Accepted for Health Systems (Published June 2014)

Abstract
Breast cancer is a complex disease and may be accompanied by multiple other health conditions. The present study investigates associations between diagnosis codes in breast cancer patients using the Nationwide Inpatient Sample data. Concomitant diagnoses codes are identified by statistically significant associations between the diagnoses codes in a given breast cancer patient. These are subsequently represented in the form of a network (BCCDN). In contrast to more classical approaches, BCCDN provides system-level insights and convenient visualization reflected by the complex wiring patterns between the diagnoses codes. Social network analysis is used to investigate highly connected codes in the BCCDN network, and their variation across three different populations: (i) the deceased breast cancer population (ii) the elderly breast cancer population (age > 65 yrs) and (iii) the adult breast cancer population (age <= 65 yrs). BCCDNs were investigated across years 2005 and 2006 in order to identify associations that are robust to the stratified sampling and population heterogeneity as well as possible errors in documentation characteristic of observational healthcare data. The results presented validate known chronic comorbidities and their persistence across the deceased and elderly breast cancer population. They also provide novel associations and potential comorbidities in breast cancer patients that may warrant a more detailed investigation.

Key words: Breast cancer, comorbidity, concomitant diagnosis network, Nationwide Inpatient Sample

Introduction
Breast cancer continues to be the most common noncutaneous cancer in American women. According to the American Cancer Society (ACS), it is estimated that approximately 232,340 new cases of invasive breast cancer will be diagnosed in 2013, with 39,620 estimated deaths from breast cancer (American Cancer Society, 2013). Breast cancer is also the leading cause of cancer death among women worldwide. The ACS estimated more than 1,000,000 new cases of breast cancer worldwide and 450,000 deaths from breast cancer in 2011 (Jemal et al., 2011). Early diagnosis can improve life expectancy but the presence of other health conditions particularly in older patients may complicate treatment and influence resource utilization. More than half of newly diagnosed breast cancer patients are above the age of 65, and with increasing life expectancy the number of older patients with breast cancer is expected to increase. Also, with increasing age the prevalence of comorbid conditions increases with possible influence on the treatment and outcome. A comorbidity is defined as a “…. clinical condition that exists before a patient’s
admission to the hospital, is not related to the principal reason for the hospitalization, and is likely to be
a significant factor influencing mortality and resource use in the hospital.........” (Elixhauser et al.,
1998). It is reported that comorbid conditions increase mortality from breast cancer twenty fold at three
years (Satariano et al., 1994). The complexity of comorbid conditions can influence breast cancer
screening and treatment but it is difficult to examine the interaction effect between diseases through
clinical studies. Analysis of hospitalizations records and observational databases has proven to be useful
in this regard. As Hurria (2011) pointed out, integrating comorbidity information can be useful in
treatment decisions and is critical for patient-tailored (personalized) interventions.

