SUPPLEMENTAL DIGITAL CONTENT 1

Additional Descriptive Figures

The figures below demonstrate differences in discontinuity in retention by age group (higher proportions at younger ages), race/ethnicity (higher proportions among black patients), and HIV acquisition risk (higher proportions among those with injection drug use risk factor). Notably, population proportions with discontinuities in care in any given calendar period are fairly stable and low across contributing cohorts (7-17%) (Figures 1a-1e).

Results and Fit Diagnostics from Additional Models

Results from Beta-binomial, zero-inflated binomial, and mixed effects logistic regressions show substantial consistency of behavior within individuals, over time (high ICC indicated by ρ between 0.2 and 0.5) (Table 1). The parameter ρ here is a measure of behavioral tracking over time, where values closer to 1 indicate tighter within-individual correlation and overdispersion with respect to a binomial distribution of outcome counts.

Predicted values derived from all models followed the observed distribution of discontinuities remarkably well (Figures 2a-2c). The ability of all models aside from standard logistic regression to capture the high proportion of individuals with no observed discontinuities in retention were similar, evidenced by good fit statistics, high ρ, tight overlap between predicted and observed probability distributions generated by linear prediction (Figures 2a-2c), and low dispersion with small deviations at the tails in quantile-quantile plots of predicted vs. observed counts and probabilities (Figures 3a-3c).

Explanation of Data Preparation and Model Selection

For the primary analysis with beta-binomial regression, the observation period was split in half to mitigate potential statistical artifacts of data truncation and misclassification bias; that

is, patients with uncaptured documentation of clinical care prior to cohort entry or after cohort exit could be misclassified as being in care (by definition) in the first and last years of data contribution. The bias could also be described as left- and right-censoring due to absence of information. By splitting the dataset before modeling, the models are only susceptible to this possible bias in a single time period (either 2000 or 2008), instead of two. In any case, estimates from a beta-binomial regression fit to the entire study period are similar to those from separate beta-binomial regressions fit to observations from 2000-2004 and 2005-2008.

With regard to the statistical models, beta-binomial, zero-inflated, and mixed effects models with random intercepts were used in this analysis because, even though the outcome is a simple binary response (either having a discontinuity in retention or not), it is a reasonable assumption and indeed an observation in this dataset that patients who are routinely in care up to some time point are actually less likely to miss future visits (and conversely, that patients with several discontinuities up to some time point are just as likely to continue experiencing discontinuities in retention). In other words, the probabilities of discontinuity are not uniform across individuals, and there is within-individual clustering of outcomes over time.

More complete descriptions of the construction and application of such methods to assess binary outcomes and within-individual tracking of behavior over time have been presented in extensive detail elsewhere [1-4].

