/\* ============================================

GROUP-BASED TRAJECTORY MODELING WORKFLOW

Using PROC TRAJ for hospital visits

============================================ \*/

LIBNAME HOME '/userdata06/room241/data\_source/youb';

/\* Load PROC TRAJ macro - adjust path as needed \*/

%INCLUDE '/path/to/traj/Proc\_Traj\_Macro.sas';

/\* ============================================

PART 0: MODEL SPECIFICATION CHECK

Compare Poisson vs ZIP to justify ZIP use

============================================ \*/

/\* Check zero-inflation in the data \*/

PROC SQL;

CREATE TABLE HOME.zero\_check AS

SELECT

COUNT(\*) AS total\_obs,

SUM(CASE WHEN hospital\_visits = 0 THEN 1 ELSE 0 END) AS zero\_count,

100.0 \* CALCULATED zero\_count / CALCULATED total\_obs AS pct\_zeros FORMAT=5.1,

MEAN(hospital\_visits) AS mean\_visits FORMAT=8.2,

VAR(hospital\_visits) AS var\_visits FORMAT=8.2

FROM HOME.lcmm\_data;

QUIT;

PROC PRINT DATA=HOME.zero\_check NOOBS;

TITLE 'Zero-Inflation Check';

TITLE2 'If pct\_zeros >> expected under Poisson, use ZIP';

TITLE3 'Expected zeros under Poisson = 100\*exp(-mean)';

RUN;

/\* Test 3-group solution with both Poisson and ZIP \*/

%PUT Testing Poisson vs ZIP model specification;

/\* Poisson model \*/

PROC TRAJ DATA=HOME.lcmm\_data OUTPLOT=HOME.test\_poisson

OUTSTAT=HOME.stats\_poisson OUT=HOME.assign\_poisson;

ID INDI\_DSCM\_NO;

VAR hospital\_visits;

INDEP time;

MODEL POISSON;

ORDER 1 2 2 2;

ORDER 2 2 2 2;

ORDER 3 2 2 2;

RUN;

/\* ZIP model \*/

PROC TRAJ DATA=HOME.lcmm\_data OUTPLOT=HOME.test\_zip

OUTSTAT=HOME.stats\_zip OUT=HOME.assign\_zip;

ID INDI\_DSCM\_NO;

VAR hospital\_visits;

INDEP time;

MODEL ZIP;

ORDER 1 2 2 2;

ORDER 2 2 2 2;

ORDER 3 2 2 2;

RUN;

/\* Compare BIC values \*/

DATA HOME.model\_spec\_comparison;

LENGTH model $20 BIC 8;

SET HOME.stats\_poisson (WHERE=(\_NAME\_='BIC') RENAME=(COL1=BIC\_poisson))

HOME.stats\_zip (WHERE=(\_NAME\_='BIC') RENAME=(COL1=BIC\_zip));

IF \_N\_ = 1 THEN DO;

model = 'Poisson';

BIC = BIC\_poisson;

OUTPUT;

END;

ELSE IF \_N\_ = 2 THEN DO;

model = 'Zero-Inflated Poisson';

BIC = BIC\_zip;

OUTPUT;

END;

KEEP model BIC;

RUN;

PROC PRINT DATA=HOME.model\_spec\_comparison NOOBS;

TITLE 'Model Specification Comparison: Poisson vs ZIP';

TITLE2 'Lower BIC indicates better fit (difference >10 is meaningful)';

VAR model BIC;

FORMAT BIC 12.2;

RUN;

/\* Clean up test datasets \*/

PROC DATASETS LIBRARY=HOME NOLIST;

DELETE test\_poisson stats\_poisson assign\_poisson

test\_zip stats\_zip assign\_zip;

QUIT;

/\* ============================================

PART 1: MODEL SELECTION (2-10 GROUPS)

Testing different group solutions systematically

============================================ \*/

/\* Create a dataset to store model fit statistics \*/

DATA HOME.model\_fit\_comparison;

LENGTH n\_groups 8 BIC AIC entropy avg\_pp min\_group\_pct 8;

DELETE;

RUN;

/\* Macro to test different group solutions \*/

%MACRO test\_trajectory\_models(min\_groups=2, max\_groups=10);

%DO n = &min\_groups %TO &max\_groups;

%PUT ========================================;

%PUT Testing &n-group solution;

%PUT ========================================;

/\* Run PROC TRAJ with &n groups \*/

/\* Using ZIP (Zero-Inflated Poisson) for zero-inflated hospital visits \*/

PROC TRAJ DATA=HOME.lcmm\_data OUTPLOT=HOME.traj\_plot\_&n.grp

OUTSTAT=HOME.traj\_stats\_&n.grp OUT=HOME.traj\_assign\_&n.grp;