Most prior comorbidity studies focused on predicting healthcare expenditures (Farley et al., 2006, Garis
and Farmer, 2002), assessing health services utilization (Dominick et al., 2005), evaluating prognostic
information (Piccirillo et al., 2004), and measuring burden of illness with comorbidities (Stier et al.,
1999). Studying the association between comorbidity and cancer is relatively recent. Yancik et al. (1996)
conducted a descriptive study to illustrate the burden of six selected comorbidities for cancer patients
(hypertension, diabetes, heart-related condition, chronic obstructive pulmonary disease, arthritis,
gastrointestinal problems) using a sample from the Surveillance, Epidemiology and End Results (SEER)
data. Hypertension was found to be the most prevalent condition. Ogle et al. (2000) examined the
relationship between comorbid conditions in cancer patients with multiple demographic and clinical
variables using a population-based cancer surveillance data with self-reported comorbidity information.
They found that the rates of comorbid conditions differed according to age, race, socioeconomic groups
and tumor sites. Other studies discussed the association between comorbidity and cancer with respect to
cancer risk and prognosis (Geraci et al., 2005, Extermann, 2007). When quantifying the effect of
comorbidity specifically for breast cancer, many studies used survival and stage of diagnosis as metrics.
Recently, Zhang et al. (2010) investigated the impact of chronic comorbidities on breast cancer inpatient
outcomes including length of stay, total charges and disposition (e.g., vital and transferring status). They
found that hospitalization records are useful as they contain diagnosis codes for both principal as well as
comorbid conditions. Traditionally, studies have emphasized the effect of comorbidities on breast cancer
detection and patient outcomes including survival. These studies implicitly assumed prior knowledge of
specific comorbidities based on extensive literature surveys and expert opinions. However, our present
study takes a different approach to the problem and aims to discover complex associations between
diagnoses codes from observational healthcare data (Nationwide Inpatient Sample data) across a
population with breast cancer as primary diagnosis. The complex wiring between the statistically
significant associations are subsequently represented in the form of a network (BCCDN). In contrast to
investigating the impact of established chronic comorbidities on the outcome, the BCCDN is expected to
validate known chronic comorbidities while revealing novel associations that may warrant further investigations and encourage novel hypothesis generation. BCCDN is also expected to reveal associations between the secondary diagnoses codes in addition to those between the primary and secondary diagnoses codes. Social network analysis metrics such as degree centrality (Barabási and Oltvai, 2004) is used subsequently to elucidate the highly connected diagnoses codes and their variation across three distinct breast cancer populations: (deceased population, elderly population: age > 65 years, and adult population: age <= 65 years). The age variable in the data refers to a patient’s age at the time of hospitalization. The BCCDN for the first group is expected to provide preliminary insights into possible comorbidities that may be related to lethality. Those across the second and third groups are expected to capture variation in comorbidity as function of age (Eisner et al., 2002). Finally, the entire analysis is repeated across years 2005 and 2006 to capture associations between diagnosis codes that are unlikely to change due to the stratified sampling and accommodate population heterogeneity inherent in the National Inpatient Sample (NIS) data. Although this study uses a hospitalization dataset in the U.S., the focus is to study the relationship between breast cancer and comorbidities. As the global incidence of breast cancer rises due to changing lifestyles and possibly increased detection through screening, the findings in this study may also apply to other countries where breast cancer becomes a rising health concern.

Materials and Methods

The 2005 and 2006 Nationwide Inpatient Sample (NIS) was developed as a part of the Healthcare Cost and Utilization Project (HCUP) by the Agency for Healthcare Research and Quality (AHRQ). It represents approximately a 20% stratified sample of US community hospitals with discharge data from about 1,000 hospitals in nearly 40 states. The survey data is well-stratified with appropriate weights and has been found to be an unbiased representation of the population (Zhang, 2011). Each admission record is accompanied by a primary diagnosis ICD-9-CM code (DX1) followed by 14 secondary diagnosis ICD9-CM codes (DX2 to DX15) representing the conditions for hospitalization. Two diagnoses codes are called concomitant (i.e. significantly associated) if they occur together in one or more patients. Only patients diagnosed with breast cancer as a primary condition are included in this study. As in an earlier study (Zhang et al., 2010), the following ICD-9-CM codes are used to identify these primary breast cancer patients (DX1): 1740-1749, 1750, 1759, 1982, 1725, 1735, 2325, 2165, 2330, V103. The 2005 (13,480 samples) and 2006 (17,914 samples) NIS data sets consist of 126 and 128 attributes respectively. Samples with missing values were first filtered out. Three different breast cancer populations were investigated, namely: (i) deceased population: NIS data attribute DIED = 1, (ii) elderly population: NIS data attribute AGE > 65, and (iii) adult population: NIS data attribute AGE <= 65. The data sets obtained for each of the cases had significant number of samples and conformed to the HCUP data user agreement.
The description of the ICD-9 diagnosis codes considered in the present study was retrieved from the Centers for Medicare and Medicaid Services online Medicare Coverage Database and other reliable data sources such as http://www.hipaaspace.com/Medical_Billing/Crosswalk.Services/. Only those ICD-9 diagnosis codes whose description was available are discussed in the present study.