SUPPLEMENTAL DIGITAL CONTENT 1

Figure 1a. "Churn" in the NA-ACCORD by Age Group (5-year Intervals), 2000-2008

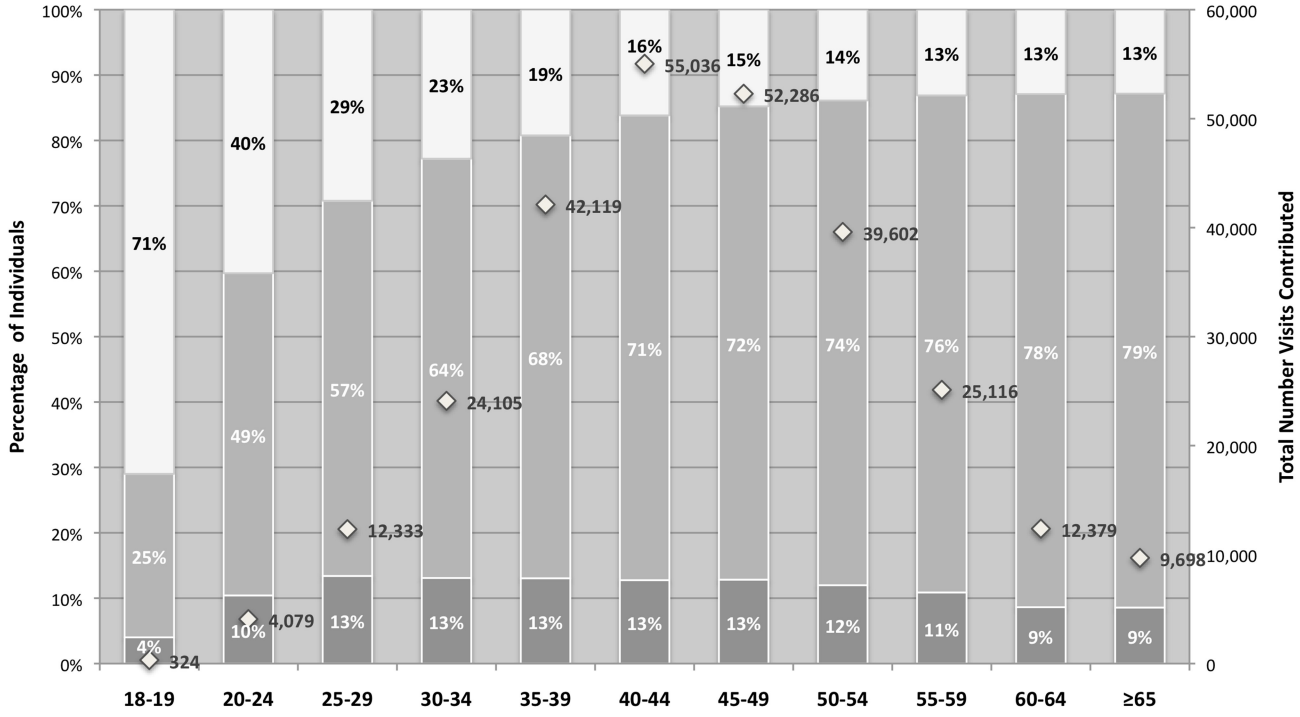


Figure 1b. "Churn" in the NA-ACCORD by Sex, 2000-2008

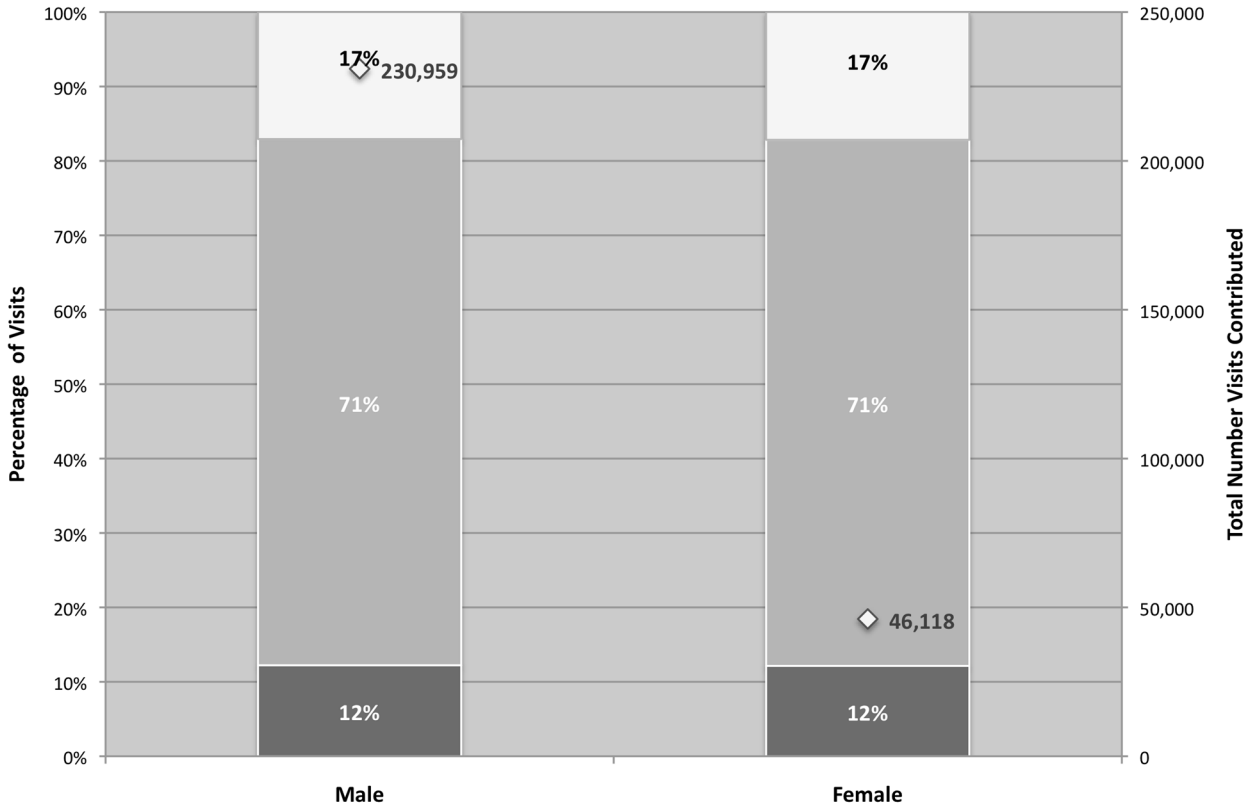


Figure 1c. "Churn" in the NA-ACCORD by HIV Acquisition Risk, 2000-2008

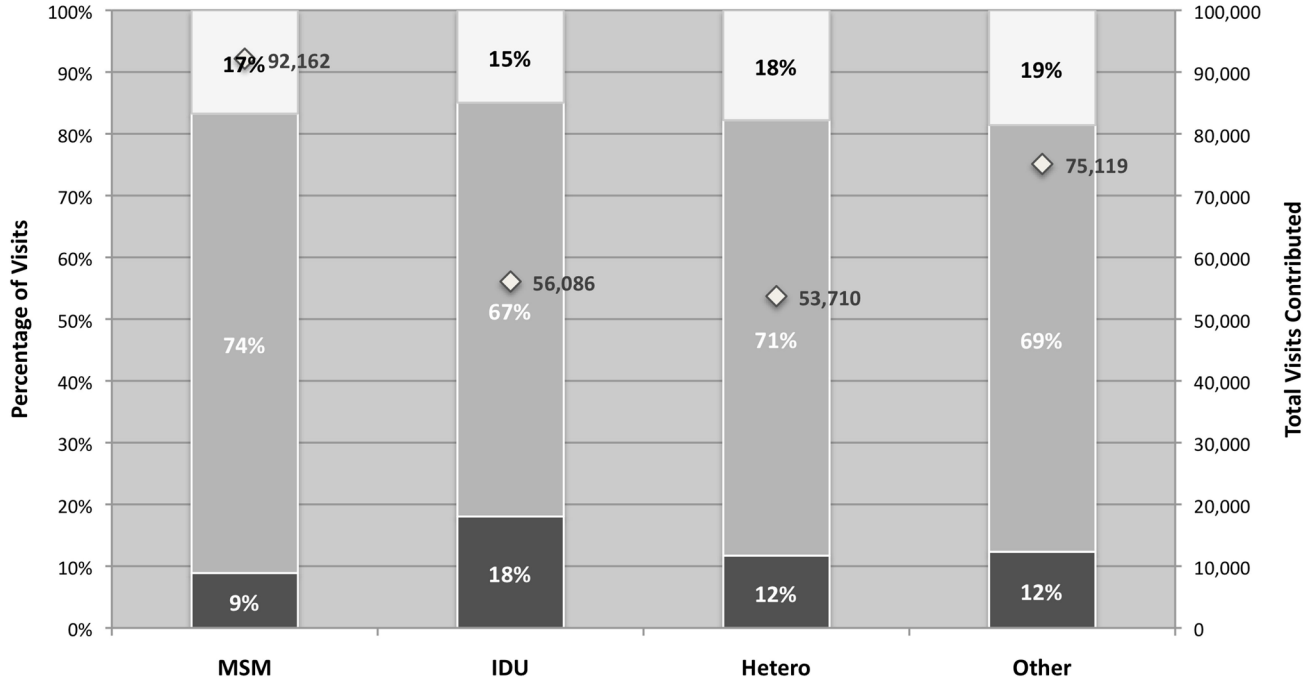
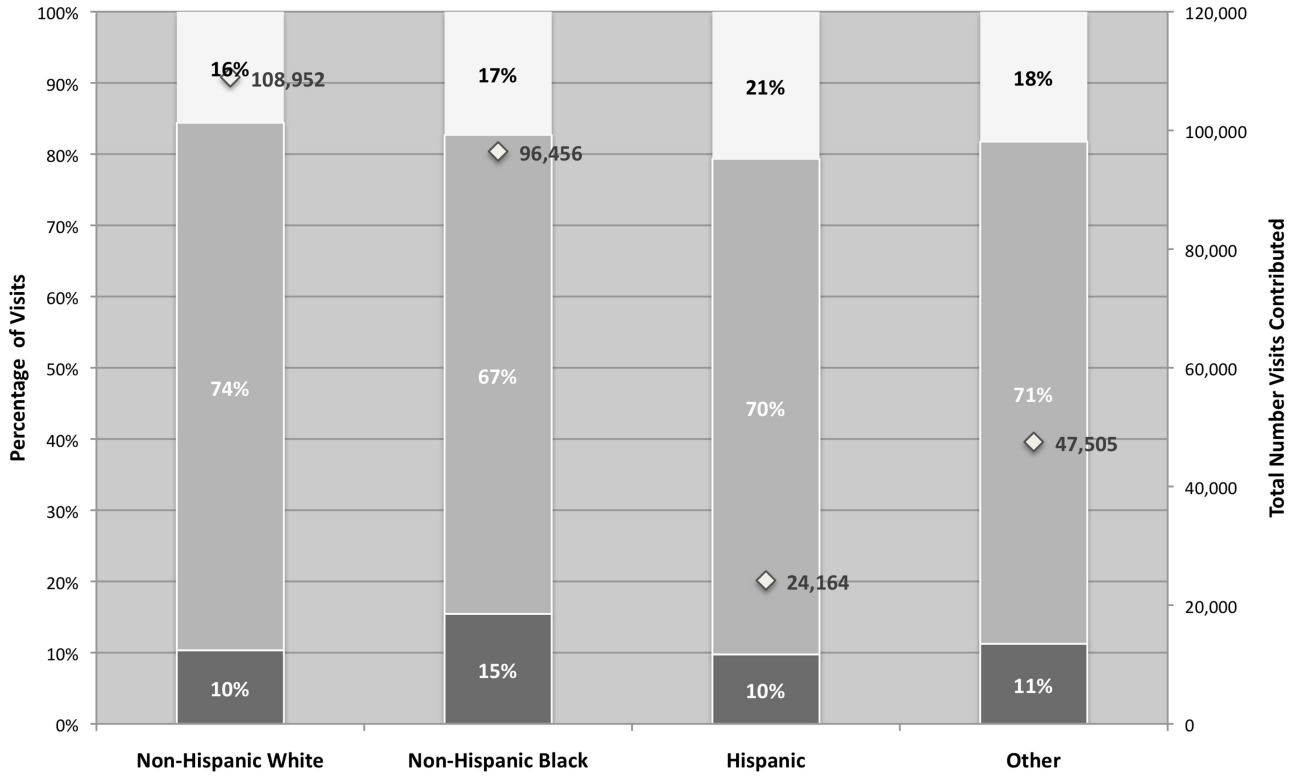
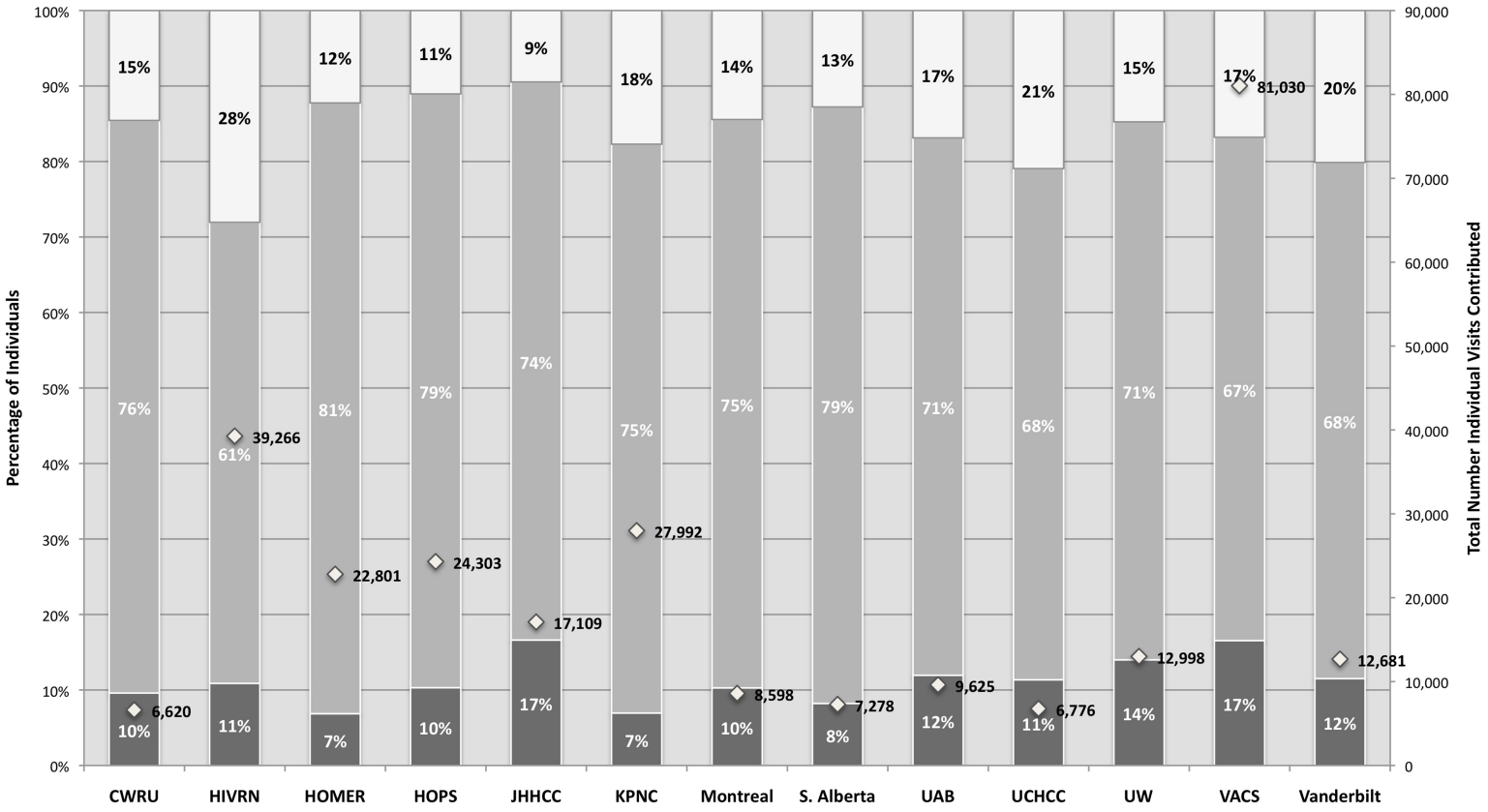


