attempted to account for cross-trial heterogeneity using MAICs, it is important to acknowledge the residual case-mix and beyond case-mix heterogeneity (Vo et al., 2019). First and foremost, clinical trials vary in terms of their inclusion/exclusion criteria, individual trial management, study design, protocol, conditioning regimes, bridging therapies, CAR T-cell engineering, and manufacturing processes. One of direct therapeutic factors, a variation in the individual CD19 and CD20 tumor antigen expression levels at baseline was not accounted in the statistical analyses as the necessary IPD was only reported by two out of all eligible studies. Moreover, ZUMA-1 did not use bridging therapy, as opposed to some of the eligible IPD trials in this study that used bridging therapy. It is unclear as to how this may have affected our results since the role of bridging chemotherapy in the CAR T-cell setting is not fully understood and subject to multiple confounding factors.
Furthermore, methodological limitation stems from the fact that the MAIC method assumes all key prognostic factors differentially distributed across studies are taken into account (Phillippo et al., 2018; Vo et al., 2019). However, we could not adjust for all important prognostic factors since eligible studies reported different patient characteristics to describe their study samples, which limited the number of common key baseline covariates reported by experimental CAR T-cell common with ZUMA-1. For example, two important key baseline covariates for R-R DLBCL that lacked in experimental CAR T-cell trials were International Prognostic Index (IPI) and Eastern Cooperative Oncology Group (ECOG) Performance Status. Even when a similar patient characteristic was reported across studies, it was often measured by different scales in different papers, which makes the adjustment by MAIC impossible, such as International Prognostic Indexes (IPI), age-adjusted IPI and Revised-IPI. More serious inconsistency was found across the trials were the differential definition of relapsed disease as a baseline characteristic between ZUMA-1 and eligible IPD trials in this study. ZUMA-1 defined refractoriness as patients who had stable disease (SD) as their best response to the last line of therapy or those who had relapsed within 12 months of a consolidative ASCT. In IPD trials, besides using the same definition as that of the ZUMA-1 trial, progression at any time after the last line of therapy is also included as a criterion for relapsed disease. Therefore, we categorized patients as either refractory or relapsed in the MAICs irrespective of the type of relapse.

Regarding the MAIC of safety outcomes, IPD trials used different grading systems for CRS and NT from ZUMA-1, which used the Lee criteria (Lee et al., 2019). Hence, only a few important and mutually reported variables were adjusted for in the analysis. Consequently, any hidden difference in other unmeasured patient characteristics across studies could invalidate the findings. Apart from simple inverse weighting, advanced statistical methods based on doubly robust estimation have been developed to adjust for between-trial heterogeneity in patient characteristics (Vo et al., 2019). These approaches require specifying one model for the weight and another model for the outcome of interest.
The advantage is that only one of these two models needs to be correctly specified to obtain valid and less biased results. Despite being more robust statistical solutions beyond the simple inverse weighting approach (used in MAIC), these methods were not used in the current study because they were not feasible in our study due to the limited sample size. Extending this study to other currently approved CAR T-cells, such as Tisagenlecleucel (Kymriah) and Lisocabtagene maraleucel (Breyanzi), may further contribute to clarify these findings.

4.5 Conclusion

In conclusion, The MAIC suggests a dual targeting approach using tandem CD19.CD20.4-1BB may have enhanced efficacy and safety compared to Yescarta. The hazard ratios of PFS were quantitatively in favor of the third generation, the sequential administration of ASCT and CD19.CD28 CAR T-cells and of the CD19. 4-1BBζ manufactured and evaluated among Chinese patients, although none of them reached statistical significance. The safety-enhanced CAR T-cell constructs included in our analysis, such as Hu19. CD8.28Z and CD19. BBz.86 demonstrated a remarkable safety profile with no occurrence of severe adverse events, yet without improvement in PFS compared to that of Yescarta.

