RIVERSIDE SPECIALIST HOSPITAL

CHRONIC DISEASES DEPARTMENT: CARDIOVASCULAR DISEASES

Compiled by: Agrippine Tobias

Introduction

Riverside Specialist Hospital in India is committed to improving patient outcomes and reducing the burden of cardiovascular diseases (CVD) through data-driven healthcare initiatives. Cardiovascular disease remains one of the leading causes of hospital admissions and mortality, and understanding the underlying patterns within our patient population is key to effective prevention and treatment.

This project analyzes clinical and demographic data collected from patients at Riverside Specialist Hospital to identify trends, risk factors, and high-risk groups associated with cardiovascular disease. By exploring indicators such as age, blood pressure, cholesterol levels, fasting blood sugar, and exercise responses, the hospital aims to pinpoint which factors contribute most to disease development.

The analysis follows a structured **data analysis process (ASK → PREPARE → PROCESS → ANALYZE → SHARE → ACT)** to ensure clarity, accuracy, and actionable insights. Data were prepared and cleaned using **Microsoft Excel**, while **Tableau** was used for visual exploration and dashboard creation.

The ultimate goal of this project is to translate analytical insights into **evidence-based prevention campaigns**, enabling Riverside Specialist Hospital to:

- Target screening and education efforts toward the most vulnerable populations.
- Develop tailored health interventions addressing modifiable risk factors.
- Monitor the effectiveness of hospital and community-based prevention strategies through defined Key Perfomance Indicators (KPIs)

This analysis demonstrates how hospitals can leverage data analytics to drive better decision-making, promote population health, and optimize resource allocation for cardiovascular disease management.

Cardiovascular Disease Data Analysis Roadmap

1. ASK

Project Objective: Analyze the hospital’s cardiovascular disease (CVD) burden, identify highest-risk subgroups and main modifiable drivers, and then recommend targeted prevention campaigns that will most reduce short-term CVD burden among patients.

Primary questions to answer:

1. What is the prevalence of CVD in our population?

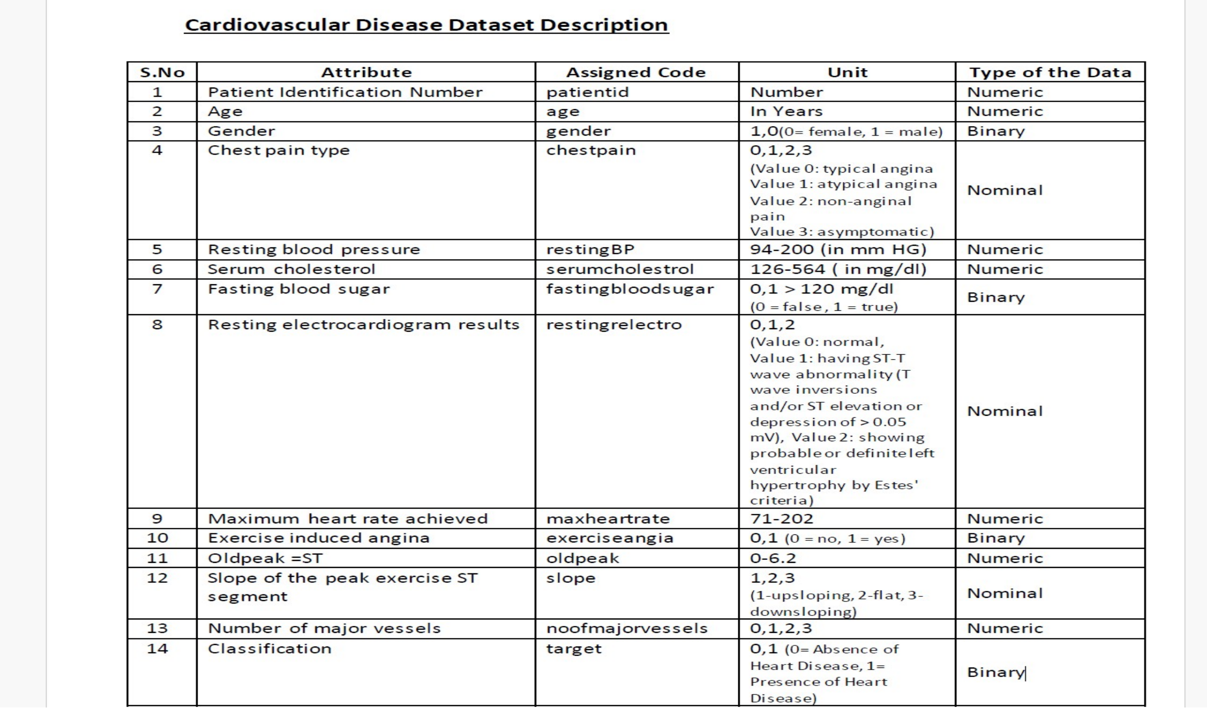
2. Which demographic groups (age, gender) are most affected?
3. Which modifiable factors (BP, cholesterol, fasting sugar, etc.) drive disease risk?
4. Which interventions will most reduce future CVD cases?
5. What KPIs should the hospital track to measure the impact of prevention campaigns?

