State-based Cardiovascular Disease Surveillance Data Pilot Project: Analyzing Cardiovascular Disease Survival in Wisconsin

Mike Yuan, Epidemiologist
Wisconsin Department of Health Services/Division of Public Health
Heart Disease and Stroke Prevention Program

BACKGROUND

Hospitalization discharge data is a claim-based record, which means it only records information by hospital visits, making it difficult to get an accurate picture of cardiovascular disease treatment and survival rates. In 2010, the Wisconsin Heart Disease and Stroke Prevention Program (HDSP) received funding from the Council of State and Territorial Epidemiologists (CSTE) and Centers for Disease Control and Prevention (CDC) to pilot a method of data de-duplication and linkage and for data extraction that could lead to a better understanding of cardiovascular disease survival rates. The de-duplication and linkage of hospitalization data to death data (2006-2008) provided an opportunity for Wisconsin to look at patient level hospital data. Furthermore, this project provided great opportunities for analysis of survival days for three major CVD events -- myocardial infarction (MI), congestive heart failure (CHF), and stroke (ST). This effective methodology will provide HDSP a tool to explore further opportunities in comparing various cardiovascular treatments and their outcomes.

INTRODUCTION

In phase I of this project Wisconsin compared two data-linking methodologies, deterministic and probabilistic, to determine if they yielded similar results. We used both methods to de-duplicate hospital data from 2006 and then linked the hospital data to 2006-2008 death data due to cardiovascular diseases (CVD.)

Both deterministic and probabilistic record linkage methods created reasonable sets of matched pairs of mortality and patient records and those sets had a high degree of common pairs. However, it was concluded that the deterministic process is likely to be more accessible, affordable and efficient. The probabilistic method requires the purchase of additional software and specific expertise among staff. The logic of identifying record matches with unique combinations of plausibly-identifying data elements is straightforward and readily programmed. However, the quality of the results is heavily dependent on the completeness and accuracy of the recorded data.

METHODS

Following completion of phase I, we began the two components of phase II, which is the focus of this report. First, we extended the time period of analysis from three to five years for both inpatient discharge and death records and focused on the years 2006-2010. We used revised deterministic linkage techniques to de-duplicate those five years of statewide inpatient discharge records and then linked the resulting patients to mortality records from the same period. This methodology created valid and more complete matches between mortality and hospital patient records. If patients were admitted more than once for a CVD condition during 2006-2010, we only used the first admission of three major conditions, MI, CHF, and stroke to link to death records. The final tally from these enhanced record linkage algorithms was data for163,673 persons. This accounts for 69% of all Wisconsin deaths during 2006-2010.

Second, we selected cases of three major CVD events for survival-rate analysis -- congestive heart failure, myocardial infarction, and stroke. Survival analysis was performed on each case type separately using SAS 9.2 PROC LIFETEST and PROC PHREG. (For more details on survival-rate analysis, please see Attachment 2.) Within a given case type, the response variable was the survival time of a patient until event (death) or right-censored until the end of the observation period. We first estimated the distribution of survival times and obtained an overall estimated survival function for each case type. Within each case type, we computed a set of estimated survival functions for different age groups and genders. Additionally, we estimated survival functions for each of the 72 counties in Wisconsin (Attachment 3.) Maps were generated using 75 percentile of patients to generate maps by counties which means that there were 75% of patients after being discharged from hospitals were still alive at particular day for each county. Using regression analysis, we investigated the relationship between survival times and potential covariates, namely, age groups (35-64, 65-74, 75-84, or 85+) and gender (male or female). In particular, a Cox proportional hazards model was fitted to examine the effect of age, gender, and age-gender interaction on hazard rates. We used partial likelihood to estimate regression coefficients and a Wald test to determine the significance of the effects.

RESULTS/FINDINGS

- Analysis of all patients with CHF showed that half of the patients with CHF survived longer than 1,363 days (about 3 years and 9 months).

Figure 1: Congestive Heart Failure Survival Rates, Wisconsin Adults Ages 35-64

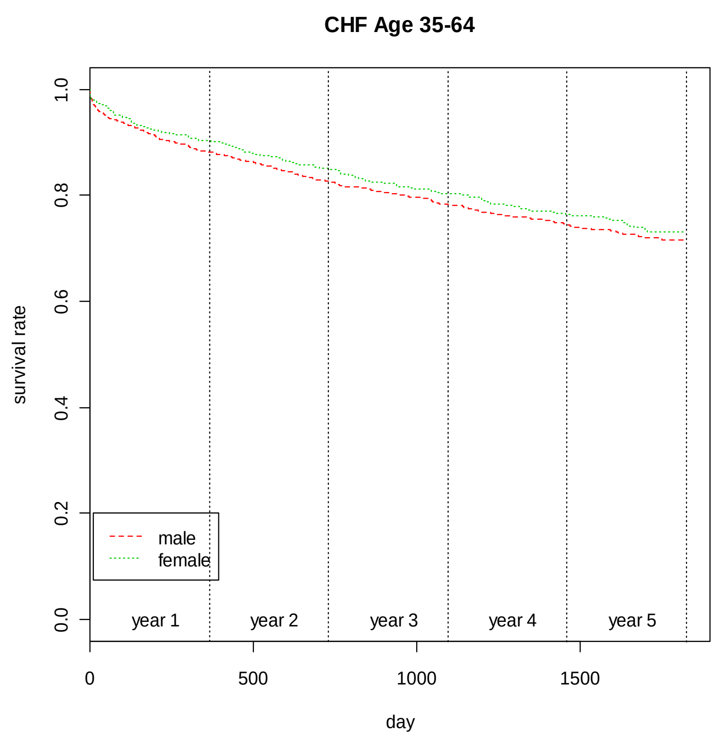
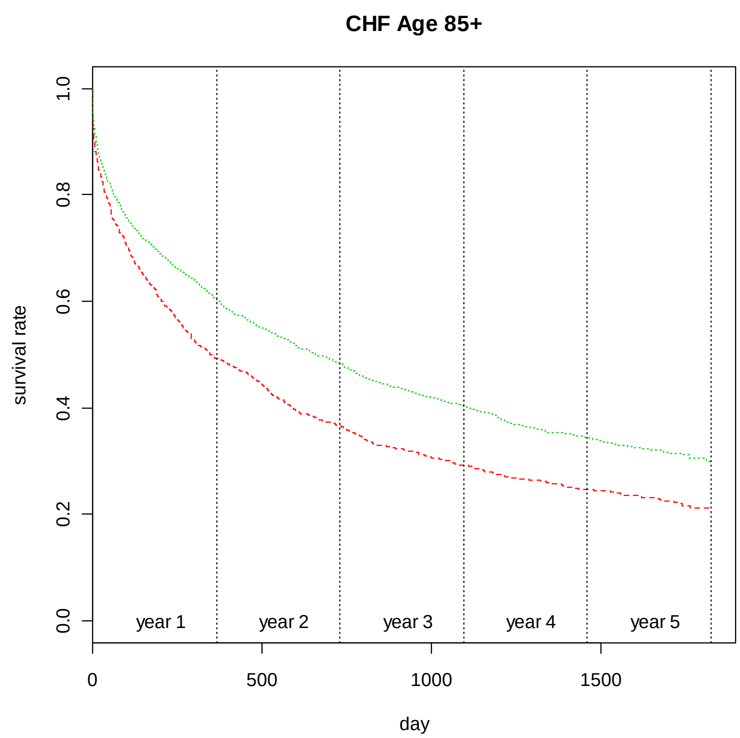


Figure 2: Congestive Heart Failure Survival Rates, Wisconsin Adults ≥Age 85



- For all patients with CHF, females had significantly longer median survival days than males. Especially, at the age group of 85 and older (662 days for females and 349 days for males, p-value<0.0001.)
- For all patients with CHF the age group of 35-64 had significantly median longer survival days than age group of ≥85 years. (1825 vs. 550 days, p-value< 0.0001)

Figure 3: Myocardial Infarction Survival Rates, Wisconsin Adults Ages 35-64

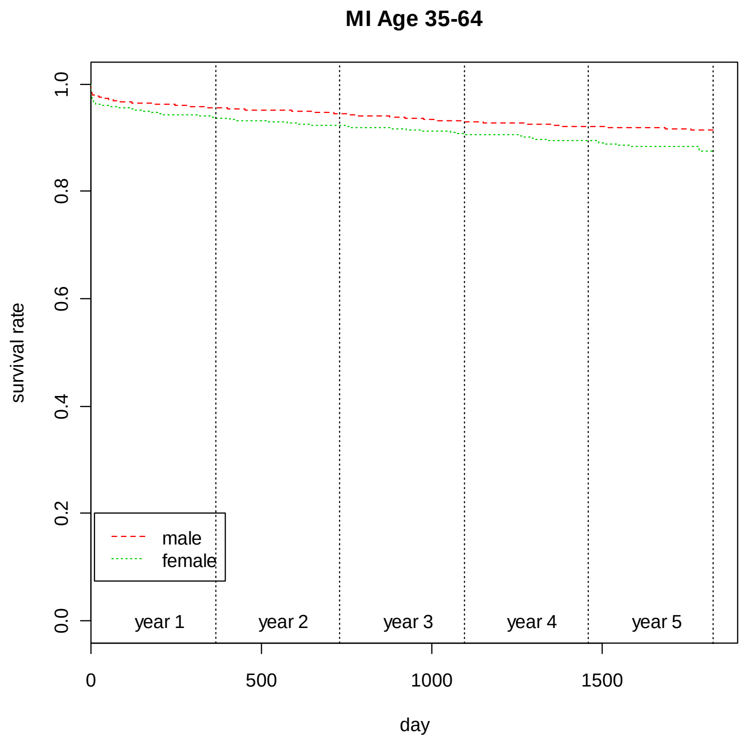
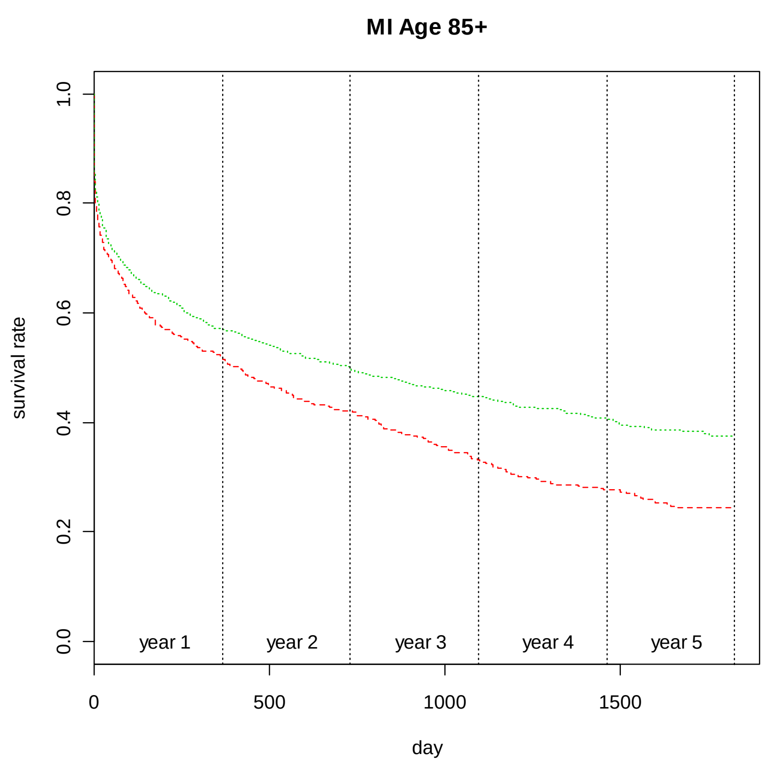


