# Form I: BTI Internship Program Mentor Application

| A.1: Mentor information | | |
|---|---|---|
| Name<br>Wei Xu<br>Principal Biostatistician | Institution<br>Princess Margaret Cancer Centre | Department<br>Biostatistics |

| A.2: Co-mentor information (if applicable) | | |
|---|---|---|
| Name<br>Geoffrey Liu<br>Senior Scientist | Institution<br>Princess Margaret Cancer Centre | Department<br>DMOH |

## A.3: Research proposal (maximum two pages)

Title

**Machine Learning and Biostatistical Longitudinal Model and Analysis on Head and Neck cancer survival and Toxicity Outcomes using Genetic Polymorphisms and Clinical Risk Factors**

Dr. Wei Xu is a Clinician Scientist and Principal Biostatistician at the Princess Margaret Cancer Centre, and an Associate Professor of Biostatistics in School of Public Health, University of Toronto. Dr. Xu's research interests focus on clinic trails design and methodology, machine learning and biostatistical methodology, statistical genetics, cancer clinical and translational studies. He has led various clinical studies and human genetic studies. As Principal-Investigator or co-Investigator on several previous Canadian government funded and NIH funded grants, he has led in study design, data integration, statistical and bioinformatics modeling and analysis on different clinic research and human genetic studies. So far, he has been published over 350 peer-reviewed papers in high impact journals of medical science, human genetics and statistics. His H-index is 61 with more than 19,000 citations.

Dr. Geoffrey Liu is the Alan B. Brown Chair in Molecular Genomics (University of Toronto), and Cancer Care Ontario Research Chair in Experimental Therapeutics and Population Studies. He is a medical oncologist at Princess Margaret Hospital (PMH), a pharmacogenetic epidemiologist, a Senior Scientist at the Ontario Cancer Institute, and a Professor of Epidemiology in School of Public Health, University of Toronto. His research interest has been in the role of biomarkers (germline, (epi)genetic, genomic, serologic, protein) as prognostic and pharmacogenetic predictive markers of treatment outcome in a variety of solid tumours, including head and neck, lung cancer, gastroesophageal, testicular and pancreatic cancers, and mesothelioma. He uses clinical trial and observational trial specimens, in addition to developing human tumour mouse xenografts for experimentation. His laboratory is one of the primary pharmacogenetic and germline sequence variant reference laboratories for several large consortia groups, including the PMH Drug Development Phase I and II Consortia, the NCIC Clinical Trials Group, and the Radiation Therapy Oncology Group.

Drs. Wei Xu and Geoffrey Liu are the co-founders and directors of Cancer Outcomes, Medicine, Biostatistics, Informatics Epidemiology and laboratory Interdisciplinary Program (COMBIEL). COMBIEL mandate is to enhance cross-disciplinary research training. COMBIEL supervisors strive to ensure that their trainees have a full experience. www.uhncombiel.com

**DESCRIPTION OF STUDY**

Head and neck squamous cell carcinomas (HNSCC) constitute the 5$^{th}$ most common malignancy worldwide; in Canada, there were 4550 new cases, and 1660 HNSCC deaths last year (www.cancer.ca). There are many challenges of mucosal HNSCC, not the least of which is its significant heterogeneity, both clinically and biologically. Oral cavity SCCs (OSCC) are managed primarily by surgery, with or without post-operative chemo-radiation (CRT). Head & neck cancers arising from the larynx, oropharynx (OPC), or hypopharynx are treated with radiation therapy (RT) alone for early stage disease, but locally-advanced diseases are also managed with a combination of CRT using

Cisplatin or 5-Fluorauracil. The majority of HNSCC patients present with locally advanced disease, with persistently disappointing outcome, with 5 year overall survival (OS) rates remaining at 40-60% over the past decades, underscoring a significant need for improvement.