Figure 1d. "Churn" in the NA-ACCORD by Race/Ethnicity, 2000-2008



SUPPLEMENTAL DIGITAL CONTENT 1

Figure 1e. "Churn" in the NA-ACCORD by Contributing Cohort, 2000-2008



- **Entered Care**: initial HIV lab under clinical observation in NA-ACCORD
- **Retained in Care**: second HIV lab was within 12 months of, but >90 days after, prior lab
- **Exited Care** (Gap Year): failed to meet "Retained in Care" definition
- **Total Contributing to Churn**

SUPPLEMENTAL DIGITAL CONTENT 1

Table 1. Odds Ratios (OR) and 95% Confidence Intervals (95% CI) of ≥1 year of discontinuity in retention among 61,438 individuals contributing ≥1 HIV-lab between 2000 and 2008, by demographic and clinical factors and statistical model type

Factor	Beta-binomial* 2000-2004		Beta-binomial† 2005-2008		Beta-binomial		Zero-Inflated Binomial		Mixed Effects, Intercepts Only		Mixed Effects, Intercepts & Slopes		Multiple Logistic, Any vs. No "Gaps"	
	OR	**(95% CI)**	**OR**	**(95% CI)**	**OR**	**(95% CI)**	**OR**	**(95% CI)**	**OR**	**(95% CI)**	**OR**	**(95% CI)**	**OR**	**(95% CI)**
Female Sex	**0.82**	(0.70, 0.95)	**1.11**	(0.91, 1.36)	**0.96**	(0.92, 1.01)	**0.92**	(0.88, 0.95)	**0.95**	(0.89, 1.02)	**0.96**	(0.89, 1.02)	**0.94**	(0.91, 0.97)
Age (per 10 y)	**0.78**	(0.74, 0.83)	**0.80**	(0.74, 0.86)	**0.79**	(0.77, 0.80)	**0.84**	(0.83, 0.85)	**0.70**	(0.69, 0.72)	**0.71**	(0.69, 0.73)	**0.77**	(0.76, 0.78)
Non-Hisp. Black (vs. Non-Hisp. White)	**1.31**	(1.16, 1.49)	**1.36**	(1.15, 1.62)	**1.26**	(1.21, 1.32)	**1.16**	(1.12, 1.20)	**1.38**	(1.30, 1.46)	**1.38**	(1.30, 1.46)	**1.26**	(1.23, 1.30)
IDU Risk (vs. non-IDU)	**1.68**	(1.49, 1.89)	**1.34**	(1.11, 1.61)	**1.58**	(1.51, 1.64)	**1.33**	(1.29, 1.38)	**1.95**	(1.83, 2.07)	**1.93**	(1.82, 2.05)	**1.59**	(1.55, 1.63)
ART (at baseline)	**0.61**	(0.54, 0.68)	**0.73**	(0.63, 0.84)	**0.69**	(0.66, 0.71)	**0.70**	(0.68, 0.72)	**0.60**	(0.57, 0.63)	**0.60**	(0.57, 0.63)	**0.65**	(0.64, 0.67)
Time Since Enrollment (per year)	**1.09**	(1.07, 1.11)	**1.22**	(1.16, 1.29)	**1.13**	(1.12, 1.14)	**1.11**	(1.11, 1.12)	**1.17**	(1.17, 1.18)	**1.18**	(1.17, 1.18)	**1.12**	(1.12, 1.12)
Intra-class Correlation (ρ)	**0.24**	(0.22, 0.26)	**0.20**	(0.17, 0.23)	**0.27**	(0.26, 0.27)	**0.28**	(0.27, 0.30)	**0.51**	(0.50, 0.52)	N/A		N/A	

All models adjusted for contributing cohort; Time since enrollment: years under clinical observation in NA-ACCORD
Non-overlapping periods of 5 & 4 years used to eliminate artifacts of truncation in separate Beta-binomial models
Mixed Effects model random intercepts are by individual and random slopes are by contributing cohort
*: Model includes individuals entering care in 2000 contributing data 2000-2004, N=5,718
†: Model includes individuals entering care in 2005 contributing data 2005-2008, N=5,071

SUPPLEMENTAL DIGITAL CONTENT 1

Figure 2a. Predicted probabilities of being "out of care" from multivariable Beta-binomial regression (mean from transformed linear predictors, including ρ) in BLACK. Observed probabilities (per individual over follow-up) in GRAY.

