PREDICTING PATIENT "COST BLOOMS" IN DENMARK: A LONGITUDINAL POPULATION-BASED STUDY

Suzanne Tamang, PhD[1*], Arnold Milstein, MD, MPH[1], Henrik Toft Sørensen MD[2], Lars Pedersen, PhD[2], Lester Mackey PhD[3], Jean-Raymond Betterton[3], Lucas Janson MA[3], Nigam Shah, MBBS, PhD[1]

[1] Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States
[2] Department of Clinical Epidemiology, Aarhus University, Aarhus, Denmark
[3] Department of Statistics, Stanford University, Stanford, CA, United States

Corresponding author: Suzanne Tamang, PhD
1265 Welch Road Suite X231
Stanford, CA, 94094
United States
Email: stamang@stanford.edu
Phone: (650)497-4392

ABSTRACT

Objectives: To compare the ability of standard vs. enhanced models to predict future high-cost patients, especially those who move from a lower to the upper decile of per capita healthcare expenditures within one year – *i.e.*, "cost bloomers".

Design: We developed alternative models to predict being in the upper decile of healthcare expenditures in Year 2 of a sample, based on data from Year 1. Our six alternative models ranged from a standard cost-prediction model with four variables (*i.e.*, traditional model features), to our largest enhanced model with 1,053 nontraditional model features. To quantify any increases in predictive power that enhanced models achieved over standard tools, we compared the prospective predictive performance of each model.

Participants and Setting: We used the population of Western Denmark between 2004 and 2011 (2,146,801 individuals) to predict future high-cost patients and examine characteristics of high-cost cohorts. Using the most recent two-year period (2010-11) for model evaluation, our whole-population prediction model included a cohort of 1,557,950 individuals with a full year of active residency in 2010. Our cost bloom prediction model excluded the 155,795 individuals who achieved a population-level high cost status in 2010, resulting in 1,402,155 individuals for prediction of cost bloomers 2011.

Primary outcome measures: Using unseen data from a future year, we evaluated each model's prospective predictive performance by calculating the ratio of predicted high-cost patient expenditures to the actual high-cost patient expenditures in Year 2 – *i.e.*, cost capture.

Results: Our best enhanced model achieved a 21 percent and 30 percent improvement in cost capture over a standard diagnosis-based model for predicting population-level high-cost patients and cost bloomers, respectively.

Conclusions: In combination with modern statistical learning methods for analyzing of large datasets, models enhanced with a large and diverse set of features led to better performance—especially for predicting future cost bloomers.

Strengths and Limitations of this Study

- We conducted a population-based study of high-cost patients, using Danish National Health Service and Civil Registration System data sources linked at the individual level and covering all residents of Western Denmark from 2004 through 2011.
- We carried out the commonly performed prediction of high-cost patients at the whole population level and developed a novel framework for predicting cost bloomers.

- We demonstrated that pairing large population health datasets and modern statistical learning methods for large-scale data analysis can improve prediction of future high-cost patients, compared to standard tools that are widely-used in the US and internationally.
- Accurate prediction of the cost bloomers is the first step in a process that must be coupled with evidence-based interventions, in order to achieve the ultimate effect we seek -- improvements in healthcare value.
- Given differences between residents, insurance status (or lack thereof), follow-up times, and other national health systems characteristics, our findings may not be generalizable to other national health systems.

INTRODUCTION

A small fraction of individuals accounts for the bulk of population healthcare expenditures in the United States, Denmark and other industrialized countries.[1-4] Although many high-cost patients show consecutive high-cost years, the majority experience a "cost bloom", or a surge in healthcare costs that propels them from a lower to the upper decile of population-level healthcare expenditures between consecutive years.[4]

Proactively identifying and managing care for high-cost patients – especially cost bloomers, who may disproportionately benefit from interventions to mitigate future high-cost years -- can be an effective way to simultaneously improve quality and reduce population health costs[5-17]. However, since the Centers for Medicare and Services (CMS) commissioned the Society of Actuaries to compare leading prediction tools more than ten years ago, scant progress has been made in improving cost-prediction tools [18 19]. To the extent that they fail to make accurate predictions, standard models leave many smaller providers, such as physician practices, vulnerable to unpredictable costs associated with cost blooms among their patient population [19-21 22]. Overcoming these and other challenges associated with the care of high-cost patients is essential to achieving a higher-value health care system.

We sought to create more accurate models for predicting high-cost patients, especially cost bloomers, who are more challenging for standard tools to predict accurately [23 24]. Also, we wished to gain new insights into the antecedents of high-cost years that distinguish cost bloomers from persistent high-cost patients, who have two or more consecutive years of high healthcare expenditures. Our hypothesis was that technological advances in the last decade allow improvement in prediction ability over current approaches. Recent progress in statistical methods for analyzing large datasets has been driven by development of new learning algorithms, and by

the ongoing explosion in the availability of large observational datasets and low-cost computation[6 10 23-26]. Paired with large and diverse health datasets available at the population-level, modern statistical learning methods may present new opportunities to advance methods underlying healthcare cost-prediction tools [27-30].

Drawing from individual-level data for the entire population of Western Denmark, we analyzed high-cost spending trends and evaluated the *prospective predictive performance* of six alternative models designed for the commonly executed prediction of high-cost patients and for our novel cost-bloom prediction task. Our models ranged in size from a baseline logistic regression model with four variables (*i.e.*, features) to a very large enhanced prediction model with over one thousand nontraditional cost-prediction features. For our larger models, we used elastic-net penalized logistic regression, which is a modern statistical learning method designed to address some of the issues associated with applying a standard stepwise regression procedure to select a best-fitting model from a plethora of choices[19 20 31 32].

Materials and Methods

Participants and Setting

Our longitudinal population-based study draws from the entire population of Western Denmark, which is representative of Denmark more broadly. The Danish National Health Service provides tax-supported health care for all Danish citizens.

For our eight-year trend analysis, we analyzed population healthcare data for the whole population from 2004 to 2011 (*N*=2,146,801). We separately analyzed trends for high-cost patients in our model evaluation year, 2011, based on data from 2009-2011. Also, for the 2011 high-cost cohort, we used each high-cost patient's Year 1 model features to quantify additional

distinctions between persistent high-cost patients and cost bloomers. Specifically, we repurposed our demographic, healthcare cost and chronic condition features to compare age distribution, mortality rates and chronic condition profile.

Our prediction study considered only those individuals with a full year of active Danish residency in Year 1, to predict high-cost patients in Year 2. Our whole-population high-cost analysis used all residents who satisfied our Danish residency criteria. For prediction of the cost bloomers, our cost-bloom analysis additionally excluded individuals who were in the upper expenditure decile in Year 1, who thus could not show a cost bloom. Similar to previous studies, we defined a "high-cost" patient as an individual in a sample who is in the upper decile of annual healthcare expenditures[12 33-35].

PATIENT INVOLVEMENT

We analyzed deidentified population healthcare data. Thus patients were not involved in the development of the research question, the outcome measures, or the study design.

DATA SOURCES

We obtained demographic information from the Danish Civil Registration System, including age, gender, and residency status, as well as geographic district of residence and social relationship data[36-40]. The Danish registries used in our study are described in more detail in the eSupplement.

The Primary and Specialist Care Files and the Danish National Patient Registry were the sources of our healthcare utilization data[18]. The Primary and Specialist Care Files specified each visit type (*i.e.*, primary care or specialist), the total cost of each visit, and whether a visit occurred during weekday business hours or during one of two off-hours time periods. The Danish National Patient Registry provided ICD-10 diagnostic codes (adopted in 1994) assigned

to each patient in the inpatient hospital setting or at a hospital outpatient clinic, NOMESCO codes for surgeries and procedures associated with inpatient visits, and healthcare costs[38 41]. Our source of prescription data was the Health Service Prescription Database[42 43] For each drug prescribed to a patient, this Database provided Anatomical Therapeutic Chemical (ATC) class information and the cost[44].

The Department of Clinical Epidemiology at Aarhus University, Denmark, provided data for our study. The Danish Data Protection Agency (Record 2013-41-1924) approved this investigation.

Alternative Prediction Models

Healthcare cost-prediction tools can be broadly categorized as diagnosis-based [*e.g.*, Ambulatory Care Groups (ACGs) and Diagnostic Cost Groups (DCGs)], pharmacy-based (*e.g.*, MedicaidRx and RxGroups), or diagnosis and pharmacy-based (*e.g.*, Episode Risk Groups (ERGs) and Impact Pro)[37]. Detailed descriptions of standard tools and their features can be found in a number of reviews of health risk assessment [45]. Standard diagnosis-based tools are the most widely-used type of cost-prediction model in the US and internationally. They consist of traditional cost-prediction features such as a diagnostic risk score, adjusted for age and gender and use regression-based learning methods [19 32 45].

Table 1 provides an overview of the types and number of traditional and nontraditional model variables – *i.e.*, model *features* – that were used to create high-cost patient prediction models. Our approach to creating enhanced models was to create a richer and more informative individual-level profile for high-cost patient prediction. We built on previous work in healthcare cost prediction, involving the development of enhanced prediction models and their evaluation [7]

[12 15 17 33 34 46 47]. Our custom features were based on those available in our Danish population health dataset.

Table 1. Year 1 model features for high-cost patient prediction are shown by data source, feature type (i.e., traditional/nontraditional) and feature category. Each row represents a unique resident and example values for a feature category. The number of features and the data type appear below each feature category; e.g., Traditional Features, "Costs 2-numerical" indicates that there are two traditional cost features in the feature category and each feature represents a numerical value.

						Clinical Registries				***Civil Reg. System***	
	TRADITIONAL FEATURES (6)					**NONTRADITIONAL FEATURES (1053)**					
Residents	**Age** 1-numeric	**Gender** 1-binary	**Disease Risk Scores** 2-numeric	**Costs** 2-numeric		**Costs** 1-numeric	**Clinical Code Sets** 894-binary	**Visits/Tx Counts and LOS** 71-numeric	**Recency** 12-numeric	**Social Relationship Status** 71-binary	**District** 4-binary
ID 1	45	F	CCS disease score and CCI chronic condition score	Inpatient and Outpatient Specialist (IOS)	Drug (Rx)	Primary Care (PC)	CCS (247), CCI (44), ICD10 (211), NOMESCO (171), ATC (221) based categories	Counts by year and quarter: IOS, Rx, PC and Surgeries; Total Inpatient Length of Stay (LOS)	Moving averages by quarter: Diagnoses, Costs, Visits, Rx, Inpatient LOS	Married-Widowed	1
ID 2	34	F								Single-Married	4
ID 3	22	M								Single	2
ID 4	32	M								Married	2
…	…	…								…	…
ID N	71	F								Widowed	1

Overall, we developed 1,053 nontraditional features. We used data from the clinical registries and medical coding systems to transform our diagnosis-based risk scores into component disease groups and chronic and non-chronic indicators by organ system, as well as to represent sparse drug and procedural information in succinct and meaningful categories, and to incorporate cost information by setting. We constructed new features to capture utilization patterns, including the number of off-hours primary care visits, total length of inpatient stays, and utilization statistics such as the quarterly moving average of ED visits and linked data from the Danish Civil Registration System (CRS). As shown in Table 1, the CRS allowed us to assign

a social relationship status to each resident that was fixed (*e.g.*, "Married" for all years) or dynamic (*e.g.*, "Married-Widowed" for widowed in the year prior to prediction).