ID INDI\_DSCM\_NO;

VAR hospital\_visits;

INDEP time;

MODEL ZIP; /\* Zero-Inflated Poisson for zero-inflated count data \*/

/\* Specify trajectory groups - all start with quadratic \*/

%DO i = 1 %TO &n;

%IF &n <= 4 %THEN %DO;

/\* For 2-4 groups, use cubic to capture complex patterns \*/

ORDER &i 3 3 3 3; /\* Cubic for intercept, linear, quadratic, cubic \*/

%END;

%ELSE %DO;

/\* For 5+ groups, use quadratic to avoid overfitting \*/

ORDER &i 2 2 2; /\* Quadratic for intercept, linear, quadratic \*/

%END;

%END;

RUN;

/\* Extract fit statistics from output \*/

DATA \_fit\_&n;

SET HOME.traj\_stats\_&n.grp;

n\_groups = &n;

/\* Extract BIC, AIC \*/

IF \_NAME\_ = 'BIC' THEN BIC = COL1;

IF \_NAME\_ = 'AIC' THEN AIC = COL1;

KEEP n\_groups BIC AIC;

RUN;

/\* Calculate entropy and minimum group size \*/

PROC SQL;

CREATE TABLE \_entropy\_&n AS

SELECT

&n AS n\_groups,

/\* Entropy calculation: -sum(p\*log(p)) \*/

-SUM(max\_pp \* LOG(max\_pp)) / (COUNT(\*) \* LOG(&n)) AS entropy,

/\* Average of maximum posterior probabilities \*/

MEAN(max\_pp) AS avg\_pp,

/\* Minimum group percentage \*/

MIN(group\_pct) AS min\_group\_pct

FROM (

SELECT

INDI\_DSCM\_NO,

MAX(GRP1PRB, GRP2PRB

%IF &n >= 3 %THEN %DO; , GRP3PRB %END;

%IF &n >= 4 %THEN %DO; , GRP4PRB %END;

%IF &n >= 5 %THEN %DO; , GRP5PRB %END;

%IF &n >= 6 %THEN %DO; , GRP6PRB %END;

%IF &n >= 7 %THEN %DO; , GRP7PRB %END;

%IF &n >= 8 %THEN %DO; , GRP8PRB %END;

%IF &n >= 9 %THEN %DO; , GRP9PRB %END;

%IF &n >= 10 %THEN %DO; , GRP10PRB %END;

) AS max\_pp,

100.0 \* COUNT(\*) / (SELECT COUNT(\*) FROM HOME.traj\_assign\_&n.grp) AS group\_pct

FROM HOME.traj\_assign\_&n.grp

GROUP BY \_GROUP

);

QUIT;

/\* Combine fit statistics \*/

DATA \_combined\_&n;

MERGE \_fit\_&n \_entropy\_&n;

BY n\_groups;

RUN;

/\* Append to comparison dataset \*/

PROC APPEND BASE=HOME.model\_fit\_comparison DATA=\_combined\_&n FORCE;

RUN;

/\* Clean up temporary datasets \*/

PROC DATASETS LIBRARY=WORK NOLIST;

DELETE \_fit\_&n \_entropy\_&n \_combined\_&n;

QUIT;

%END;

%MEND;

/\* Run the model comparison \*/

%test\_trajectory\_models(min\_groups=2, max\_groups=10);

/\* Display model comparison table \*/

PROC PRINT DATA=HOME.model\_fit\_comparison NOOBS;

TITLE 'Model Fit Comparison: 2-10 Group Solutions';

VAR n\_groups BIC AIC entropy avg\_pp min\_group\_pct;

FORMAT BIC AIC 12.2 entropy avg\_pp 8.3 min\_group\_pct 8.2;

RUN;

/\* Export for supplemental materials \*/

PROC EXPORT DATA=HOME.model\_fit\_comparison

OUTFILE='/path/to/output/supplemental\_model\_fit\_table.xlsx'

DBMS=XLSX REPLACE;

RUN;

/\* ============================================

PART 2: SELECT OPTIMAL MODEL

Based on criteria: BIC, entropy >?, avg PP >0.70, min group >5%

============================================ \*/

/\* Flag models meeting all criteria \*/

DATA HOME.model\_selection;

SET HOME.model\_fit\_comparison;

