Electronic health records (EHRs) were designed to share data and information with authorized clinicians, caregivers, stakeholders, and the patients themselves to provide more coordinated, patient-centered care. Recently, there has been increasing interest in using this data for scientific research with many opportunities for scientific discoveries in real-world settings, particularly for rare diseases. However, significant challenges include bias, noisy and missing data, and the need for resources for data management and computation. In this talk we give an overview of some lessons learned regarding strategies for dealing with EHR data management, cleaning, causal inference, and prediction.