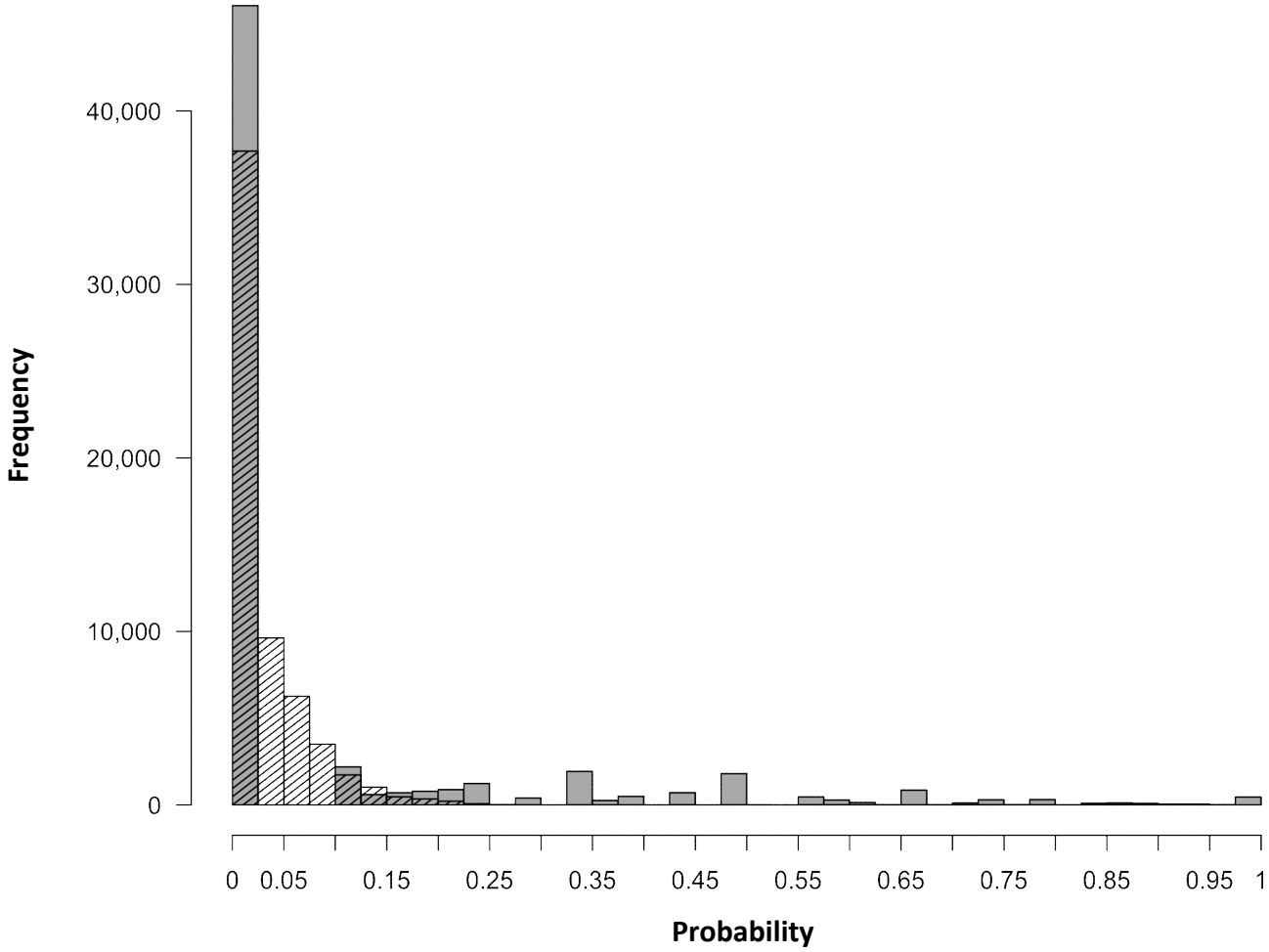


Figure 2b. Predicted probabilities of being "out of care" from multivariable Zero-inflated binomial regression (mean from transformed linear predictors, including □) in BLACK. Observed probabilities (per individual over follow-up) in GRAY.

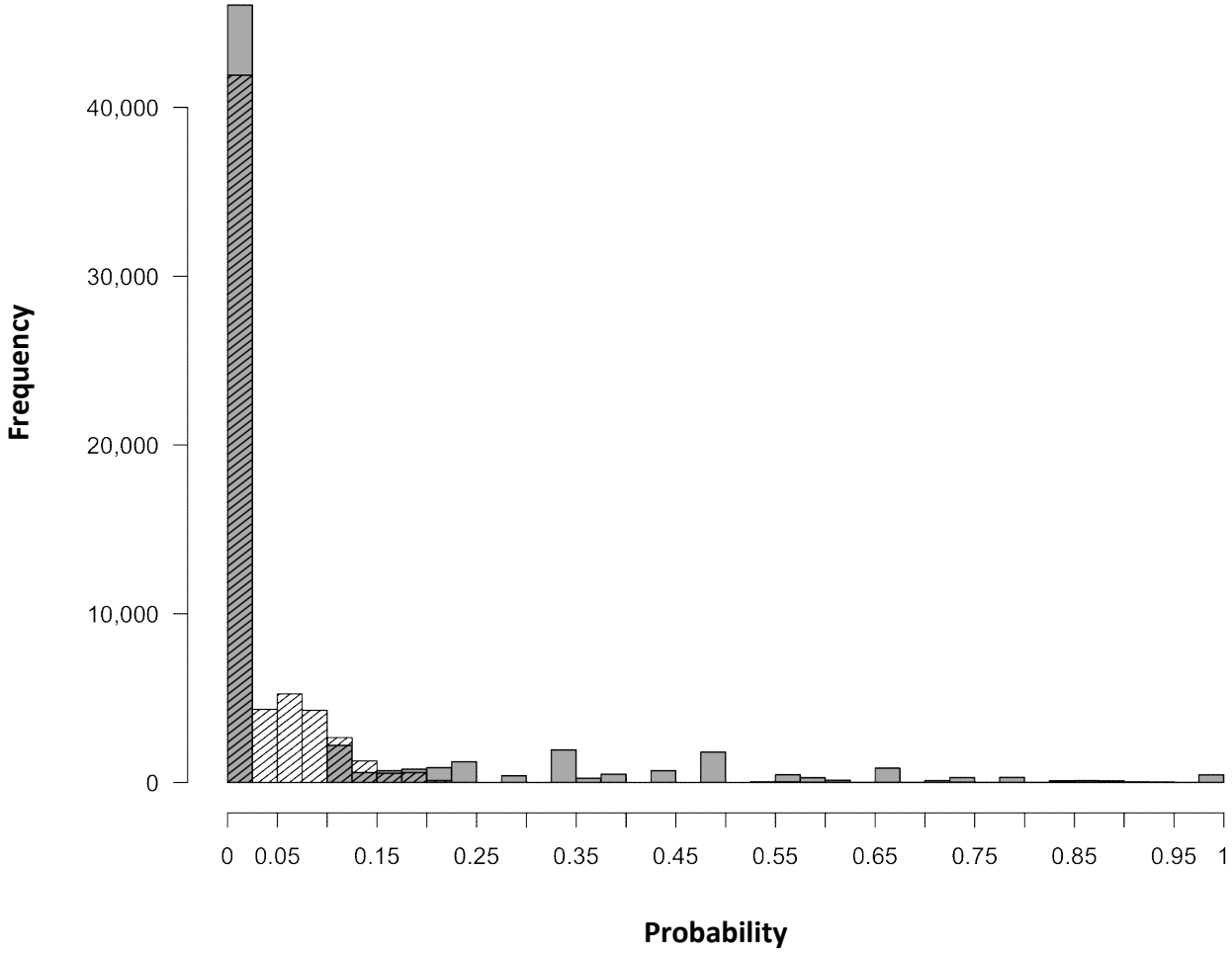
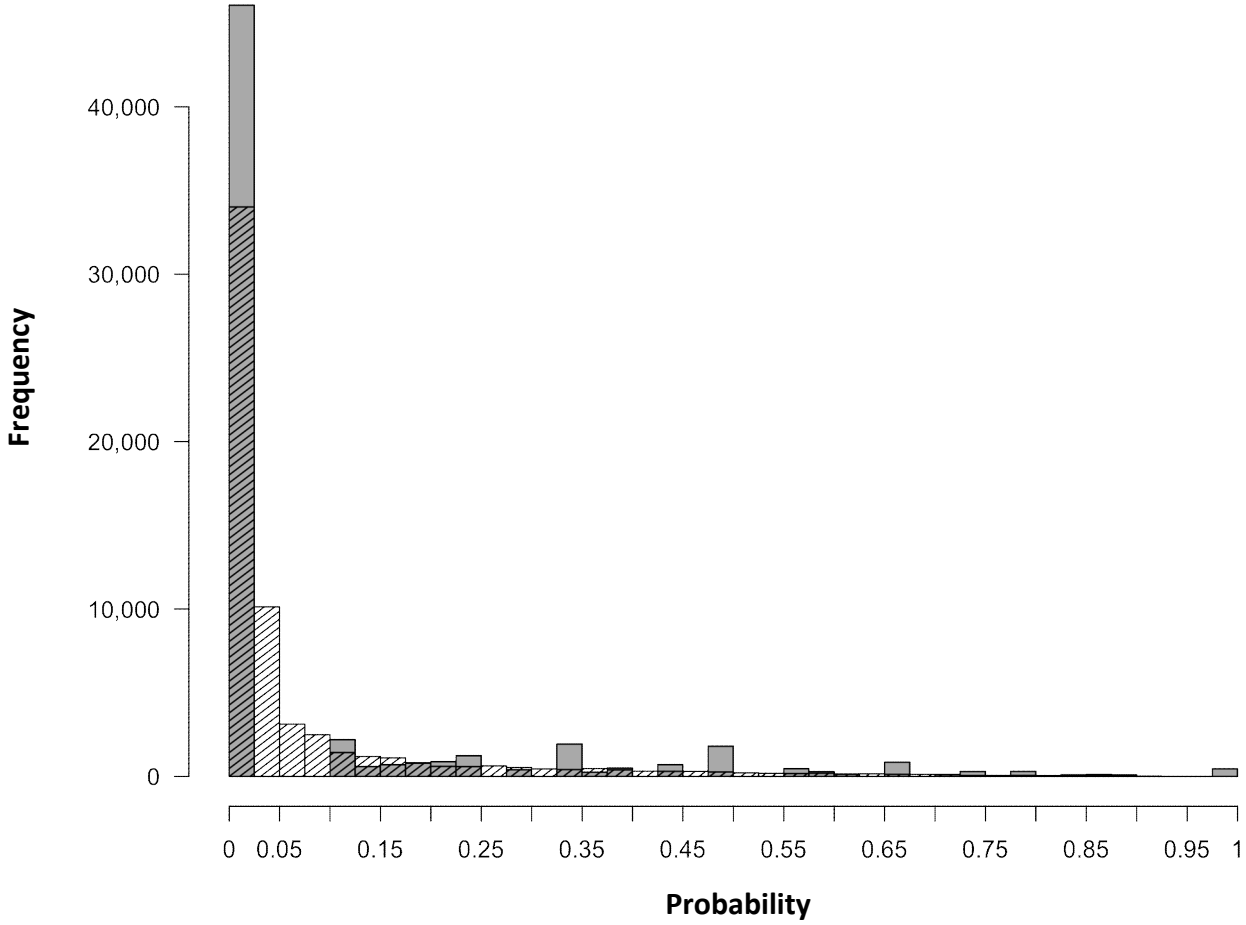


Figure 2c. Predicted probabilities of being "out of care" from multivariable mixed effects logistic regression (mean from transformed linear predictors, with random

effects) in BLACK. Observed probabilities (per individual over follow-up) in GRAY.