We created a total of six alternative prediction models: *two standard models* with traditional features and *four enhanced models* with traditional and nontraditional features. Table 2 provides a description of each model's feature types, logistic regression method and the number of traditional and nontraditional features. We developed our standard models based on their description in the literature[31-33]. Standard Model 1, our baseline model, is representative of a standard diagnosis-based cost-prediction model that includes age, gender, diagnostic risk score, and chronic condition risk score. We estimated disease risk scores for each resident, based on the Agency for Healthcare Research and Quality (AHRQ) Clinical Classification Software (CCS) and Chronic Condition Indicator (CCI) coding systems [33 48-50]. Building on the baseline model, Standard Model 2 is representative of a diagnosis and pharmacy-based prediction tool and included inpatient and outpatient specialist costs and drug costs as features.

Table 2. Description of alternative standard and enhanced high-cost patient prediction models, presenting the feature types included, the statistical method used for prediction and the number of traditional, nontraditional and total model features.

Model	Feature Description	Regression Method	Feature Count		
			Traditional	Nontraditional	Total
Standard Model 1	Age + Gender + Disease Risk Scores	standard	4	0	4
Standard Model 2	Age + Gender + Disease Risk Scores + Hospital Inpatient & Specialist + Rx Costs	standard	6	0	6
Enhanced Model 1	Age + Gender + Disease Risk Scores + Hospital Inpatient & Specialist + Rx Costs + Primary Care Costs	standard	6	1	7
Enhanced Model 2	Age + Gender + Disease Risk Scores + Hospital Inpatient & Specialist + Rx Costs + Social Relationship Status	penalized	6	71	77
Enhanced Model 3	Full Feature Set without Costs	penalized	6	1028	1034
Enhanced Model 4	Full Feature Set	penalized	6	1053	1059

Our simplest enhanced model, Enhanced Model 1, builds on Standard Model 2 with the addition of primary care costs. Enhanced Model 2 extends Enhanced Model 1 with an additional 71 social relationship features. To quantify performance of an enhanced model in lieu of cost data, we used our full feature set and excluded all cost features (25 in total) to create Enhanced Model 3. Finally, all 1,059 traditional and nontraditional healthcare utilization, diagnostic, prescription and civil registry derived features were used to create Enhanced Model 4.

With inclusion of many features, regression problems require statistical model selection to identify a parsimonious model. For Enhanced Models 2 through 4, ranging from 77 to 1,059 cost-prediction features, we used elastic-net *penalized logistic regression*, which addresses some of the issues associated with applying a standard stepwise regression procedure to select a best-fitting model from a plethora of choices[26 51-53]. Penalized regression is a prominent statistical learning methods for analyzing large high-dimensional datasets and has been successfully used in scientific and business applications [13 54-56]. For our larger enhanced models, the main advantage penalized logistic regression offered over a standard approach was the ability to simultaneously conduct feature selection and model fitting[25 30 53]. A detailed description of stepwise and penalized regression can be found in the work of Taylor and Tibshirani[53].

An overview of our model development and evaluation framework appears in

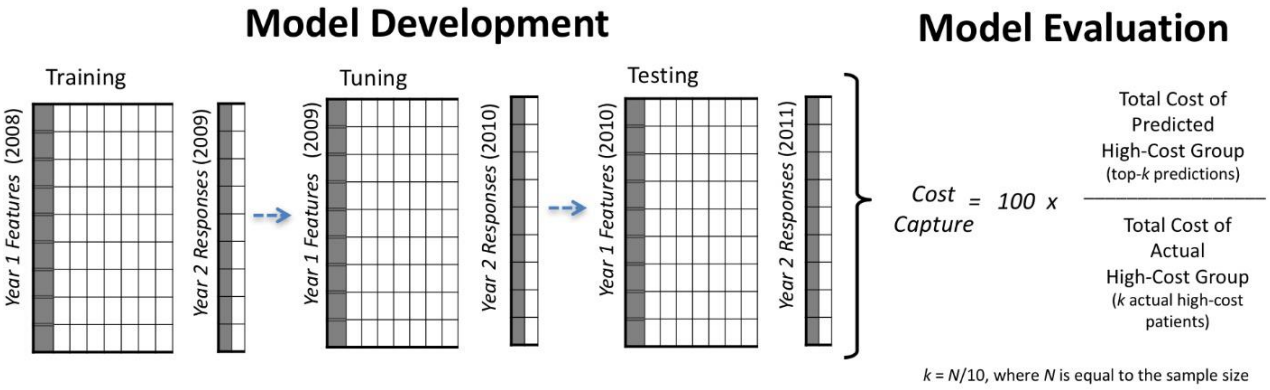

Figure 1. For our penalized regression models, the process involved three steps: Step 1: *training* on 2008 data to predict high-cost patients for the year 2009; Step 2: *tuning* on 2009 data to predict high-cost patients for the year 2010, and Step 3: *testing* our model from Step 2 on unseen data from 2010 to predict high-cost patients in 2011 (i.e., prospective model validation). We learned the initial parameters for each model in the training step, commonly called *model calibration* in the health risk-assessment literature. For penalized logistic regression models, tuning was used to refine the final model based on the 2010 classification error of predicted to actual high-cost patients. Since standard regression models cannot be refined by tuning--as there are no free parameters to set beyond the initial parameters learned in training--the tuning step was not performed. For standard regression models, the process involved two steps: Step 1: *training* on 2009 data to predict high-cost patients for year 2010; and Step 2 *testing* on the model fitted in Step 1 using 2010 data for prediction of high-cost patients in 2011.

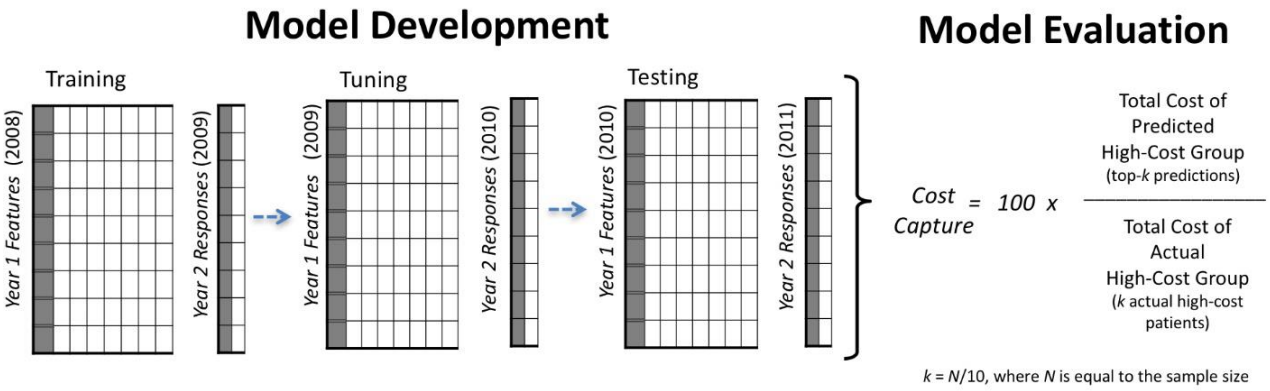

Figure 1. Overview of our model development and evaluation framework. Three independent panel datasets were used for training (model fitting), tuning and testing steps. To evaluate alternative models, we calculated the ratio of predicted high-cost patient expenditures to actual high-cost patient expenditures in Year 2.

Model Evaluation

Using unseen data from the most recent two-year period in our dataset (2010-11), we evaluated models by calculating the ratio of predicted high-cost patient expenditures to the actual high-cost patient expenditures in Year 2 – *i.e.*, cost capture. Cost capture has been reported in previous studies and is based on the "predictive ratio", commonly used to evaluate cost-prediction models in the health risk-assessment literature and in actuarial reports.[12 21 32 34]

The formula for cost capture is shown in

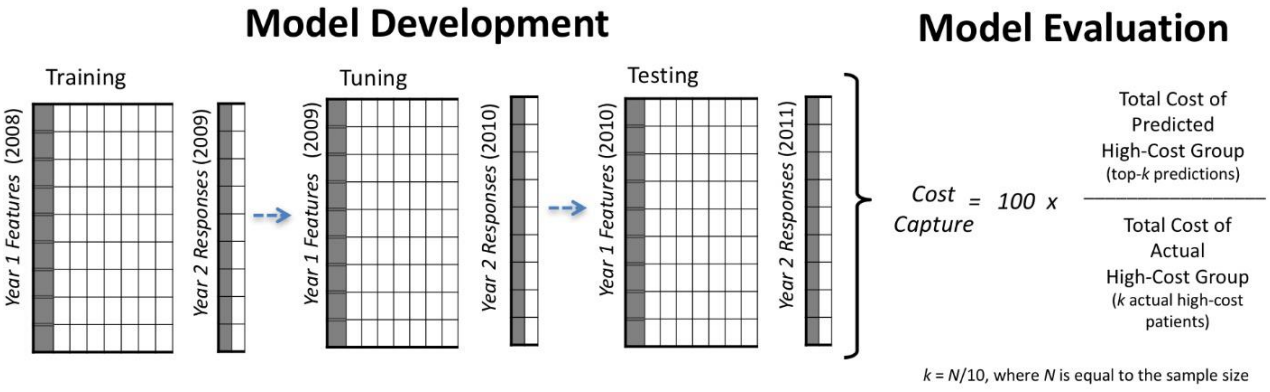

Figure 1. Given a model with a prediction sample of 10,000 individuals, cost capture is calculated from Year 2 data by: Step 1: estimating the size of the upper decile, k, where $k=N/10$, and N is the sample size, Step 2: identifying the predicted high-cost group by selecting the 1,000

(k=10,000/10) individuals predicted to be high-cost in one year with the highest probability (i.e., the top-k predictions), Step 3: aggregating the Year 2 expenditures accrued by the 1,000 individuals in the predicted high-cost group and the 1,000 individuals in the actual high-cost group, and Step 5: dividing the Year 2 healthcare expenditures of the predicted high-cost group by that of the actual high-cost group's.

We report the area under the receiver operating characteristic (AUC) as a secondary outcome for comparing our population-level models with prior studies that do not report cost capture or a comparable measure. However, given the highly skewed nature of per capita spending in the upper decile of it is important to consider relative costliness. For example, the upper decile accounts for 65 percent and the upper centile accounts for 22 percent of US healthcare spending.[53] From a cost-prediction perspective, it is better to correctly predict one high-cost patient that accrued $350,000 than three high-cost patients with $35,000. The main limitation of AUC as a sole evaluation metric for high-cost patient prediction is that, unlike cost capture and standard predictive ratios, is does not impose a penalty proportional to the misclassified individual's future costliness, which is key for performance characterization.