Figure 4: Myocardial Infarction Survival Rates, Wisconsin Adults ≥Age85



- More than half of patients with MI survived longer than 1,825 days (5 years, study limit.)
- For all patients with MI, females had significantly median longer survival days than males. Especially, at the age group of 85 and older (729 days for females and 411days for males, p-value<0.0001.)
- For all patients with MI, the age group of 35-64 had significantly longer survival days than age group of ≥85 years. (1825 days vs. 611 days, p-value<0.0001)

Figure 5: Stroke Survival Rates, Wisconsin Adults Ages 35-64

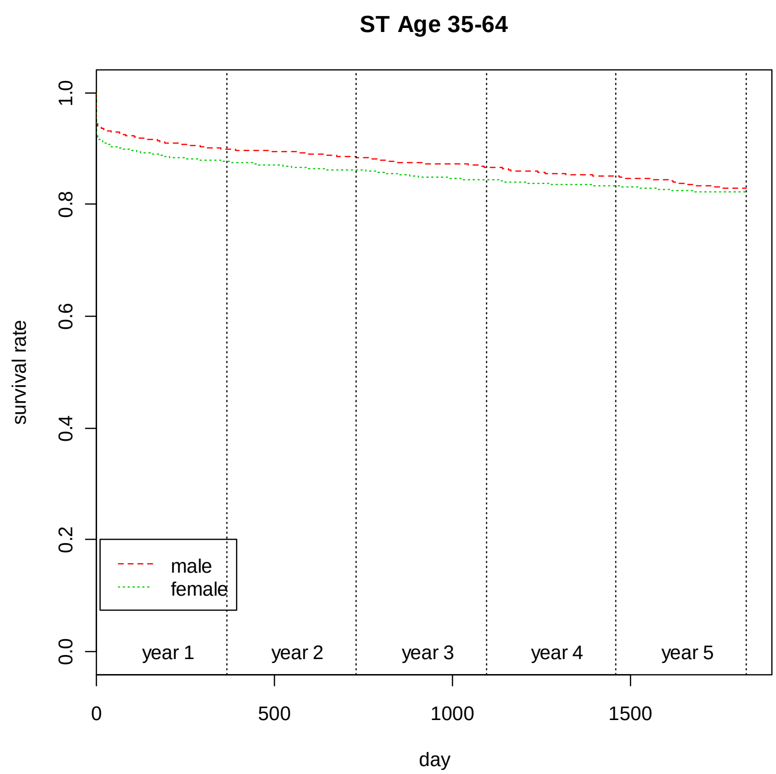
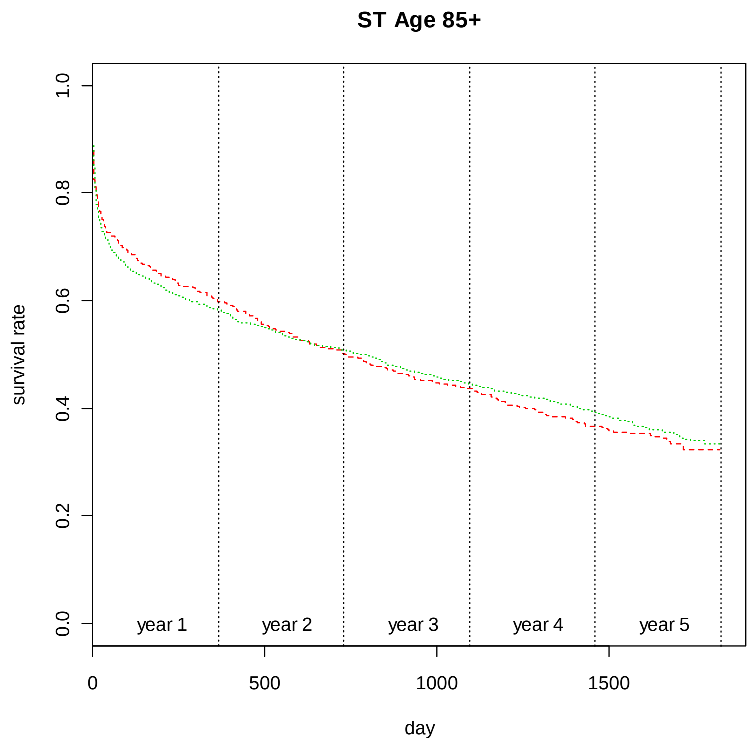
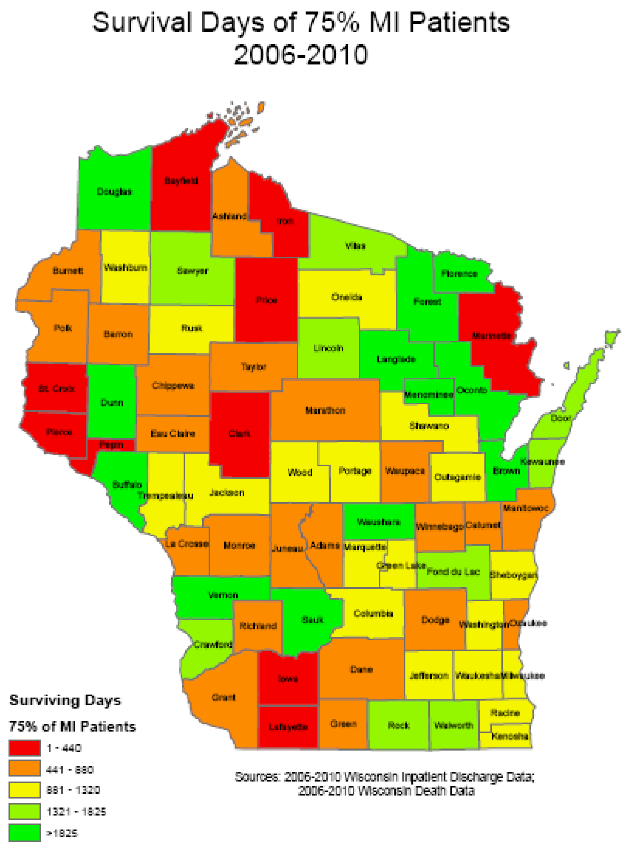


Figure 6: Stroke Survival Rates, Wisconsin Adults Ages ≥ 85



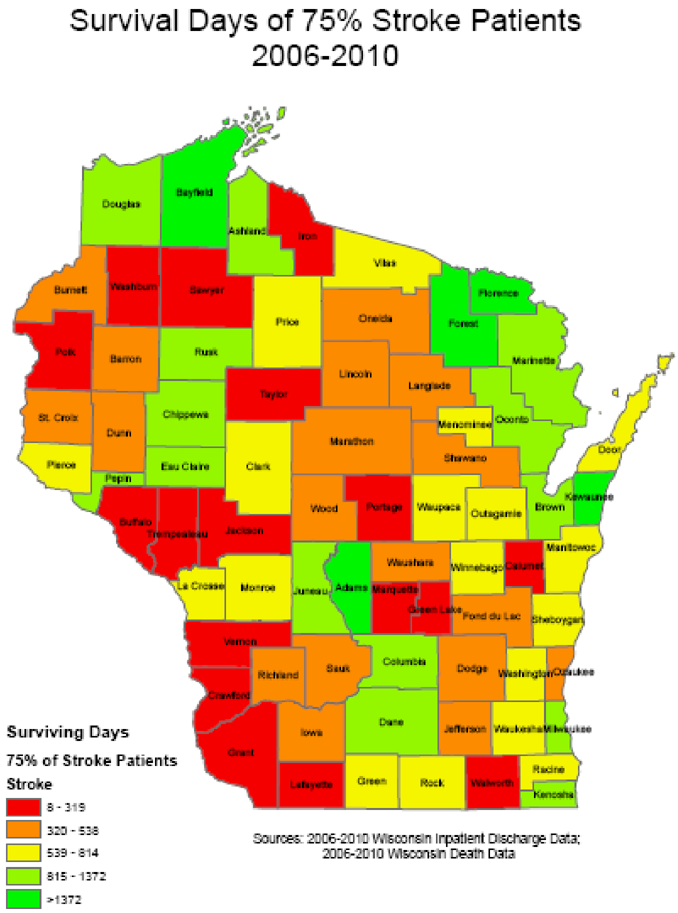
- More than half of patients with stroke survive longer than 1,825 days (5 years, study limit.)
- For all patients with stroke, younger age group of 35-64 had significant longer median surviving days than older age group ≥ 85. (1825 days vs. 763 days, p-value<0.0001)
- For all patients with stroke, there was no significant difference between males and females overall and at any age groups.

Figure 7: MI Survival Rates by Counties



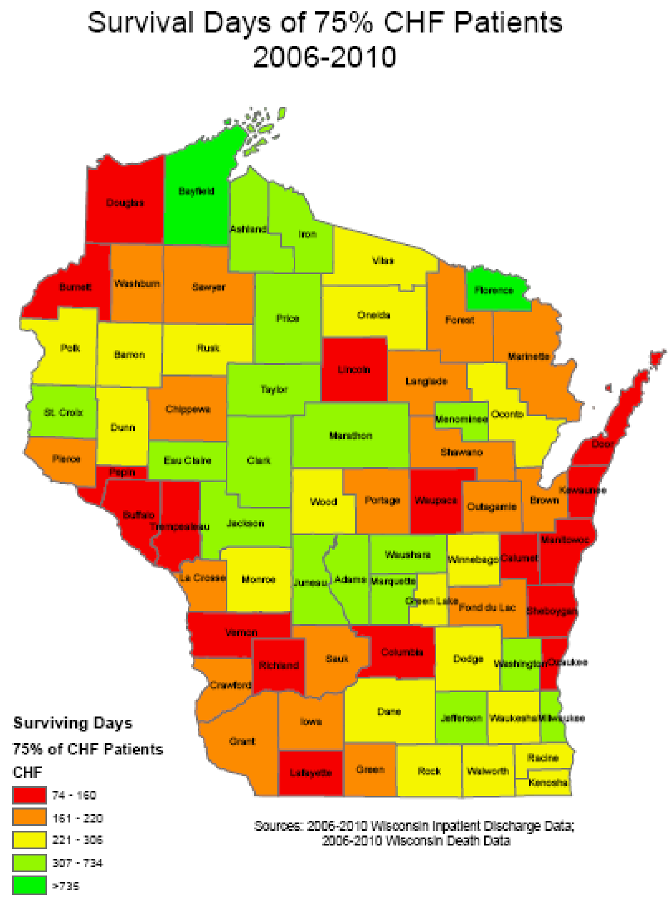
- For MI, patients in the northwestern part of Wisconsin had shorter surviving days..

Figure 8: Stroke Survival Rates by Counties



- For stroke, patients in the western part of Wisconsin had shorter surviving days.