Key stakeholders: Litu Tobias (Chief Program Officer) and Riverside Hospital surveillance team

2. PREPARE

2.1 **Dataset used**

The data source used for this project is Cardiovascular disease dataset. This dataset is downloaded from Kaggle (https://www.kaggle.com/code/jocelyndumlao/cardiovascular-health-analysis/input) and was made available through Jocelyn Munlao**.**



Cardiovascular Disease Dataset Description

S.No	Attribute	Assigned Code	Unit	Type of the Data
1	Patient Identification Number	patientid	Number	Numeric
2	Age	age	In Years	Numeric
3	Gender	gender	1,0(0= female, 1 = male)	Binary
4	Chest pain type	chestpain	0,1,2,3 (Value 0: typical angina Value 1: atypical angina Value 2: non-anginal pain Value 3: asymptomatic)	Nominal
5	Resting blood pressure	restingBP	94-200 (in mm HG)	Numeric
6	Serum cholesterol	serumcholestrol	126-564 (in mg/dl)	Numeric
7	Fasting blood sugar	fastingbloodsugar	0,1 > 120 mg/dl (0 = false , 1 = true)	Binary
8	Resting electrocardiogram results	restingrelectro	0,1,2 (Value 0: normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)	Nominal
9	Maximum heart rate achieved	maxheartrate	71-202	Numeric
10	Exercise induced angina	exerciseangia	0,1 (0 = no, 1 = yes)	Binary
11	Oldpeak =ST	oldpeak	0-6.2	Numeric
12	Slope of the peak exercise ST segment	slope	1,2,3 (1-upsloping, 2-flat, 3-downsloping)	Nominal
13	Number of major vessels	noofmajorvessels	0,1,2,3	Numeric
14	Classification	target	0,1 (0= Absence of Heart Disease, 1= Presence of Heart Disease)	Binary

2.2 **Accessibility and privacy of data**:

Data used is from Kaggle, a public domain, thus it is open source. The owner has dedicated the work to the public domain by waiving all of his or her rights to work worldwide under copyright law for users to modify, distribute and perform the work, even for commercial purposes, all without asking permission. Patients id column was removed to protect patient privacy.

2.3 **Information about our dataset:**

This heart disease dataset is acquired from one of the multispecialty hospitals in India. Over 14 common features make it one of the heart disease datasets available so far for research purposes. This dataset consists of 1000 participants aged from 20 to 80 years, both males and females randomly sample at the chronic disease department.

2.4 **Data limitations**

While the dataset provides valuable information for understanding cardiovascular disease patterns among patients at **Riverside Specialist Hospital**, there are several limitations that should be acknowledged when interpreting the findings:

1. **Undefined Data Collection Period**

- The dataset does not specify the time frame during which patient records were collected. Without a defined collection period, it is difficult to determine whether the data reflects current trends or historical patterns.

2. **Lack of Lifestyle and Behavioral Information**

- Important risk factors such as smoking status, alcohol consumption, diet, and physical activity levels are not included.These lifestyle factors play a major role in cardiovascular risk, so their absence limits the ability to assess the full range of modifiable contributors.

3. **No Data on Underlying or Coexisting Conditions**

The dataset does not include information on other chronic diseases (e.g., diabetes duration, obesity, chronic kidney disease) that can significantly influence cardiovascular risk.This limits the ability to adjust for comorbidities when identifying independent risk factors.

