In electronic health record research it is common to have variables based on billing codes or on automated computation. Expert audit of a validation sample of physician notes is valuable for assessing or improving the accuracy of these variables. When the inferential goal is to fit a regression model, the optimal design of a validation sample can be quite different for model-assisted and model-based inference. I will talk about the reasons for the difference and the robustness/precision tradeoffs involved in choosing an approach.

Keywords: two-phase sampling, optimal design, generalised linear model, measurement error