SUPPLEMENTAL DIGITAL CONTENT 1

Figure 3a. Quantile-quantile plot of observed vs. predicted numbers of "out of care" years (or "gaps" in care) over follow-up from multivariable Beta-binomial regression

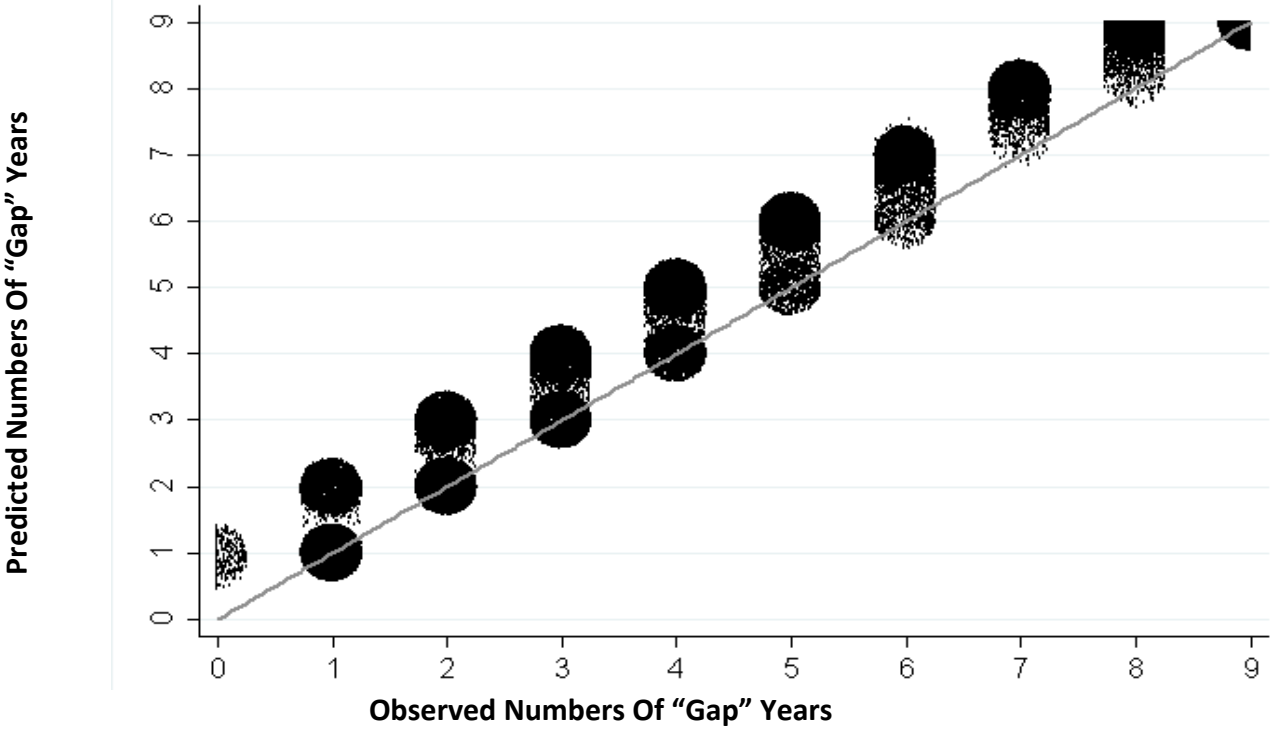
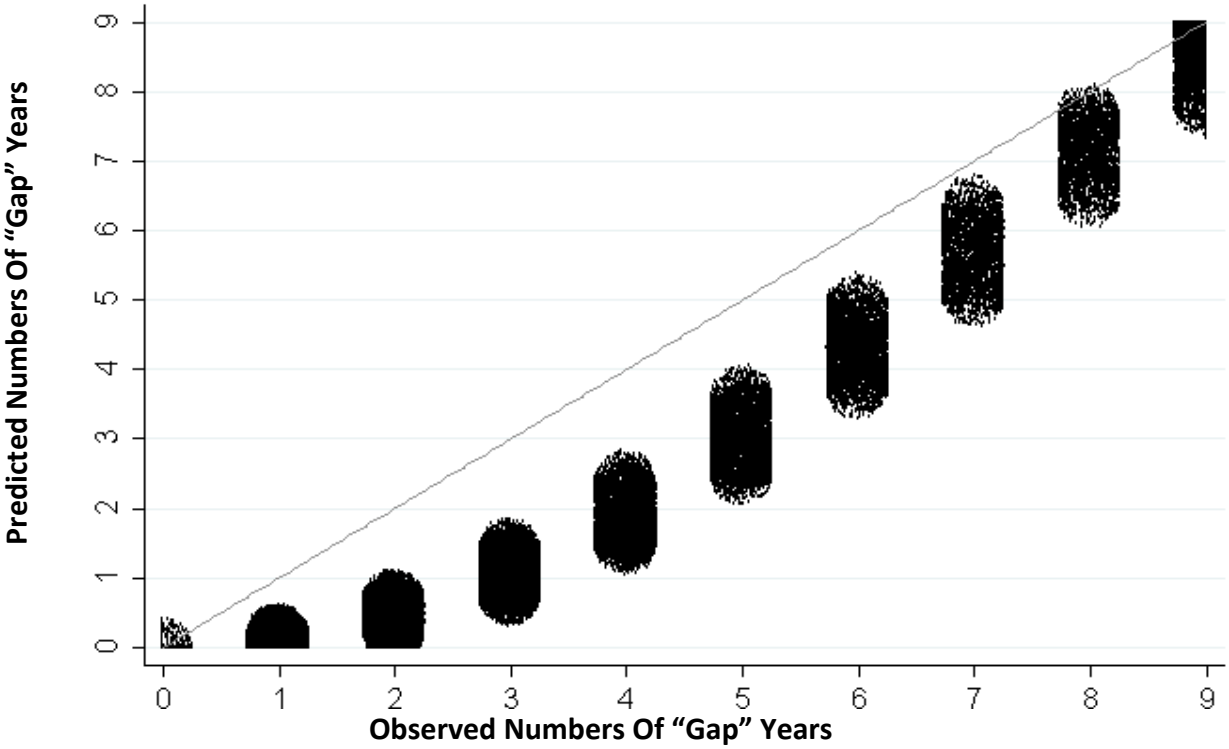
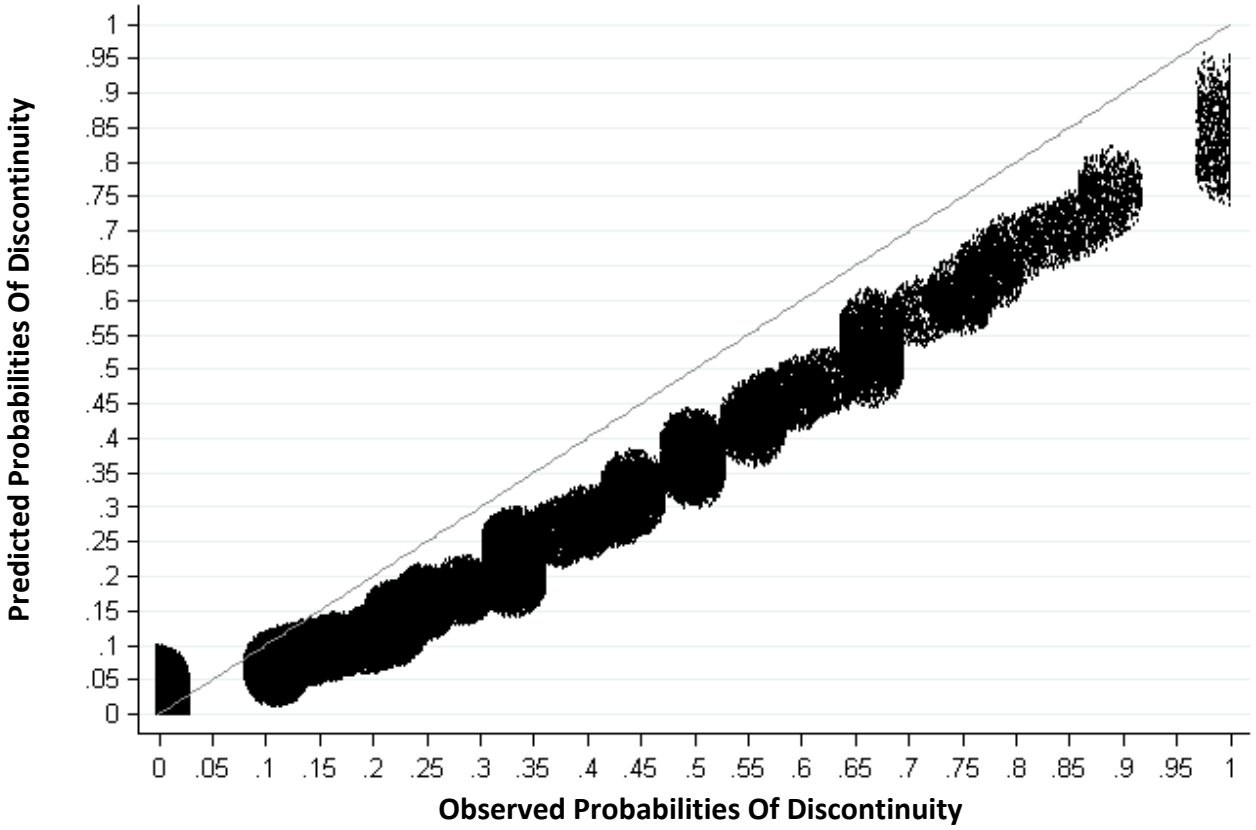