RESULTS

Our analysis of 2,146,801 individuals in our eight-year trend analysis, showed that 314,989 had one or more years of high-cost spending from 2004 through 2011. Within this group, **Error! Reference source not found.** shows the percent of patients (y-axis) by their total high-cost years (x-axis) and their longest duration of consecutive high-cost persistence (saturation scale). The majority (51%) showed only one high-cost year. Among the individuals with multiple high-cost years, many did not experience them consecutively. However, the more

consecutive high-cost years a patient experienced, the more likely they were to remain high-cost the following year.

Our trend analysis of high-cost patients in our evaluation year, 2011, included 155,795 high-cost patients, who collectively accrued 73 percent of Western Denmark's total healthcare expenditures in 2011. Among the high-cost group, 68 percent (105,904) were cost bloomers in 2011 and half (77,897) did not have a high-cost year in either 2009 or 2010. The remaining 32 percent (49,855) of high-cost patients in 2011 also had a high-cost year in 2010; in this group, 21 percent (10,470) had a third year of high-cost persistence in 2009.

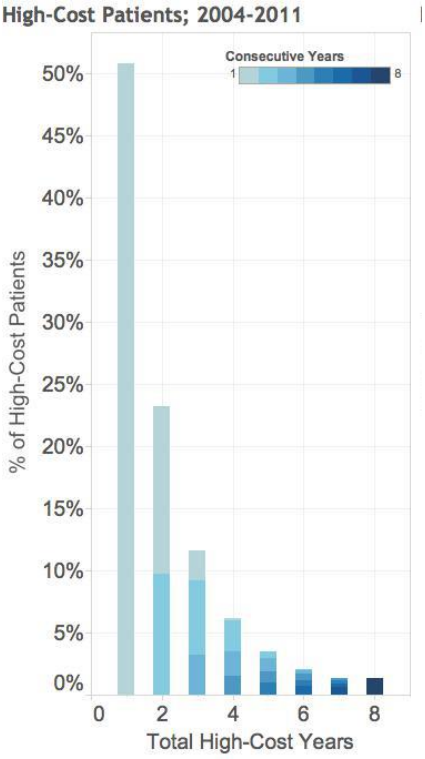

Figure 2. High-cost persistence in Western Denmark (N=2,146,801). Among the 314,989 individuals with any high-cost years, the bars show the percent of high-cost patients by total high-cost years; color saturation increases proportionally to the longest duration of consecutive high-cost years for each individual from 2004 through 2011.

Using Year 1 features to examine differences between the cost bloomers and the persistent high-cost patients, our analysis revealed that cost bloomers in 2011 were more likely to have zero inpatient hospital costs than persistent high-cost patients (47 percent vs. 7 percent). We also found that relative to persistent high-cost patients, cost bloomers showed more than four time fewer chronic conditions and were less likely to be diagnosed with chronic conditions related to

the circulatory system, neoplasms or the respiratory system.

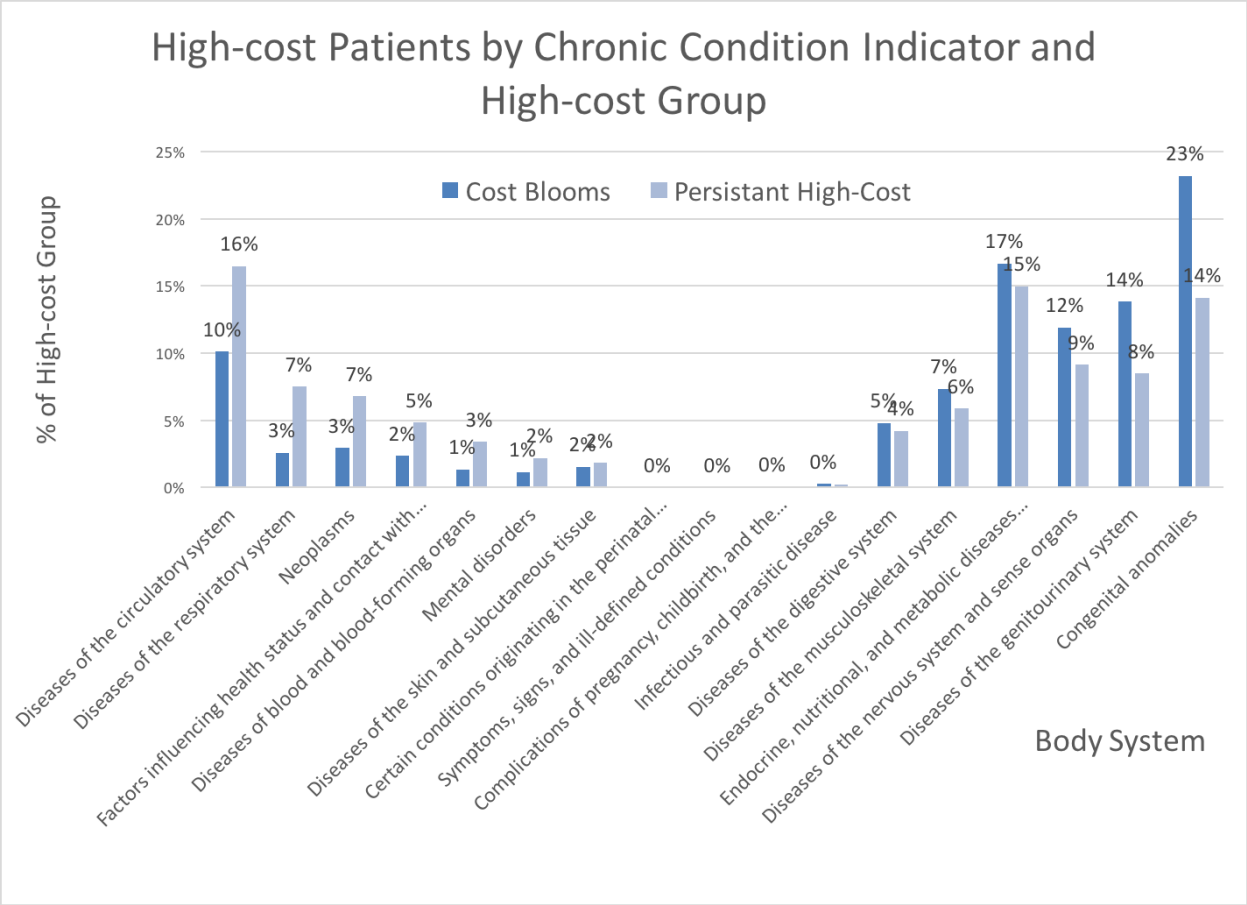

Figure 3 illustrates the proportion of AHRQ CCI chronic condition indicators among cost bloomers and persistent high-cost patients in 2010.

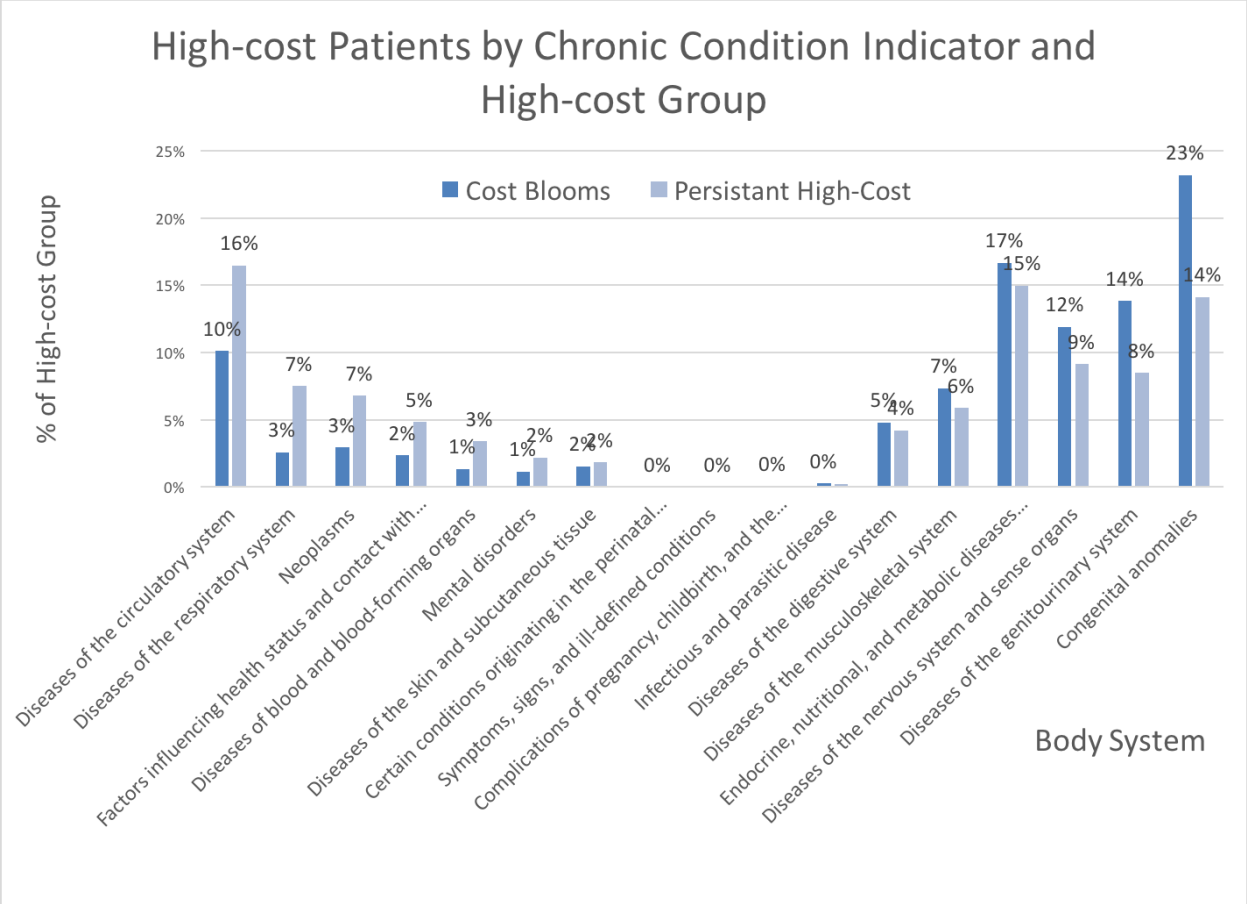

Figure 3. Proportion of chronic condition indicators among persistent high-cost patients (N= 49,855) and cost bloomers (N=105,904). Bars show the percent of patients with each indicator in the prior year, 2010; color identifies the high-cost group.

Lastly, we found that cost bloomers in 2011 were on average younger (55 vs. 59 years) and had a lower median age than persistent high-cost patients (58 vs. 62).