4.6 References


CHAPTER 5. DISCUSSION

5. 1 Ecological Study Design in Environmental Health and Cancer Research

5.1.1 Summary

Chapter 2 presented an ecological study comparing individuals living in areas near polluted waste sites with individuals not leaving near such sites. During 18 years, 1996-2004, New York State, excluding New York City, was studied and 28,941 individuals had been discharged from hospital after a pancreatic cancer diagnosis. Individuals having a ZIP code near a polluted waste site had somewhat higher pancreatic cancer risk than others, RR 1.06 (95%CI 1.03-1.09) for VOC sites and 1.05 (1.01-1.08) for POPs. Living near a site with ethylbenzene pollution showed a higher ratio, 1.34. Potential confounders were managed by adjustment for the race (whites, blacks), age (four groups), and area-poverty status. The results indicate that residential exposure to both volatile and semi-volatile organic compounds coming from identified hazardous waste sites is associated with an elevated risk of being hospitalized for exocrine pancreatic cancer. Accounting for the limitations inherent in the ecological study design, these results warrant the need for further study of the contribution of exposure to organic chemicals to pancreatic cancer, at a minimum.

5.1.2 Methodological Challenges and Opportunities

The ecological studies are particularly useful for testing hypotheses about the disease and unknown risk factors. However, the ecological study design is often not appropriate to infer causal causality as observed associations could be biased by the ecological fallacy (Schwartz, 1994).

The ecological fallacy refers to the limitations to make inferences about individuals based on the
analyses of a group of people. Individuals significantly vary in terms of the important risk factors that may modify or confound the relationship between the health outcomes and exposure under the study. In the current analysis which lacked individual patient-level data, it was impossible to account for major lifestyle risk factors, such as BMI, smoking, and excessive alcohol consumption known to be associated with pancreatic cancer as well as potentially important comorbidities such as, diabetes mellitus and higher blood pressure. The individual patient level comorbidity and behavioral risk data are often not available from large administrative databases, the hospital discharge data in this study. For a better assessment of potential confounding risk, researchers attempt to clarify the distribution of key risk factors in the study population by using the best available population-based surveillance data, such as The Behavioral Risk Factor Surveillance System (BRFSS) (https://www.cdc.gov/brfss/about/index.htm).

The current study demonstrated common challenges that population-based environmental health studies face including lack of certainty in environmental exposure assignment to the study population and lack of population mobility data to help determine the duration of exposure. These problems stem from a lack of primary and dedicated data sources that provide ambient air pollutants concentration for organic chemical exposures concerned in the study. The additional problem arises as these organic chemicals occur in a mixture with other pollutants that make measuring and tracking these chemicals in ambient air near residential settings almost impossible (Carpenter et al., 1998). Lack of ambient air pollution data availability combined with their mixed distribution reduces the specificity of exposure classifications in ecological studies as demonstrated in this chapter.

Historically, the necessity for an adequate technique of detecting cancer risk factors motivated the development of the case-control research design (Cornfield, 1951). For a better assessment of disease and exposure relationship case-control and cohort studies have been further developed. However, in the study of rare diseases such as pancreatic cancer concerning the exposure to specific environmental pollutants, both study designs have their unique advantages and limitations. Major
strengths of prospective cohort studies, when properly designed, include the possibility to infer causality given researchers can identify whether exposure is seen to occur before outcome and also the opportunity to measure if any changes in both exposure and outcome due to study subjects are followed up longitudinally. Nevertheless, prospective cohort studies can be extremely costly, especially in the study of cancer, the cohort population may need to be followed for decades. Also, when studying rare outcomes, such as pancreatic cancer, a very large sample size would be required to derive confident inferences which are often further complicated by the increased dropout of subjects over a long study period. Also, selection bias can occur if the difference between those who participated and those who did not participate in the study were associated with the study outcome.

In comparison to cohort studies, case-control studies are more affordable and are particularly suited to the study of rare diseases, such as pancreatic cancer as the diseased are selected at the outset of the study. However, the validity of the case-control study is prone to a successful selection of both cases and controls that adequately represent their respective populations and most importantly its inability to infer causality given the absence of chronology of disease and exposure (Crombie, 1981; Levin, 2006a, 2006b). To circumvent these methodological problems, several alternative research designs have been suggested.