Figure 3b. Quantile-quantile plot of observed vs. predicted numbers of "out of care" years (or "gaps" in care) over follow-up from multivariable Zero-inflated binomial regression



SUPPLEMENTAL DIGITAL CONTENT 1

Figure 3c. Quantile-quantile plot of observed vs. predicted probabilities of being "out of care" over follow-up from multivariable mixed effects logistic regression, random intercepts and slopes by cohort



SUPPLEMENTAL DIGITAL CONTENT 1

Code for Regression Models

Separate beta-binomial models for those entering in 2000 and 2005 (in R):

```
tmp <- read.table("/Users/…/betabinomial_2005Split_revised.csv", header=TRUE,
 sep=",")

smallset00 <- subset(tmp, yearmin==2000)
smallset05 <- subset(tmp, yearmin==2005)

## N=5,718 for 2000
## N=5,071 for 2005

#bb00 <- vglm(cbind(zi,ni-zi)~1,family=betabinomial,data=smallset3)
#vglm model gives two linear predictors:
#   logit(mu) = linear predictor1
#   logit(rho) = linear predictor2
# basic model fits the linear predictors to mu
# mu is interpretable as prob(event)
# rho is the correlation
# (these interpretations based on betabinomial function)
# when looking at output
# Intercept:1 is the intercept for logit(mu) = ~1 + ...
# Intercept:2 is the intercept for logit(rho) = ~1
# To get value of rho, expit(Intercept:2)
# Other parmeters interpretable just like logistic regression

bb00.1 <- vglm(cbind(z2i,ni-
 z2i)~1+sex+age+as.factor(race)+idu+haart+timesinceenroll+as.factor(cohortid),
 family=betabinomial,data=smallset00)

bb05.1 <- vglm(cbind(z2i,ni-
 z2i)~1+sex+age+as.factor(race)+idu+haart+timesinceenroll+as.factor(cohortid),
 family=betabinomial,data=smallset05)

#note: country will not fit if cohort is included!

#you can look at summary(bb00.1) and summary(bb05.1) to see summary of the
 coefficients

summary (bb00.1)

# From model fit, can get estimates and 95% CI
#2000 model
#sex
exp(-0.201065 + 1.96*c(-0.0768908, 0, 0.0768908))
#age per 10 years
exp(-0.24372 + 1.96*c(-0.030322, 0, 0.030322))
#race2 vs 1
exp(0.273188 +1.96*c(-0.0650849, 0, 0.0650849))
#IDU vs. non-IDU risk
exp(0.516331 + 1.96*c(-0.0609316, 0, 0.0609316))
#haart
exp(-0.502012 + 1.96*c(-0.0556544, 0, 0.0556544))
#time in care
exp(0.084815 + 1.96*c(-0.0102848 , 0, 0.0102848))
```

SUPPLEMENTAL DIGITAL CONTENT 1

```
#rho - intraclass correlation
expit(-1.165717 + 1.96*c(-0.0559595, 0, 0.0559595))

summary(bb05.1)

#2005 model
#sex
exp(0.1044675 + 1.96*c(-0.1018003,0, 0.1018003))
#age per 10 years
exp(-0.224282 + 1.96*c(-0.036714,0, 0.036714))
#race2 vs 1
exp(0.3095516 +1.96*c(-0.0877028 ,0, 0.0877028))
#IDU vs. non-IDU risk
exp(0.2901161 + 1.96*c(-0.0942344,0, 0.0942344))
#haart
exp(-0.3206143 + 1.96*c(-0.0747905, 0, 0.0747905))
#time in care
exp(0.2022460 + 1.96*c(-0.0258403, 0, 0.0258403))
#rho - intraclass correlation
expit(-1.3944666 + 1.96 * c(-0.0896366,0, 0.0896366))


# predicted values
#this computes the predicted value of mu and rho for each
#person and then uses these to get the probability of 0,1,2,3
#missing visits.  I then average over all individuals to get the
#model estimate probabilities

bb00.pred <- expit(predict(bb00.1))
bb00.preddist0 <- dbetabin(0,5,bb00.pred[,1],bb00.pred[,2])
bb00.preddist1 <- dbetabin(1,5,bb00.pred[,1],bb00.pred[,2])
bb00.preddist2 <- dbetabin(2,5,bb00.pred[,1],bb00.pred[,2])
bb00.preddist3 <- dbetabin(3,5,bb00.pred[,1],bb00.pred[,2])
bb00.preddist4 <- dbetabin(4,5,bb00.pred[,1],bb00.pred[,2])
mean(bb00.preddist0)
mean(bb00.preddist1)
mean(bb00.preddist2)
mean(bb00.preddist3)
mean(bb00.preddist4)

bb05.pred <- expit(predict(bb05.1))
bb05.preddist0 <- dbetabin(0,4,bb05.pred[,1],bb05.pred[,2])
bb05.preddist1 <- dbetabin(1,4,bb05.pred[,1],bb05.pred[,2])
bb05.preddist2 <- dbetabin(2,4,bb05.pred[,1],bb05.pred[,2])
bb05.preddist3 <- dbetabin(3,4,bb05.pred[,1],bb05.pred[,2])
mean(bb05.preddist0)
mean(bb05.preddist1)
mean(bb05.preddist2)
mean(bb05.preddist3)

#observed values
#look to see how many people have perfect followup (ni = 4)
# and then look at the distribution of zi = 0, 1, 2, 3, 4
 table(smallset00$ni,smallset00$z2i)
 c(4134,764,447,223,150)/5718
 table(smallset05$ni,smallset05$z2i)
 c(4255,518,241,57)/5071
```

SUPPLEMENTAL DIGITAL CONTENT 1

```
smallset00$bbpred <- predict(bb00.1, type="response")
smallset05$bbpred <- predict(bb05.1, type="response")

bbh00 <- hist(smallset00$bbpred, breaks=seq(0,1,by=0.025))
obsh00 <- hist(smallset00$p2i, breaks=seq(0,1,by=0.025))

bbh05 <- hist(smallset05$bbpred, breaks=seq(0,1,by=0.025))
obsh05 <- hist(smallset05$p2i, breaks=seq(0,1,by=0.025))

par(mar=c(4, 5, 2, 0.5))

plot(obsh00, col="darkblue", las=0, ylim=c(0,4500), cex.lab=1.5,
 main="Observed vs. Predicted Prob. of Discontinuity by Beta-binomial, 2000-
 2004", xlab="Probability of Discontinuity", ylab="Frequency")
par(new=TRUE)
plot(bbh00, col="red", xaxt="n", yaxt="n", xlab="", ylab="", las=0,
 ylim=c(0,4500), main=" ")

plot(obsh05, col="darkblue", las=0, ylim=c(0,4500), cex.lab=1.5,
 main="Observed vs. Predicted Prob. of Discontinuity by Beta-binomial, 2005-
 2008", xlab="Probability of Discontinuity", ylab="Frequency")
par(new=TRUE)
plot(bbh05, col="red", xaxt="n", yaxt="n", xlab="", ylab="", las=0,
ylim=c(0,4500), main=" ")
```