Figure 9: CHF Survival Rates by Counties



- For CHF, patients in the northeastern and southwestern parts of Wisconsin had shorter surviving days.

LIMITATIONS

- The maximum period of analysis for this project was 1,825 days (5 years); therefore, survival times greater than 1,825 days were not available for analysis.
- Only the first admission of three conditions, MI, CHF, and stroke in 2006-2010 hospitalization records was selected to use in the survival analysis. Patients only had any minor CVD events in study period or patients had those conditions before 2006 were not available for analysis. Therefore, the survival day could be longer than 1,825 days if more years of data could be linked or all CVD conditions will be included for analysis.
- Patients could have had any of the three major conditions (MI, CHF, and stroke) prior to 2006 and not be included in this analysis.
- Patients who were not hospitalized primarily due to CVD (primary diagnosis is not CVD related) but died of CVD later were not included in this analysis, i.e. hospitalized diabetic patients (primary diagnosis was diabetes) after being discharged died of CVD were not included in this analysis.
- Surviving days are not age-adjusted by county populations therefore, may not reflect the real demographics in each County.

AREAS FOR FUTURE STUDY/ANALYSIS

- Using similar methods, an analysis could be done to determine if particular procedures and or treatments received in a hospital will affect the length of survival for different conditions. This can be used for quality improvement (QI) purpose for acute care of cardiovascular disease.
- A further analysis could be performed for critical conditions other than MI, CHF, and stroke to examine surviving days.
- Survival could also be explored by insurance status and payer type to determine if different payers will affect survival days for each selected conditions or not.
- There was no particular pattern for geographic distribution of all three conditions therefore, further geographic factors such as locations of primary stroke centers,

critical access hospitals, trauma hospitals etc., could also be examined to determine if they had any effect on length of survival.

- Overall, the results promise new data resources for the surveillance and analysis of CVD and other chronic diseases. For example, survival rates for patients with different demographics, co-morbidities or procedures could be compared and evaluated.
- We can age-adjust survival rates for each county to re-examine geographical distribution to see any significant change or not.
- Linking hospital data to emergency medical services (EMS) data can be used to analyze pre-hospital situations such as symptom recognition, CPR, driving time to hospital etc. will affect surviving outcome or not.

ACKNOWLEDGEMENT

Dr. Mark Wegner: Chronic Disease Medical Director/DPH/WI DHS

Brian Weaver: Section Chief/DPH/WI DHS

Julie Bauman: HDSP Program Director/DPH/WI DHS

Richard Miller: Research Scientist/DPH/WI DHS

Dr. Jun Zhu: Professor of Statistics/UW-Madison

Contact Information:

Herng-Leh (Mike) Yuan

Herngleh.yuan@wisconsin.gov

608-267-2487

The Wisconsin CVD Surveillance Data Pilot Project

Phase II
Refinements to Methods for Matching Mortality and Inpatient Records Over Five Years

Richard Miller
Office of Health Informatics
Wisconsin Department of Health Services, Division of Public Health

The Heart Disease and Stroke Prevention Program
Wisconsin Department of Health Services, Division of Public Health

1. Project Summary

This project builds on an earlier study that developed new data resources for the surveillance of cardiovascular disease (CVD) and analyses of treatment outcomes. That project explored methodological alternatives for linking hospital inpatient records for the same patient and then linking those patients to mortality records. This extension applied lessons from that study to improve the record linkage processes and results.

The initial study revealed limitations to using ZIP code and Social Security number (SSN) as parts of combinations of data elements for deterministic record linkage of patients' discharge records to death records. This project explored how the regional ZIP code (the first three digits of a ZIP code or "ZIP3") and the last four digits of the SSN ("SSN4") might be used in linking combinations, instead of the complete fields.

As before, I worked with all patients and all deaths. I extended the time period from three to five years of inpatient discharge and mortality records: 2006-10. I revised the deterministic linkage techniques to de-duplicate those five years of statewide inpatient discharge records and then to link the resulting patients to mortality records from the same period. Valid and more complete matches between mortality and hospital patient records were created by these enhancements: 69 percent of all 2006-10 deaths were matched to a patient record.

The results promise new data resources for the surveillance and analysis of CVD and other chronic diseases. For example, survival rates for patients with different demographics, co-morbidities or procedures can be compared and evaluated.

2. Linkage Methodologies: Lessons from Phase I

If information uniquely identifying the same individual in two datasets were perfectly captured and recorded, then a "**deterministic** record linkage methodology" would be reliable. Pairs of records could be compared for exactly matching indentifying information and exact matches would *determine* true record matches.

In real world data systems, of course, uniquely identifying data elements are not often available. Further, recorded data on names, dates, addresses and other identifying information are subject to many influences that may cause differences between records and missing values in records. A "**probabilistic** record linkage methodology" recognizes that a pair of records with some matching elements – or even nearly-matching elements - has a *probability* of being a "true match." The methodology uses statistical tools to estimate and evaluate that probability; pairs with a high probability (90%+) of being true matches are identified for research use.

<u>Deterministic record linkage</u>. How well can "a patient" be defined or "determined" as a unique value of available combined data elements? Before inpatient records can be linked to mortality reports, their patients must be de-duplicated so that one record per

patient is compared to death reports. The Phase I project worked with all inpatient discharge and mortality records for the three-year period 2006-08.

I developed that inpatient record linkage with alternative deterministic algorithms and evaluated the results. The experiments and evaluations demonstrated that, using deterministic record linkages, several alternative combinations of available patient-identifying data elements are reasonable functional definitions of "patients" in the 2006-08 Wisconsin hospital inpatient discharge files.

One combination of elements was found to be rarely missing and had the most unique values (and thus the least false positives). This combination is the best general choice as a linking patient identifier for these Wisconsin inpatient records: Inpatient records with exactly matching values of the patient's initials, date of birth, gender and ZIP code were 'determined' to define the same patient. The evaluation noted that "The known false positives should be excluded by dropping links that have an admission before the previous discharge date or after a recorded in-hospital death."

This combination of data elements identifies 1,279,286 unique values or "patients" among the 2,017,460 inpatient discharge records from 2006-08; the number of patients is 63% of the total number of discharges. Only a few percent of inpatients are residents of other states, so I estimated that about one-in-five Wisconsin residents were hospitalized in a Wisconsin hospital at least once during this period. Of course, this includes almost all newborns and their mothers.

Next I applied the deterministic methodology to matching patients and mortality records. Two sets of iterative deterministic algorithms were developed to link (1) in-hospital death and inpatient records and (2) other deaths and inpatient records. The iterative algorithms first identify records with exact matches on the strongest combination of data elements and set them aside. The residual records then are tested for matches on the next-strongest combination, and so on. Ultimately about 66 percent of 2006-08 Wisconsin resident and occurrent mortality records were matched to a Wisconsin resident patient in a Wisconsin hospital.

<u>Probabilistic record linkage</u>. I used the *LinkSolv* product to develop probabilistic linkage models for Phase I. The advantages of probabilistic linkage are its statistical rigor and its ability to identify record matches based on less-than-perfect data quality.

The *LinkSolv* software applies and extends rigorous statistical procedures by taking into account the many real-world situations which complicate the assumptions of more elementary probabilistic algorithms, such as missing data, correlated agreements and disagreements between data fields, recording errors, and so on. *LinkSolv* identifies those record matches with a user-defined high probability of being a true match but also imputes a small number of possible matches with a lower probability due to missing and incorrect data.

Following experimentation with more elaborate models, I settled on a fairly streamlined set of models that used initials, encrypted last name, gender, last four SSN digits, date of birth, and first three ZIP digits. I simplified the task by using only patients whose last hospitalization was in 2006 and by excluding birth-related hospitalizations.

Comparative results

As expected, the probabilistic record linkage process yields more matched pairs of death/patient records than the deterministic process. However, the difference is not dramatically large. Overall, the more rigorous probabilistic method validated the results of the deterministic linkage. Comparisons of the results indicated strengths and weaknesses in the combinations of data elements used for deterministic matching.

> Initials, date of birth and sex. Patients' and deceased' data were generally reliable and consistent for these data elements.
>
> ZIP code of residence. ZIP code was less reliable as a matching element because small moves within a region will often result in different ZIP codes. Among the deceased, older patients may be particularly likely to make such residential moves and of course are a large share of the deceased. The probabilistic linkage models only used the more stable first three ZIP digits. This is the major reason that the exact-match deterministic method failed to identify some matches that the probabilistic process found.
>
> SSN. Similarly, using full social security numbers limited the success of exact matching. For patients, the SSNs were teased out of policy numbers when possible but are often missing or reflect assigning a spouse's SSN for their partner's Medicare ID. The probabilistic models used only the last four digits of the SSN, after experimentation. This reduced false negative matches due to recording errors on either record.

Analyses of discrepant results indicated ways that the more-approachable deterministic models could be improved. The report on Phase I observed that "further work should evaluate deterministic algorithms using the last four SSN digits instead of the full SSN and first three ZIP digits instead of the full ZIP." This Phase II effort explores how that could be done and describes the results.

3. Revised Deterministic Record Linkage of Inpatient Hospital Records

Wisconsin collects a record on every inpatient discharge from a Wisconsin hospital. These records provide a rich and population-based source of information on each inpatient stay. Additionally, it is very useful to link discharge records for the same patient. Many studies require a patient-based or episode-of-care level of analysis that combines information from inpatient stays by the same person. A patient-level organization of the dataset also facilitates record linkage to other data sets which are person-based, such as mortality and birth records.

Wisconsin's Hospital Inpatient Discharge Data System does not contain a single unique and definitive patient identifier in the current medical records or billing systems that is perfectly consistent for the same person over time and between hospital systems. The Phase I project developed and evaluated alternative combinations of data elements as patient identifiers to answer the question: How well can a "patient" be defined or "determined" as *a unique value of available combined data elements*?

It should be noted that '*patient identifier'* simply means some code or value that is, ideally, unique to each patient, analogous to a social security number. It does **not** mean that the actual name or address of any patient is known or can be determined. Indeed, Wisconsin statutes prohibit such personal identification.

The initial work found that the most efficient and complete linking combination among the six combinations considered is the patient's initials, date of birth, gender and ZIP code. For example, 'CK' + '-753' + '2' + '53711' is concatenated into "CK-753253711." (In the discharge records, the date of birth is expressed in the SAS format "number of days before or after January 1, 1960" and the gender is coded 1 or 2.) This combination identified the number of patients as about 57% of the total number of discharges.

When matching patient records to death records with probabilistic methods, I found that the three-digit ZIP3 was a useful alternative to the full five-digit ZIP code. The reason is that recently-deceased patients often changed their residence within a few months or years of death by moving into skilled care facilities or the homes of relatives. Such moves are likely to be relatively local, within a metropolitan area or a ZIP3 region.

Using ZIP3 in the patient-death deterministic matching requires that patients be defined using ZIP3. Some number of patients will be uniquely identified when defined with initials+DOB+gender+ZIP but will not have a unique value when ZIP3 is substituted. Therefore, using ZIP3 indiscriminately would create some false positive matches among inpatient records. I proceeded in two steps.

<u>Step 1</u>. The 3,298,116 inpatient discharges from Wisconsin hospitals with non-missing ZIP codes were linked with both the ZIP and ZIP3 combinations.