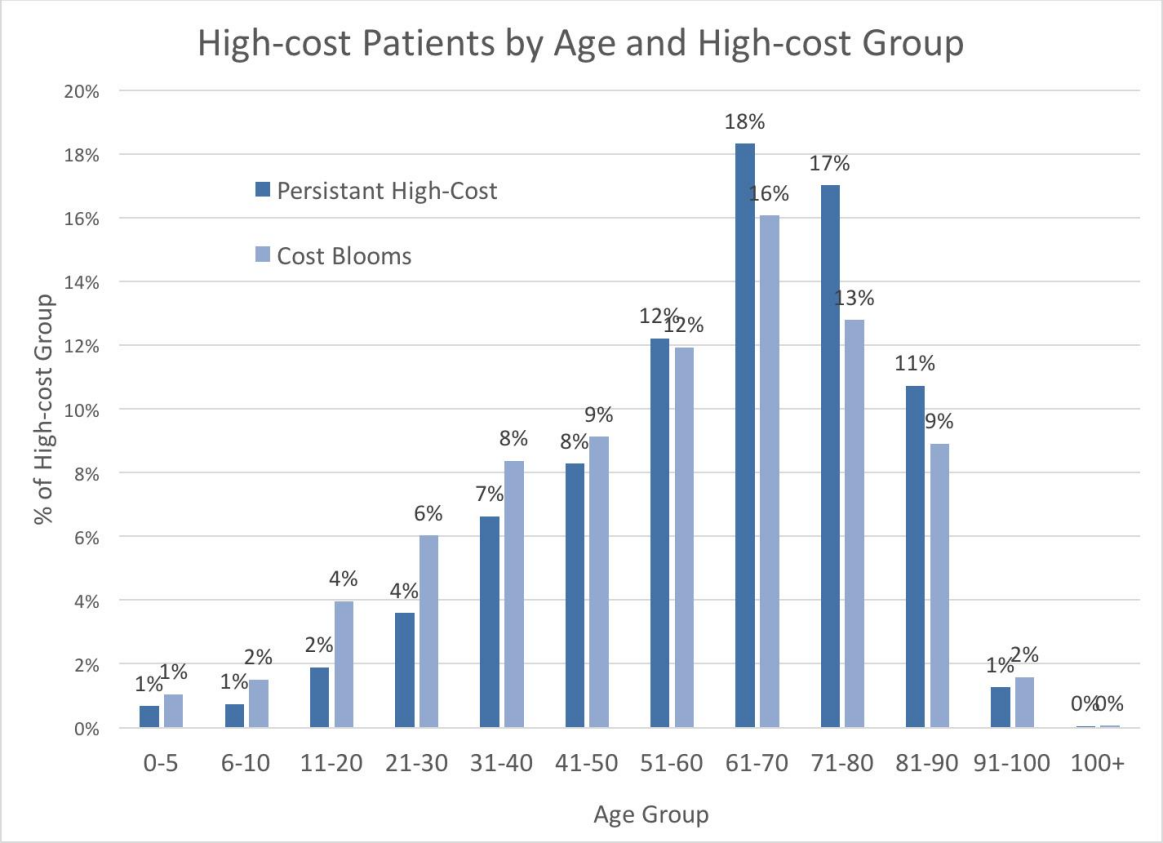

Figure shows the age distribution among high-cost patients by high-cost status. Cost bloomers had lower one-year mortality rates (5 percent vs. 9 percent) and two-year mortality rates (8 percent vs. 16 percent).

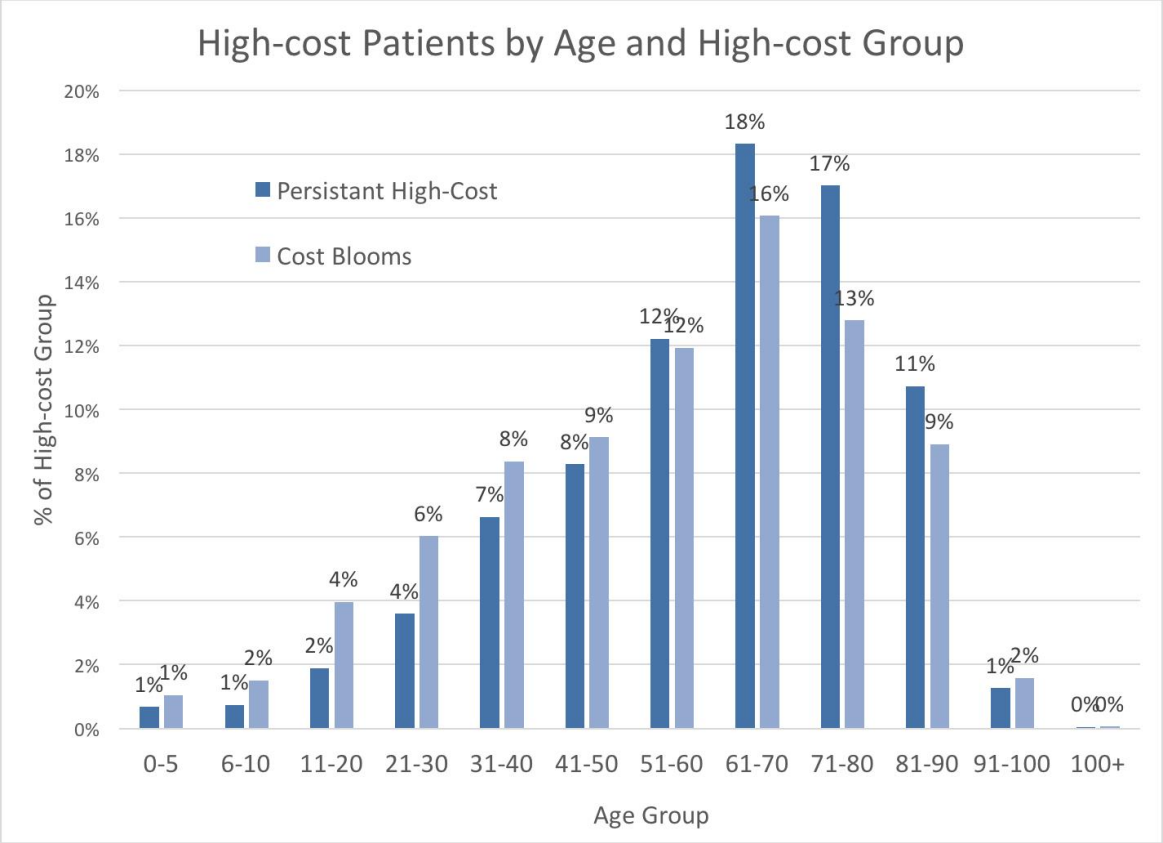

Figure 4. Age distribution of 2011 high-cost patients by high-cost status (N=155,756). Lines show the percent of patients by age; color distinguishes persistent high-cost or cost bloom status. Persistent high-cost patients and cost bloomers had mean and median interquartile age ranges of 30 and 34, respectively.

Prediction Performance

Our whole-population analysis included 1,557,950 individuals with a full year of active residency in 2010. After excluding the 155,795 individuals who already had high-cost status in 2010, our cost bloom model included 1,402,155 individuals. Table 3 compares the performance of alternative models among the whole population (155,795 individuals) and among cost bloomers (140,216 persons) in 2011. Our best-performing model captured 60 percent of the costs attributed to high-cost patients at the population-level and 49 percent of costs attributed to cost bloomers. Overall, we observed a 21 percent and 30 percent improvement in cost capture over baseline for population-level high cost and cost bloom prediction, respectively.

Table 3. Comparison of alternative models for predicting future high-cost patients at the population level and cost bloomers. Column headers indicate each model and the number of model features appears in parentheses. Results with the highest cost capture value are shown in bold.

		Alternative High Cost Patient Prediction Models					
Prediction Sample	**Metric**	Standard Model 1	Standard Model 2	Enhanced Model 1	Enhanced Model 2	Enhanced Model 3	Enhanced Model 4
		Number of Model Features					
		4 (Baseline)	6	7	77	1034	1059
Whole-population Analysis (N=1,557,950)	AUC	0.775	0.814	0.825	0.823	0.823	0.836
	Cost Capture	0.495	0.559	0.577	0.579	0.578	**0.600**
Cost Bloom Analysis (N=1,402,155)	AUC	0.719	0.748	0.772	0.765	0.771	0.786
	Cost Capture	0.376	0.443	0.455	0.461	0.466	**0.487**

Focusing on the cost-blooming population, in

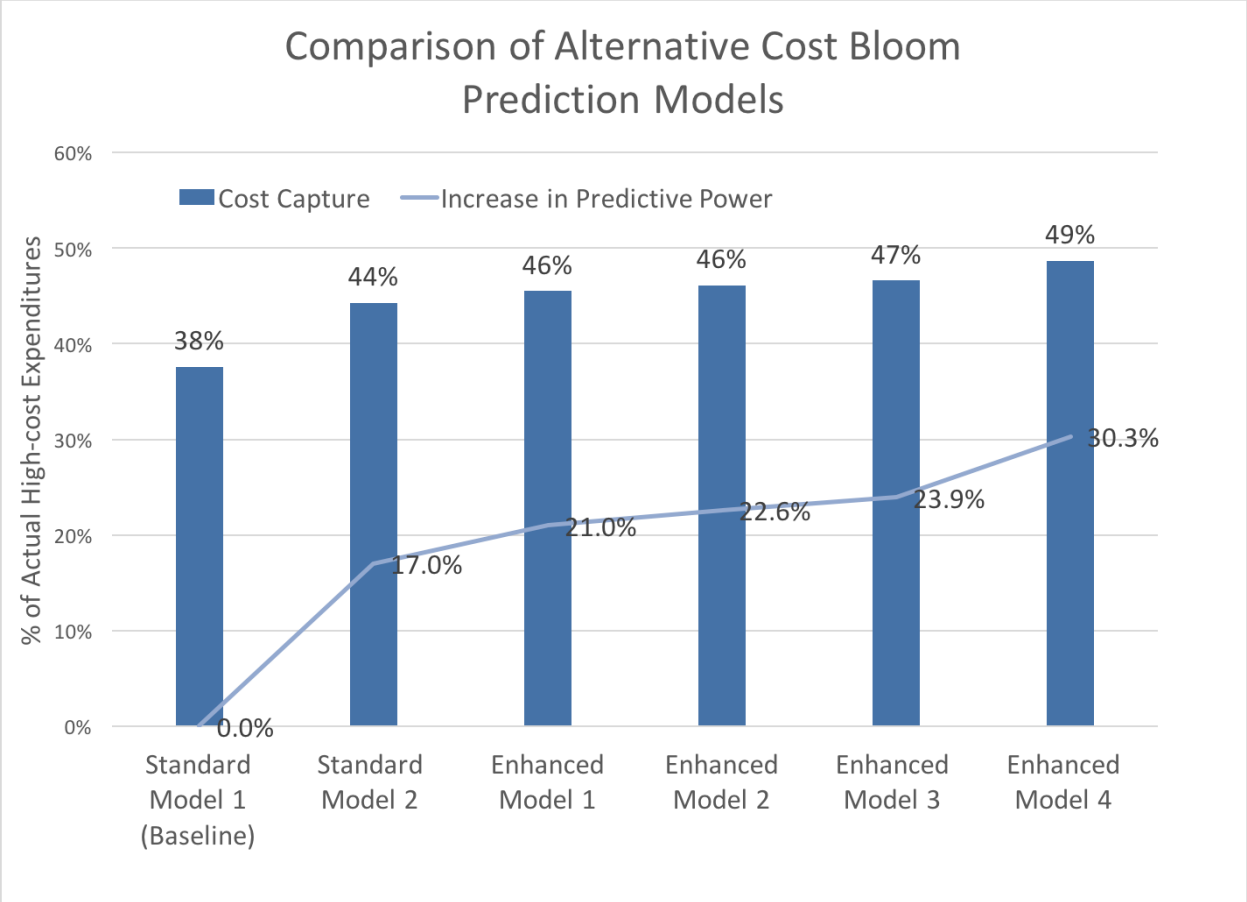

Figure , we show the percent increase in predictive performance that each model achieved over our baseline model, Standard Model 1, with a total of four features. Standard Model 2 achieved a 17 percent increase over the baseline by adding two standard features, inpatient and outpatient specialist costs and drug costs. Predictive power continued to increase to 21 percent over the baseline with the addition of our first nontraditional feature, primary care costs, in Enhanced Model 1, and again to 23 percent in Enhanced Model 2, which also included 71 social relationship status features. Enhanced Model 3 used all available features (with the exception of 25 features derived from cost data) and showed a 24 percent increase in predictive power over the baseline. Our full feature set of 1,059 features, Enhanced Model 4, achieved a 30 percent increase over the baseline model and was consistently our best model, independent of prediction task.