SUPPLEMENTAL DIGITAL CONTENT 1

Beta-binomial and Zero-inflated binomial models (in R):

```
tmp <- read.table("/Users/…/Beta_and_ZIbinomial.csv", header=TRUE, sep=",")

## PACKAGE MANAGER: Load "locfit" and "VGAM"

#bb00 <- vglm(cbind(zi,ni-zi)~1,family=betabinomial,data=smallset3)
#vglm model gives two linear predictors:
#   logit(mu) = linear predictor1
#   logit(rho) = linear predictor2
# basic model fits the linear predictors to mu
# mu is interpretable as prob(event)
# rho is the correlation
# (these interpretations based on betabinomial function)
# when looking at output
# Intercept:1 is the intercept for logit(mu) = ~1 + ...
# Intercept:2 is the intercept for logit(rho) = ~1
# To get value of rho, expit(Intercept:2)
# Other parmeters interpretable just like logistic regression

bbage <- vglm(cbind(z2i,ni-
 z2i)~1+age+as.factor(cohortid),family=betabinomial,data=tmp)
exp(-0.14189 + 1.96*c(-0.0086586 , 0, 0.0086586 ))
pt(-16.38770, 122861)

bbsex <- vglm(cbind(z2i,ni-
 z2i)~1+sex+as.factor(cohortid),family=betabinomial,data=tmp)
exp(0.132131 + 1.96*c(-0.023819 , 0, 0.023819 ))
pt( 5.54734, 122861, lower.tail=FALSE)

bbrace <- vglm(cbind(z2i,ni-
 z2i)~1+as.factor(race)+as.factor(cohortid),family=betabinomial,data=tmp)
exp(0.290334 + 1.96*c(-0.020353 , 0, 0.020353 ))
pt( 14.26529, 122861, lower.tail=FALSE)

bbidu <- vglm(cbind(z2i,ni-
 z2i)~1+idu+as.factor(cohortid),family=betabinomial,data=tmp)
exp(0.526956 + 1.96*c(-0.019868 , 0, 0.019868))
pt(26.52237, 122861, lower.tail=FALSE)

bbhaart <- vglm(cbind(z2i,ni-
 z2i)~1+haart+as.factor(cohortid),family=betabinomial,data=tmp)
exp(-0.408673  + 1.96*c(-0.017422 , 0, 0.017422))
pt(-23.4573, 122861)

bbcd4 <- vglm(cbind(z2i,ni-
 z2i)~1+CD4N+as.factor(cohortid),family=betabinomial,data=tmp)
exp(-0.082111 + 1.96*c(-0.0033794, 0, 0.0033794))
pt(-24.29768, 122861)

bbvl <- vglm(cbind(z2i,ni-
 z2i)~1+logvload+as.factor(cohortid),family=betabinomial,data=tmp)
exp(0.2472194 + 1.96*c(-0.006806  , 0, 0.006806 ))
pt(36.3237, 122861, lower.tail=FALSE)
```

SUPPLEMENTAL DIGITAL CONTENT 1

```
bbcountry <- vglm(cbind(z2i,ni-z2i)~1+country,family=betabinomial,data=tmp)
exp(-0.54384 + 1.96*c(- 0.0289388, 0,  0.0289388))
pt(-18.793, 122861)

bbtime <- vglm(cbind(z2i,ni-
 z2i)~1+timesinceenroll+as.factor(cohortid),family=betabinomial,data=tmp)
exp(0.104535 + 1.96*c(-0.002518 , 0, 0.002518))
pt(41.51458, 122861, lower.tail=FALSE)

#note: country will not fit if cohort is included!

#you can look at summary(bb00.1) to see summary of the coefficients
bb00.1 <- vglm(cbind(z2i,ni-
 z2i)~1+sex+age+as.factor(cohortid)+idu+timesinceenroll,family=betabinomial,da
 ta=tmp)
# From model fit, can get estimates and 95% CI
summary(bb00.1)

#sex
exp(-0.036201 + 1.96*c(-0.02459664, 0, 0.02459664))
#age per 10 years
exp(-0.23915 + 1.96*c(-0.0094905, 0, 0.0094905))
#race2 vs 1
exp(0.232643 +1.96*c(-0.02118001, 0, 0.02118001))
#IDU vs. non-IDU risk
exp(0.455505 + 1.96*c(-0.02061255, 0, 0.02061255))
#haart
exp(-0.376982 + 1.96*c(-0.01791416, 0, 0.01791416))
#timeincare per year
exp(0.122830 + 1.96*c(-0.00260939, 0, 0.00260939))
#rho - intraclass correlation
expit(-1.025137 + 1.96*c(-0.01607362, 0, 0.01607362))
#mu - predicted mean prob
expit(-1.573719 + 1.96*c(-0.04864572, 0, 0.04864572))

bb00.1 <- vglm(cbind(z2i,ni-
 z2i)~1+sex+age+as.factor(race)+as.factor(cohorttype)+as.factor(risk)+timesinc
 eenroll,family=betabinomial,data=tmp)

# From model including cohort-type (US Centr., US Decentr., Canadian)
#sex
exp(-0.167316 + 1.96*c(-0.0340949, 0, 0.0340949))
#age per 10 years
exp(-0.21847  + 1.96*c(-0.011302, 0, 0.011302))
#race2 vs 1
exp(0.253462 +1.96*c(-0.0264523, 0, 0.0264523))
#risk2 vs 1
exp(0.766652 + 1.96*c(-0.0337535, 0, 0.0337535))
#timeincare per year
exp(0.194235 + 1.96*c(-0.0042959, 0, 0.0042959))
#Canadian vs. US Centr.
exp(-0.910223 + 1.96*c(-0.0499774, 0, 0.0499774))
#rho - intraclass correlation
expit(-1.191908 + 1.96*c(-0.0216113, 0, 0.0216113))

# predicted values
#this computes the predicted value of mu and rho for each
```

```
#person and then uses these to get the probability of 0,1,2,3
#missing visits.  I then average over all individuals to get the
#model estimate probabilities

bb00.pred <- expit(predict(bb00.1))
bb00.preddist0 <- dbetabin(0,9,bb00.pred[,1],bb00.pred[,2])
bb00.preddist1 <- dbetabin(1,9,bb00.pred[,1],bb00.pred[,2])
bb00.preddist2 <- dbetabin(2,9,bb00.pred[,1],bb00.pred[,2])
bb00.preddist3 <- dbetabin(3,9,bb00.pred[,1],bb00.pred[,2])
bb00.preddist4 <- dbetabin(4,9,bb00.pred[,1],bb00.pred[,2])
bb00.preddist5 <- dbetabin(5,9,bb00.pred[,1],bb00.pred[,2])
bb00.preddist6 <- dbetabin(6,9,bb00.pred[,1],bb00.pred[,2])
bb00.preddist7 <- dbetabin(7,9,bb00.pred[,1],bb00.pred[,2])
bb00.preddist8 <- dbetabin(8,9,bb00.pred[,1],bb00.pred[,2])

bb00.denspred <- dbetabin(tmp$zi,tmp$ni,bb00.pred[,1],bb00.pred[,2])
bb00.distpred <- pbetabin(tmp$zi, tmp$ni, bb00.pred[,1], bb00.pred[,2])
bb00.countpred <- bb00.distpred*tmp$ni

mean(bb00.preddist0)
mean(bb00.preddist1)
mean(bb00.preddist2)
mean(bb00.preddist3)
mean(bb00.preddist4)
mean(bb00.preddist5)
mean(bb00.preddist6)
mean(bb00.preddist7)
mean(bb00.preddist8)

#observed values
#look to see how many people have perfect followup (ni = 9)
# and then look at the distribution of zi = 0, 1, 2, 3, 4, etc.
 table(tmp$ni,tmp$z2i)
 c(46078,7083,3649,1947,1201,684,403,218,116,59)/61438

bbh <- hist(bb00.denspred, breaks=seq(0,1,by=0.025))
obsh <- hist(tmp$p2i, breaks=seq(0,1,by=0.025))

par(mar=c(4, 5, 2, 0.5))
xpos<- seq(0,1, by=0.05)

plot(obsh, col="black", las=0, xaxt="n", ylim=c(0,45000), cex.lab=1.5,
 main="Observed vs. Predicted Prob. of Discontinuity from Beta-binomial",
 xlab="Probability of Discontinuity", ylab="Frequency")
par(new=TRUE)
plot(bbh, col="gray", xaxt="n", yaxt="n", xlab="", ylab="", las=0,
 ylim=c(0,45000), main=" ")
axis(1, at=xpos, labels=xpos)

qqplot(tmp$p2i, bb00.denspred, ylim=c(0,1), xlab="Observed Prob.
 Discontinuity", ylab="Predicted Prob. Discontinuity", main="Q-Q Plot,
 Observed vs. Predicted Prob. of Discontinuity from Beta-binomial")
par(new=TRUE)
abline(0,1)

bbcounth <- hist(bb00.countpred, breaks=seq(0,9,by=1))
obscounth <- hist(tmp$zi, breaks=seq(0,9,by=1))
```