The objective of this study is to evaluate molecular prognostic markers in head and neck cancer (HNC) outcomes (1), as clinical prognostic factors are imprecise, and improving our ability to predict clinical outcomes is the cornerstone to individualized patient management. Inherited genetic variations (typically single nucleotide polymorphisms, SNPs) are being studied as prognostic factors in many cancers (2-12). SNPs can be measured by a simple blood test using available technologies in most clinical diagnostic laboratories (13). The overall objective is to test and validate a comprehensive list of SNPs as predictors of HNC outcomes, utilizing the latest SNP selection, bioinformatic and multivariate modeling strategies, multiple replication datasets, multistage pathway approaches, large well characterized patient populations, and clearly defined outcomes. The goal is to adapt the latest machine learning and biostatistical methods, and apply them to study large-scale cancer survival and toxicity data patterns coming from clinical studies and large institutional databases such as those held at Princess Margaret Cancer Centre.

There are multiple aims for this proposed study. Aim 1: To validate the association between HNC outcomes and 30 SNPs/SNP pathways identified from the published literature in the three datasets. We hypothesize that each SNP/SNP pathway is independently associated with HNC outcome. Aim 2: To apply a genome-wide association (GWA) analysis between HNC outcomes and 600,000 SNPs genotyped for all QUE patients (Aim 2A). From Aim 2A, we will then identify the genome wide significant ($p<5x10-7$) SNPs to validate in PMH1 (Aim 2B) and explore in PMH2 (Aim 2C). In Aims 2B/2C, we hypothesize that each SNP is independently associated with HNC outcome. Aim 3: DNA repair dominates the published literature for SNPs and outcome in both HNCs and other aerodigestive tract cancers. We hypothesize that DNA repair capacity, as measured by the COMET assay, is associated with HNC outcomes in a prospective dataset (n=156), PMH3. Aim 4: To perform integrative prognostic modeling using machine learning and statistical learning algorithms. We will test the combined role of SNPs and serum prognostic markers (being validated in a separate NCIC grant, co-PIs Meyer/Bairati) in QUE, hypothesizing that SNPs of genes associated with these serum markers modify, interact, or are correlated with each other in HNC outcomes (Aim 4A). We will also test the combined role of significant SNPs identified in Project 4 and significant RNA signatures identified in Projects 1 & 2, as independent predictors of survival outcomes in a subset of PMH1/PMH2 (Aim 4B). Finally, we will test the association between various DNA repair gene SNPs and DNA repair capacity, as measured by the COMET assay (Aim 4C).

There are several research topics for the intern student to work based on this study depending on the student's research interest. He/she can work on exploratory multivariate prognostic modeling of time to event data (18, 19). A long term goal of the study is to identify a prognostic tool that incorporates clinical and molecular factors, firstly by identifying the best individual markers within each discipline (RNA, SNP, dosimetry), with a future goal to validate across disciplines (20). The study integration provides opportunity to perform initial exploratory multi-disciplinary analyses. QUE will have both serologic and SNP data, and we will explore different serologic markers and associated SNPs within the same pathway. Likewise, a subset of PMH1 and all of PMH2 will have mRNA and miRNA data, thus allowing exploration of different SNP-gene expression signatures in relation to HNC outcomes. Finally, we will explore the relationship between the COMET assay and DNA repair SNPs that are associated with outcomes on validation or replication.

Besides that, the internship student can work on the utility of bioinformatics & integrative databases. Using protein structure predictive software and sequence analysis programs, we can evaluate polymorphic variants in EGFR

pathway genes. Hundreds of pathway SNPs can be chosen for further evaluation in an upcoming Phase III Intergroup study of advanced non-small cell lung cancer. In a separate analysis of colon cancer, OPHID (U of Toronto protein-protein interaction database) was used to identify Met-signaling targets that were then verified experimentally (21). The same database has also been used to help narrow gene expression-related prognostic biomarkers in ovarian cancer (22). Adapted bioinformatics and higher order statistical approaches can be applied for continuous variable outcomes, including cluster analysis to detect gene-gene interactions, neural networks/artificial intelligence, and adaptations of classification and regression tree (CART) analysis (23-25).

Another topic is the analysis of survival and toxicity. The impact of single SNP on each of the time to event outcomes (FFS, OS, recurrence-free rate, time to second primary event) will be analyzed independently, using Kaplan-Meier method to estimate survival curves; Log-rank test will be used to detect SNP effect on survival. Cox Proportional-Hazards (CPH) models will be developed for multivariate analysis. Logistic regression will be applied on toxicity rate. Covariates being considered for adjustment will include age, gender, disease stage, histologic site, performance status, treatment and ethnicity. Some variables can be treated as stratification variables instead of adjustment variables, as appropriate. For example, QUE treatments fell into 3 groups: α-tocopherol + β-carotene supplementation; α-tocopherol supplementation; and placebo (14, 15). To adjust for population heterogeneity, we can apply a stratified proportional hazard model to each SNP, which allows the baseline hazards to differ between strata and tests whether our results are simply driven by strata-specific baseline hazards (26).

