I aspire to pursue a Ph.D. in Biostatistics at Harvard University to conduct research on developing statistical inference methods to address methodological needs under a public health context. Living in a time of immense uncertainty has sparked my curiosity to explain the rationale of societal phenomena by deriving and testing models with data. Although data is easy to access in our digital world, finding information that overlaps in time, space, sampling parameters, and so on isn’t so simple. Estimating models’ uncertainty in order to provide more information and expand on our own capabilities excites me and motivates me to pursue a Ph.D. in Biostatistics. My long-term research goals are to polish methodologic research in statistical inference methods, and to direct these advances towards addressing public health problems. Harvard’s strengths in these fields align very well with my academic interests and long-term goals. This program is my first choice for its interdisciplinary focus and resources in biostatistics.

**Academic & Research Background**

My enthusiasm for science and research first took root when I was a Physics-major student at Sun Yat-sen University. I took many advanced mathematics courses and sought to integrate the principles employed in both math and physics to address a specific research project conducting MCMC simulations of Ising Models on lattices in all regimes of temperature. Focusing on the spin-spin correlation for Ising Models, I realized the Many-body system is inherently noisy and high dimensional: diverse spins react to a large number of environmental factors and interact with neighbors, while measurements are subject to different types of uncertainty. The qualitative idea of direct and indirect interactions could be described by conditional and marginal dependence in statistical language so that they can be testable hypotheses. The statistical model describing such a system should have multiple correlated responses and multiple predictors. Realizing modeling can solve broader questions not restricted to statistical physics, my interest shifted to statistics.

At Harvard Chan, I consolidated my theoretical statistics skills by taking PhD’s core curriculums consisting of Probability, Inference and Methods. I also worked with Dr. Leo Anthony Celi to apply regression adjustment and factor analytical models in clinical data. In my first-author work presented at APHA 2022, I conducted a retrospective cohort study of patients using the MIMIC-IV database to evaluate the strength of association between specific organ dysfunction and mortality for each subpopulation cohort. I was fascinated not only by inferencing the inequality in clinical data but also by the methodological challenges that these large data sets provide. In the public health context, different statistical models are proposed almost every day. How do we properly choose models and measure our uncertainty and robustness on the choice? How do we choose which predictors are important and how do we measure the uncertainty of these choices? In complicated models, how do we properly fit and do inference that is scalable and reliable?

These questions commonly arise in health inequities studies. To investigate, I conducted a project under supervision of Dr. Rachel Nethery and Dr. Brent Coull to consider the impacts of new US Census differential privacy(DP) algorithms on health inequity among different racialized groups. Through a multi-level variant of the spatial Poisson regression model, I conducted both simulation studies and real data analyses to model the standardized mortality rate (SMR) at the census tract (CT) level in Massachusetts. Herein, both a CT-specific random effect with a conditionally autoregressive spatial covariance structure and an unstructured group-specific error term that models overdispersion in the disease count are incorporated. Subsequently, I computed and evaluated metrics for SMR estimates to investigate whether small-area spatial patterns in model-smoothed disease/mortality risk estimates are preserved when using the DAS-protected denominators in standard models and assessed extent and direction of biases related to DP algorithm. This work enhanced my statistical methodology background in spatial methods and disease mapping models, and it also drove me to contemplate the spatial component in the models: why not build on that foundation of spatial methods from our model to expand the methodology to a Bayesian nonparametric setting?

I then took that foundational knowledge developed from the DP project and expanded it by actually developing a new spatial method that can be more flexible for modern data types: the Bayesian Nonparametric Ensemble (BNE) model. Under the supervision of Dr. Coull and Dr. Jeremiah Liu, I utilized BNE framework that used a dependent random measure to adaptively combine models based on their predictive accuracy in the feature space and nonparametrically modelled the ensemble’s predictive cumulative density function so that the model’s quantification of the predictive uncertainty is consistent with observed data. My first application of this novel method was to generate a spatial prediction model and estimate the associated prediction uncertainties for fine particle (PM2.5) levels in eastern New England, USA. Through this air pollution work, I implemented our ensemble framework to integrate information on PM2.5 levels from three distinct PM2.5 exposure models’ out-of-sample prediction for 51 monitors, along with the other 4 BNE approaches to performed posterior inference by our spatial prediction models using Hamiltonian Monte Carlo. By decomposing the predictive uncertainty into model selection uncertainty and model prediction uncertainty, I visualized low uncertainties around 51 monitors. The fabulous adaptation and flexibility drove me to leveraging it to more type of data, allowing from point data to aggregated ones: I then worked on a 2nd project to help the BNE method expand to handle population level area aggregated data, and to generate posterior predictive distribution for population estimation. Even though this project is in its early stages, it is my most exciting project, using cutting edge methods to gain mechanistic understandings that are relevant to public health in an unexpected way.

**Career Goal & Harvard**

In the future, I seek to explore the paths to methodology improvement for estimation of the health impacts and leverage existing data to provide more cost-effective ways of getting at larger health questions. Facing mountains of fragmented data, I am eager to develop statistical methods that enable maximally rigorous and impactful uses of data to answer environmental health questions. It’s also challenged for biostatistics researchers to handle the balance between complexity of their models and the uncertainty quantifications. From my reading and my research experience, I would argue that there are plenty of methods in a cutting edge that can help, e.g, functional space variational inference, Gaussian Process, Bayesian neural network etc. If I master these advanced methods, I hope to contribute to improve the methodology development in biostatistics, and to help promote more effective statistical methods.

Given my master training in biostatistics, continuing in the PhD program perfectly matches my career goal of entering academia to conduct statistical methodological research. I am inspired and excited by the prospect of continuing my work with Dr. Coull and Dr. Rachel on [] research.