/\* Apply selection criteria \*/

meets\_entropy = (entropy >= 0.60); /\* Adjust threshold \*/

meets\_avg\_pp = (avg\_pp >= 0.70);

meets\_min\_group = (min\_group\_pct >= 5.0);

/\* Combined flag \*/

meets\_all\_criteria = (meets\_entropy AND meets\_avg\_pp AND meets\_min\_group);

/\* BIC improvement over previous model \*/

IF \_N\_ > 1 THEN bic\_change = BIC - LAG(BIC);

RUN;

PROC PRINT DATA=HOME.model\_selection NOOBS;

TITLE 'Model Selection Criteria Applied';

VAR n\_groups BIC bic\_change entropy avg\_pp min\_group\_pct meets\_all\_criteria;

FORMAT BIC bic\_change 12.2 entropy avg\_pp 8.3 min\_group\_pct 8.2;

RUN;

/\* ============================================

PART 3: RUN FINAL MODEL WITH OPTIMAL GROUPS

Manually set optimal\_n after reviewing results

============================================ \*/

%LET optimal\_n = 4; /\* UPDATE THIS after reviewing model\_selection table \*/

%PUT ========================================;

%PUT Running final &optimal\_n-group solution;

%PUT ========================================;

PROC TRAJ DATA=HOME.lcmm\_data

OUTPLOT=HOME.final\_trajectories

OUTSTAT=HOME.final\_stats

OUT=HOME.final\_assignments;

ID INDI\_DSCM\_NO;

VAR hospital\_visits;

INDEP time;

MODEL ZIP; /\* Zero-Inflated Poisson \*/

/\* Adjust polynomial orders based on optimal solution \*/

%DO i = 1 %TO &optimal\_n;

ORDER &i 3 3 3 3; /\* Adjust as needed \*/

%END;

RUN;

/\* ============================================

PART 4: CREATE TRAJECTORY VISUALIZATION

With descriptive labels and group percentages

============================================ \*/

/\* Calculate group sizes and percentages \*/

PROC SQL;

CREATE TABLE HOME.group\_sizes AS

SELECT

\_GROUP AS group,

COUNT(\*) AS n\_patients,

100.0 \* COUNT(\*) / (SELECT COUNT(\*) FROM HOME.final\_assignments) AS pct FORMAT=5.1

FROM HOME.final\_assignments

GROUP BY \_GROUP

ORDER BY group;

QUIT;

/\* Add descriptive labels based on trajectory patterns \*/

/\* MANUAL STEP: Review trajectories and assign meaningful labels \*/

DATA HOME.group\_labels;

INPUT group $ label $50.;

DATALINES;

1 Low-Stable

2 Moderate-Increasing

3 High-Decreasing

4 Persistently High

;

RUN;

/\* Merge labels with sizes \*/

PROC SQL;

CREATE TABLE HOME.trajectory\_descriptions AS

SELECT

s.group,

l.label,

s.n\_patients,

s.pct,

CATX(' ', l.label, '(', PUT(s.pct, 5.1), '%)') AS group\_label

FROM HOME.group\_sizes AS s

LEFT JOIN HOME.group\_labels AS l

ON s.group = l.group

ORDER BY s.group;

QUIT;

/\* Create publication-ready trajectory plot \*/

PROC SGPLOT DATA=HOME.final\_trajectories NOAUTOLEGEND;

TITLE 'Trajectory Groups of Hospital Visit Patterns';

SERIES X=time Y=\_TRAJ\_GROUP1 / LINEATTRS=(THICKNESS=2 COLOR=blue)

LEGENDLABEL='Group 1' NAME='g1';

SERIES X=time Y=\_TRAJ\_GROUP2 / LINEATTRS=(THICKNESS=2 COLOR=red)

LEGENDLABEL='Group 2' NAME='g2';

SERIES X=time Y=\_TRAJ\_GROUP3 / LINEATTRS=(THICKNESS=2 COLOR=green)

LEGENDLABEL='Group 3' NAME='g3';

%IF &optimal\_n >= 4 %THEN %DO;

SERIES X=time Y=\_TRAJ\_GROUP4 / LINEATTRS=(THICKNESS=2 COLOR=orange)

LEGENDLABEL='Group 4' NAME='g4';

%END;

%IF &optimal\_n >= 5 %THEN %DO;

SERIES X=time Y=\_TRAJ\_GROUP5 / LINEATTRS=(THICKNESS=2 COLOR=purple)

LEGENDLABEL='Group 5' NAME='g5';

%END;

XAXIS LABEL='Quarter Since Treatment Completion' VALUES=(0 TO 20 BY 4);

YAXIS LABEL='Mean Number of Hospital Visits' MIN=0;