The earliest hospitalization for a patient can be identified by sorting the records by a patient identifier, admission date, discharge date and whether the discharge status was 'expired.' (The latter element was added so that the hospital stays for a patient who was admitted, discharged, transferred to another hospital or readmitted, and then died, all on the same day, would have the hospital stays correctly ordered.) I calculated the number of days between each linked hospitalization: the days between an admission and the previous discharge date.

Two significant indicators of false positive patient matches were explored in Phase I:

- Linked records with any admission date preceding the previous admission's discharge date.
- Linked records with a record indicating the patient died during the stay linked to subsequent hospitalizations with later admission dates.

In Phase II patient record matching I treated both indicators as fatal flaws for that patient's linkage. I collected all records with either condition, then identified the "patients" and all of their other records and excluded them all from the linked patient file. These situations involved relatively small numbers of patients (0.4% of records with ZIP-defined patients and 0.9% of ZIP3-defined patients' records).

Table 1 shows that the ZIP3 combination identifies fewer "patients," but not dramatically so. There are fewer because ZIP3 is less precise as a tie-breaker among persons with identical initials, date of birth and gender. ZIP3 also creates more false positive matches, which are often identified by inconsistencies among dates and dropped from the file. Over five data years, the number of discharges represents somewhat fewer unique patients than over three data years (63% unique).

Table 1.
Alternative Patient Identifiers
Wisconsin Inpatients Discharged 2006-10

"Patient" Identifier	Usable Records	Number of unique values or "patients"	Percent with unique values
Initials + DOB + sex + ZIP	3,283,873	1,910,441	58.2%
Initials + DOB + sex + ZIP3	3,265,635	1,793,283	54.9%

Source: Wisconsin Department of Health Services, Wisconsin Hospital Inpatient Data Archive.

Step 2. I examined the "patients" identified by the two combinations for all 3,283,873 records matched by the ZIP combination. The two identifiers yield identical results for 93.3% of all records, they disagree on patient assignment among 6.1% of records (200,865 discharges), and the ZIP3 combination is not useable for 0.6% of records (18,238 discharges with discrepant dates).

The "patients" associated with all records with different ZIP and ZIP3 patient linkages were identified. They involved 91,019 ZIP3-defined "patients" and 194,711 ZIP-defined "patients" and a total of 405,513 discharge records. I then explored the insurance policy number field as a way to resolve these discrepancies in patient assignment. Examination of a sample of these patients' records showed that many ZIP3-defined patients had consistent policy numbers, evidence of residential moves by the same patient between ZIP codes. Therefore, linking records within this small group using the policy number in combination with initials, date of birth and gender could reasonably confirm ZIP3-defined patients.

This linking combination with policy number confirmed 36,428 ZIP3-defined patients with 154,283 records, about 40% of all patients with discrepant links. These patients may be added to the large set of ZIP3-defined patients already confirmed. The combined set has 1,738,692 "patients" with 3,014,405 discharge records. Matching these confirmed ZIP3-defined patients to mortality records may now take advantage of ZIP3 as a more effective matching element than ZIP.

The residual records either were not confirmed as ZIP3-defined patients or had no ZIP3-defined patient links at all. These 269,468 records total 132,201 ZIP-defined patients. These patients may be linked to mortality records with the same algorithms used in the Phase I project.

The total of 1,870,893 patients is 43,135 fewer than is defined by using only ZIP in the linking combination. This reduction of 2% in the number of defined patients represents the consolidation of false negative matches.

4. Linking Patients to In-Hospital Deaths

As in Phase I, my strategy was to apply an iterative matching procedure in two steps. The first was to match records that had the most data elements in common: the records for deaths that occurred in a Wisconsin hospital. The second step examined all of the remaining records for matches.

Since the patients are defined two ways, it is necessary to follow this strategy for each set of patients separately. I began with the much larger ZIP3-defined set of patients.

A. Linking ZIP3-defined patients

In-hospital deaths are indicated by a specific discharge status code in the discharge records and by a "patient type" code in the death records. If the coding conventions capture the same situation in both data sources, then we should efficiently find matched pairs of records.

The first step is to locate each ZIP3-defined patient's most recent discharge among the 3,014,405 inpatient records, which produces one record for all ZIP3-defined 1,738,654 patients. Then those records that indicate the patient expired are identified; there are 51,054 such patients who were also Wisconsin residents. (Residents are much more likely to have had all of their hospital experiences in a Wisconsin hospital.)

Similarly, among the 237,176 deaths registered for 2006-10, I selected those occurring in Wisconsin to Wisconsin residents in a hospital setting. I also excluded those occurring in one of the federal Veterans Administration hospitals, since the VA hospitals are not part of the inpatient discharge data system. There were 56,962 deaths meeting these screens.

As shown in Table 2, I applied the same five successive linking combinations of data elements used in Phase I, except that:

1. ZIP3 could be used instead of ZIP in some combinations. This allows more pairs of records to be potential matches.
2. The last four digits of SSN (SSN4) were used, instead of SSN. This avoids the false negative matches caused by not-infrequent transcription errors in the full SSN (death records).

This procedure matched 95% of the in-hospital deaths among the ZIP3-defined patients to 85% of all in-hospital deaths.

Table 2.
Linked In-Hospital Deaths and ZIP3-defined Patients
Wisconsin Inpatient and Mortality Records, 2006-10

		Matched Records		**Remaining Unmatched Records**	
Linker	**# Pairs Matched**	**% of inpatient records**	**% of mortality records**	**# of inpatient records**	**# of mortality records**
				51,054	**56,962**
Initials + DOB + Sex + ZIP3	45,375	88.9%	79.6%	5,679	11,587
Initials + DOB + Sex + SSN4	1,103	2.2	1.9	4,576	10,454
Initials + Sex + ZIP3 + DOD	652	1.3	1.1	3,924	9,832
Hospital + DOD + DOB	1,333	2.6	2.3	2,589	8,499
Initials + Sex + DOB	25	<0.1	<0.1	**2,564**	**8,474**
All Linked Pairs	**48,488**	**95.0%**	**85.1%**		

<u>B. Linking ZIP-defined patients</u>

Among the 132,201 ZIP-defined patients, there were 2,212 in-hospital deaths. These were compared to the residual 8,474 in-hospital death records from the previous process. I used the same five successive linking combinations of data elements used in Phase I.

As shown in Table 3, this procedure matched 89% of the in-hospital deaths among the relatively small number of ZIP-defined patients to 23% of the remaining in-hospital deaths.

Table 3.
Linked In-Hospital Deaths and ZIP-defined Patients
Wisconsin Inpatient and Mortality Records, 2006-10

Linker	# Pairs Matched	Matched Records: % of inpatient records	Matched Records: % of mortality records	Remaining Unmatched Records: # of inpatient records	Remaining Unmatched Records: # of mortality records
				2,212	**8,474**
Initials + DOB + Sex + ZIP	1,234	55.8%	14.6%	978	7,240
Initials + DOB + Sex + SSN4	300	13.6	3.5	678	6,940
Initials + Sex + ZIP + DOD	0	0	0	678	6,940
Hospital + DOD + DOB	423	19.1	5.0	255	6,517
Initials + Sex + DOB	17	0.8	0.2	**238**	**6,500**
All Linked Pairs	**1,974**	**89.2%**	**23.3%**		

5. Linking Patients to Other Deaths

A. Linking ZIP3-defined patients

Now that we've done what we can to match patients known to have expired in-hospital to their death records, the next step is to look for matches between the remaining ZIP3-defined patients and all remaining unmatched death records.

The Phase I work showed that only two linking combinations were effective:

- Initials + data of birth + gender + ZIP, and
- Initials + data of birth + gender + SSN.

Now we can really take advantage of the lessons from the probabilistic linkage in Phase I. This demonstrated that ZIP3 and SSN4 might be important enhancements to these combinations of data elements. Thus this step uses:

- Initials + date of birth + gender + ZIP3, and
- Initials + date of birth + gender + SSN4.

As Table 4 shows, 61% of the remaining mortality records are matched to a patient by the combination that includes ZIP3 and another 2% are matched with the SSN4 combination.

Table 4.
Linked Residual Deaths to Residual ZIP3-defined Inpatients
Wisconsin Inpatient and Mortality Records, 2006-10

Linker	**# Pairs Matched**	**Matched Records**		**Remaining Unmatched Records**	
		% of inpatient records	**% of mortality records**	**# of inpatient records**	**# of mortality records**
				1,608,660	**174,919**
Initials + DOB + Sex + ZIP3	106,926	6.6%	61.1%	1,501,734	67,993
Initials + DOB + Sex + SSN4	3,004	0.2	1.7	**1,498,730**	**64,989**
All Residual Linked Pairs	**109,930**	**6.8%**	**62.8%**		

B. Linking ZIP-defined patients

The same two Phase I combinations are examined to match the final 64,989 residual mortality records to any of the 129,476 ZIP-defined patients not matched in the in-hospital deaths process.

Table 5 shows the modest results among these remaining sets of records. A final 3,651 pairs of matched patient-death records are identified.

Table 5.
Linked Residual Deaths and Residual ZIP-defined Inpatients
Wisconsin Inpatient and Mortality Records, 2006-10

Linker	**# Pairs Matched**	**Matched Records**		**Remaining Unmatched Records**	
		% of inpatient records	**% of mortality records**	**# of inpatient records**	**# of mortality records**
				129,476	**64,989**
Initials + DOB + Sex + ZIP3	3,431	2.6%	5.3%	126,045	61,558
Initials + DOB + Sex + SSN4	220	0.2	0.3	**125,845**	**61,338**
All Residual Linked Pairs	**3,651**	**0.3%**	**5.6%**		

6. All Linked Pairs of Patient – Mortality Records

When the four sets of matched record pairs are combined, they total 164,043 record pairs. A final edit compared the last inpatient discharge date to the date of death in each pair. A small number of record pairs (319) had death dates that preceded discharge dates by more than two days. This is strong evidence of a false positive match and these records were dropped.

The final tally from these enhanced deterministic record linkage algorithms was the linkage of 163,724 patients to mortality records over the five-year period 2006-10. These matched records account for 69% of all Wisconsin deaths and 9% of all hospital inpatients over this period.

This result is a marginal improvement over a deterministic record-matching procedure that uses only ZIP-defined patients. The combined ZIP3 + ZIP approach adds 5,163 matches, a 3.2% increase.

7. Case-Selection for Analyses Using Linked Patient-Death Records

Selecting linked patient and death records for analyses must consider the lapse of time on matching rates. Patients whose hospital episodes occurred earlier in the period have more observed time to expire and thus have higher rates of matches. Similarly, deaths occurring later in the period have more identified patients to be matched against.

The percent of records matched over calendar quarters is shown in Table 6. Fully 12.3% of patients whose last hospitalization was in the first quarter of 2006 were matched to a death occurring between then and the end of 2010. Even in the last quarter of 2010, 4% of patients expired in the same quarter and were matched to a mortality record. Conversely, 44.6% of deaths in the first quarter of 2006 were matched to a patient whose last hospitalization was in that same quarter, but 76.7% of last quarter 2010 deaths were matched to patients whose last inpatient stay was sometime in the previous five years.