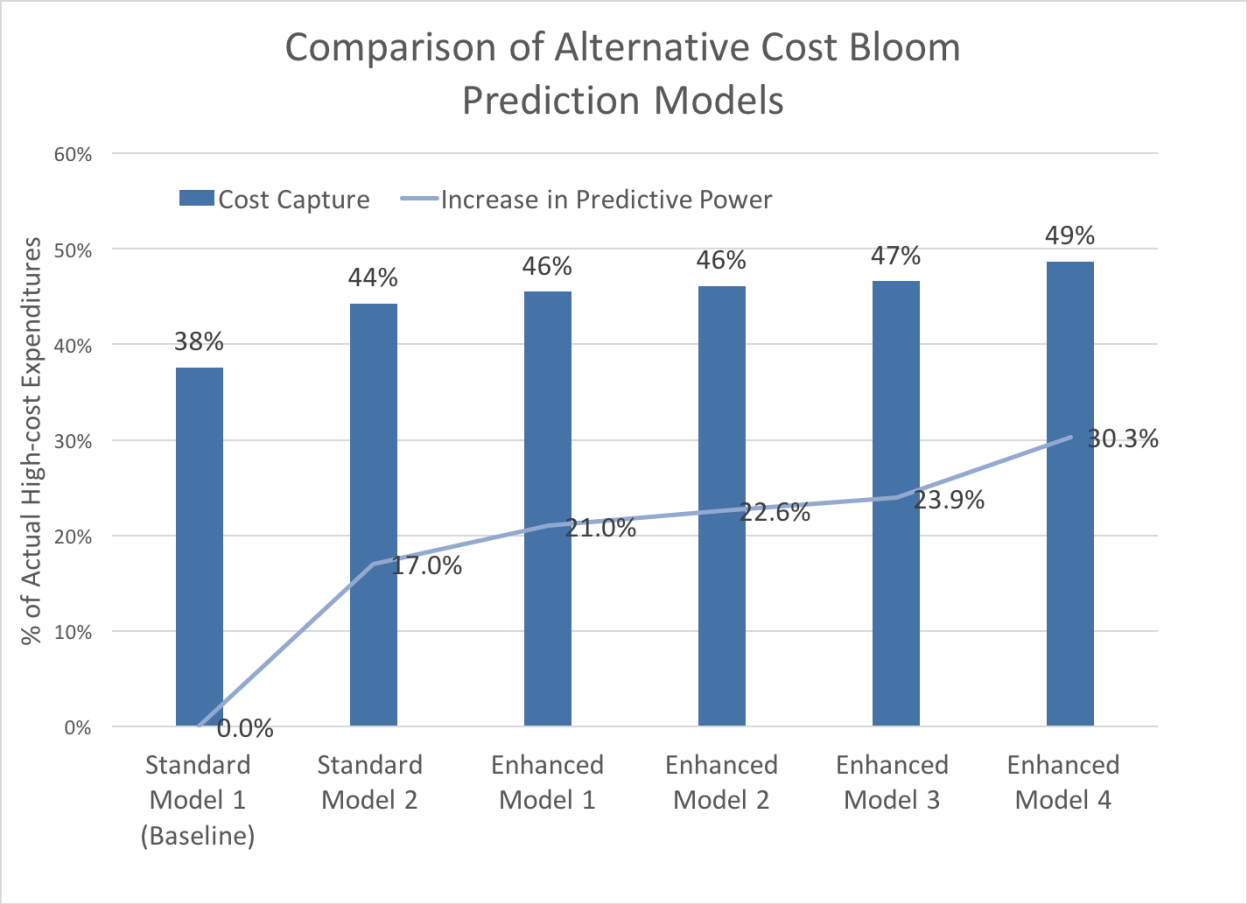

Figure 5. Performance of alternative cost bloom prediction models by cost capture and relative improvement over the baseline. Bars show cost capture for each model; lines show the percent increases in predictive power. More details on each model are provided in Table 2.

DISCUSSION

Our study makes three key contributions. Taken together, they provide future directions for improving prediction of high-cost patients, who drive the majority of population healthcare spending in the United States, Denmark and other industrialized countries.

First, we provide additional evidence for the importance of accurately identifying cost bloomers, which is underscored by their prevalence among high-cost patients, and their potential for intervention. Similar to the US, we found that cost bloomers represented the majority of all high-cost patients. Compared to individuals with more persistent high-cost years, we found that

cost bloomers were younger, showed less morbidity, lower mortality and, based on their Year 1 data, had fewer chronic conditions on average; also, cost bloomers were less likely to be diagnosed with the types of chronic condition indicators commonly associated with high healthcare costs – *e.g.*, diseases of the circulatory system and neoplasms.

Second, we demonstrate the ability of modern statistical learning methods and diverse population healthcare data to advance methods underlying healthcare cost-prediction tools. Our best enhanced model reported 21 percent and 30 percent improvement in cost capture over baseline for population-level high cost and cost bloom prediction, respectively. Our framework for the development and evaluation of enhanced models can be described as a machine learning approach to prediction. Machine learning is a field at the intersection of computer science and statistics, with a fast growing literature on modern statistical learning methods for analyzing large and complex datasets[27-29 57]. A typical prediction framework involves the use of a training set of data, in which the outcome and feature measurements for a set of objects are observed, to build a prediction model, or "learner", which will enable prediction of the outcome for new unseen objects[25]. A good learner is one that accurately predicts the outcome of interest[25]. The notable improvement our best model achieved over standard tools suggests that the assessment of other types of data-intensive machine learning methods warrant further study.

Our third contribution is an enhanced model for prediction of cost blooms, which produced a 30 percent improvement in cost capture over a standard diagnosis-based model. Since our cost-bloom prediction task is novel, we have no external model comparison. However, prior studies have been conducted on prediction of high-cost patients at the population-level[12 33 34 35 58]. For a nationally representative dataset, the top model was developed by Fleishman *et al*., using an enhanced model that was developed using the AHRQ's Medical Expenditure Panel

Survey (MEPS) dataset, reporting an AUC of 0.84. While we also achieved an AUC of 0.84, there is an important distinction -- Fleishman *et al*.'s AUC measure is not a prospective measure of predictive performance like the AUC reported in our work. Retrospective measures of model fit, such as reported by Fleishman *et al*., are known to be overly optimistic relative to predictions based on out-of-sample data[25]. Also, the MEPS data collection process consists of multiple face-to-face interviews conducted with participants and their family members over a two-year period. Although this enables rich longitudinal data to be collected for research purposes, such an extensive primary data collection process can be resource prohibitive for providers to administer for their own population. Based on unseen data from the most recent two-year period in our dataset, we evaluated our models with a prospective measure of predictive performance, our enhanced models were developed with only secondary data sources available at the population-level and our best enhanced model showed higher PPV (33 percent vs 29 percent).

There are several policy and practice implications for our work. More accurate cost-prediction tools can be used by providers to proactively identify patients at high risk of a cost bloom. However, many providers lack access to the type of comprehensive healthcare and cost data available in Denmark. In the US, recent legislation in support of data-sharing among Accountable Care Organizations and the growth of population registries will facilitate individual-level linkages across settings and providers; however, this now remains impossible for most practices. To provide utility in a setting where only some of our feature categories are available for prediction of cost bloomers, we demonstrated that our simplest enhanced model achieved a 21

percent increase in predictive performance over the baseline (see

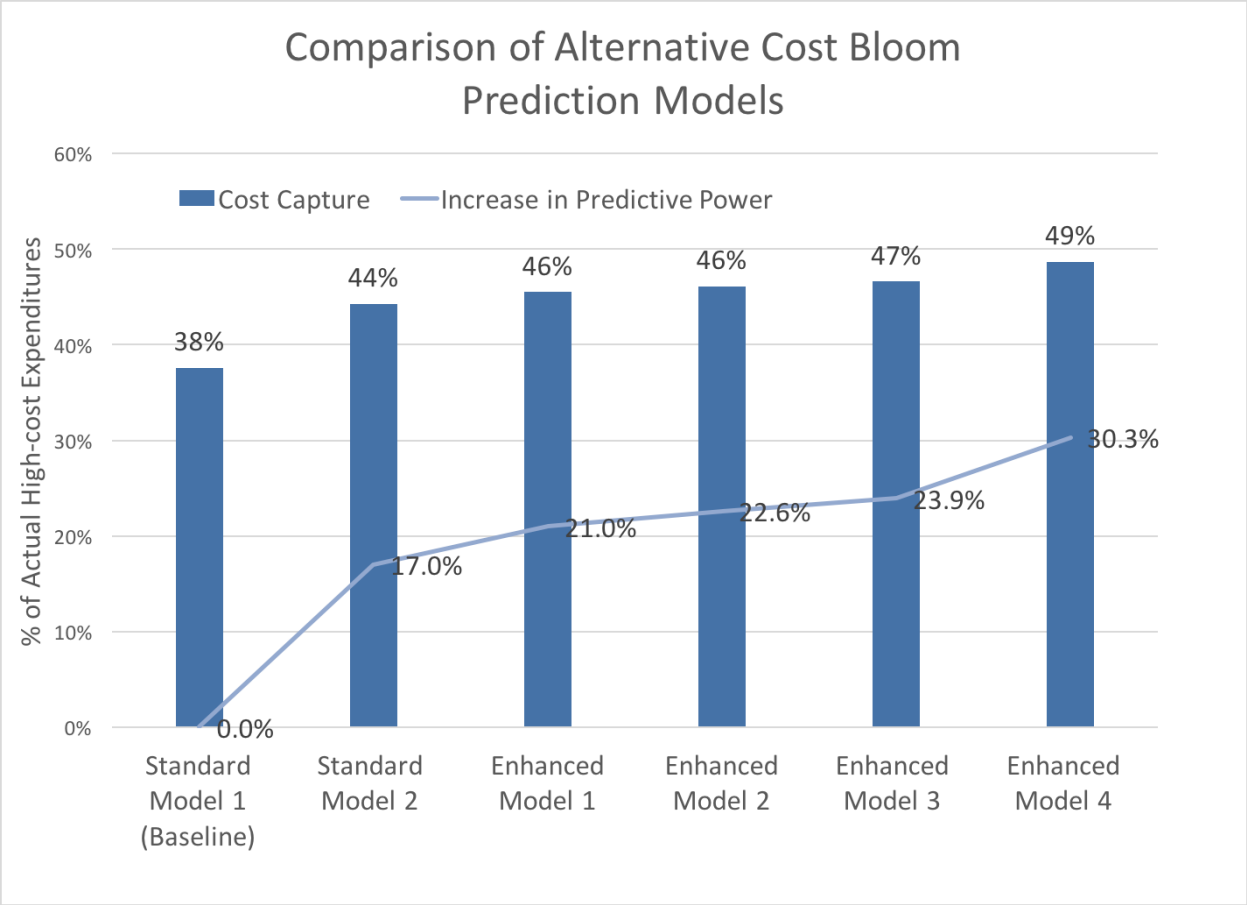

Figure) with only seven model features. Also, we found that our full feature set without cost features resulted in a 24 percent increase over the baseline model, suggesting the benefit of our enhanced models for providers who cannot link cost information.