**Breast Cancer Concomitant Diagnosis Network (BCCDN)**

A majority of the breast cancer population had multiple diagnosis codes (concomitant diagnoses), characteristic of a complex disease phenotype. This in turn provided the rationale to investigate their association. The statistical metric (phi-correlation) is the counterpart of Pearson correlation for dichotomous variables (Hidalgo et al., 2009). In the present study, phi-correlation was used to identify significant associations between the diagnoses codes. The phi-correlation as defined by Hidalgo et al. (2009) between two ICD-9 codes $i$ and $j$ is given by

$$
\phi_{ij} = \frac{C_{ij}N-P_iP_j}{\sqrt{P_iP_j(N-P_i)(N-P_j)}}
$$

where $C_{ij}$ represents the number of subjects who have concomitant diagnoses, i.e. co-occurrence of $(i,j)$, and $P_i$ and $P_j$ represent the isolated occurrences of codes $i$ and $j$ in the population of size $N$. By definition, phi-correlation is a symmetric measure, and therefore the resulting network is an undirected graph and does not provide any directional information between the diagnosis codes.

By performing a $t$-test, the null hypothesis ($H_0$: $\phi_{ij} = 0$) was rejected at a significance level of 0.05 if the $t$-statistic satisfies:

$$
t = \frac{\phi_{ij}\sqrt{n-2}}{\sqrt{1 - \phi_{ij}^2}} \geq 1.96,
$$

where $n = \max(P_i, P_j)$. Since the construction of BCCDN requires repeated statistical testing for significant associations across multiple pairs it is possible to accentuate false-discoveries. In order to minimize false-discoveries, multiple-testing corrections such as those controlling for family-wise error rate (e.g., the Bonferroni Correction) or false-discovery rate (e.g., the Benjamini-Hochberg) are often encouraged (Hochberg and Tamhane, 1987; Benjamini and Hochberg, 1995). The NIS data corresponding to the elderly and adult population was reasonably larger however that of the deceased breast cancer population was considerably small across the years (2005, 2006). Therefore, no multiple-testing correction was imposed. Significant pairs of diagnosis codes identified using the above statistical test was connected by an edge in order to realize the BCCDN. Degree centrality of the nodes in the BCCDN was subsequently investigated.
Results

**BCCDN of the Deceased Population**

The deceased breast cancer population across the years 2005 (~2.7% of the total samples) and 2006 (~2.1% of the total samples) were comparable but relatively small compared to the entire population. The age distribution was around 59±15 years indicating that age might not necessarily be a major cause of mortality. The rationale for investigating the deceased breast cancer population can be attributed to the fact that comorbidity in this patient population may have a higher potential to be related to lethality.

The BCCDN of the significant pairs of the diagnosis codes across the deceased population using the data from 2005 as well as 2006 is shown in Fig. 1. The primary diagnosis code 1749 (*Malignant neoplasm of breast not otherwise specified*) is one of the highly connected nodes as expected. The diagnoses codes shown in bold, Fig. 1, 25000 (*Diabetes Mellitus Type II without complications*), 3310 (*Alzheimer’s disease*), 4019 (*Hypertension*), 4280 (*Congestive heart failure*), 29410 (*Dementia*), 42731 (*Atrial fibrillation*) and 2989 (*Psychosis*) correspond to common chronic comorbid conditions ([http://www.hcup-us.ahrq.gov/reports/factsandfigures/2007/exhibit2_4.jsp](http://www.hcup-us.ahrq.gov/reports/factsandfigures/2007/exhibit2_4.jsp)) among general admissions. Of interest is to note that the core sub-network includes diagnosis codes that correspond to secondary neoplasms characteristic of metastatic breast cancer. These include 1970 (*secondary malignant neoplasm of the lung*), 1977 (*malignant neoplasm of the liver secondary*), 1983 (*secondary malignant neoplasm of brain and spinal cord*) and 1985 (*secondary malignant neoplasm of bone and bone marrow respectively*). On the other hand, the following diagnoses codes 99592 (*severe sepsis without septic shock*), 99591 (*sepsis unspecified organism*) and 78552 (*sever sepsis with septic shock*) correspond to possible infection during the course of the treatment. These codes are associated with each other as expected as shown in Fig. 1. SIRS (Systematic inflammatory response syndrome) infection and severe sepsis are related and have been shown to contribute significantly to mortality. Severe sepsis has also been argued to be accentuated by the use of chemotherapeutic agents and in post-operative settings (Vogel et al., 2010). The pair 2762 (*acidosis*) and 5849 (*acute renal failure*) are again associated medical conditions that were identified as significant and shown in Fig. 1. Similarly, the pairs 3310 (*Alzheimer’s disease*) and 29410 (*Dementia*) correspond to related mental conditions and are associated as expected and shown in Fig. 1. The pair 4111 (*bacterial infections staphylococcus aureus infections*) and V090 (*infections resistant to penicillin*) are related to bacterial infections and are associated as expected and shown in Fig. 1.
### Node Description and Degree