As expected, patients whose most recent discharge was in 2006 are the most likely to have a matched death record. The 2010 quarters show the significant short-term relation of hospitalization and the probability of death.

The pattern of the percents of deaths matched to a patient in Table 6 indicates the necessary look-back period from date of death to find deceased's hospitalizations. About three-fourths of the 2009-10 deaths were matched to a patient, that rate is fairly stable over those eight quarters, and the rate for 2008 deaths is slightly lower. That suggests that researchers must look to a period of at least three years prior to death to find last hospitalizations. Conversely, researcher should look out at least three years after a hospitalization to relate it to the probability and characteristics of subsequent mortality.

These observations are based on all patients and deaths. The situation for more specific demographic, disease or surgical conditions will vary, probably significantly.

Table 6.
Percents of Records Matched, by Quarter
Wisconsin Inpatient and Mortality Records, 2006-10

Year	Quarter	Percent of Patients Matched to a Death Record	Percent of Deaths Matched to a Patient Record
2006	1	12.3%	44.6%
	2	12.0	56.6
	3	11.0	61.3
	4	11.7	62.2
2007	1	11.5	64.3
	2	10.6	66.9
	3	9.8	67.3
	4	10.3	68.0
2008	1	10.7	69.9
	2	9.5	72.3
	3	8.6	72.1
	4	9.1	71.8
2009	1	9.0	74.0
	2	8.0	74.7
	3	7.4	74.2
	4	7.6	75.4
2010	1	7.1	76.0
	2	6.2	76.6
	3	5.3	76.1
	4	4.0	76.7
Total		**9%**	**69%**

Note: The patients' quarter is that of their last, most recent inpatient discharge.

Survival analyses should sample from patients whose last hospitalization was earlier in the five-year period, so that the probability of a mortality event being observed is greater.

8. Summary

This second phase project elaborated deterministic record-matching procedures for identifying hospital patients in discharge records and linking patients to mortality records over a five-year period. The lessons of the first phase, which compared deterministic to probabilistic techniques, were applied to improve the deterministic algorithms. Those lessons involved substituting ZIP3 for ZIP and SSN4 for SSN where possible in key deterministic combinations of data elements.

This project demonstrated how most Wisconsin hospital patients can reliably be identified using their ZIP3 residence, rather than their full ZIP code. The project then showed how to use ZIP3 and SSN4 in enhanced deterministic linkage procedures to identify patient matches to mortality records over a five-year period.

Statistical Methods

Three case types, chronic heart failure, myocardial infarction, and stroke, are studied. Survival analysis is performed on each case type separately. Within a given case type, the response variable is the survival time of a patient until event (death) or right-censored until the end of the observation period. First, the distribution of the survival times is estimated. An overall estimated survival function is obtained for each case type. Within each case type, a set of estimated survival functions is computed for different age classes and gender. Furthermore, such estimated survival functions are computed for each of the 72 counties of Wisconsin. Second, regression analysis is carried out to investigate relationship between survival times and potential covariates. Two coviarates, age classes (35-64, 65-74, 75-84, or 85+) and gender (male or female), are considered. In particular, a Cox proportional hazards model is fitted to examine the effect of age, gendear, and age-gender interaction on hazard rates. Partial likelihood is used for estimation of the regression coefficients and a Wald test is implemented to determine the significance of the effects. The survival analysis is carried out using SAS 9.2 PROC LIFETEST and PROC PHREG.

Results and Interpretation

Figure 1 gives the overall estimated survival function for chronic heart failure and Figure 2 the estimated survival functions for 8 combinations of age class and gender. Similarly, Figures 3-4 are estimated survival functions for myocardial infarction, whereas Figures 5-6 for ST. Table 1 summarizes the number of events and total number of patients overall for each case type and by age class and gender. Table 2 gives the test results of individual effects overall for each case type and by age class.

For chronic heart failure E, there is weak evidence of an interaction between age class and gender (p=0.09). However, both age class and gender are highly significant (p<0.0001). Overall, as the age of a patient increases, the expected survival time decreases. As for gender, female patients survive longer than male patients. The gender effect, however, depends on the age class. Gender is not significant for patients in the 35-64 age class and the evidence is moderate in the 65-74 age class, whereas the gender effect is highly significant for patients older than 75.

For myocardial infarction, there is strong evidence of both age and gender main effects, as well as an interaction between age class and gender (p<0.0001). The gender effect is not significant for patients in the age range of 65 to 84, but is highly significant in the age group 35-64 or older than 85.

For stroke, there is strong evidence of an age effect, but there is no effect of gender nor age-gender interaction. Overall, as the age of a patient increases, the expected survival time decreases.

CHRONIC HEART FAILURE

Age	Cases (#)			Deaths (#)		
Age Class	**Cases**	**Gender**	**Cases**	**Deaths**	**Gender**	**Deaths**
35-64	2,131	M	1301	565	M	353
		F	830		F	212
65-74	2,057	M	1142	959	M	557
		F	915		F	402
75-84	3,720	M	1829	2,144	M	1135
		F	1891		F	1009
85+	3,163	M	1128	2,230	M	860
		F	2035		F	1370
Total Age 35+	11,071			5,898		
All Age	11,159			6,064		

MYOCARDIAL INFARCTION

Age	Cases (#)			Deaths (#)		
Age Class	**Cases**	**Gender**	**Cases**	**Deaths**	**Gender**	**Deaths**
35-64	3,443	M	2,565	313	M	212
		F	878		F	101
65-74	1,671	M	1,005	418	M	253
		F	666		F	165
75-84	2,118	M	1,057	894	M	459
		F	1,061		F	435
85+	1,578	M	585	1,034	M	433
		F	993		F	601
Total Age 35+	8,810			2,659		
All Age	8,878			2,724		

STROKE

Age	Cases (#)			Deaths (#)		
Age Class	**Cases**	**Gender**	**Cases**	**Deaths**	**Gender**	**Deaths**
35-64	2,789	M	1,542	464	M	248
		F	1,247		F	216
65-74	2,346	M	1,304	603	M	346
		F	1,042		F	257
75-84	3,340	M	1,605	1,359	M	690
		F	1,735		F	669
85+	1,963	M	636	1,256	M	413
		F	1,327		F	843
Total Age 35+	10,438			3,682		
All Age	10,607			3,807		

Table 1: A summary of the number of cases and deaths, by age class and gender.

CHRONIC HEART FAILURE

Factors	**Test Statistic**	**DF**	**P-value**
Age	442.24	3	<0.0001
Gender	43.66	1	<0.0001
Age*Gender	6.49	3	0.0901
Gender effect by Age Class			
35-64	0.96	1	0.3269
65-74	4.47	1	0.0344
75-84	27.81	1	<0.0001
85+	41.33	1	<0.0001

MYOCARDIAL INFARCTION

Factors	**Test Statistic**	**DF**	**P-value**
Age	484.16	3	<0.0001
Gender	21.83	1	<0.0001
Age*Gender	24.07	3	<0.0001
Gender effect by Age Class			
35-64	8.20	1	0.0042
65-74	0.04	1	0.8356
75-84	1.15	1	0.2834
85+	21.58	1	<0.0001

SROKE

Factors	**Test Statistic**	**DF**	**P-value**
Age	651.13	3	<0.0001
Gender	0.18	1	0.6703
Age*Gender	3.83	3	0.2796
Gender effect by Age Class§			
35-64	0.91	1	0.3408
65-74	0.93	1	0.3357
75-84	4.10	1	0.0429
85+	0.18	1	0.6724

Table 2: Test statistics, degrees of freedom (DF), and p-values for the effect of age, gender, age*gender interaction, and gender effect within each age class. § further tests are not needed, as the overall age*gender interaction is not significant.

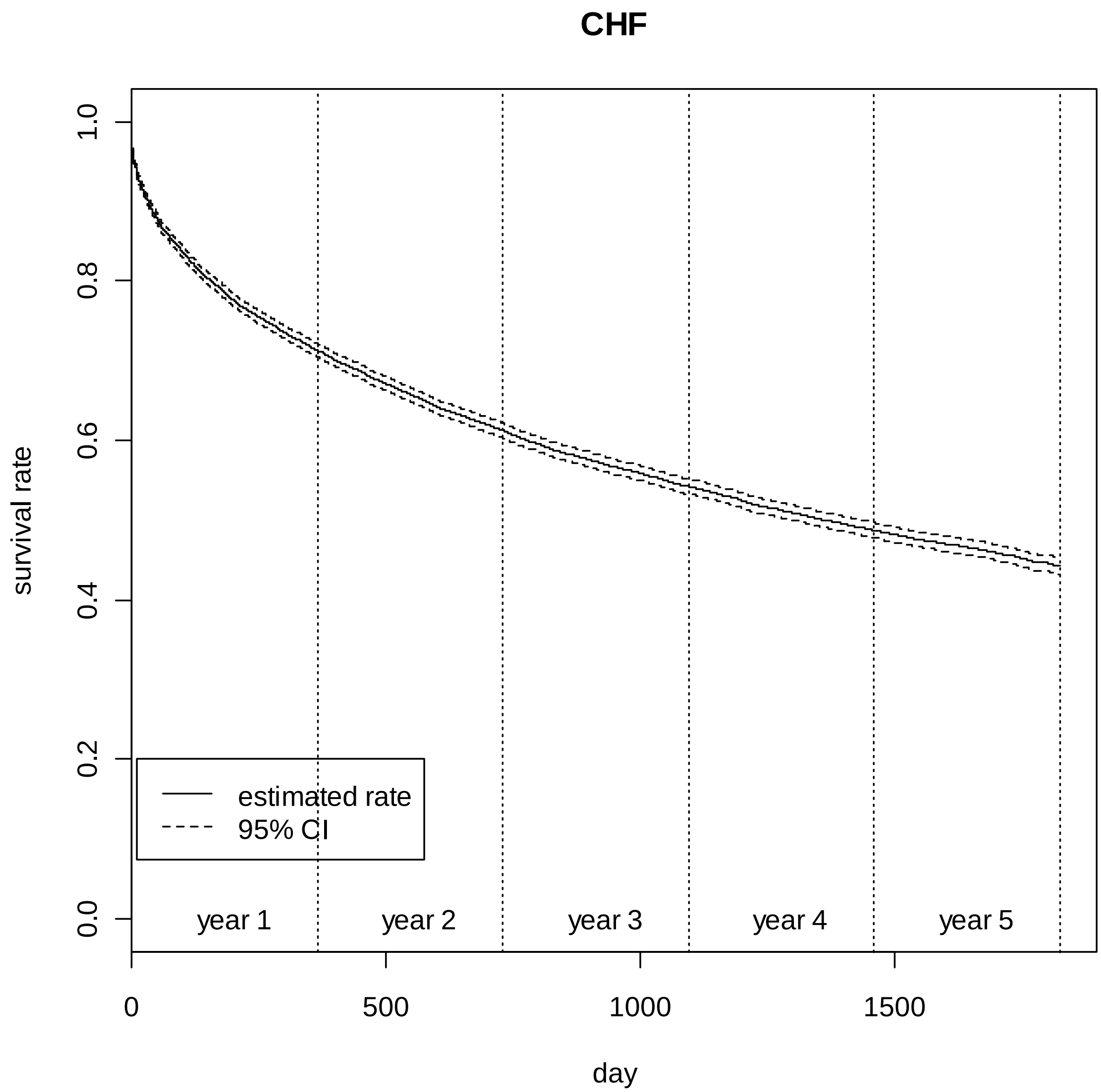


Figure 1: Estimated survival function of patients with chronic heart failure (CHF) and 95% confidence intervals (CI). Median survival time is 1,363 days and 75% survival time is 262 days.