Certain research designs, such as nested case-control and case-cohort studies, are especially useful for investigating rare diseases, such as cancers in general. Both study designs combine the strengths of both cohort and case-control studies while reducing the costs of exposure assessment in a prospective setting. The primary distinction between nested case-control and a case-cohort studies is how the controls are selected (Kirch, 2008). In a case-cohort study, cases are identified from the cohort members who developed the disease while controls are ascertained prior to the cases identified. In other words, controls are randomly chosen from all cohort participants despite they developed the
disease under the study or not, and that baseline data can be collected at the beginning of the study. Whereas nested case-control study identifies the cases that occurred in the study cohort and ascertain the number of matched control(s) from among those in the cohort who are free from disease at the time of disease occurrence in the case (Ernster, 1994).

The primary benefit of a case-cohort design over a nested case-control design is that in case-cohort research, the same control group may be utilized for comparison with multiple different case groups. However, nested case-control design is more efficient than case-cohort design in a study of cancer with a long follow-up time. In addition, nested case-control studies don’t require complex statistical techniques as often needed for case-cohort studies (Ernster, 1994; Kirch, 2008).

The applied example of these advanced study designs is the recently published study that used a case-control design nested within two prospective, general population cohorts, such as the Kaiser Permanente Multiphasic Health Checkup (MHC) cohort from the United States and the Janus Serum Bank cohort from Norway (Engel et al., 2019). The study examined the liver cancer risk in relation to the serum organochloride pesticides measured among the subjects of two cohorts back in the 1960s and 1970s, respectively. In addition to a robust study design, the strength of this study was the direct individual patient level biological measurement of exposure collected prior to liver cancer diagnosis and adjustment for important potential confounders, including smoking history, BMI, alcohol consumption, and race. Also, it is noteworthy that the study excluded the cases diagnosed within one year of blood collection in order to avoid potential bias due to chemical interaction between the chemotherapies and serum OC concentrations (Engel et al., 2019). Multiple serum measurement studies reported that chemotherapy reduces the serum concentration of organochlorines including in NHL cases (Baris et al., 2000).
5.2 Mapping the COVID-19 Vulnerable Locales by Epidemiological and Socioeconomic Risk Factors

5.2.1 Summary

Chapter 3 presented the study to map COVID-19 vulnerable locales by epidemiological and socioeconomic factors. We demonstrated that vulnerable areas in terms of healthcare access, health behavior, crowding, area morbidity, education, difficulty to social distancing, population mobility were associated with increased COVID-19 incidence rates (Weinstein et al., 2021).

We chose South Korea to learn from their successful anti-pandemic practice and exemplary COVID-19 data transparency. Considering previous pandemics’ impacts, COVID-19 is comparable to the health and economic damage caused by the Spanish Flu in 1918 (Javelle & Raoult, 2021). The South Korean anti-pandemic strategy was considered one of the most effective strategies worldwide. The country accomplished to break the virus transmission without business shutdowns and lockdowns as many other countries used (J. R. Kim & Choi, 2021). Moreover, South Korea had its presidential election during the severe pandemic period yet without a surge in new cases afterward. This made the country’s anti-pandemic efforts a subject to study and learn from.

COVID-19 has corroborated insights gained from SARS, H1N1 influenza, and MERS pandemics showing socioeconomically disadvantaged populations are more severely impacted by pandemics (de Lusignan et al., 2020; S. Kim & Kim, 2018; Quinn et al., 2011; Stone et al., 2010). The South Korea gained pandemic mitigation experience from MERS which hit the country hard in 2016. During COVID-19 epidemic, the country prioritized the vulnerable population and promptly provided social and mediclas supports which could have contributed positively to their overall anti-pandemic success. Our study results corroborate this finding by showing that the strength of the association between the COVID-19 incidence and socioeconomic risk factors reduced over time potentially due to the protective effects of public health and medical interventions with a strong focus on the vulnerable populations. In addition, the current analysis chose South Korea because its COVID-19 incidence data presented extremely high
overdispersion, temporality, and spatial clustering, being more complex than typical infectious disease data. This allowed us to check our framework’s functionality to address the dramatic spatial and temporal dynamic of COVID-19.