**References**

1. Forastiere, A.A., Trotti, A., Pfister, D.G., and Grandis, J.R. Head and neck cancer: recent advances and new standards of care. J Clin Oncol *24:* 2603-2605, 2006.
2. Toffoli, G., and Cecchin, E. Clinical implications of genetic polymorphisms on stomach cancer drug therapy. Pharmacogenomics J *7:* 76-90, 2007.
3. Jamroziak, K., and Robak, T. Pharmacogenomics of MDR1/ABCB1 gene: the influence on risk and clinical outcome of haematological malignancies. Hematology *9:* 91-105, 2004.
4. Heist, R.S., Zhou, W., Chirieac, L.R., Cogan-Drew, T., Liu, G., Su, L., Neuberg, D., Lynch, T.J., Wain, J.C., and Christiani, D.C. MDM2 polymorphism, survival, and histology in early stage non-small-cell lung cancer. J. Clin. Oncol. *25:* 2243-2247, 2007.
5. Park, S.Y., Hong, Y.C., Kim, J.H., Kwak, S.M., Cho, J.H., Lee, H.L., and Ryu, J.S. Effect of ERCC1 polymorphisms and the modification by smoking on the survival of non-small cell lung cancer patients. Med. Oncol. *23:* 489-498, 2006.
6. Suk, R., Gurubhagavatula, S., Park, S., Zhou, W., Su, L., Lynch, T.J., Wain, J.C., Neuberg, D., Liu, G., and Christiani, D.C. Polymorphisms in ERCC1 and grade 3 or 4 toxicity in non-small cell lung cancer patients. Clin. Cancer Res. *11:* 1534-1538, 2005.
7. Zhou, W., Gurubhagavatula, S., Liu, G., Park, S., Neuberg, D.S., Wain, J.C., Lynch, T.J., Su, L., and Christiani, D.C. Excision repair cross-complementing group 1 polymorphism predicts overall survival in advanced non-small cell lung cancer patients treated with platinum-based chemotherapy. Clin. Cancer Res. *10:* 4939-4943, 2004.
8. Gurubhagavatula, S., Liu, G., Park, S., Zhou, W., Su, L., Wain, J.C., Lynch, T.J., Neuberg, D.S., and Christiani, D.C. XPD and XRCC1 genetic polymorphisms are prognostic factors in advanced non-small-cell lung cancer patients treated with platinum chemotherapy. J. Clin. Oncol. *22:* 2594-2601, 2004.
9. Martinez-Balibrea, E., Mazano, J.L., Martinez-Cardus, A., Moran, T., Ciragui, B., Catot, S., Taron, M., and Abad, A. Combined analysis of genetic polymorphisms in thymidylate synthase, uridine diphosphate glucoronosyltransferase and x-ray cross complementing factor 1 genes as a prognostic factor in advanced

colorectal cancer patients treated with 5-fluorouracil plus oxaliplatin or irinotecan. Oncol. Rep. *17:* 637-645, 2007.