KEYLEGEND 'g1' 'g2' 'g3' %IF &optimal\_n >= 4 %THEN 'g4';

%IF &optimal\_n >= 5 %THEN 'g5'; / POSITION=topright;

RUN;

/\* Export high-resolution figure \*/

ODS GRAPHICS / RESET IMAGENAME='Figure1\_Trajectories' IMAGEFMT=tiff

WIDTH=8in HEIGHT=6in DPI=300;

ODS LISTING GPATH='/path/to/figures/';

PROC SGPLOT DATA=HOME.final\_trajectories NOAUTOLEGEND;

/\* Same plot code as above \*/

RUN;

ODS LISTING CLOSE;

/\* ============================================

PART 5: BASELINE CHARACTERISTICS TABLE

Comparing all trajectory groups

============================================ \*/

/\* Merge trajectory assignments with baseline data \*/

PROC SQL;

CREATE TABLE HOME.analysis\_dataset AS

SELECT

a.INDI\_DSCM\_NO,

a.\_GROUP AS trajectory\_group,

d.\*

FROM HOME.final\_assignments AS a

INNER JOIN HOME.lcmm\_data AS d

ON a.INDI\_DSCM\_NO = d.INDI\_DSCM\_NO

WHERE d.time = 0; /\* Baseline only \*/

QUIT;

/\* Create formatted baseline characteristics table \*/

%MACRO baseline\_table(varlist=, vartype=);

/\* For categorical variables \*/

%IF &vartype = CAT %THEN %DO;

PROC FREQ DATA=HOME.analysis\_dataset;

TABLES trajectory\_group \* (&varlist) / CHISQ NOCOL NOPERCENT;

TITLE "Baseline Characteristics by Trajectory Group - Categorical";

RUN;

%END;

/\* For continuous variables \*/

%IF &vartype = CONT %THEN %DO;

/\* Test for normality first \*/

PROC MEANS DATA=HOME.analysis\_dataset N MEAN STD MEDIAN Q1 Q3;

CLASS trajectory\_group;

VAR &varlist;

TITLE "Baseline Characteristics by Trajectory Group - Continuous";

RUN;

/\* ANOVA (if normal) \*/

PROC GLM DATA=HOME.analysis\_dataset;

CLASS trajectory\_group;

MODEL &varlist = trajectory\_group;

MEANS trajectory\_group / TUKEY;

TITLE "ANOVA: Comparing Trajectory Groups";

RUN;

QUIT;

/\* Kruskal-Wallis (if non-normal) \*/

PROC NPAR1WAY DATA=HOME.analysis\_dataset WILCOXON;

CLASS trajectory\_group;

VAR &varlist;

TITLE "Kruskal-Wallis Test: Comparing Trajectory Groups";

RUN;

%END;

%MEND;

/\* Run for your baseline variables \*/

%baseline\_table(varlist=SEX\_TYPE EPISODE GAIBJA\_TYPE SIDO RSLT\_FN KITDIV, vartype=CAT);

%baseline\_table(varlist=AGE\_NHIS hospital\_visits, vartype=CONT);

/\* Create comprehensive Table 1 \*/

PROC TABULATE DATA=HOME.analysis\_dataset FORMAT=8.1;

CLASS trajectory\_group SEX\_TYPE EPISODE GAIBJA\_TYPE RSLT\_FN KITDIV;

VAR AGE\_NHIS hospital\_visits;

TABLE

(SEX\_TYPE EPISODE GAIBJA\_TYPE RSLT\_FN='Treatment Success' KITDIV='Treatment Round') \* (N PCTN)

(AGE\_NHIS hospital\_visits) \* (MEAN STD MEDIAN),

trajectory\_group ALL / RTS=40;

TITLE 'Table 1: Baseline Characteristics by Trajectory Group';

RUN;

/\* Export Table 1 \*/

ODS EXCEL FILE='/path/to/output/Table1\_Baseline\_Characteristics.xlsx';

PROC TABULATE DATA=HOME.analysis\_dataset FORMAT=8.1;

/\* Same code as above \*/

RUN;

ODS EXCEL CLOSE;

/\* ============================================

PART 6: MULTINOMIAL LOGISTIC REGRESSION

Predictors of trajectory membership

============================================ \*/

/\* Set reference group (choose clinically meaningful, e.g., Low-Stable) \*/

%LET ref\_group = 1; /\* UPDATE based on your group labels \*/

/\* Univariable analyses \*/

%MACRO univariable\_mlogit(predictor=);

PROC LOGISTIC DATA=HOME.analysis\_dataset;

