Appendix: Annotated STATA code

Data Organization and Variable Definitions

Start by defining a baseline period, an index date, and an outcome period. The index date separates the baseline period from the outcome period and indicates when the patient received either the study treatment or a comparator. Data in the baseline period should be organized to have each observation represent a patient-event (either a prescription, a lab value or an outcome). Data in the outcome period should be organized with each observation representing a patient. The following variables are used in the code below.

Txrx: Indicator variable equal to 1 if baseline prescription event is study treatment and zero otherwise.
Allrx: Indicator variable equal to 1 if baseline event is a prescription and zero otherwise.
provider: Unique provider ID.
facility: Unique facility ID.
year: Year of index date.
treatment: Indicator variable equal to 1 if index treatment is study treatment and zero otherwise.
outcomedays: Number of days from index date to first outcome.
censored: Number of days from index date to censoring.

Step One: Choose and Specify IV

In practice pattern IV applications, the investigator often will be computing a provider- or facility-level mean in the baseline period, excluding the particular patient whose record is being processed. Thus, the patient's value or values are excluded from both the numerator and denominator of the rate used to predict his or her treatment. This can be done efficiently by calculating the overall numerator and overall denominator for all records in the baseline data and then subtracting the patient-specific values on each line before combining to form the rate. The rate is then saved as a patient-level variable and added to the outcome data.

```
egen numer1 = sum(Txrx), by(provider)
egen denom1 = sum(Allrx), by(provider)
egen numer2 = sum(Txrx), by(patient)
egen denom2 = sum(Allrx), by(patient)
gen numer3 = numer1 – numer2
gen denom3 = denom1 – denom2
gen IVrate = numer3/denom3
```

Step Two: Choose and Specify Control Variables

Control variables typically include standard demographics, risk-adjustment variables (based on diagnosis codes), and baseline medications and lab values if available. If provider-level process quality variables are included, they can be constructed excluding the individual patient using the same coding technique as Step One.

Step Three: Choose Falsification Sample and Outcomes

The falsification sample should ideally include the same control variables as the study sample. Falsification outcomes should be as close as possible to study outcomes without being affected by the study treatment.

Step Four: Estimate IV Model

The IV model is estimated on the outcome period data, using two equations bootstrapped together. The first is the treatment equation, which can be estimated by logistic regression or probit if treatment is binary. Results of this regression are used to calculate the predicted residual, which is included in the outcome equation. The following treatment equation example includes fixed effects for facilities and years:

```
xi: probit treatment IVrate {control variables} i.facility i.year
predict Txprob
gen Txres = treatment – Txprob
```

The following outcome equation example estimates a Cox proportional hazards model including fixed effects for years and random effects for facilities:

```
stset outcomedays, failure(censored)
xi: stcox treatment Txres {control variables} i.year, shared(facility)
```

Step Five: Compute Falsification Test

The falsification test is computed using an alternative formulation of the outcome equation. In this example, a falsification sample is used with the study outcome. If the IVrate variable has a statistically significant effect on the outcome, the instrument is rejected.

```
stset outcomedays, failure(censored)
xi: stcox IVrate {control variables} i.year, shared(facility)
```