<table>
<thead>
<tr>
<th>Node</th>
<th>Description</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1749</td>
<td>MALIGN NEOPL BREAST NOS</td>
<td>5</td>
</tr>
<tr>
<td>1977</td>
<td>SECOND MALIG NEO LIVER</td>
<td>4</td>
</tr>
<tr>
<td>1970</td>
<td>SECONDARY MALIG NEO LUNG</td>
<td>4</td>
</tr>
<tr>
<td>1985</td>
<td>SECONDARY MALIG NEO BONE</td>
<td>7</td>
</tr>
<tr>
<td>1983</td>
<td>SEC MAL NEO BRAIN/SPINE</td>
<td>6</td>
</tr>
<tr>
<td>5990</td>
<td>URIN TRACT INFECTION NOS</td>
<td>1</td>
</tr>
<tr>
<td>78009</td>
<td>OTHER ALTER CONCIOUSNESS</td>
<td>5</td>
</tr>
<tr>
<td>2378</td>
<td>DIS PLAS PROTEIN MET NEC</td>
<td>1</td>
</tr>
<tr>
<td>41400</td>
<td>COR ATH UNSP VSL NTW/GFT</td>
<td>1</td>
</tr>
<tr>
<td>V090</td>
<td>INF MCRG RSTN PNCLLINS</td>
<td>1</td>
</tr>
<tr>
<td>7823</td>
<td>EDEMA</td>
<td>1</td>
</tr>
<tr>
<td>5728</td>
<td>OTH SEQUELA, CHR LIV DIS</td>
<td>1</td>
</tr>
<tr>
<td>5119</td>
<td>PLEURAL EFFUSION NOS</td>
<td>1</td>
</tr>
<tr>
<td>2875</td>
<td>THROMBOCYTOPENIA NOS</td>
<td>1</td>
</tr>
<tr>
<td>78039</td>
<td>CONVULSIONS NEC</td>
<td>1</td>
</tr>
<tr>
<td>4254</td>
<td>PRIM CARDIOMYOPATHY NEC</td>
<td>1</td>
</tr>
<tr>
<td>7994</td>
<td>CACHEXIA</td>
<td>4</td>
</tr>
<tr>
<td>2762</td>
<td>ACIDOSIS</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Node</th>
<th>Description</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>7895</td>
<td>ASCITES</td>
<td>1</td>
</tr>
<tr>
<td>78552</td>
<td>SEPTIC SHOCK</td>
<td>2</td>
</tr>
<tr>
<td>99592</td>
<td>SIRS-INFECT W ORG DYSF</td>
<td>3</td>
</tr>
<tr>
<td>99591</td>
<td>SIRS-INFECT W/O ORG DYSF</td>
<td>1</td>
</tr>
<tr>
<td>V4581</td>
<td>AORTOCORONARY BYPASS</td>
<td>1</td>
</tr>
<tr>
<td>4111</td>
<td>INTERMED CORONARY SYND</td>
<td>1</td>
</tr>
<tr>
<td>2769</td>
<td>ELECTROLYT/FLUID DIS NEC</td>
<td>2</td>
</tr>
<tr>
<td>2765</td>
<td>HYPOVOLEMIA</td>
<td>2</td>
</tr>
<tr>
<td>5849</td>
<td>ACUTE RENAL FAILURE NOS</td>
<td>1</td>
</tr>
<tr>
<td>51881</td>
<td>ACUTE RESPIRATORY FAILURE</td>
<td>1</td>
</tr>
<tr>
<td>4019</td>
<td>HYPERTENSION NOS</td>
<td>1</td>
</tr>
<tr>
<td>4280</td>
<td>CONGES HEART FAIL NOS</td>
<td>2</td>
</tr>
<tr>
<td>25000</td>
<td>DIABETES MELITTUS II W/O CMP</td>
<td>1</td>
</tr>
<tr>
<td>42731</td>
<td>ATRIAL FIBRILLATION</td>
<td>1</td>
</tr>
<tr>
<td>3310</td>
<td>ALZHEIMER'S DISEASE</td>
<td>1</td>
</tr>
<tr>
<td>29410</td>
<td>DEMENTIA W/O BEHAV DIST</td>
<td>1</td>
</tr>
<tr>
<td>2989</td>
<td>PSYCHOSIS NOS</td>
<td>4</td>
</tr>
</tbody>
</table>