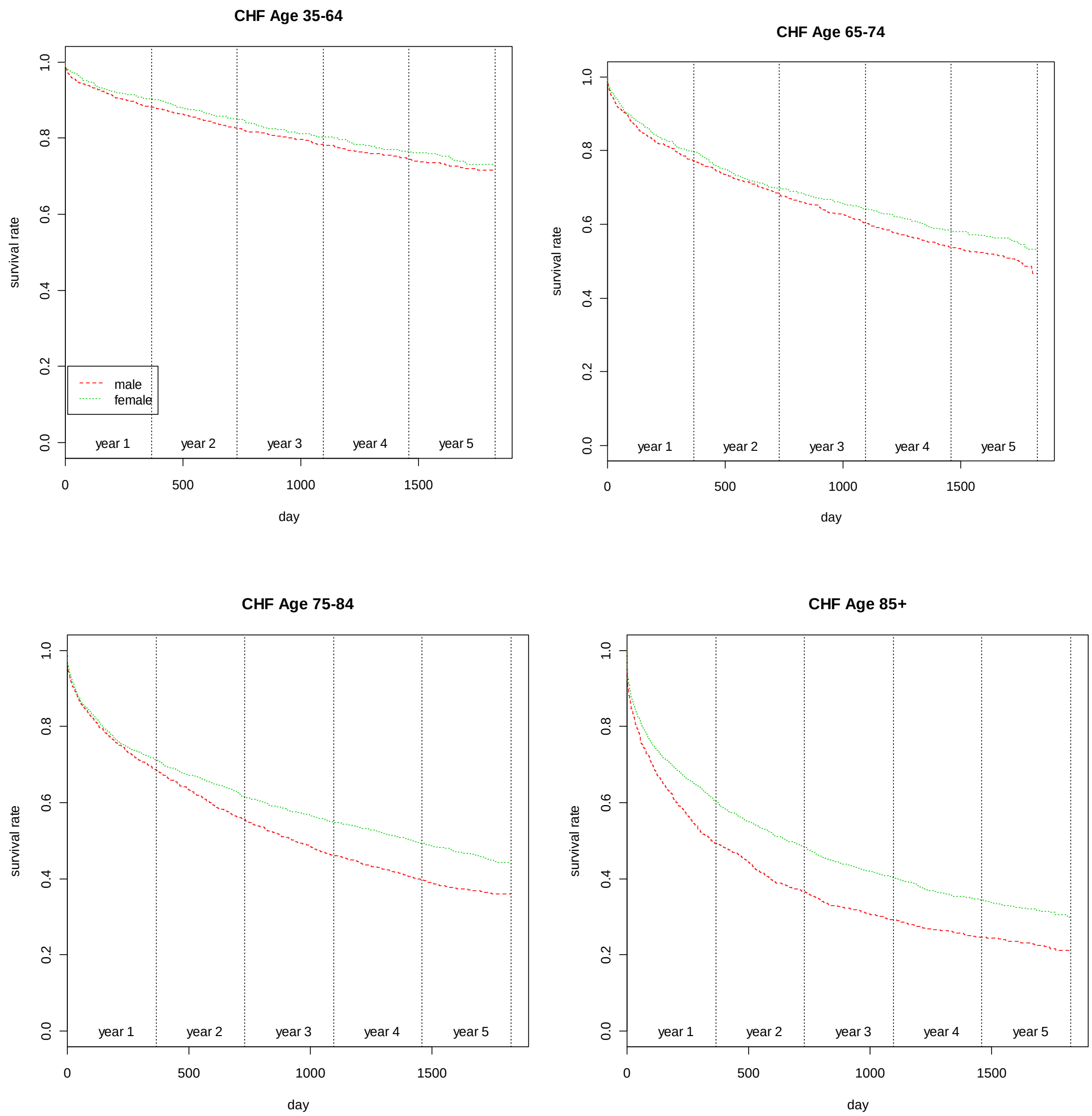


Figure 2: Estimated survival function of patients with chronic heart failure (CHF) by age class and gender. Median survival time is 1747 days for 65-74 males, 932 days for 75-84 males, 1422 days for 75-84 females, 349 days for 85+ male, 662 days for 85+ female; otherwise more than 1,825 days. 75% survival time is 1418 days for 35-64 males, 1629 days for 35-64 males, 451 days for 65-74 males, 483 days for 65-74 females, 219 days for 75-84 males, 233 days for 75-84 females, 64 days for 85+ male, 108 days for 85+ female.

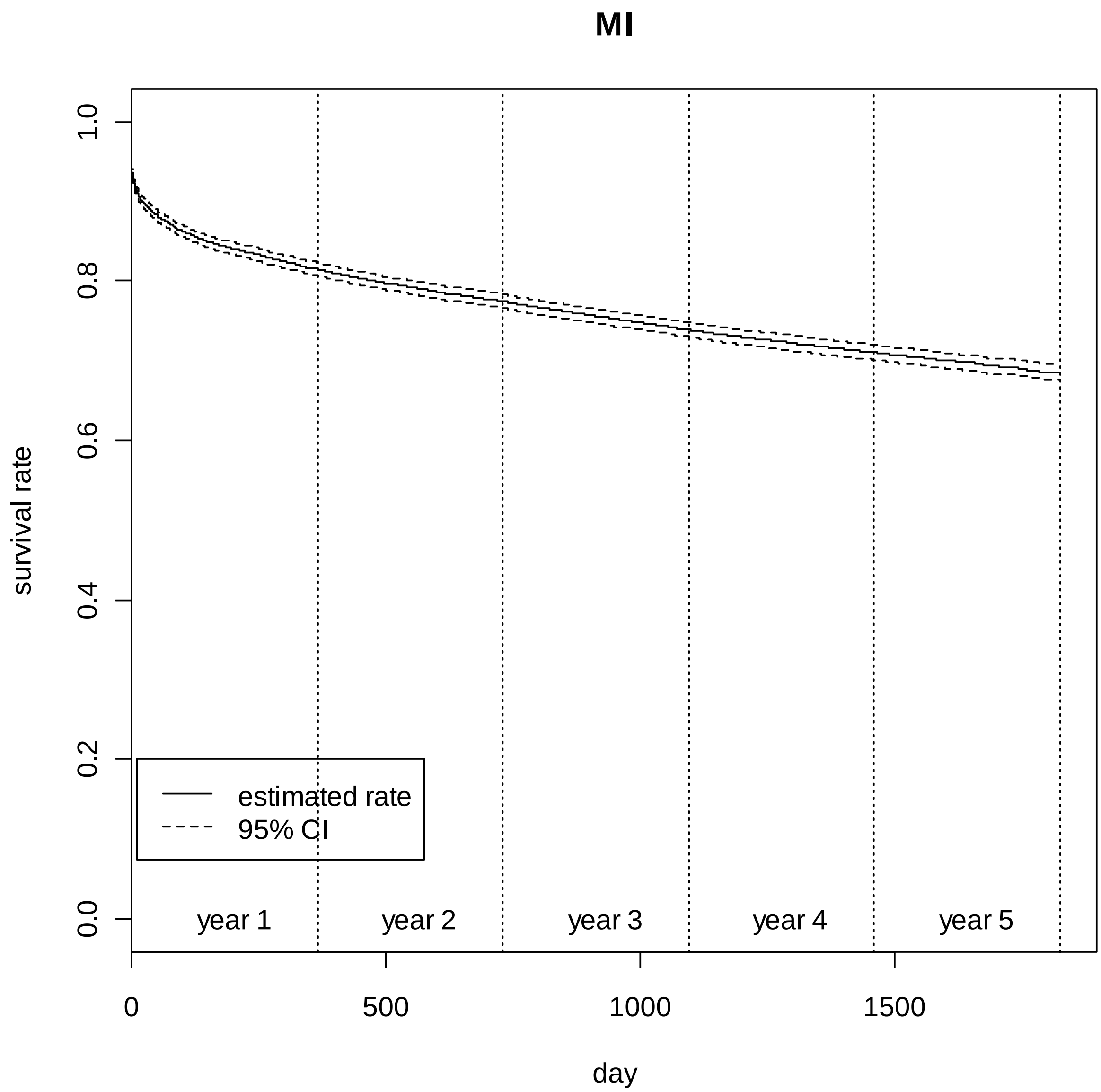


Figure 3: Estimated survival function of patients with myocardial infarction (MI) and 95% confidence intervals (CI). Median survival time is more than 1,825 days and 75% survival time is 977 days.