5.2.2 Methodological Highlights

The highlight of the analytical approach used in the current analysis as follows:

- The focus is on providing the robust SES measurement based on the established social theories, namely Coleman’s Social Theory and Blumenshine’s mechanistic framework. Guided but not arbitrary selection of epidemiological and socioeconomic factors enabled to select the risk factors that are causally relevant to the COVID-19 health outcome.

- The combined use of global and spatial statistical methods increased the accuracy as global models verified the geographical models' model fit. Secondly, we collaborated with Alan da Silva in this study, the author GWNBR technique used in the current analysis. Having guidance from the method’s author had optimized the use of method in the study data and provided interactive learning experience.

- We showcased the use of multiple composite indexes that clarifies whether certain SES and epidemiologic determinants independently contributed to the COVID-19 risk after accounting for the effects of other SES factors. To the best of our knowledge, existing studies were limited to a bi-variate model, which does not reflect real-life scenarios as individuals are affected by multiple socioeconomic factors but not one at a time.

- GWNBR used in this study has been considered as a robust modeling technique for non-stationary spatially counting data with overdispersion. Detailed information regarding the strengths of GWNBR can be found in the Introduction section (see 1.3.2.2 Model extensions, GWNBR).
Researchers considering using GWNBR, however, need to account for the current limitations of the technique. As GWNBR is relatively recently developed the limitations of this technique have not been fully understood and are currently being investigated by Dr. Silva, who co-authored the peer-reviewed article presented in this dissertation. However, currently known methodological restrictions of GWNBR, as reviewed by Dr. Silva and Gomes colleagues, the method has limited test options for the assessment of the goodness of fit of the models. Currently, it uses Akaike information criterion (AIC), whereas the use of AICc would give more accurate estimates regarding model fit, particularly for smaller samples. In addition, GWNBR is prone to the risk of overfitting due to the use of the cross-validation method to find the optimum bandwidth. This leads to the computation of bandwidths that strongly reflect the existing data at the cost of the reduction of its predictive capacity to fit for a new dataset. Finally, developing semi-parametric GWNBR models is currently not possible which would increase the versatility of GWNBR to handle situations when the variables have no distinct characteristics of spatial variations (Gomes et al., 2017). The methodology to support precision targeting of high-risk populations can bolster preventive measures to reduce the healthcare burden. The methodological framework showcased in this study can be applied to other nations or future pandemics with a minor modification to the local context.

5.3 Indirect Comparisons in Comparative Effectiveness Study

5.3.1 Summary

Chapter 4 demonstrated the application of the MAIC method in comparing the effectiveness of novel yet experimental therapies to the currently approved drug in the market. The anticipated implication of this study is to inform and inspire clinicians, CAR T-cell developers, and patients with relapsed and refractory LBCL in their journey of search for even better CAR T-cell therapies. Although
currently approved CAR T-cells demonstrated unprecedently high response in relapsed / refractory LBCL in the salvage setting, lack of outcome durability and toxicity remain to be key challenges. Various experimental CAR T-cell therapies have been proposed to overcome these limitations. Yet, to date, no data exist on how currently clinically evaluated innovative CAR T-cell approaches compare to the FDA-approved CAR T-cells.

The study demonstrated how comparative effectiveness study allows measuring the current successes against the current standard which is necessary for the advancement in the field. The study demonstrated that the highly anticipated tandem CAR T-cells may meet with the expectation in the field with higher efficacy and safer profile than the currently approved drug, Yescarta. Specific CAR T-cells that hold a high promise for reduced toxicity were shown to present remarkable safety in this study. However, there was no evidence of better clinical efficacy than currently approved drugs based on Progression-Free Survival (PFS). Even though the study results are not stable due to the small sample size, the study provides important insights based on the direction of the results combined with the biological plausibility.