The generalizability of our findings to other national health systems is a limitation of our study. Similar to the US, the bulk of Denmark's annual health cost is driven by hospital-based services and annual healthcare costs are highly concentrated among a small fraction of the population. Since the distribution of national health costs, medical visits, and disease profiles in Denmark is consistent with that of other industrialized countries (see Figures e1 through e4 in the eSupplement), our findings should be relevant in other similar settings. However, we acknowledge that there are differences among residents, insurance status (or lack thereof),

follow-up times, and other national health system characteristics. In the absence of access to a comparably large sample, representative of the US population, for studying high-cost patients, a future direction for our work is the external validation of our models using private insurance market data from the US.

Finally, the ability to more accurately predict future high-cost patients is an important first step to improving the value of their care. However, high-performing models are only as beneficial as the evidence-based practices in place for managing the care of future high-cost patients[5 6 10]. Our diverse set of cost-prediction features resulted in improvements over standard models and allowed us to characterize some distinctions between persistent high-cost patients and cost bloomers; however, our approach to prediction emphasizes performance over interpretability – *i.e.*, a key limitation of our models is that they are not designed to provide a meaningful "explanation" of why someone will bloom. Accurate prediction of the cost bloomers is the first step, but to inform the development of interventions or polices related to compensation for the care and management of cost bloomers, additional analyses to characterize more specific disease and utilization profiles are warranted, and may help to better understand the potential for providers and payers to effectively address the factors underlying cost blooming.

CONCLUSIONS

We carried out the commonly performed prediction of high-cost patients at the population-level and described a new framework for predicting cost bloomers. We demonstrate that diverse population health data, in conjunction with modern statistical learning methods for analyzing large datasets, can improve prediction of future high-cost patients over standard diagnosis-based tools, especially for our cost-bloom prediction task. Our best performing

enhanced model captured 60 percent of high-cost patient expenditures and 49 percent of cost bloomer expenditures. It also achieved achieved 21 percent and 30 percent improvements in cost capture over a standard diagnosis-based claims model for predicting future high-cost patients and cost bloomers, respectively. We expect our study to inform providers and payers, who need better strategies to address the differential risks posed by the small fraction of patients who account for the bulk of population healthcare spending.

ACKNOWLEDGEMENTS

This study is supported by the Aarhus University Research Foundation.

COMPETING INTERESTS

All authors have completed the ICMJE uniform disclosure. Henrik Toft Sørensen and Lars Pedersen are supported by the Program for Clinical Research Infrastructure (PROCRIN) established by the Lundbeck Foundation and the Novo Nordisk Foundation. Suzanne Tamang, Arnold Milstein, Lester Mackey, Jean-Raymond Betterton, Lucas Janson and Nigam Shah have no competing interests to disclose.

CONTRIBUTORS

AM and NHS had the original idea for the study. ST, HTS, LP, LM and LJ contributed to the design of the study and ST is the guaranteer. AM and HTS facilitated the data use agreement; LP provided ST with cleaned and deidentified study data. ST preformed the descriptive analyses and developed the data matrices for the prediction study. ST, JK, LM and LJ conducted the prediction study. All authors reviewed the manuscript. ST, LM, JB, LJ and NHS reviewed the analysis. ST and NHS declare that they had full access to all of the data in the

study, and can take responsibility for the integrity of the data and the accuracy of the data analysis, and controlled the decision to publish. The authors grant an exclusive international license to the publisher.

DATA SHARING AGREEMENT

No additional study data are available.

TRANSPARENCY STATEMENT

This statement affirms that our manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

REFERENCES

1. National Institute for Health Care Management. The Concentration of Health Care Spending. NIHCM Data Brief: NIHCM Foundation, 2012.

2. Joynt KE, Gawande AA, Orav EJ, et al. Contribution of preventable acute care spending to total spending for high-cost Medicare patients. JAMA 2013;**309**(24):2572-8.

3. Cohen S, Uberoi N. Differentials in the Concentration in the Level of Health Expenditures across Population Subgroups in the U.S., 2010. Statistical Brief: Agency for Healthcare Research and Quality, 2013.

4. Cohen SB, Yu W. The Concentration and Persistence in the Level of Health Expenditures over Time: Estimates for the U.S. Population, 2008-2009: Agency for Healthcare Research and Quality, 2012.

5. Bates DW, Saria S, Ohno-Machado L, et al. Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. Health Affairs 2014;**33**(7):1123-31.

6. Douglas McCarthy JR, Sarah Klein. Models of Care for High-Need, High-Cost Patients: An Evidence Synthesis: The Commonwealth Fund, 2015.

7. Ash AS, Zhao Y, Ellis RP, et al. Finding future high-cost cases: comparing prior cost versus diagnosis-based methods. Health services research 2001;**36**(6 Pt 2):194-206.

8. Feder JL. Predictive Modeling And Team Care For High-Need Patients At HealthCare Partners. Health Affairs 2011;**30**(3):416-18.

9. Haas LR, Takahashi PY, Shah ND, et al. Risk-stratification methods for identifying patients for care coordination. The American journal of managed care 2013;**19**(9):725-32.

10. Hong C, Siegel A, Ferris T. Caring for High-Need, High-Cost Patients: What Makes for a Successful Care Management Program? Issue Brief: The Commonwealth Fund, 2014.

11. Kansagara D, Englander H, Salanitro A, et al. Risk Prediction Models for Hospital Readmission. JAMA 2011;**306**(15):1688.

12. Meenan RT, Goodman MJ, Fishman PA, et al. Using risk-adjustment models to identify high-cost risks. Medical care 2003;**41**(11):1301-12.

13. Rose S. A Machine Learning Framework for Plan Payment Risk Adjustment. Health services research 2016.

14. Roski J, Bo-Linn GW, Andrews TA. Creating Value In Health Care Through Big Data: Opportunities And Policy Implications. Health Affairs 2014;**33**(7):1115-22.

15. Wherry LR, Burns ME, Leininger LJ. Using Self-Reported Health Measures to Predict High-Need Cases among Medicaid-Eligible Adults. Health services research 2014;**49 Suppl 2**:2147-72.

16. Yong PL, Saunders RS, Olsen LA. The Healthcare Imperative: Lowering Costs and Improving Outcomes. Workshop Series Summary. Washington (DC), 2010.

17. Zook CJ, Moore FD. High-Cost Users of Medical Care. New England Journal of Medicine 1980;**302**(18):996-1002.

18. The Commonwealth Fund Commission on a High Performance Health System. The Performance Improvement Imperative: Utilizing a Coordinated, Community-Based Approach to Enhance Care and Lower Costs for Chronically Ill Patients: The Commonwealth Fund, 2012.

19. Schone E, Brown R. Risk Adjustment: What is the current state of the art, and how can it be improved?: Robert Wood Johnson Foundation, 2013.

20. Lodh M, Raleigh ML, Uccello CE, et al. Risk Assessment and Risk Adjustment. Issue Brief: American Academy of Actuaries, 2010.

21. Winkelman R MS. A comparative analysis of claims-based tools for health risk assessment: Society of Actuaries., 2007:1-70.

22. Centers for Medicare & Medicaid Services. HHS-Operated Risk Adjustment Methodology Meeting. Discussion Paper: Centers for Medicare & Medicaid Services, Center for Consumer Information & Insurance Oversight, 2016.

23. Goodson JD, Bierman AS, Fein O, et al. The future of capitation: the physician role in managing change in practice. Journal of general internal medicine 2001;**16**(4):250-6.

24. Asthana S, Gibson A. Setting health care capitations through diagnosis-based risk adjustment: a suitable model for the English NHS? Health Policy 2011;**101**(2):133-9.

25. T. Hastie RT, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction.: Springer, 2011.

26. glmnet: Lasso and elastic-net regularized generalized linear models [program], 2009.

27. Deo RC. Machine Learning in Medicine. Circulation 2015;**132**(20):1920-30.

28. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science 2015;**349**(6245):255-60.

29. Bibault JE, Giraud P, Burgun A. Big Data and machine learning in radiation oncology: State of the art and future prospects. Cancer Lett 2016.

30. Passos IC, Mwangi B, Kapczinski F. Big data analytics and machine learning: 2015 and beyond. Lancet Psychiatry 2016;**3**(1):13-5.

31. Winkelman R, Mehmud S. A Comparative Analysis of Claims-Based Tools for HealthRisk Assessment: Society of Actuaries, 2007.

32. Cummings RB, Cameron BA. A Comparative Analysis of Claims-based Methods for Health Risk Assessment for Commercial Populations: Society of Actuaries, 2002.

33. Fleishman JA, Cohen JW. Using Information on Clinical Conditions to Predict High-Cost Patients. Health services research 2010;**45**(2):532-52.

34. Moturu ST, Johnson WG, Liu H. Predictive risk modelling for forecasting high-cost patients: a real-world application using Medicaid data. International Journal of Biomedical Engineering and Technology 2010;**3**(1/2):114.

35. DeSalvo KB, Fan VS, McDonell MB, et al. Predicting mortality and healthcare utilization with a single question. Health services research 2005;**40**(4):1234-46.

36. Saunders MK. In Denmark, big data goes to work. Health Aff (Millwood) 2014;**33**(7):1245.

37. Johannesdottir SA, Horvath-Puho E, Ehrenstein V, et al. Existing data sources for clinical epidemiology: The Danish National Database of Reimbursed Prescriptions. Clinical epidemiology 2012;**4**:303-13.

38. Pedersen CB. The Danish Civil Registration System. Scandinavian journal of public health 2011;**39**(7 Suppl):22-5.

39. Pedersen CB, Gotzsche H, Moller JO, et al. The Danish Civil Registration System. A cohort of eight million persons. Dan Med Bull 2006;**53**(4):441-9.