4. **Cross-sectional Nature of Data**

- The dataset represents a single snapshot of each patient rather than longitudinal follow-up data.As a result, causal relationships or disease progression over time cannot be determined, only associations.

Despite these limitations, the dataset still provides meaningful insights into clinical and demographic patterns of cardiovascular disease. However, future analyses should aim to incorporate:lifestyle and behavioral factors,longitudinal (follow-up) data, and broader clinical histories to strengthen understanding of risk dynamics and improve the design of prevention campaigns.

3. PROCESS

DATA QUALITY

Microsoft Excel is used for data cleaning

3.1 . Import and Initial Audit

- Opened CSV in Excel . Save working copy as Cardiovascula_Disease_Dataset.xlsx.
- **Total rows**: **=COUNTA(A:A)-1, thus there are 1000 participants.**
- **Duplicates check**: Conditional Formatting > Highlight Cells Rules > Duplicate Values, no duplicate found
- **Missing values per column**: use =COUNTBLANK(B2:B1001), 0. No missing values found

3.2 B. Standardize types

- Ensured numeric columns are numeric. Used VALUE().
- Converted categorical numeric codes to readable labels

✓ Gender: =IF(C2=0,"Female","Male")
✓ Chest pain: =CHOOSE(D2+1,"typical angina","atypical angina","non-anginal","asymptomatic")

- Changed title “target” in the last column to “disease_outcome”

3.3 **Create useful derived columns**

- **Age group**

=MIN(B:B), 20

Divided age groups as follows

✓ **=IF(B2<30,"20-29",IF(B2<40,"30-39",IF(B2<50,"40-49",IF(B2<60,"50-59",IF(B2<70,"60-69","70+")))))**

- **Hypertension flag**

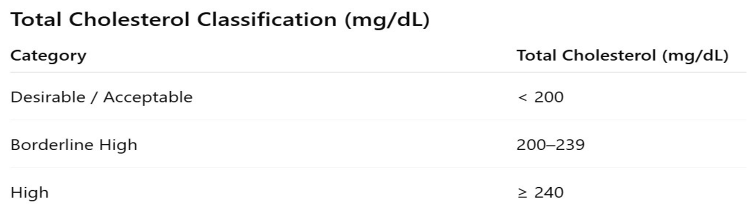
Divided the blood pressure into 2 categories namely: 90-119 is stage 2 hyptenstion and above 120 is severe hypertension as recommended by the American Heart Association (https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings)

- Serum cholesterol

➢ 53/1000 records were found to have 0 as a value, which is not physiologically possible, thus a median serum cholesterol was calculated to minimize bias.

- =MEDIAN(IF(F2:F1000<>0,F2:F1000)), =326.5, thus all records with zero initially have this new value.

Serum cholesterol was grouped as follows.



Total Cholesterol Classification (mg/dL)

Category	Total Cholesterol (mg/dL)
Desirable / Acceptable	< 200
Borderline High	200–239
High	≥ 240

3.4 **Missing values**

During data validation, 53 out of 1,000 patient records (5.3%) were found to have serum cholesterol values recorded as **0 mg/dL**.
Since a value of zero is **physiologically impossible**, these entries were identified as **missing or erroneous data** likely due to incomplete records or data entry issues.

To maintain the overall integrity of the dataset without reducing the sample size, the invalid cholesterol values were **replaced with the median serum cholesterol** computed from valid entries. This method minimizes bias while preserving the distribution of the data for subsequent analysis.

This cleaning step was documented to ensure transparency and reproducibility of the analysis process.

4. ANALYZE

SUMMARIES USING EXCEL

4.1 Prevalence of heart disease by age groups

Count of disease_outcome	**disease_outcome**		
age_group	**Absence of heart disease**	**Presence of heart disease**	**Grand Total**
20-29	8.40%	9.70%	18.10%
30-39	6.90%	9.80%	16.70%
40-49	5.90%	9.90%	15.80%
50-59	6.70%	9.80%	16.50%
60-69	5.70%	9.10%	14.80%
70+	8.40%	9.70%	18.10%
Grand Total	**42.00%**	**58.00%**	**100.00%**

Table 1: Prevalence of heart disease by AgeGroups

As shown in Table 1 above, the prevalence of heart disease in our population is 58%, of which 13.30% are females and 44.70% are males. According to research, heart disease increase with age, however our data shows that all groups have a similar prevalence of disease with values ranging from 9.10% in those aged 60-69 as lowest to 9.90% in those aged 40-49 as highest.

