**Identifying Healthy Normal Patients**

Griffin M Weber, MD, PhD

weber@hms.harvard.edu

July 7, 2014

**ABSTRACT**

When conducting a clinical study, a “normal” control population is often desired. Identifying normal patients in a clinical data repository can be challenging because healthy patients do not go the hospital; and, because patients receive care from multiple hospitals, the absence of a diagnosis in one hospital’s electronic health record does not mean a patient does not have the disease. We define a set of 10 simple heuristic filters to eliminate patients who might be sick or abnormal. Out of 2,019,774 patients at two large academic hospitals, these filters excluded all but 31,352 (1.55%), which illustrates how difficult it can be to identify normal controls.

**METHODS**

The dataset included 2,019,774 patients at Partners HealthCare System (Brigham and Women’s Hospital and Massachusetts General Hospital) with facts between January 1, 2001, and July 11, 2009. Of these, 886,547 are male, 1,132,920 are female, and 307 do not have a recorded gender. The dataset included 1,034,746,555 clinical data facts, such as diagnoses, medications, laboratory tests, and demographic codes. We used AHRQ’s Chronic Condition Indicator (CCI) for ICD-9-CM (<http://www.hcup-us.ahrq.gov/toolssoftware/chronic/chronic.jsp>) to group diagnoses into one of 18 body systems and classify them as either chronic or non-chronic. Out of 92,073,962 diagnoses, 40,271,611 (43.7%) were chronic and 51,803,351 (56.3%) were non-chronic.

We defined “healthy normal” patients as those meeting the following ten criteria:

1. Age 40-69
2. Alive
3. Has a recorded gender
4. Has at least one fact in the past year
5. Has at least one diagnosis
6. Has no chronic diagnoses
7. Has no diagnoses in select body systems (endocrine, nutritional, metabolic; blood or blood-forming organs; mental disorders; congenital anomalies; and certain conditions originating in the perinatal period)
8. Has no diagnoses with a prevalence less than 0.01% (partitioned by 5 year age group and gender)
9. Number of total data facts (demographics, diagnoses, medications, laboratory test results, etc.) within 10th to 90th percentile of patients (partitioned by 5 year age group and gender)
10. Number of diagnosis data facts (counting duplicates of the same code) within 0 to 90th percentile of patients (partitioned by 5 year age group and gender)

**RESULTS**

Figure 1 shows the age distribution of the patients in the study dataset. The plateau of patients between 0 and 8 years old mostly represents patients born during the 8.5 year data collection period. There are more women than men between the ages of 20 and 65, but approximately the same number in other age groups. Filter #1 selects only patients between 40 and 69 years old, which represents 43% of the population. The rationale for starting at 40 years old is that there may be younger patients who appear healthy now, but will develop chronic conditions later in life. This gives these chronic diseases a chance to emerge.



**Figure 1. Number of patients by current age at Partners.**

Normal does not mean healthy. Figure 2 shows that a “normal” (50th percentile) 95 year old patient is deceased. A normal 50 year old has at least 2 chronic disease diagnoses, and only one in four have no chronic diagnoses. Most newborns have several diagnoses; however, these are typically non-chronic and do not necessarily represent a disease. Filter #1 eliminated patients age 70 years or older. Beyond this age, being completely healthy is quite rare and could indicate the presence of protective genes not seen in a normal population.

****

**Figure 2. Health status by age as defined by types of diagnoses.**

Filter #4 selects only patients seen in the past year. This is to eliminate patients who may be receiving treatment for diseases at other hospitals. It also eliminates patients who are deceased, but the hospital has no record that those patients died. Unfortunately, it probably eliminates many healthy patients too. In the Partners Healthcare data, only 43% of patients have been seen in the past year.

Filter #8 considers the prevalence of the rarest diagnosis patients have. Diagnosis prevalence was calculated from all patients with at least one diagnosis code, partitioned by sex and 5 year age group (e.g., 40-44, 45-49, etc.). The filter eliminated patients who have a diagnosis with a prevalence less than 0.01% in their demographic group. This is approximately 10% of the patients who have diagnoses at Partners Healthcare.

Filter #9 counts a patient’s total number of data facts. A data fact is any demographic code, diagnosis, medication, laboratory test result, etc. In the same way that being in the highest or lowest percentiles of height, weight, or other physiologic measures could indicate an underlying medical condition, it is also abnormal to have an extremely high or low number of facts for a given sex and age group. Patients with too many facts are seen at the hospital often, which suggests a problem, regardless of whether they have any chronic diagnoses. Patients with too few facts may be receiving treatment for diseases at other hospitals, or they may be or unable to (choosing not to) obtain healthcare. Thus, Filters #9 eliminates patients at the upper and lower extremes in number of facts. Filter #10 does the same, but just for the upper extreme in number of diagnosis facts.