5.3.2 Challenges and Opportunities in Matching Adjusted Indirect Comparisons

Matching Adjusted Indirect Comparison (MAIC) provides more robust estimates than traditional network meta-analysis methods in comparing the effectiveness of more than two medical interventions. This improved precision and accuracy stem from its use of individual patient data for the intervention arm compared to traditional aggregate models that entirely rely on the comparison of aggregate data for both arms. Commonly, researchers often have access to individual patient data for their own product while lacking the individual patient data for the other drug/intervention in the market. This hinders head-to-head comparison of drugs/interventions. However, unanchored MAIC allows to use the
of aggregate published data for the comparator arm. This helps to fill the knowledge gap to the best available data sources.

Unanchored MAIC, on the other hand, is vulnerable to residual bias due to unobserved prognostic variables and effect modifiers because of the lack of a common comparator across trials. Given MAIC is a relatively new statistical technique, studies will be required to determine the degree of error in unanchored estimates. Despite the opportunities provided with the unanchored MAIC, it is important to acknowledge this technique doesn’t eliminate the residual case-mix and beyond case-mix heterogeneity (T. T. Vo et al., 2019b). Any statistical techniques including MAIC, drugs/interventions for comparison are derived from separate studies. Independent clinical trials are inevitably different in terms of trial management, study design, protocol, therapeutic regimes, supporting therapies, drug manufacturing processes. The accuracy of MAIC depends on the availability of key prognostic factors and the extent of cross-trial difference in the data collection, measurement, and definitions of the prognostic variables (Phillippo et al., 2018; T. T. Vo et al., 2019a).

Consequently, any unknown difference in patient characteristics across studies could invalidate the findings (Phillippo et al., 2018). To improve the MAIC further, advanced statistical methods based on doubly robust estimation have been developed to adjust for between-trial heterogeneity in patient characteristics (T. T. Vo et al., 2021). These approaches require specifying one model for the weight and another model for the outcome of interest. The advantage is that only one of these two models needs to be correctly specified to obtain valid and less biased results. Despite being more robust statistical solutions beyond the simple inverse weighting approach (used in MAIC), these methods were not used in the current study because they were not feasible due to the limited sample size. The study demonstrated in Chapter 4 provides a methodological framework that can be applied to similar research problems, particularly, to the comparison of novel and experimental CAR T-cells to other FDA-approved CAR T-cells, such as Tisagenlecleucel (Kymriah) and Lisocabtagene maraleucel (Breyanzi).
5.4 References

Begg, C. B., Group, for the G. S., Hummer, A. J., Group, for the G. S., Mujumdar, U., Group, for the G. S., Armstrong, B. K., Group, for the G. S., Kricke, A., Group, for the G. S., Marrett, L. D., Group, for the G. S., Millikan, R. C., Group, for the G. S., Gruber, S. B., Group, for the G. S., Culver, H. A., Group, for the G. S., Zanetti, R., ... Group, for the G. S. (2006). A design for cancer case–control studies using only incident cases: Experience with the GEM study of melanoma. *International Journal of Epidemiology*, 35(3), 756–764. https://doi.org/10.1093/ije/dyl044


APPENDIX

Appendix A: Chapter 3 Supplementary Materials

Supplementary Materials

Table S1. Confirmed COVID-19 cases diagnosed during January 20 - July 1, 2020 in South Korea

<table>
<thead>
<tr>
<th>Province</th>
<th>Early phase</th>
<th>Middle phase</th>
<th>Late Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Busan</td>
<td>99</td>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>Chungcheongbuk</td>
<td>34</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Chungcheongnam</td>
<td>85</td>
<td>17</td>
<td>28</td>
</tr>
<tr>
<td>Daegu</td>
<td>6358</td>
<td>529</td>
<td>35</td>
</tr>
<tr>
<td>Daeguon</td>
<td>21</td>
<td>17</td>
<td>80</td>
</tr>
<tr>
<td>Gangwon</td>
<td>30</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td>Gwangju</td>
<td>48</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Gyeonggi</td>
<td>324</td>
<td>325</td>
<td>567</td>
</tr>
<tr>
<td>Gyeongsangbuk</td>
<td>1098</td>
<td>111</td>
<td>45</td>
</tr>
<tr>
<td>Gyeongsangnam</td>
<td>86</td>
<td>26</td>
<td>17</td>
</tr>
<tr>
<td>Incheon</td>
<td>38</td>
<td>47</td>
<td>253</td>
</tr>
<tr>
<td>Jeju</td>
<td>3</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Jeollabuk</td>
<td>8</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>Jeollanam</td>
<td>4</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Sejong</td>
<td>34</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Seoul</td>
<td>298</td>
<td>278</td>
<td>656</td>
</tr>
<tr>
<td>Ulsan</td>
<td>32</td>
<td>8</td>
<td>15</td>
</tr>
</tbody>
</table>