CLASS trajectory\_group (REF="&ref\_group") &predictor / PARAM=REF;

MODEL trajectory\_group = &predictor / LINK=glogit;

TITLE "Univariable Analysis: &predictor";

RUN;

%MEND;

%univariable\_mlogit(predictor=SEX\_TYPE);

%univariable\_mlogit(predictor=AGE\_NHIS);

%univariable\_mlogit(predictor=EPISODE);

%univariable\_mlogit(predictor=GAIBJA\_TYPE);

%univariable\_mlogit(predictor=INC5);

%univariable\_mlogit(predictor=SIDO);

%univariable\_mlogit(predictor=RSLT\_FN);

%univariable\_mlogit(predictor=KITDIV);

/\* Multivariable model \*/

PROC LOGISTIC DATA=HOME.analysis\_dataset;

CLASS trajectory\_group (REF="&ref\_group")

SEX\_TYPE EPISODE GAIBJA\_TYPE INC5 SIDO RSLT\_FN KITDIV / PARAM=REF;

MODEL trajectory\_group =

SEX\_TYPE AGE\_NHIS EPISODE GAIBJA\_TYPE INC5 SIDO RSLT\_FN KITDIV

/ LINK=glogit EXPB;

/\* Output odds ratios and confidence intervals \*/

ODS OUTPUT ParameterEstimates=HOME.multivariable\_results;

TITLE 'Multivariable Multinomial Logistic Regression';

TITLE2 'Predictors of Trajectory Group Membership';

RUN;

/\* Format multivariable results for publication \*/

DATA HOME.multivariable\_formatted;

SET HOME.multivariable\_results;

/\* Calculate odds ratio and 95% CI \*/

OR = EXP(Estimate);

OR\_lower = EXP(Estimate - 1.96 \* StdErr);

OR\_upper = EXP(Estimate + 1.96 \* StdErr);

/\* Format for presentation \*/

OR\_CI = CATX(' ', PUT(OR, 5.2), '(', PUT(OR\_lower, 5.2), '-', PUT(OR\_upper, 5.2), ')');

KEEP Variable trajectory\_group OR\_CI ProbChiSq;

RUN;

PROC PRINT DATA=HOME.multivariable\_formatted NOOBS;

TITLE 'Table 2: Multivariable Predictors of Trajectory Membership';

VAR Variable trajectory\_group OR\_CI ProbChiSq;

RUN;

/\* Export results \*/

PROC EXPORT DATA=HOME.multivariable\_formatted

OUTFILE='/path/to/output/Table2\_Multivariable\_Results.xlsx'

DBMS=XLSX REPLACE;

RUN;

/\* ============================================

PART 7: SUPPLEMENTAL MATERIALS

Additional model diagnostics and sensitivity analyses

============================================ \*/

/\* Average posterior probability by group \*/

PROC SQL;

CREATE TABLE HOME.avg\_pp\_by\_group AS

SELECT

\_GROUP AS group,

COUNT(\*) AS n,

MEAN(CASE

WHEN \_GROUP = 1 THEN GRP1PRB

WHEN \_GROUP = 2 THEN GRP2PRB

WHEN \_GROUP = 3 THEN GRP3PRB

WHEN \_GROUP = 4 THEN GRP4PRB

ELSE .

END) AS avg\_posterior\_prob FORMAT=5.3

FROM HOME.final\_assignments

GROUP BY \_GROUP;

QUIT;

PROC PRINT DATA=HOME.avg\_pp\_by\_group NOOBS;

TITLE 'Supplemental Table S1: Average Posterior Probabilities by Group';

RUN;

/\* Odds of correct classification \*/

PROC SQL;

CREATE TABLE HOME.classification\_odds AS

SELECT

\_GROUP AS group,

MEAN(CASE

WHEN \_GROUP = 1 THEN GRP1PRB / (1 - GRP1PRB)

WHEN \_GROUP = 2 THEN GRP2PRB / (1 - GRP2PRB)

WHEN \_GROUP = 3 THEN GRP3PRB / (1 - GRP3PRB)

WHEN \_GROUP = 4 THEN GRP4PRB / (1 - GRP4PRB)

ELSE .

END) AS avg\_odds FORMAT=8.2

FROM HOME.final\_assignments

GROUP BY \_GROUP;

QUIT;

PROC PRINT DATA=HOME.classification\_odds NOOBS;

TITLE 'Supplemental Table S2: Average Odds of Correct Classification';

RUN;

%PUT ========================================;

%PUT GBTM Analysis Complete!;

%PUT Review results and update group labels;

%PUT ========================================;