10. Sternlicht, M.D., Dunning, A.M., Moore, D.H., Pharoah, P.D., Ginzinger, D.G., Chin, K., Gray, J.W., Waldman, F.M., Ponder, B.A., and Werb, Z. Prognostic value of PAI1 in invasive breast cancer: evidence that tumor-specific factors are more important than genetic variation in regulating PAI1 expression. Cancer Epidemiol. Biomarkers and Prev. *15:* 2107-2014, 2006.
11. Wu, X., Gu, J., Wu, T.T., Swisher, S.G., Liao, Z., Correa, A.M., Liu, J., Etzel, C.J., Amos, C.I., Huang, M., Chiang, S.S., Milas, L., Hittelman, W.N., and Ajani, J.A. Genetic variation in radiation and chemotherapy drug action pathways predict clinical outcome in esophageal cancer. J. Clin. Oncol. *24:* 3789-3798, 2006.
12. Izzo, J.G., Wu, T.T., Wu, X., Ensor, J., Luthra, R., Pan, J., Correa, A., Swisher, S.G., Chao, C.K.S., Hittelman, W.N., and Ajani, J.A. Cyclin D1 guanine/adenine 870 polymorphism with altered protein expression is associated with genomic instability and aggressive clinical biology of esophageal adenocarcinoma. J. Clin. Oncol. *25:* 698-707, 2007.
13. Liu, G., Zhou, W., and Christiani, D.C. Molecular epidemiology of non-small cell lung cancer. Semin. Respir. Crit. Care Med. *26:* 265-272, 2005.
14. Bairati I, Meyer F, Gelinas M, Fortin A, Nabid A, Brochet F, Mercier JP, Tetu B, Harel F, Abdous B et al: Randomized trial of antioxidant vitamins to prevent acute adverse effects of radiation therapy in head and neck cancer patients. J Clin Oncol 2005, 23(24):5805-5813.
15. Bairati I, Meyer F, Gelinas M, Fortin A, Nabid A, Brochet F, Mercier JP, Tetu B, Harel F, Masse B et al: A randomized trial of antioxidant vitamins to prevent second primary cancers in head and neck cancer patients. J Natl Cancer Inst 2005, 97(7):481-488.
16. Bairati I, Meyer F, Jobin E, Gelinas M, Fortin A, Nabid A, Brochet F, Tetu B: Antioxidant vitamins supplementation and mortality: a randomized trial in head and neck cancer patients. Int J Cancer 2006, 119(9):2221-2224.
17. Meyer F, Bairati I, Fortin A, Gelinas M, Nabid A, Brochet F, Tetu B: Interaction between antioxidant vitamin supplementation and cigarette smoking during radiation therapy in relation to long-term effects on recurrence and mortality: a randomized trial among head and neck cancer patients. Int J Cancer 2008, 122(7):1679-1683.
18. Therneau TM, Grambsh PM. Modeling survival data: extending the Cox model. Springer-Verlag, New-York, 2000.
19. Klein JP, Moeschbergher ML. Survival analysis: techniques for censored and truncated data. Springe, New-York, 2nd Ed, 2003.
20. Yu, Z., Li, Z., Jolicoeur, N., Zhang, L., Fortin, Y., Wang, E., Wu, M., and Shen, S.-H. Aberrant allele frequencies of the SNPs located in microRNA target sites are potentially associated with human cancers. Nucleic Acids Res. *35:* 4535-4541, 2007.
21. Seiden-Long, I.M., Brown, K.R., Shih, W., Wigle, D.A., Radulovich, N., Jurisica, I., Tsao, M.S. Transcriptional targets of hepatocyte growth factor signaling and Ki-ras oncogene activation in colorectal cancer. Oncogene *25:* 91-102, 2006.
22. Motamed-Khorasani, A., Jurisica, I., Letarte, M., Shaw, P.A., Parkes, R.K., Zhang, X., Evangelou, A., Rosen, B., Murphy, K.J., and Brown, T.J. Differentially androgen-modulated genes in ovarian epithelial cells from BRCA mutation carriers and control patients predict ovarian cancer survival and disease progression. Oncogene *26:*198-214, 2007.
23. Xu, W., Taylor, C., Veenstra, J., Bull, S.B., Corey, M., Greenwood, C.M. Recursive partitioning models for linkage in COGA data. BMC Genet. *30:* Suppl 1:S38. Epub ahead of print, 2005.
24. Xu, W., Schulze, T.G., DePaulo, J.R., Bull, S.B., McMahon, F.J., Greenwood, C.M. A tree-based model for allele-sharing-based linkage analysis in human complex diseases. Genet. Epidemiol. *30:*155-169, 2006.

25. Xu, W., Hui, L., Hu, P., Bull, S., and Greenwood, C.M.T. Linkage analysis on chromosome 1 for rheumatoid arthritis NARAC data: gene-gene and gene-environment interactions. BMC Genetics *in press*, 2007.
26. Prentice, R.L., Williams, B.J., and Peterson, A.V. On the regression analysis of multivariate failure time data. Biometrika *62:*373-379, 1981.