Form I: BTI Internship Program Mentor Application

A.1: Mentor information		
Name Osvaldo Espin-Garcia Principal Biostatistician	Institution Princess Margaret Cancer Centre, University Health Network	Department Biostatistics

A.2: Co-mentor information (if applicable)		
Name	Institution	Department
Title		

A.3: Research proposal

Title:
Characterizing Genetic Susceptibility to Smoking and Drinking Behaviors in Patients with Head and Neck Cancer via Multi-Trait Polygenic Risk Scores

Research proposal, including background, aims and research methodology; training objectives and environment including mentorship plan; necessary intern background/strengths. Outline how the research proposal is in line with OICR's Strategic Plan.

Large-scale genomic studies often collect multiple traits of interest across thousands (or even hundreds of thousands) of subjects to answer diverse research questions. Traditional analytical approaches often focus on one-trait at-a-time; however, phenomena such as pleiotropy (one gene influencing two or more traits) make multivariate analysis of great interest for researchers. Moreover, it is believed that multi-trait analysis can improve statistical power to detect genetic associations by borrowing information from correlated traits.

On the other hand, genome-wide polygenic risk scores (PRS), which are weighted sums of the allelic states at single-nucleotide polymorphisms (SNPs), have become increasingly used to quantify (some level of) the genomic susceptibility of a given trait. The weights to construct these PRS are typically obtained from previously reported summary statistics from well-powered independent genome-wide association studies. The use of summary statistics is a crucial issue in PRS construction, as confidentiality concerns usually limit individual-level data sharing. Moreover, although PRS are typically calculated for a single trait at a time, recent work has extended their construction for multiple traits, e.g. Turley et al, 2018, Nature Genetics; Tanigawa et al, 2019, Nature Communications; Aguirre et al, 2020, bioRxiv [preprint]. Briefly, Turley et al use an efficient generalized method of moments to combine summary information among multiple non-necessarily overlapping summary statistics. In contrast, Tanigawa et al and Aguirre et al use the Decomposition of Genetic Associations (DeGAs) approach, which implements a truncated singular value decomposition to identify latent components of genetic associations.

Recently, a large-scale genome-wise association study on non-cancer populations has identified loci that are associated to an increased risk of smoking and drinking traits, shedding some light on the genetic predisposition to these behaviors (Liu et al, 2019, Nature Genetics). These associations remain unexplored in populations of patients with cancer. If confirmed, however, these loci may be useful for patient characterization in prognostic models.

The purpose of this project is to evaluate use of the multi-trait polygenic risk scores and their prognostic performance in identifying subpopulations with a higher mortality among populations of patients with cancer. To this end, the intern will investigate the use of current multi-trait PRS approaches in a large-scale

international consortium of patients with head and neck cancer (VOYAGER), which consists of over 8,000 participants.

The specific aims of the project are as follows:

1. Perform a supervisor-curated literature review on univariate and multivariate (genome-wide) PRS construction and utilization, with a special emphasis on methods
2. Identify differences and commonalities as well as strengths and limitations of existing multi-trait PRS methods
3. Familiarize with genotyping information data formats as well as available implementations for multi-trait PRS construction
4. Perform data analysis on the available VOYAGER data using reproducible pipelines
5. Summarize and report the findings for scientific dissemination

The intern is expected to have a good understanding of linear models/generalized linear models and maximum likelihood estimation. In addition, the following computing skills will be useful/desirable: R code for standard methods, custom R code for non-standard methods, BASH scripts to process and analyze high-throughput data (in UNIX environment); familiarity with Python and/or command line software (or willingness to gain experience with these) would be an asset.

Mentorship plan: Individual meetings with the intern will be held on a weekly basis to discuss progress and inquiries on the research project and next steps. These meetings intend to support independent, critical, and inquisitive thinking as well as to promote ownership of the intern work. In addition, lab meetings, possibly in combination with other faculty members' labs, will be held once per month; these lab meetings allow the trainees to present ongoing work to peers and open opportunities to give and receive feedback. Lastly, the intern will be encouraged to attend a variety of talks, seminars, and workshops as available. *COVID19 Considerations*: depending on the pandemic situation by the beginning of the internship, the meetings will take place virtually or live.

About the site: The Department of Biostatistics at the Princess Margaret Cancer Centre collaborate with researchers on a broad spectrum of oncology studies by providing expertise on statistical design, biostatistics and bioinformatics modeling, data analysis, and grant application. The Department consists of principle, senior, and junior biostatisticians along with research analysis/associates and post-doctoral fellows. We lend our expertise on a number of grant panels, research ethics boards, and serve as statistical reviewers for a number of peer-reviewed journals.