40. Schmidt M, Pedersen L, Sorensen HT. The Danish Civil Registration System as a tool in epidemiology. European journal of epidemiology 2014;**29**(8):541-9.

41. Schmidt M, Pedersen L, Sørensen HT. The Danish Civil Registration System as a tool in epidemiology. European journal of epidemiology 2014;**29**(8):541-49.

42. Hansen RP, Olesen F, Sorensen HT, et al. Socioeconomic patient characteristics predict delay in cancer diagnosis: a Danish cohort study. BMC health services research 2008;**8**:49.

43. Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. Scandinavian journal of public health 2011;**39**(7 Suppl):30-3.

44. NOMESCO Nordic Medico Statistical Committee. NOMESCO: Classification of Surgical Procedures, 2007.

45. WHO Collaborating Centre for Drug Statistics Methodology. ATC classification index with DDDs, 2013.

46. Agency for Healthcare Research and Quality. Medical Expenditure Panel Survey (MEPS): AHRQ, Rockville, MD; 2015 [Available from: http://www.ahrq.gov/research/data/meps/index.html.

47. Zhao Y, Ash AS, Ellis RP, et al. Predicting pharmacy costs and other medical costs using diagnoses and drug claims. Medical care 2005;**43**(1):34-43.

48. Elixhauser A, Steiner C, Harris DR, et al. Comorbidity measures for use with administrative data. Medical care 1998;**36**(1):8-27.

49. Elixhauser A, Steiner C, Palmer L. Clinical Classifications Software (CCS): U.S. Agency for Healthcare Research and Quality, 2014.

50. Robinson JW. Regression tree boosting to adjust health care cost predictions for diagnostic mix. Health services research 2008;**43**(2):755-72.

51. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software 2010;**33**(1):1-22.

52. Hastie HZaT. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 2005;**67**(Part 2):301–20.

53. Taylor J, Tibshirani RJ. Statistical learning and selective inference. Proc Natl Acad Sci U S A 2015;**112**(25):7629-34.

54. Eldar E, Hauser TU, Dayan P, et al. Striatal structure and function predict individual biases in learning to avoid pain. Proceedings of the National Academy of Sciences 2016;**113**(17):4812-17.

55. Ebrahimi M, Boughorbel S, Al-Ali R, et al. Model Comparison for Breast Cancer Prognosis Based on Clinical Data. Plos One 2016;**11**(1):e0146413.

56. Guo Y, Wei Z, Keating BJ, et al. Machine learning derived risk prediction of anorexia nervosa. BMC Med Genomics 2016;**9**:4.

57. Darcy AM, Louie AK, Roberts LW. Machine Learning and the Profession of Medicine. JAMA 2016;**315**(6):551-2.

58. Bertsimas D, Bjarnadóttir MV, Kane MA, et al. Algorithmic Prediction of Health-Care Costs. Operations Research 2008;**56**(6):1382-92.

eSUPPLEMENT

The Danish Healthcare System

The Danish healthcare system has three administrative levels: (1) the state, which is responsible for legislation, national guidelines, surveillance, and health financing through the Ministry of Health and Prevention; (2) the five regions, which are responsible for the delivery of primary and secondary care; and (3) the 98 municipalities, which are responsible for school health, child dental care, home nursing, public health, prevention, and rehabilitation.

Population Health Data Sources

Primary Care and Specialist Care Files[1]

In the Danish National Health Service more than 98% of patients (Group 1) are registered with a general practitioner (GP) of their choice. The patients have the right to free treatment from their GP. Treatment by a specialist is available after referral from a GP. The GPs or specialists are paid for each consultation and supplementary diagnostic and therapeutic procedures, recorded according to patients' Civil Registration Number. The Health Service records the number of patients on the doctor's list, and the number of doctors in practice. Patients in Group 2 can select a specialist as well as a GP, but must pay a part of the GP's and specialist's fee.

The National Health Service Prescription Database[2]

Pharmacists in Denmark are equipped with an electronic accounting system primary used to secure reimbursement from the National Health Service. The database includes all prescriptions redeemed since 2004. For each redeemed prescription, information on patients' Civil Registration Number, the amount and type of drug prescribed according to the Anatomical Therapeutic Chemical (ATC) Classification System, and the day the drug was dispensed is transferred electronically from all pharmacies in Denmark to the database.

The Danish National Patient Registry[3]

The Danish National Patient Registry, covering all Danish hospitals, contains data on admissions and discharge dates and discharge diagnoses from all Danish non-psychiatric hospitals since 1977 and on emergency room and outpatient clinic visits since 1995. Each hospital discharge is assigned one primary diagnosis and up to 19 secondary diagnoses, classified according to the International Classification of Diseases, 10th edition.

The Danish Civil Registration System[4]

The Danish Civil Registration System (CRS), established in April 1968, records all changes in vital status and emigration in the entire Danish population, with daily electronic updates. Upon registration in the CRS, each resident receives a unique Civil Registration Number, which is used in all Danish registries.

DESCRIPTIVE STATISTICS

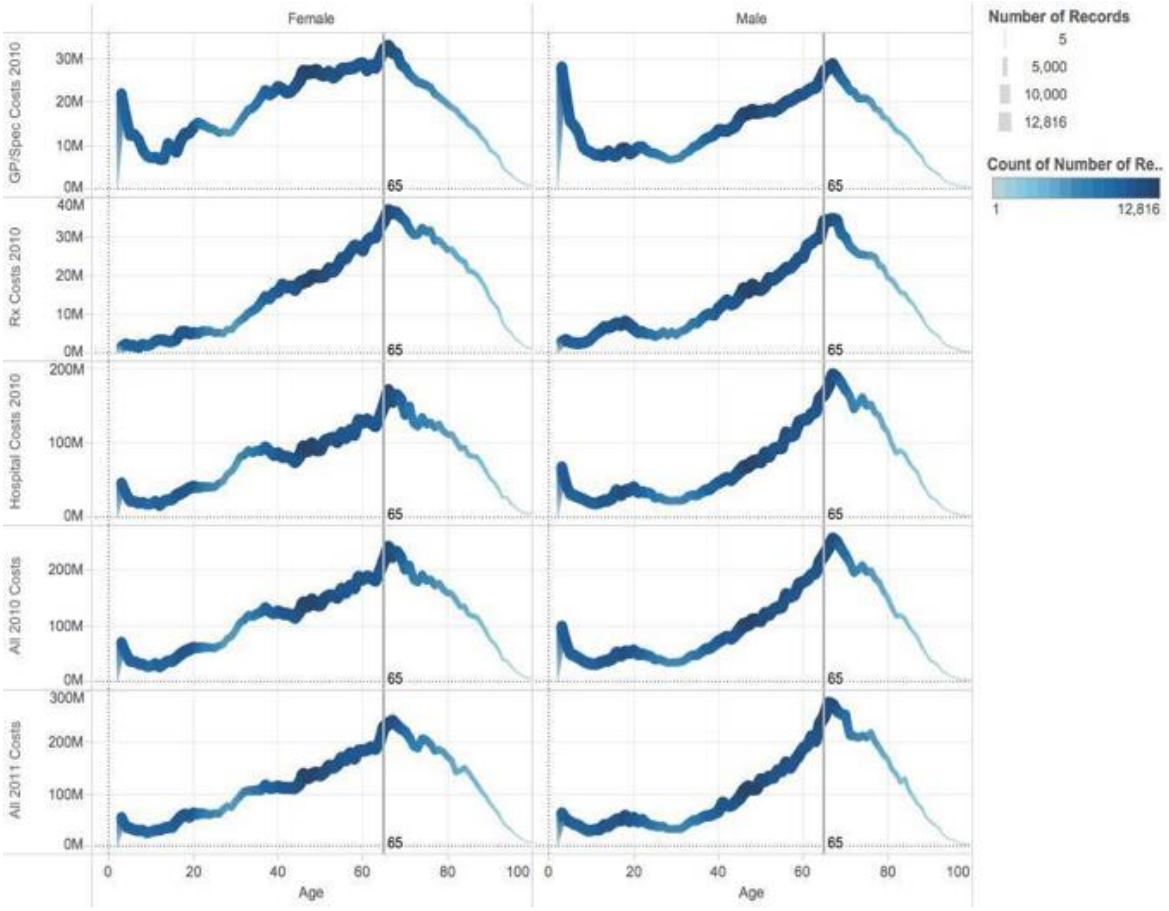

Exhibit E1. healthcare costs in 2010 and total cost in 2011 by patient. Primary care and specialist (GP/Spec), medication (Rx), hospital and hospital clinic (Hospital), and total (All) costs in 2010; total costs in 2011. Residents are shown by gender and age; color and thickness of the lines indicate the number of records; drop lines mark individuals who are 65 years old.

