

Multimedia Appendix 4: Log-Linear model of primary care office visits (without telephone visits) based on demographic variables and baseline utilization phenotype

Model predictors	Adjusted $R^2$	AIC <sup>a</sup>
Age-sex <sup>b</sup>	.135	49,638
Payer <sup>c</sup>	.112	49,989
Naïve phenotypes (NP) <sup>d</sup>	.241	46,261
Utilization phenotype (UP) <sup>e</sup>	.296	44,344
Age-sex and payer	.173	48,505
Age-sex, payer, and NP	.306	44,147
Age-sex, payer, and UP	.344	42,695

<sup>a</sup>Akaike Information Criterion

<sup>b</sup>Age-sex bins are categorical variables of the combination of male or female with the following age groups: 18-34, 35-49, 50-64, 65-69, 70-84,85-115.

<sup>c</sup>Payer were defined as commercial, Medicare/Medical, or Other

<sup>d</sup>The naïve phenotype is a categorical variable that was obtained by summing to the total number of healthcare encounters in the baseline year. These values were rank ordered and divided into seven percentiles

<sup>e</sup>The utilization phenotype is a categorical variable encoding one of the seven phenotypes clusters created by our algorithm.