4.2 Prevalence of heart disease by gender

Females

gender	female

Count of disease_outcome	**disease_outcome**		
age_group	**Absence of heart disease**	**Presence of heart disease**	**Grand Total**
20-29	8.51%	9.36%	17.87%
30-39	5.53%	9.36%	14.89%
40-49	6.81%	9.36%	16.17%
50-59	7.23%	9.36%	16.60%
60-69	7.66%	8.94%	16.60%
70+	7.66%	10.21%	17.87%
Grand Total	**43.40%**	**56.60%**	**100.00%**

Table 2: Prevalence of heart disease in females

As shown in Table 2, 56% of females had heart disease in the study. The rate of disease is almost similar between age groups however the 70+ had the highest prevalence.

Males

gender	male

Count of disease_outcome	disease_outcome		
age_group	**Absence of heart disease**	**Presence of heart disease**	**Grand Total**
20-29	8.37%	9.80%	18.17%
30-39	7.32%	9.93%	17.25%
40-49	5.62%	10.07%	15.69%
50-59	6.54%	9.93%	16.47%
60-69	5.10%	9.15%	14.25%
70+	8.63%	9.54%	18.17%
Grand Total	**41.57%**	**58.43%**	**100.00%**

Table 3: Prevalence of heart disease in males

As shown above in Table 3, 58% of males had heart disease. The rate of disease is consistent in different age groups with age 40-49 with the highest prevalence of 10%. Compared to females, males recorded a slightly higher rates in disease within different age groups.

ANALYSIS IN TABLEAU SOFWARE:

- Opened **Tableau Desktop** → **Connect** → Text file → choose file.

- Ensured that Age Group and Gender are dimensions (string), target and Flags are numbers (whole).

4.3 **Calculating prevalence**

- Used a calculated field: *{ FIXED [Age Group], [Gender] : SUM(IF [Target] = 1 THEN 1 ELSE 0 END) }*

=580, thus 580/1000 is 58%

4.4 Risk factor prevalence among diseased

A. Hypertension

- Used a calculated field: *SUM(IF [Target] = 1 AND [hypertensionParamFlag] = 1 THEN 1 ELSE 0 END) /SUM(IF [Target] = 1 THEN 1 ELSE 0 END)*

= 96.55% of those diseased had severe hypertension.

B. **Serum Cholesterol**

- Used a calculated field (cholesterol_category) to categorize cholesterol:

=*IF [Serumcholestrol] < 200 THEN "Desirable" ELSEIF [Serumcholestrol] >= 200 AND [Serumcholestrol] <= 239 THEN "Borderline High" ELSE "High Cholesterol" END*

- Used a calculated field (cholesterolFlag) to create a high cholesterol flag:

=IF [Serumcholestrol] >= 240 THEN 1 ELSE 0 END

- Used calculated field to create the prevalence of high cholesterol among diseased :

=*SUM(IF [Target] = 1 AND [highcholesterolFlag] = 1 THEN 1 ELSE 0 END)/ SUM(IF [Target] = 1 THEN 1 ELSE 0 END)*