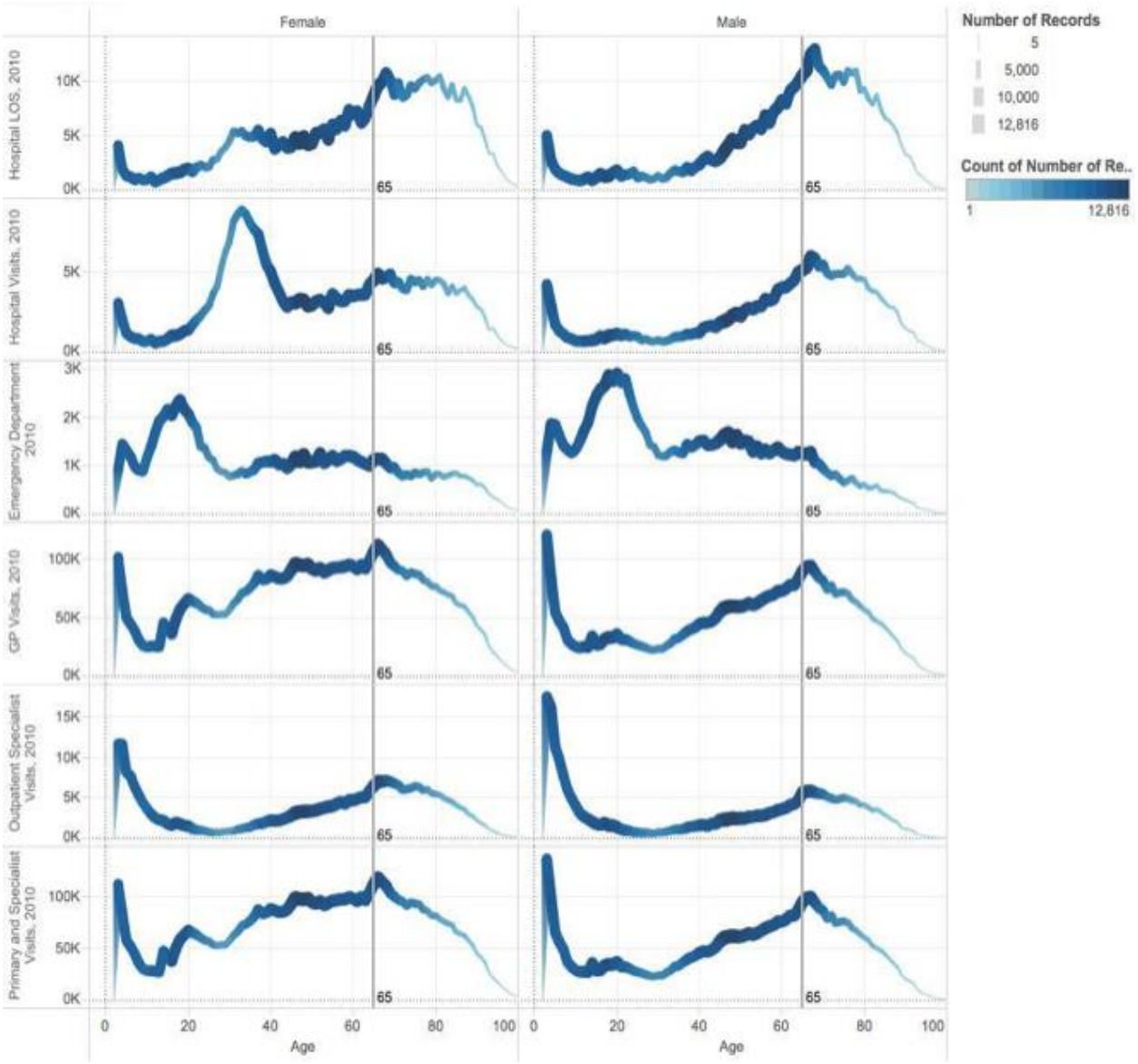

Figure E2. Total healthcare utilization in 2010 by patient. Inpatient hospital days (LOS), hospital and hospital clinic (Hospital) visits, emergency visits, primary care (GP) visits, outpatient specialist visits, and combined primary care and specialist (Primary and Specialist) visits in 2010. Residents shown by gender and age; color and thickness of the line indicate the number of records; drop line marks individuals who are 65 years old.

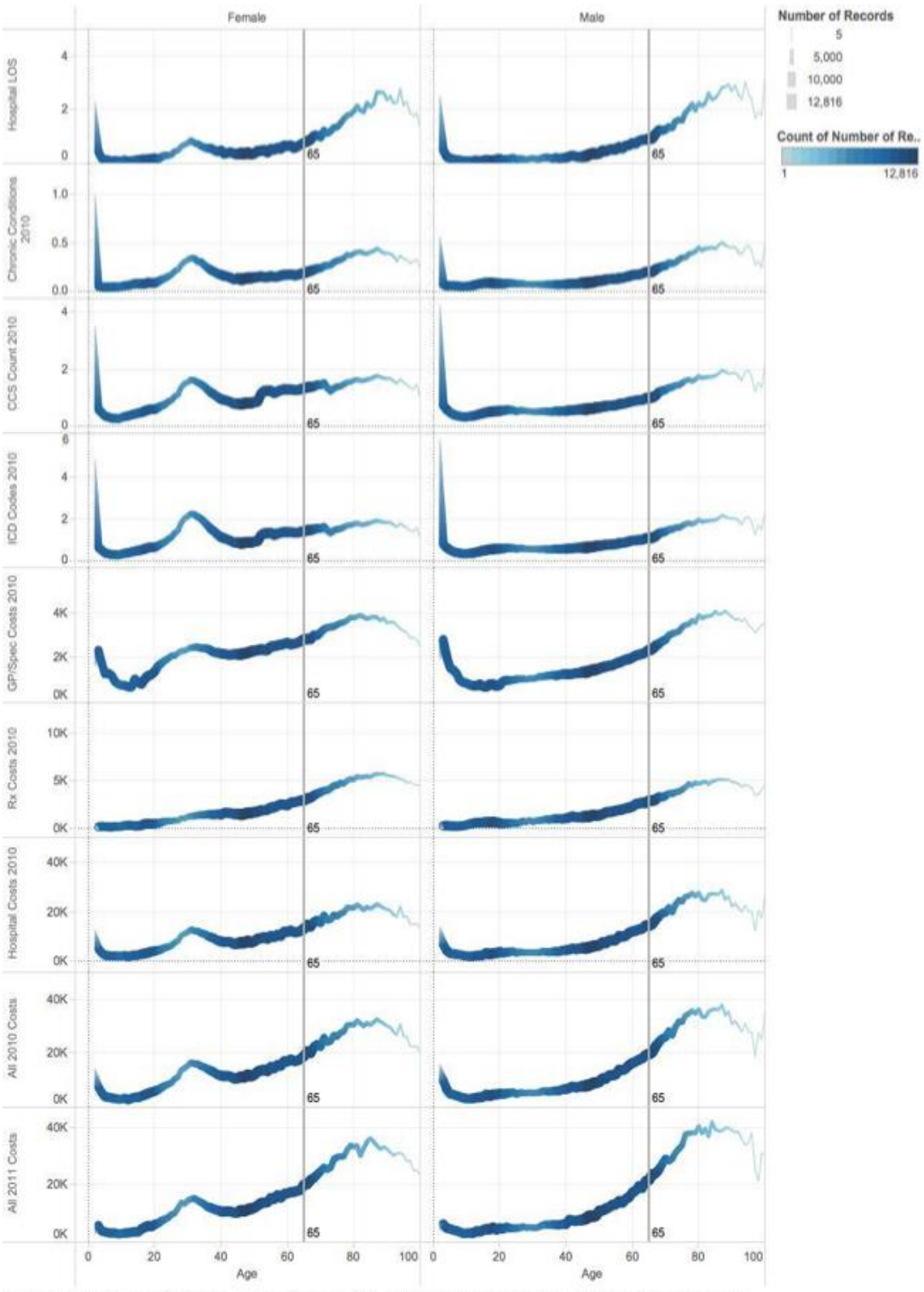

Figure E3. Average healthcare utilization in 2010 and average cost in 2011 by patient. Hospital days (LOS), chronic conditions (CCI), clinically significant diseases (CCS), ICD codes, primary care and specialist visits (GP/Spec), medications filled (Rx), hospital visits, and total (All) costs in 2010; average total costs in 2011. Residents are shown by gender and age; color and thickness of the line indicate the number of records; drop line marks individuals who are 65 years old.

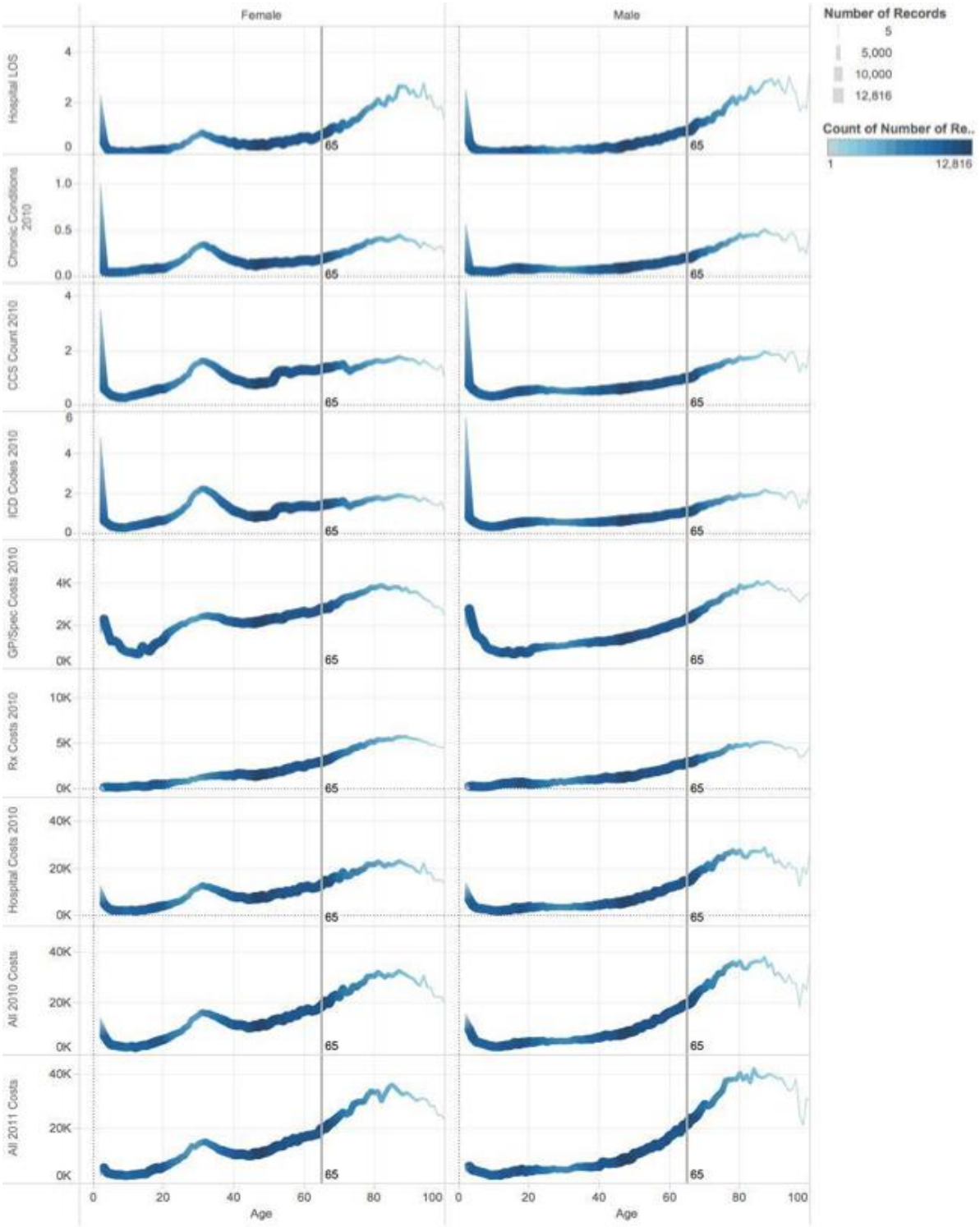

Figure E4. Median healthcare utilization in 2010 and median cost in 2011 by patient. Hospital days (LOS), chronic conditions (CCI), clinically significant diseases (CCS), ICD codes, primary care and specialist visits (GP/Spec), medications filled (Rx), hospital visits, and total (All) costs in 2010; median total costs in 2011. Residents are shown by gender and age; color and thickness of the line indicate the number of records; drop line marks individuals who are 65 years old.

REFERENCES

1. Hansen RP, Olesen F, Sorensen HT, et al. Socioeconomic patient characteristics predict delay in cancer diagnosis: a Danish cohort study. BMC health services research 2008;**8**:49.

2. Johannesdottir SA, Horvath-Puho E, Ehrenstein V, et al. Existing data sources for clinical epidemiology: The Danish National Database of Reimbursed Prescriptions. Clinical epidemiology 2012;**4**:303-13.

3. Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. Scandinavian journal of public health 2011;**39**(7 Suppl):30-3.

4. Schmidt M, Pedersen L, Sørensen HT. The Danish Civil Registration System as a tool in epidemiology. European journal of epidemiology 2014;**29**(8):541-49.