**Figure 1.** BCCDN Deceased Population: Dual-circle layout representing the association between the diagnoses codes for the deceased breast cancer population across 2005 and 2006 is shown. The size of the node labels is proportional to the degree with highly connected nodes in the inner circle. The degree centrality of the diagnoses codes and its description are shown right below with codes corresponding to chronic comorbid conditions in bold.
BCCDN of the elderly breast cancer population (>65 years) and comparison to the deceased population

A similar analysis was used to identify statistically significant (Hidalgo et al., 2009) co-occurring diagnosis codes for the elderly breast cancer population (> 65 years) across the years 2005 (~38.7% of the total samples) and the 2006 (~39% of the total samples). The corresponding BCCDN network for 2006 is shown in Fig. 2 (left pane). The degree centrality distribution (Barabási and Oltvai, 2004) of these networks as expected was positively skewed with only a few highly connected diagnosis codes comprising the tail of the distribution with a majority of the codes sparsely associated. The top ten diagnosis codes ranked by their degree centrality identified from the BCCDN at 2006 were (4019, 1985, 41401, 4280, 25000, 1749, 1977, 42731, 27651, 53081), Fig. 2 (left pane) whereas those identified from the BCCDN at 2005 were (1749, 1985, 4019, 1970, 1977, 27651, 78009, 25000, 4280 and 42731). Eight out of the top ten diagnoses codes persisted across the years, these were: (1749, 4019, 1985, 4280, 25000, 1977, 42731, 27651). Seven out of these eight codes shared considerable overlap with those across the deceased population shown in Fig. 1. These consisted of the primary diagnoses code 1749 (malignant neoplasm of breast not otherwise specified), those corresponding to secondary malignancies such as 1977 (secondary malignant neoplasm of liver) and 1985 (secondary malignant neoplasm of bone). They also included chronic comorbidities 4019 (hypertension), 4280 (congestive heart failure), 25000 (diabetes mellitus type 2 without complications) and 42731 (atrial fibrillation).

BCCDN of the adult population (<= 65 years) and comparison to the deceased population

In order to understand the variation of the BCCDN as a function of age, statistically significant (Hidalgo et al., 2009) co-occurring diagnosis codes were identified for the adult breast cancer population (<= 65 years) across 2005 (~61% of the total samples) and 2006 (~61% of the total samples). BCCDN of the adult breast cancer population in 2006 is shown in Fig. 2 (right pane). Similar to earlier observation across the elderly population, the degree centrality distribution (Barabási and Oltvai, 2004) was positively skewed with only a few highly connected diagnosis codes. The top ten diagnosis codes ranked by their degree centrality across the BCCDN for the year 2005 were (1985, 1749, 4019, 1977, 1970, 1983, 2330, 25000, 7994, 53081) whereas those for 2006 were (4019, 1985, 1977, 1749, 25000, 53081, 1970, V163, 27651, 2330), Fig. 2 (right pane). Eight out of the top ten highly connected diagnosis codes persisted across 2005 as well as 2006 and consisted of (1749, 1985, 1977, 1970, 4019, 25000, 2330, 53081). Seven out of these eight codes also exhibited overlap with the deceased population, Fig. 1. These included primary diagnosis code 1749 (malignant neoplasm of breast not otherwise specified), codes corresponding to secondary malignancies 1970 (secondary malignant neoplasm of lung), 1977 (secondary malignant neoplasm of liver), 1985 (secondary malignant neoplasm of bone), and those corresponding to chronic comorbidities 4019 (hypertension) and 25000 (diabetes mellitus type 2 without complications).
Unlike the elderly population, the BCCDN of the adult population across the 2005 and 2006 also had primary diagnosis code (2330: Carcinoma in-situ of breast) and relatively smaller overlap with the chronic comorbidities identified across the deceased population.