= 84.48% of those with heart diseases have high serum cholesterol

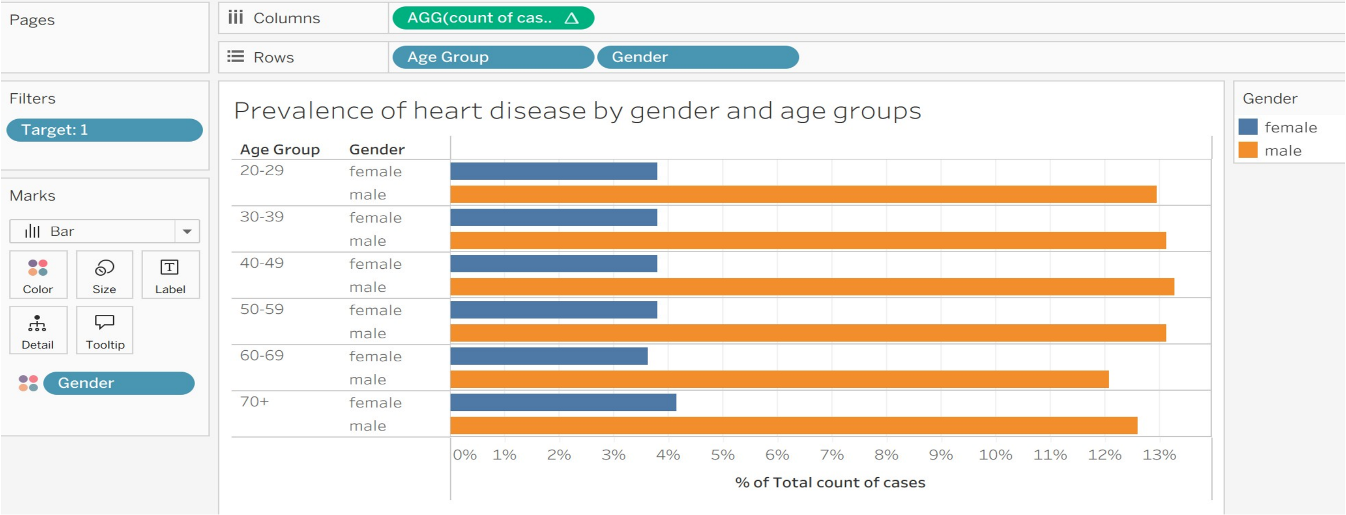
C. **Fasting blood sugar**

- Used a calculated field to calculate prevalence of high blood sugar(diabetes) in diseased

=*SUM(IF [Target] = 1 AND [Fastingbloodsugar] = 1 THEN 1 ELSE 0 END)/SUM(IF [Target] = 1 THEN 1 ELSE 0 END)*

= 41.38% of those with heart diseases are diabetic

4.5 Dashboard 1: Risk and Demographics



Columns: AGG(count of cas..

Rows: Age Group, Gender

Filters: Target: 1

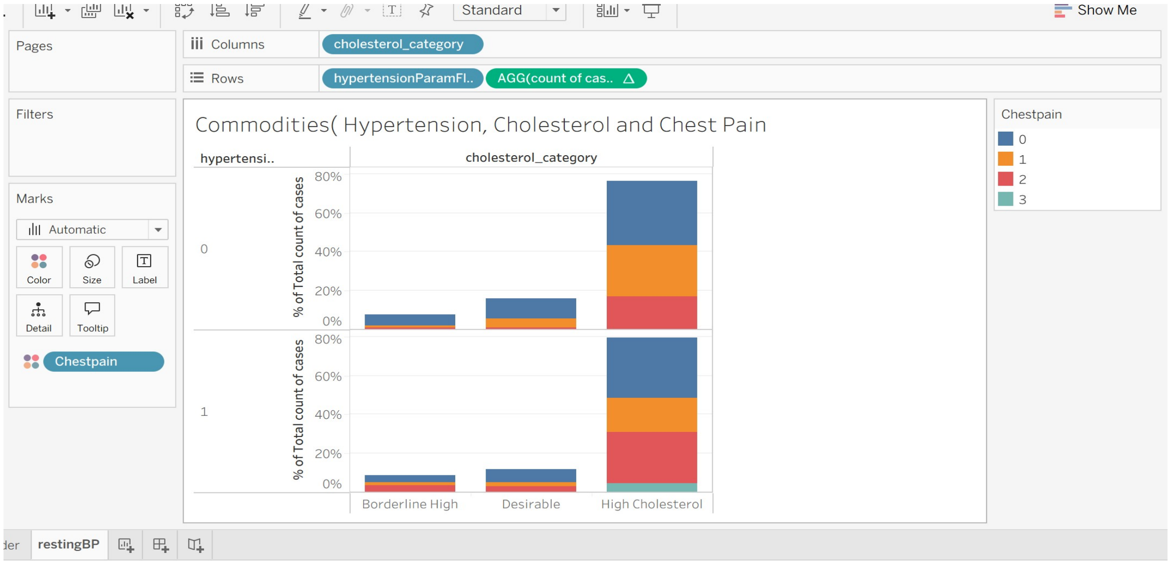
Marks: Bar

Color: Gender

Figure 1: The prevalence of heart diseases by gender and age groups

As shown in the above figure, females had a lower rate of heart diseases than males in all age groups. Further, in males the rates was more higher in the younger population, however in females, disease risk increased with age.