COVID-19 cases are sorted by South Korea’s provinces (comprised of 250 districts) and the pandemic phase. Early phase: January 20 to March 20, 2020; Middle phase: March 21 to April 15, 2020; Late phase: April 16 to July 1, 2020.
Table S2. Data sources reviewed and used for SES measurement

<table>
<thead>
<tr>
<th>Data sources</th>
<th>Data items</th>
</tr>
</thead>
</table>
| Korean Community Health Survey 2018, Korea Centres for Disease Control and Prevention | (% people with obesity, by measurement  
(% people drunk alcohol <1 per month  
(% people who currently smoke  
(% people obese, self-reported  
(% of people who used health care last year  
(% people who could not use healthcare when needed last year  
(% people with depression based of PHQ-9 screening  
(% of the person who is all aligned for the early symptoms of stroke  
(% of the person on insulin and other treatment specific to diabetes mellitus) |
| Health Insurance Statistics 2018, National Health Insurance Corporation       | (% people with health insurance                                                                                                           |
| Disability Status 2018, Ministry of Health and Welfare                        | (% people with severe disability                                                                                                       |
| Death Cause Statistics 2018, National Statistics Agency                      | Age adjusted mortality rate due to neoplasm  
Age adjusted mortality rate due to circulatory system disease  
Age-adjusted mortality rate from infectious parasitic diseases  
Overall, age-adjusted mortality rate  
Age adjusted mortality rate due to respiratory diseases                        |
| Korean Census Bureau 2015                                                     | (% people with high school education  
(% of foreign registered people  
Number of people per household                                                                 |
| Office of Statistics 2015, Regional Statistics                               | GDP per capita in million won                                                                                                           |
| Internal Migration Statistics 2018, Statistics Korea                           | Internal net migration between regions                                                                                                   |
| State of Urban Planning 2018, Ministry of Land, Infrastructure, Transport and Tourism | Area per capita  
Urban area per capita                                                                                                                  |

*PHQ-9: Patient Health Questionnaire - 9, standard survey tool.
Table S3. PCA details showing factor scores with their weights, and selected area health and SES variables and thematic composite indices