**Figure 2.** BCCDN network for the elderly (> 65yrs) and adult breast cancer (<= 65 yrs) population in 2006.

The corresponding dual circle layouts are shown in (left pane, elderly, > 65 yrs) and (right pane, adult, <= 65 yrs) respectively. The font size of the node labels is proportional to the degree with more highly connected nodes (top 10 nodes) shown in the inner circle. The codes in the perimeter (outer-circle) have been suppressed for enhanced clarity. The degree centrality of the top ten diagnoses codes is enclosed under each of the plots respectively.

**Discussion**

The underlying mechanism of the interaction among breast cancer and comorbidities are still not fully understood. In contrast to majority of the literature on outcome analysis of comorbidity influence, this study investigated statistically significant associations between concomitant diagnosis codes obtained from the Nationwide Inpatient Sample data. Subsequently, network abstractions (BCCDN) were used to
obtain system-level understanding revealed by the complex associations between these diagnosis codes. Social network analysis metrics such as degree centrality were used to identify dominant players in the BCCDN and their variations across time and distinct breast cancer population. Three distinct breast cancer populations were considered, namely (i) deceased population, (ii) elderly breast cancer population (> 65 yrs) and adult breast cancer population (<= 65 yrs). The entire analysis was repeated across two time slots in order to identify robust association’s immune to the stratified sampling, possible errors in documentation and population heterogeneity inherent in observational healthcare data. For the elderly population, the highly connected diagnosis codes across 2005 and 2006 consisted of \(1749, 1977, 1985, 4019, 4280, 25000, 42731, 27651\). This included primary diagnoses, secondary malignancies and chronic comorbidities prevalent in breast cancer population. A similar analysis of the adult population resulted in the diagnoses codes: \(1749, 2330, 1970, 1977, 1985, 4019, 25000, 53081\). Our analysis also revealed that diagnoses codes corresponding to chronic comorbidities were more prevalent in the elderly as opposed to the adult population. Also, the diagnosis codes corresponding to the chronic comorbidities exhibited a greater overlap between the elderly and the deceased population. While the present study focused on observational healthcare data (NIS data) from the healthcare providers in the United States, the results and approach presented are generic and can be extended to other settings.

There are some limitations of the present study due to the nature of the dataset. Since the dataset is completely de-identified, it is impossible to determine multiple visits from the same patient. However, we checked for possible duplicate situations by identifying patient age, race, hospital number, location of the patient, median household income for patient’s zip code, and payer information, but did not conclusively identify any duplication. In addition, there is no information regarding pre-existing conditions, complications during the course of the treatment or diagnoses codes of patients as a function of time. Therefore, the resulting BCCDN is an undirected graph with no explicit directional information between the diagnosis codes. Furthermore, there is no information about breast cancer staging in this dataset. Some information about the cancer is captured by the diagnosis codes. It is anticipated that different breast cancer stages may result in different BCCDN. Finally, we assume the recording sequence of the diagnosis coding follows a standard guideline that the primary reason for admission in this case being breast cancer, (primary diagnosis) followed by secondary conditions. Since the number of samples in the deceased population was considerably small, we did not impose any multiple testing corrections to control for false-discovery rate or family-wise error rate. However, we believe establishing the robustness of the results across two distinct time stamps (2005, 2006) will alleviate some of these concerns.
Acknowledgements

Have been deleted to assist in the blind review process

References