4.6 Dashboard 2: Resting blood pressure and cholesterol


Columns: cholesterol_category

Rows: hypertensionParamFl.. AGG(count of cas.. △

Marks: Automatic

Chestpain

Figure 2: Hypertension vs. Cholesterol

Looking at Figure 2 above, all the patients involved in this study were hypertensive. However, the severe hypertensive also were having high cholesterol. Furthermore, the most common chest pain types in people with both hypertension and high cholesterol is typical angina. Those with both severe hypertension and high cholesterol are also likely to experience non-anginal pain.

5. SHARE AND ACT - Communicate Insights and Recommendations

After data analysis, the main findings to project objectives are:

1. Prevalence of CVD in our population and which demographics are affected?

- The prevalence of heart disease in our population is **58%**, of which 13.30% are females and 44.70% are males. There is no much difference in rates between the age groups, however the rate of disease in males in higher in the younger population while in females it increased with age. The prevalence rate is quite high, with such a huge gap between the rates of males as compared to females, thus compaigns will target efforts to alliviate the burden in the male population.

2. Which modifiable factors (BP, cholesterol, fasting sugar, etc.) drive disease risk?

- 96.55% of those diseased had severe hypertension. 84.48% of those with heart diseases have high serum cholestero 41.38% of those with heart diseases are diabetic. All the patients involved in this study were hypertensive. However, the severe hypertensive also were having high cholesterol. Furthermore, the most common chest pain types in people with both the less severe and severe hypertension and high cholesterol is typical angina. Those with both severe hypertension and high cholesterol are also likely to experience non-anginal pain.

- Thus this study concludes that **males, hypertension, high cholesterol, typical angina chest pain and diabetes are the main risk to heart diseases in the population**.

3. Which interventions will most reduce future CVD cases?

Based on our findings, the following interventions would likely have the highest impact:

A. Targeted Screening and Early Detection

- **Focus population:** Males, especially younger males, and older females because data shows high risk in this population.
- **Screenings to prioritize:** Blood pressure, cholesterol levels, fasting glucose, and cardiovascular risk assessment.
- **Implementation:** Scale up routine check-ups and introduce a community screening programs especially in areas where males are likely to show up or mobile clinics to increase accessibility.

B. Lifestyle and Behavioral Interventions

- **Hypertension control:** 96% of the cases were severly hypertensive, thus we need to promote healthy dietary habits including promoting low-sodium diets, weight management and adherence to antihypertensive medications.
- **Diabetes prevention/control:** Screening for pre-diabetes, dietary education, exercise programs, and glycemic control for diabetics.

C. Targeted Education Campaigns

Focus: Raising awareness about typical angina symptoms and non-anginal chest pain as warning signs of combined commodities that increase heart disease risk.

Mediums: Community meetings for the elderly, pamphlets distribution at the hospital, social media campaigns to increase accessibility of information, especially for the youth.

Goal: Encourage early healthcare seeking behavior to reduce late-stage CVD presentations.

D. Integrated Care Programs

Approach: The hospital to combine management of hypertension, diabetes, and hyperlipidemia to reduce cumulative risk.

Improve Care coordination: Robust patient follow-ups system, medication adherence reminders for patients , and multidisciplinary clinics (cardiology, endocrinology, nutrition).

4. What KPIs should the hospital track to measure impact of prevention campaigns?

KPI Category	KPI	Measurement / Metric
Population Health	Prevalence of hypertension, high cholesterol, diabetes	% of screened population with each condition
	High-risk individuals identified	% of population classified as high-risk
	Hospital admissions for CVD	Number of CVD-related admissions
Clinical Outcomes	Blood pressure control	% of hypertensive patients with BP <target
	Cholesterol control	% of patients with LDL cholesterol < target
	Blood glucose control	% of diabetic patients with fasting glucose < target
	Incidence of acute cardiovascular events	Number of heart attacks, strokes, angina episodes
Program Engagement	Screening program participation	Number of individuals screened
	Educational session attendance	Number of participants in workshops or campaigns
	Enrollment in lifestyle modification programs	Number of high-risk individuals enrolled