<table>
<thead>
<tr>
<th>Health/SE themes</th>
<th>Selected variables from national surveys</th>
<th>PCA Factor</th>
<th>Weights</th>
<th>Quartiles of composite indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthcare access</td>
<td>Healthcare utility rate</td>
<td>0.892</td>
<td>0.448</td>
<td>18.0 21.0 29.6 44.2</td>
</tr>
<tr>
<td></td>
<td>Insurance coverage rate</td>
<td>0.869</td>
<td>0.437</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Healthcare needs met when needed</td>
<td>0.226</td>
<td>0.115</td>
<td></td>
</tr>
<tr>
<td>Health behaviour</td>
<td>% people with obesity, by measurement</td>
<td>0.955</td>
<td>0.438</td>
<td>41.6 44.0 45.7 49.9</td>
</tr>
<tr>
<td></td>
<td>% people drunk alcohol, &lt; 1/month</td>
<td>0.936</td>
<td>0.429</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% people who currently smoke</td>
<td>0.220</td>
<td>0.100</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% people obese, self-reported</td>
<td>0.069</td>
<td>0.033</td>
<td></td>
</tr>
<tr>
<td>Crowding</td>
<td>* Area per capita</td>
<td>0.900</td>
<td>0.402</td>
<td>11.6 12.9 14.6 22.6</td>
</tr>
<tr>
<td></td>
<td>* Urban area per district</td>
<td>0.734</td>
<td>0.329</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N of households per capita</td>
<td>0.602</td>
<td>0.269</td>
<td></td>
</tr>
<tr>
<td>Area morbidity</td>
<td>Overall, AAMR</td>
<td>0.919</td>
<td>0.190</td>
<td>88.1 96.7 102.7 116.9</td>
</tr>
<tr>
<td></td>
<td>Respiratory, AAMR</td>
<td>0.838</td>
<td>0.173</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Circulatory, AAMR</td>
<td>0.743</td>
<td>0.153</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Infectious and parasitic diseases, AAMR</td>
<td>0.702</td>
<td>0.145</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% people with severe disability</td>
<td>0.689</td>
<td>0.142</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% people on diabetes treatment</td>
<td>0.483</td>
<td>0.099</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% people with stroke symptoms</td>
<td>0.436</td>
<td>0.090</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% people with mental health diseases</td>
<td>0.020</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>Difficulty to social distancing</td>
<td>% people living in apartment buildings</td>
<td>0.794</td>
<td>0.286</td>
<td>9.2 13.3 15.4 27.6</td>
</tr>
<tr>
<td></td>
<td>% workers in retail services</td>
<td>0.749</td>
<td>0.269</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N of students per class, high school</td>
<td>0.677</td>
<td>0.244</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% workers in health and social services</td>
<td>0.558</td>
<td>0.201</td>
<td></td>
</tr>
<tr>
<td>Population mobility</td>
<td>Net migration between districts</td>
<td>0.732</td>
<td>0.500</td>
<td>0.7 1.1 1.9 7.2</td>
</tr>
<tr>
<td></td>
<td>% foreign residents in the area</td>
<td>0.732</td>
<td>0.500</td>
<td></td>
</tr>
</tbody>
</table>

The factor scores and weights of each contributing variable associated with the first PCA-identified component and the quartiles of the health/SE themes are shown. Education is not included since it is already a single variable. Abbreviations: Principal component analysis (PCA); Socioeconomic (SE); age-adjusted mortality rate (AAMR). Superscripts: * km².
Table S4. GNBR model estimates with and without the estimated data for Daegu’s subparts throughout study period (January 20 through July 1, 2020)

<table>
<thead>
<tr>
<th>Health/SE themes</th>
<th>Included estimated data</th>
<th>Excluded estimated data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate*</td>
<td>SE</td>
</tr>
<tr>
<td>Area-morbidity</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>Education</td>
<td>-0.11</td>
<td>0.04</td>
</tr>
<tr>
<td>Crowding</td>
<td>0.22</td>
<td>0.12</td>
</tr>
<tr>
<td>Difficulty to social distancing</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>Population mobility</td>
<td>-0.38</td>
<td>0.09</td>
</tr>
<tr>
<td>Healthcare access</td>
<td>-0.14</td>
<td>0.03</td>
</tr>
<tr>
<td>Health behaviour</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>Dispersion</td>
<td>3.68</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>1527</td>
<td></td>
</tr>
</tbody>
</table>

Superscripts: *Parameter estimates; Abbreviations: Standard error (SE); Relative Risk (RR); lower boundary of 95% confidence interval (LCL); upper boundary of 95% confidence interval (UCL); Global negative binomial regression (GNBR); Akaike’s information criterion (AIC).

Table S5. Parameter estimates and the Relative Risk of the COVID-19 incidence associated with health and SES determinants by three time periods corresponding with the early, middle and late phases

