Project Overview: Causal Modeling of Player Readiness

Context: KU Leuven & OHL Partnership

1. Project Scope & Objectives

The primary objective is to develop a predictive model to estimate the **'Readiness to Train'** for individual football players. The project aims to move beyond simple prediction into **prescriptive analytics** (Causal ML) by recommending optimal sequential training loads to maximize match-day performance while minimizing injury risk.

2. Problem Statement

Professional football involves high-dimensional data for a small number of subjects (Small N, Large T). Standard 'data-hungry' Deep Learning methods risk overfitting.

- **The Challenge:** Identifying patterns in sparse injury data.
- **The Opportunity:** Combining *Domain Expert Knowledge* (Sport Science rules) with *Data-Driven Approaches*.

3. Data Asset Overview

The dataset is a longitudinal time-series (~23,000 rows) consisting of daily entries per player, combining objective and subjective metrics.

Key Input Features:

- **External Load (GPS):** Total Distance, HSD (>19.8 km/h), Decelerations (>3ms^2), Sprints.
- **Subjective Wellbeing (McLean):** Fatigue, Soreness, Sleep Quality, Stress, Mood (normalized via Z-scores).
- **Transformed Metrics:** Acute:Chronic Workload Ratio (ACWR 7:42), Exponential Moving Averages.

4. Theoretical Framework

The project is grounded in established Sport Science principles (referencing Gabbett, Impellizzeri, Hecksteden):

- **Supercompensation:** Modeling how biological processes respond to stress and recovery.
- **The 'Sweet Spot':** Identifying optimal acute load to maximize fitness without spiking injury risk.
- **Individual Profiling:** Using Z-scores to acknowledge unique physiological baselines.

5. Methodology: Causal Machine Learning

Given that training is sequential (Training today affects readiness tomorrow), this project will utilize **Sequential Treatment Effect** estimation.

- **Proposed Method:** SURVITE (Time-varying treatment effects).
- **Treatment:** Training Intensity (e.g., Recovery, Low, Moderate, Hard).
- **Goal:** Evaluate the impact of a training sequence (e.g., 5 days) on the outcome at Day 6 (Match).

6. Research Benchmarking

The study will perform a tripartite comparison:

- **1. Expert-Based:** Decisions based purely on sport science rules and staff intuition.
- **2. Data-Driven:** Decisions derived purely from raw data (GPS/Subjective).
- **3. Synergistic (Hybrid):** A Causal ML approach combining raw data with transformed domain knowledge to find the optimal synergy.