SUPPLEMENTAL DIGITAL CONTENT 1

```
par(mar=c(4, 5, 2, 0.5))
xpos<- seq(0,9, by=1)

plot(obscounth, col="black", las=0, xaxt="n", ylim=c(0,45000), cex.lab=1.5,
 main="Observed vs. Predicted # of Discontinuous Visits from Beta-binomial",
 xlab="# of Discontinuous Visits", ylab="Frequency")
par(new=TRUE)
plot(bbcounth, col="gray", xaxt="n", yaxt="n", xlab="", ylab="", las=0,
 ylim=c(0,45000), main=" ")
axis(1, at=xpos, labels=xpos)

qqplot(tmp$zi, bb00.countpred, ylim=c(0,9), xlab="Observed # Discontinuous
 Visits", ylab="Predicted # Discontinuous Visits", main="Q-Q Plot, Observed
 vs. Predicted #s of Discontinuous Visits from Beta-binomial")
par(new=TRUE)
abline(0,1)

#Application of zero-inflated binomial model to account for large proportion
#of patients with no "out of care" labs during study period

zib00.1 <- vglm(cbind(z2i,ni-
 z2i)~1+sex+age+as.factor(race)+idu+haart+timesinceenroll+as.factor(cohortid),
 family=zibinomial(zero=1),data=tmp)

summary(zib00.1)

#linear predictors are 1.logit(phi) and 2.logit(mu)

#sex
exp(-0.087135 + 1.96*c(-0.01971286, 0, 0.01971286))
#age per 10 years
exp(-0.17338 + 1.96*c(-0.0076311, 0, 0.0076311))
#race2 vs 1
exp(0.145587 +1.96*c(-0.01720420, 0, 0.01720420))
#IDU vs. non-IDU risk
exp(0.287578 + 1.96*c(-0.01734329, 0, 0.01734329))
#haart
exp(-0.355817 + 1.96*c(-0.01467876, 0, 0.01467876))
#timeincare per year
exp(0.107724 + 1.96*c(-0.00219808, 0, 0.00219808))
#mu - mean prob of missing visit, analog of "rho" from beta-binomial model
expit(-0.736107 + 1.96*c(-0.04078547, 0, 0.04078547))
#phi - predicted prob for missing visit, ignoring binomial portion
expit(0.162653 + 1.96*c(-0.01018835, 0, 0.01018835))

zib00.pred <- expit(predict(zib00.1))
summary (zib00.pred)

#Predict probability of missing visit for each individual, given their
#number of "in care" visits and total visits (years in study).

#For prediction against ZIbinomial distribution, must include predicted mean
 probability FIRST, then predicted phi SECOND, which is reverse order from
 prediction vector.

zib00.denspred <- dzibinom(tmp$zi,tmp$ni,zib00.pred[,2],log=FALSE,
```

```
 zib00.pred[,1])
zib00.distpred <- pzibinom(tmp$z2i,tmp$ni,zib00.pred[,2],log=FALSE,
 zib00.pred[,1])
zib00.countpred <- zib00.distpred*tmp$ni

zibh <- hist(zib00.denspred, breaks=seq(0,1,by=0.025))
obsh <- hist(tmp$p2i, breaks=seq(0,1,by=0.025))

zibcounth <- hist(zib00.distpred, breaks=seq(0,9,by=1))
obscounth <- hist(tmp$zi, breaks=seq(0,9,by=1))

par(mar=c(4, 5, 2, 0.5))
xpos<- seq(0,1, by=0.05)

plot(obsh, col="black", las=0, xaxt="n", ylim=c(0,45000), cex.lab=1.5,
 main="Observed vs. Predicted Prob. of Discontinuity from Zero-inflated-
 binomial", xlab="Probability of Discontinuity", ylab="Frequency")
par(new=TRUE)
plot(zibh, col="gray", xaxt="n", yaxt="n", xlab="", ylab="", las=0,
 ylim=c(0,45000), main=" ")
axis(1, at=xpos, labels=xpos)

qqplot(tmp$p2i, zib00.denspred, ylim=c(0,1), xlab="Observed Prob.
 Discontinuity", ylab="Predicted Prob. Discontinuity", main="Q-Q Plot,
 Observed vs. Predicted Prob. of Discontinuity from Zero-inflated-binomial")
par(new=TRUE)
abline(0,1)

par(mar=c(4, 5, 2, 0.5))
xpos<- seq(0,9, by=1)

plot(obscounth, col="black", las=0, xaxt="n", ylim=c(0,45000), cex.lab=1.5,
 main="Observed vs. Predicted Prob. of Discontinuity from Zero-inflated-
 binomial", xlab="Probability of Discontinuity", ylab="Frequency")
par(new=TRUE)
plot(zibcounth, col="gray", xaxt="n", yaxt="n", xlab="", ylab="", las=0,
 ylim=c(0,45000), main=" ")
axis(1, at=xpos, labels=xpos)

qqplot(tmp$zi, zib00.countpred, ylim=c(0,9), xlab="Observed Prob.
 Discontinuity", ylab="Predicted Prob. Discontinuity", main="Q-Q Plot,
 Observed vs. Predicted Prob. of Discontinuity from Zero-inflated-binomial")
par(new=TRUE)
abline(0,1)

tmp$bbprobpred <- bb00.denspred
tmp$zibprobpred <- zib00.denspred
tmp$bbcountpred <- bb00.countpred
tmp$zibcountpred <- zib00.countpred

write.csv(tmp, file="/Users/…/Beta_and_ZIbinomial.csv")
```

SUPPLEMENTAL DIGITAL CONTENT 1

Mixed Effects Logistic with random intercepts and slopes by cohort (in Stata):

```
use "C:\Documents and Settings\...\linearmixedeffects_revised.dta", clear

/*Run GLLAMM mixed effects logistic model to get Odds of Discontinuity,
 including random effect for individual intercepts by ID=naid*/
xi: gllamm outcome sex age10 i.racei.risktimesinceenrolli.cohortid , i(naid)
 link(logit) family(binom) nip(30) adapt

gllamm, eform
estimates store gllabase

/*Create equations to model random intercepts AND random slopes by
 contributing cohort=cohortID*/
gen con=1
eqint: con
eq slope: cohortid

/*use estimates from GLLAMM model with random intercepts alone as a starting
 point for model with BOTH random intercepts and slopes and define starting
 values for slope variance and covariance matrix =0*/
mat a= e(b)
mat a= (a,0,0)
xi: gllamm outcome sex age10 i.racei.risktimesinceenrolli.cohortid, i(naid)
 link(logit) family(binom) nrf(2) eqs(int slope) ip(m) nip(15) adapt from(a)
 copy

gllamm, eform
estimates store gllafull

/*Perform Likelihood-Ratio Test to see improved performance in model with
 both random intercepts & slopes vs. model with just random intercepts*/
lrtestgllabasegllafull

estimates restore gllafull

/*Get linear predictors from full model & transform to obtain predicted
 probabilities of discontinuity at each observation, then take mean for each
 individual */
gllapredglpredict, linpredict
gen float glprob = invlogit(glpredict)
bynaid, sort : egen float meangllafull_prob = mean(glprob)

/*Plot distribution of probabilities of discontinuity, observed=p2i
 predicted=meangllafull_prob */
twoway histogram p2i, start(0) bfcolor(black) blcolor(black) bin(40) freq ||
 histogram meangllafull_prob, start(0) bfcolor(gs10) blcolor(black) bin(40)
 legend(off) freq

/*Produce Q-Q plot to visualize ability of Mixed Effects logistic model with
 random intercepts & slopes to predict observed discontinuity patterns*/
qqplotmeangllafull_prob p2i, mcolor(black) msize(tiny) msymbol(circle)
 mlwidth(none) jitter(7)
```

SUPPLEMENTAL DIGITAL CONTENT 1

```
/* Fit Comparison Model accounting for Cohort Types: US Centralized Care
 (KPNC & VACS), US DeCentralized Care, Canadian Care */
xi: xtlogit outcome sex age10 i.racei.riski.cohort_typetimeincare, i(naid)
 quad(30)
xtlogit, or
estimates store xtfull
```

SUPPLEMENTAL DIGITAL CONTENT 1

References

1. Gange SJ, Muñoz A, Saez M, Alonso J. Use of the beta-binomial distribution to model the effect of policy changes on appropriateness of hospital stays. J R Stat Soc Ser C Appl Stat. 1996;45(3):371-82.
2. Xue X, Gange SJ, Zhong Y, Burk RD, Minkoff H, Massad LS, et al. Marginal and mixed-effects models in the analysis of human papillomavirus natural history data. Cancer Epidemiol Biomarkers Prev. 2010 Jan;19(1):159-69.
3. Skrondal A, Rabe-Hesketh S. Some applications of generalized linear latent and mixed models in epidemiology: Repeated measures, measurement error and multilevel modeling. Nor Epidemiol.2003;13(2):265-78
4. Yee, TW. A User's Guide to the vgam Package.http://www.stat.auckland.ac.nz/~yee/VGAM (2007)