<table>
<thead>
<tr>
<th></th>
<th>Early phase</th>
<th></th>
<th>Middle phase</th>
<th></th>
<th>Late phase</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>RR (LCL - UCL)</td>
<td>P-value</td>
<td>Estimate</td>
<td>RR (LCL - UCL)</td>
<td>P-value</td>
</tr>
<tr>
<td>Healthcare access</td>
<td>-0.14</td>
<td>0.87 (0.82 - 0.93)</td>
<td>&lt;.0001</td>
<td>-0.13</td>
<td>0.88 (0.84 - 0.93)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Health behaviour</td>
<td>0.04</td>
<td>1.04 (1.00 - 1.08)</td>
<td>0.028</td>
<td>0.03</td>
<td>1.03 (1.00 - 1.06)</td>
<td>0.080</td>
</tr>
<tr>
<td>Crowding</td>
<td>0.22</td>
<td>1.25 (0.99 - 1.57)</td>
<td>0.058</td>
<td>-0.05</td>
<td>0.96 (0.83 - 1.10)</td>
<td>0.512</td>
</tr>
<tr>
<td>Area morbidity</td>
<td>0.05</td>
<td>1.05 (1.03 - 1.06)</td>
<td>&lt;.0001</td>
<td>0.04</td>
<td>1.04 (1.02 - 1.06)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Education</td>
<td>-0.11</td>
<td>0.90 (0.83 - 0.97)</td>
<td>0.005</td>
<td>-0.08</td>
<td>0.92 (0.88 - 0.97)</td>
<td>0.003</td>
</tr>
<tr>
<td>Difficulty to social distancing</td>
<td>0.07</td>
<td>1.07 (1.01 - 1.14)</td>
<td>0.021</td>
<td>0.06</td>
<td>1.06 (1.02 - 1.11)</td>
<td>0.006</td>
</tr>
<tr>
<td>Population mobility</td>
<td>-0.38</td>
<td>0.69 (0.57 - 0.82)</td>
<td>&lt;.0001</td>
<td>-0.02</td>
<td>0.98 (0.84 - 1.14)</td>
<td>0.791</td>
</tr>
</tbody>
</table>

Early phase: January 20 to March 20, 2020; Middle phase: March 21 to April 15, 2020; Late phase: April 16 to July 1, 2020. Abbreviations: Relative Risk (RR); lower boundary of 95% confidence interval (LCL); upper boundary of 95% confidence interval (UCL).
<table>
<thead>
<tr>
<th></th>
<th>Healthcare access</th>
<th>Health behaviour</th>
<th>Crowding</th>
<th>Area morbidity</th>
<th>Education</th>
<th>Difficulty to social distancing</th>
<th>Population mobility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthcare access</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health behaviour</td>
<td>0.547</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crowding</td>
<td>0.568</td>
<td>0.124</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area morbidity</td>
<td>0.604</td>
<td>0.534</td>
<td>0.415</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>0.127</td>
<td>0.390</td>
<td>-0.080</td>
<td>0.510</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulty to social distancing</td>
<td>-0.032</td>
<td>0.370</td>
<td>-0.047</td>
<td>0.311</td>
<td>0.315</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Population mobility</td>
<td>-0.096</td>
<td>0.086</td>
<td>-0.019</td>
<td>0.018</td>
<td>0.163</td>
<td>0.129</td>
<td>1</td>
</tr>
</tbody>
</table>
Appendix B: Chapter 4 Supplementary Materials

Figure S.4.1. PRISMA Flow Diagram, Large B-Cell Lymphoma (LBCL)

Identification

N=308 records identified through e-database searching

N=12 records from ChiCTR and clinicaltrials.gov

N=161 records after duplicates removed

Screening

N=55 full-text articles screened

N=25 excluded
- 3 Pivotal trials
- 12 Case reports
- 5 Case-series
- 3 Letter to editor
- 2 No IPD data

N=30 full-text articles assessed for eligibility

N=8 excluded
- 3 Sample size <10
- 1 Only 28-day follow-up
- 2 No IPD available
- 2 sub-cohorts of pivotal trials

N=24 studies used in feasibility assessment (n = 20)

Eligibility

N=16 studies used for quantitative synthesis

Included
Figure S.4.2. MAIC of experimental CAR T-cells and Yescarta regarding PFS among patients who received infusion. Kaplan Meier survival curves. Hazard Ratios and 95% Confidence Intervals computed through Cox Proportional Hazards Models.

A. Dual targeting
   Tandem CD19, CD20 with 4-1BB

B. Third generation
   CD19 with CD28 & 4-1BB

C. ASCT + CAR T-cell
   Sequential ASCT and CD19, CD28