# Development of Scalable, Electronic Health Record (EHR)-Based Screening for Undiagnosed Systemic Mastocytosis: PREDICT-SM

## Background

Systemic mastocytosis (SM) is a rare, chronic mast cell disorder (MC) driven by activating mutations in the KIT or D816V mutation (~95% cases). Uncontrolled proliferation and activation of MCs in patients with SM can cause severe, debilitating, and unpredictable symptoms which may result in reduced quality of life and early mortality in advanced SM.

The heterogeneity of phenotypic presentation coupled with the low specificity of symptoms spread across multiple organ systems can make diagnosing SM challenging. Specifically, diagnoses can be delayed up to 7-10 years from the onset of symptoms in patients with non-advanced SM.

Earlier diagnosis of SM can decrease SM-associated symptoms, improve quality of life, and decrease secondary organ damage. Adoption of electronic health records (EHRs) along with rapid improvement in computational methods has created opportunities to apply machine learning and artificial intelligence (AI) to clinical data from multiple health systems to learn how to identify patients with rare diseases.

Here, in the first of a series of reports describing findings from PREDICT-SM, we describe the initial findings of PREDICT-SM, a study that aims to develop a pragmatic, accurate, and scalable approach to screen for undiagnosed SM by applying AI tools to EHR data.

## Study objectives

1. **Define patients with rare diseases and artificial intelligence (AI) to clinical data from multiple health systems to learn how to identify patients with rare diseases.**

2. **To develop a single- and multi-site EHR prediction model for undiagnosed SM.**

3. **To validate the single- and multi-site EHR prediction models.**

4. **To evaluate the performance of the EHR prediction models in real-world settings.**

## Results

### Study design

- For Aim 1 and Aim 2, we plan to use longitudinal EHR data from 13 health systems.
- Data will be collected for patients who were ≥18 years of age as of January 1, 2022, and had ≥5 ambulatory care encounters including ≥2 encounters at a single primary care, dermatology, gastroenterology, or pulmonology practice.
- Patients with ≥2 SM diagnostic codes (SM Dx code), elevated serum tryptase (Tryptase) and/or documented blood testing for SM will be included in the study.
- Possible SM cohort: Patients with ≥2 SM diagnostic codes (SM Dx code), elevated serum tryptase (Tryptase) and/or documented blood testing for SM who were not definitively diagnosed with SM.
- Indicated SM cohort: Patients who were definitively diagnosed with SM.

### Study population

- **Possible SM cohort (n=75):** Patients with ≥2 SM diagnostic codes (SM Dx code), elevated serum tryptase (Tryptase) and/or documented blood testing for SM who were not definitively diagnosed with SM.
- **Indicated SM cohort (n=585,449):** Patients who were definitively diagnosed with SM.
- **Study cohort (n=585,111):** Patients who were ≥18 years of age and had ≥5 ambulatory care encounters as of January 1, 2022.

### Key findings

- **Possible SM cohort:**
  - Median tryptase concentration: 572 (90% CI: 2.50–9.90)
  - Diagnostic accuracy of the EHR prediction model: 57.2%

- **Indicated SM cohort:**
  - Median tryptase concentration: 572 (90% CI: 2.50–9.90)
  - Diagnostic accuracy of the EHR prediction model: 57.2%

### Summary

This study highlighted the importance of the need to address observed demographic biases in patient populations with SM to ensure equity of trained prediction models. There was considerable variability in how often tryptase testing was performed, but the variability did not result in large differences in the frequency of SM diagnosis across care practices. Preliminary findings suggest there is potentially a substantial subset of patients with SM from which to draw a SM prediction model, which may help discover SM at an earlier stage, allow earlier intervention, and improve outcomes for patients with SM.