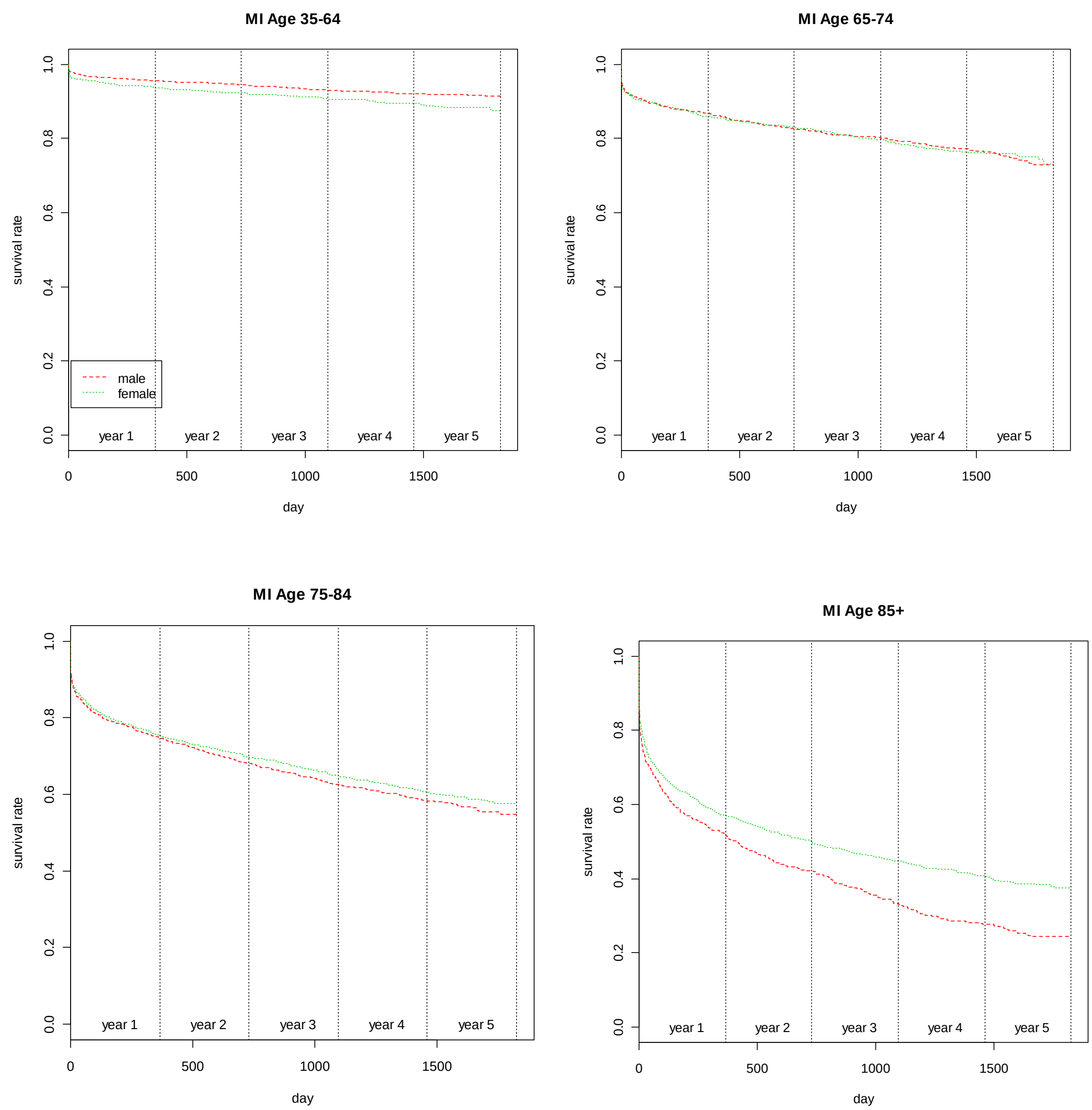


Figure 4: Estimated survival function of patients with myocardial infarction (MI) by age class and gender. Median survival time is 411 days for 85+ male, 729 days for 85+ female; otherwise more than 1,825 days. 75% survival time is 1628 days for 65-74 males, 1681 days for 65-74 females, 343 days for 75-84 males, 380 days for 75-84 females, 18 days for 85+ male, 31 days for 85+ female; otherwise more than 1,825 days.

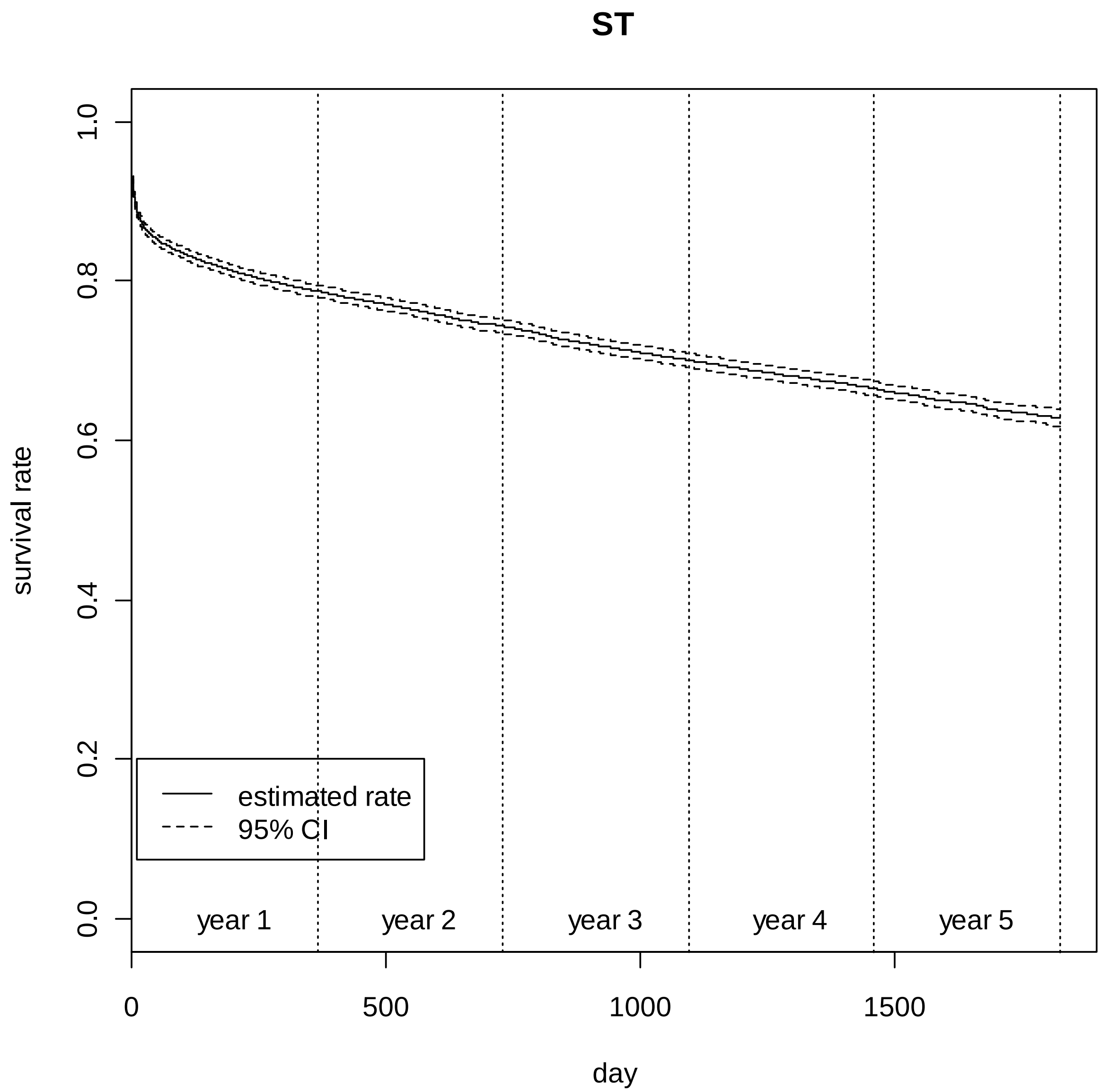


Figure 5: Estimated survival function of patients with stroke (ST) and 95% confidence intervals (CI). Median survival time is more than 1,825 days and 75% survival time is 665 days.

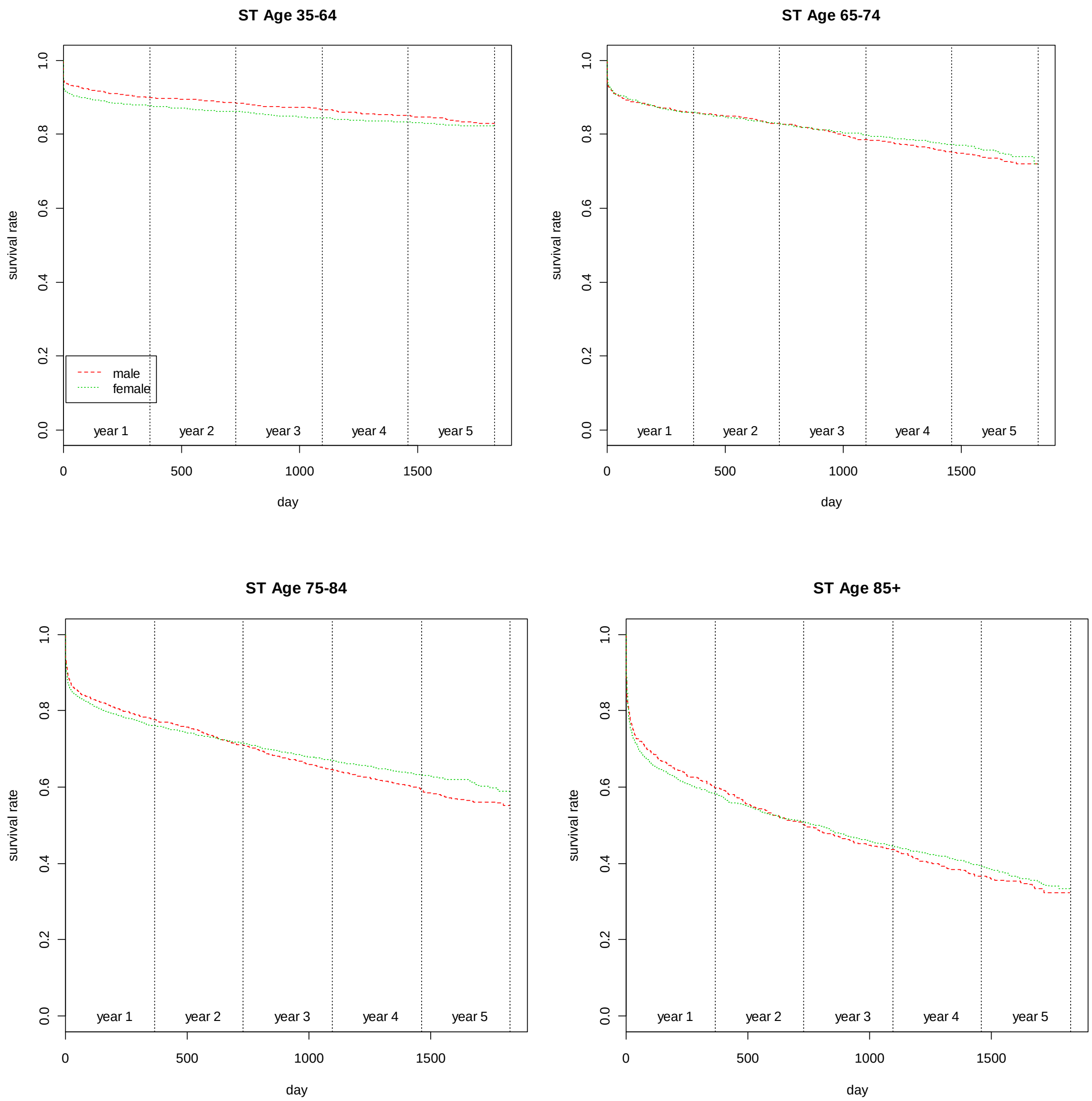
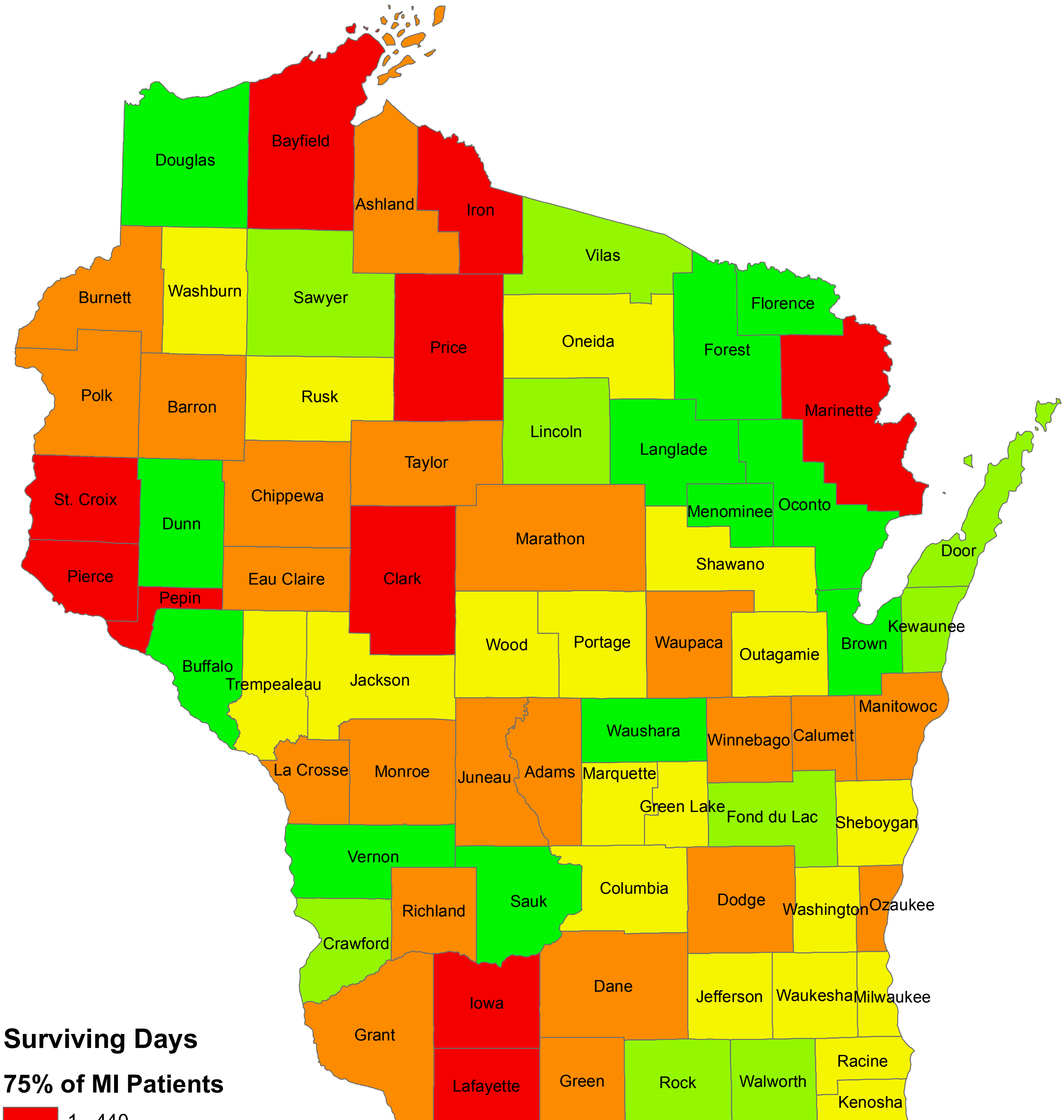