Table 1 lists the number of patients who passed or were eliminated by each of the 10 filters when applied independently to the entire population. In Table 2, the filters were applied sequentially until a final group of 31,598 patients remained. This is just 1.56% of the Partners Healthcare patient population.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Filter** | **Description** | **# Patients****Who Pass****Filter** | **% Patients****Who Pass Filter** | **# Patients Eliminated****By Filter** | **% Patients****Eliminated****By Filter** |
| 0 | All Patients | 2,019,774 | 100.00% | 0 | 0.00% |
| 1 | Age 40-69 | 859,791 | 42.57% | 1,159,983 | 57.43% |
| 2 | Alive | 1,884,565 | 93.31% | 135,209 | 6.69% |
| 3 | Has gender | 2,019,467 | 99.98% | 307 | 0.02% |
| 4 | Recently seen | 774,044 | 38.32% | 1,245,730 | 61.68% |
| 5 | Has diagnoses | 1,971,716 | 97.62% | 48,058 | 2.38% |
| 6 | No chronic diagnoses | 659,844 | 32.67% | 1,359,930 | 67.33% |
| 7 | No select body systems | 1,110,823 | 55.00% | 908,951 | 45.00% |
| 8 | No rare diagnoses | 1,799,174 | 89.08% | 220,600 | 10.92% |
| 9 | Normal # of facts | 1,615,808 | 80.00% | 403,966 | 20.00% |
| 10 | Normal # of diagnoses | 1,817,826 | 90.00% | 201,948 | 10.00% |

**Table 1. Number of patients and percent of total patients remaining after each filter is applied independently to the entire population.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Filter** | **Description** | **# Patients****Who Pass****Filter** | **% Patients****Who Pass Filter** | **# Patients Eliminated****By Filter** | **% Patients****Eliminated****By Filter** |
| 0 | All Patients | 2,019,774 | 100.00% | 0 | 0.00% |
| 1 | Age 40-69 | 859,791 | 42.57% | 1,159,983 | 57.43% |
| 2 | Alive | 814,813 | 40.34% | 44,978 | 5.23% |
| 3 | Has gender | 814,781 | 40.34% | 32 | 0.00% |
| 4 | Recently seen | 362,659 | 17.96% | 452,122 | 55.49% |
| 5 | Has diagnoses | 358,225 | 17.74% | 4,434 | 1.22% |
| 6 | No chronic diagnoses | 41,264 | 2.04% | 316,961 | 88.48% |
| 7 | No select body systems | 40,404 | 2.00% | 860 | 2.08% |
| 8 | No rare diagnoses | 39,384 | 1.95% | 1,020 | 2.52% |
| 9 | Normal # of facts | 31,617 | 1.57% | 7,767 | 19.72% |
| 10 | Normal # of diagnoses | 31,598 | 1.56% | 19 | 0.06% |

**Table 2. Number of patients and percent of patients remaining after each filter is applied sequentially to the entire population.** The bottom row indicates the final number of healthy normal patients.

Tables 3 and 4 indicate the number and percentage of healthy normal patients by demographic group. A greater percentage of female than male, younger than older, and minority than white patients are normal. Overall, 3.68% of patients between the ages of 40-69 are healthy and normal. Table 5 lists the top 10 diagnoses of healthy normal women and men. Most of these diagnoses are ICD-9 codes for normal screening and evaluation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Age** | **# Female** | **% Female** | **# Male** | **% Male** | **Total** |
| 40-44 | 4,277 | 4.46% | 2,989 | 4.60% | 7,266 |
| 45-49 | 4,129 | 4.44% | 2,584 | 3.83% | 6,713 |
| 50-54 | 3,928 | 4.45% | 2,222 | 3.40% | 6,150 |
| 55-59 | 3,197 | 3.93% | 1,650 | 2.67% | 4,847 |
| 60-64 | 2,687 | 3.63% | 1,321 | 2.20% | 4,008 |
| 65-69 | 1,787 | 3.15% | 827 | 1.64% | 2,614 |
| All | 20,005 | 4.09% | 11,593 | 3.13% | 31,598 |

**Table 3. The number of normal patients by age and gender.** The number in parentheses represents the percentage of total patients in that demographic group who are normal.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Race** | **# Female** | **% Female** | **# Male** | **% Male** | **Total** |
| Asian | 706 | 6.01% | 404 | 5.72% | 1,110 |
| Black | 1,339 | 4.33% | 700 | 3.27% | 2,039 |
| Hispanic | 1,207 | 4.51% | 872 | 4.04% | 2,079 |
| White | 13,289 | 4.01% | 7,713 | 3.03% | 21,002 |
| Other | 3,464 | 3.91% | 1,904 | 2.89% | 5,368 |
| All | 20,005 | 4.09% | 11,593 | 3.13% | 31,598 |

**Table 4. The number of normal patients by race and gender.** The number in parentheses represents the percentage of total patients in that demographic group who are normal

|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **Rank** | **ICD9** | **Name** |
| F | 1 | V76.12 | Other screening mammogram |
| F | 2 | V76.2 | Screening for malignant neoplasms of the cervix |
| F | 3 | V70.0 | Routine general medical examination at a health care facility |
| F | 4 | 729.5 | Pain in limb |
| F | 5 | V72.31 | Routine gynecological exam |
| F | 6 | 793.80 | Abnormal mammogram, unspecified |
| F | 7 | 611.72 | Lump or mass in breast |
| F | 8 | 793.89 | Other abnormal findings on radiological examination of breast |
| F | 9 | V76.10 | Breast screening, unspecified |
| F | 10 | 216.5 | Benign neoplasm of skin of trunk, except scrotum |
| M | 1 | V70.0 | Routine general medical examination at a health care facility |
| M | 2 | 729.5 | Pain in limb |
| M | 3 | 786.50 | Unspecified chest pain |
| M | 4 | 606.9 | Male infertility, unspecified |
| M | 5 | 211.3 | Benign neoplasm of colon |
| M | 6 | V04.81 | Need for prophylactic vaccination and inoculation, Influenza |
| M | 7 | 719.41 | Pain in joint involving shoulder region |
| M | 8 | 724.2 | Lumbago |
| M | 9 | 216.5 | Benign neoplasm of skin of trunk, except scrotum |
| M | 10 | 724.5 | Backache, unspecified |

**Table 5. Top diagnoses of normal patients.**