Figure 6: Estimated survival function of patients with stroke (ST) by age class and gender. Median survival time is 732 days for 85+ male, 778 days for 85+ female; otherwise more than 1,825 days. . 75% survival time is 1479 days for 65-74 males, 1660 days for 65-74 females, 543 days for 75-84 males, 450 days for 75-84 females, 26 days for 85+ male, 19 days for 85+ female; otherwise more than 1,825 days.

Survival Days of 75% MI Patients
2006-2010



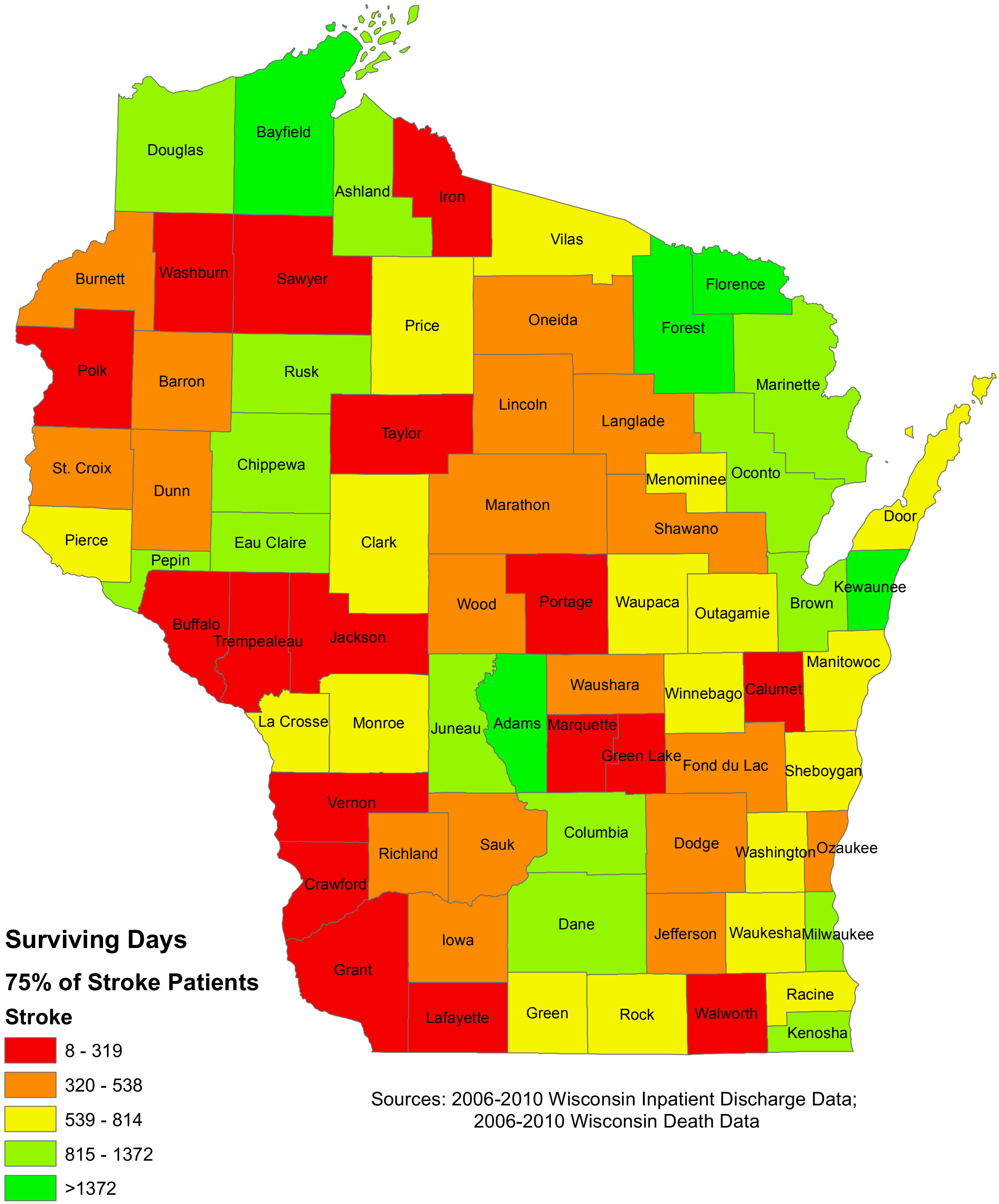
Surviving Days

75% of MI Patients

- 1 - 440
- 441 - 880
- 881 - 1320
- 1321 - 1825
- >1825

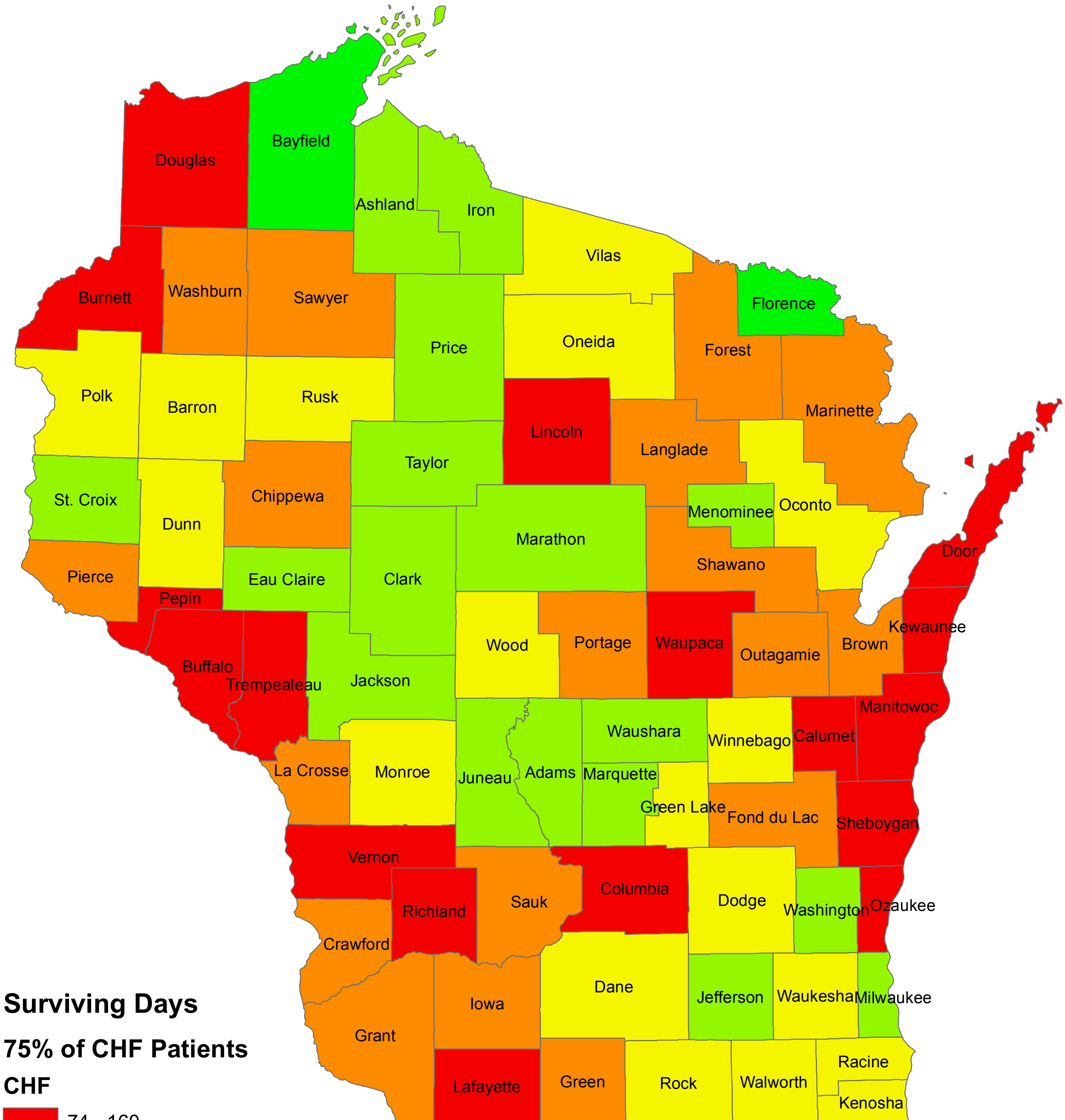
Sources: 2006-2010 Wisconsin Inpatient Discharge Data;
2006-2010 Wisconsin Death Data

Survival Days of 75% Stroke Patients 2006-2010



Sources: 2006-2010 Wisconsin Inpatient Discharge Data;
2006-2010 Wisconsin Death Data

Survival Days of 75% CHF Patients
2006-2010



Surviving Days

75% of CHF Patients

CHF

- 74 - 160
- 161 - 220
- 221 - 306
- 307 - 734
- >735

Sources: 2006-2010 Wisconsin Inpatient Discharge Data;
2006-2010 Wisconsin Death Data