Data Analytics on Diabetes Dataset to Predict and to Take Up Preventive Measures

**Chandrala, Deeksha, H Rameshwari, NikhilaSomu**

(USN:1MS12CS020) Computer Science and Engineering, MSRIT, Banglore-560054

(USN:1MS12CS025) Computer Science and Engineering, MSRIT, Banglore-560054

(USN:1MS12CS036) Computer Science and Engineering, MSRIT, Banglore-560054

(USN:1MS12CS066) Computer Science and Engineering, MSRIT, Banglore-560054

**Abstract**: Management of hyperglycaemia in hospitalized patients has a significant bearing on outcome, in terms of both morbidity and mortality. However, there are few national assessments of diabetes care during hospitalization which could serve as a baseline for change. This analysis of a large clinical database (74 million unique encounters corresponding to 17 million unique patients) was undertaken to provide such an assessment and to find future directions which might lead to improvements in patient safety. Almost 70,000 inpatient diabetes encounters were identified with sufficient detail for analysis. Multivariable logistic regression was used to fit the relationship between the measurement of HbA1c and early readmission while controlling for covariates such as demographics, severity and type of the disease, and type of admission. Results show that the measurement of HbA1c was performed infrequently (18.4%) in the inpatient setting. The statistical model suggests that the relationship between the probability of readmission and the HbA1c measurement depends on the primary diagnosis. The data suggest further that the greater attention to diabetes reflected in HbA1c determination may improve patient outcomes and lower cost of inpatient care.

Keywords: HbA1c test, R programming language, Regression, statistical analysis

1. **Introduction**

It is increasingly recognized that the management of hyperglycaemia in the hospitalized patient has a significant bearing on outcome, in terms of both morbidity and mortality. This recognition has led to the development of formalized protocols in the intensive care unit (ICU) setting with rigorous glucose targets in many institutions. However, the same cannot be said for most non-ICU inpatient admissions. Rather, anecdotal evidence suggests that inpatient management is arbitrary and often leads to either no treatment at all or wide fluctuations in glucose when traditional management strategies are employed. Although data are few, recent controlled trials have demonstrated that protocol-driven inpatient strategies can be both effective and safe. As such, implementation of protocols in the hospital setting is now recommended. However, there are few national assessments of diabetes care in the hospitalized patient which could serve as a baseline for change. The present analysis of a large clinical database was undertaken to examine historical patterns of diabetes care in patients with diabetes admitted to a US hospital and to inform future directions which might lead to improvements in patient safety. In particular, we examined the use of HbA1c as a marker of attention to diabetes care in a large number of individuals identified as having a diagnosis of diabetes mellitus. We hypothesize that measurement of HbA1c is associated with a reduction in readmission rates in individuals admitted to the hospital.

Databases of clinical data contain valuable but heterogeneous and difficult data in terms of missing values, incomplete or inconsistent records, and high dimensionality understood not only by number of features but also their complexity. Additionally, analysing external data is more challenging than analysis of results of a carefully designed experiment or trial, because one has no impact on how and what type of information was collected. Nonetheless, it is important to utilize these huge amounts of data to find new information/knowledge that is possibly not available anywhere.

1. **Design And Implementation**

The data collected from the UCI Machine Learning Repository for diabetes. The following process was undertaken:

1. Data Collection

The Health Facts data we used was an extract representing 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States: Midwest (18 hospitals), Northeast (58), South (28), and West (16). Most of the hospitals (78) have bed size between 100 and 499, 38 hospitals have bed size less than 100, and bed size of 14 hospitals is greater than 500.

Information was extracted from the database for encounters that satisfied the following criteria:

(1)It is an inpatient encounter (a hospital admission).

(2)It is a “diabetic” encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.

(3)The length of stay was at least 1 day and at most 14 days.

(4)Laboratory tests were performed during the encounter.

(5)Medications were administered during the encounter.

Criteria 3-4 were applied to remove admissions for procedures and so forth, which were of less than 23 hours of duration and in which changes in diabetes management were less likely to have occurred. It should be noted that the diabetic encounters are not all encounters of diabetic patients but rather only these encounters where diabetes was coded as an existing health condition. 101,766 encounters were identified to fulfil all of the above five inclusion criteria and were used in further analysis. Attribute/feature selection was performed by our clinical experts and only attributes that were potentially associated with the diabetic condition or management were retained. From the information available in the database, we extracted 55 features describing the diabetic encounters, including demographics, diagnoses, diabetic medications, number of visits in the year preceding the encounter, and payer information.

1. Data Cleaning

The original database contains incomplete, redundant, and noisy information as expected in any real-world data. There were several features that could not be treated directly since they had a high percentage of missing values. These features were weight (97% values missing), payer code (40%), and medical specialty (47%). Weight attribute was considered to be too sparse and it was not included in further analysis. Payer code was removed since it had a high percentage of missing values and it was not considered relevant to the outcome. Medical specialty attribute was maintained, adding the value “missing” in order to account for missing values.

The preliminary dataset contained multiple inpatient visits for some patients and the observations could not be considered as statistically independent, an assumption of the logistic regression model. We thus used only one encounter per patient; in particular, we considered only the first encounter for each patient as the primary admission and determined whether or not they were readmitted within 30 days. Additionally, we removed all encounters that resulted in either discharge to a hospice or patient death, to avoid biasing our analysis. After performing the above-described operations, we were left with 69,984 encounters that constituted the final dataset for analysis.

1. Selecting a Programming Language for Statistical Analysis

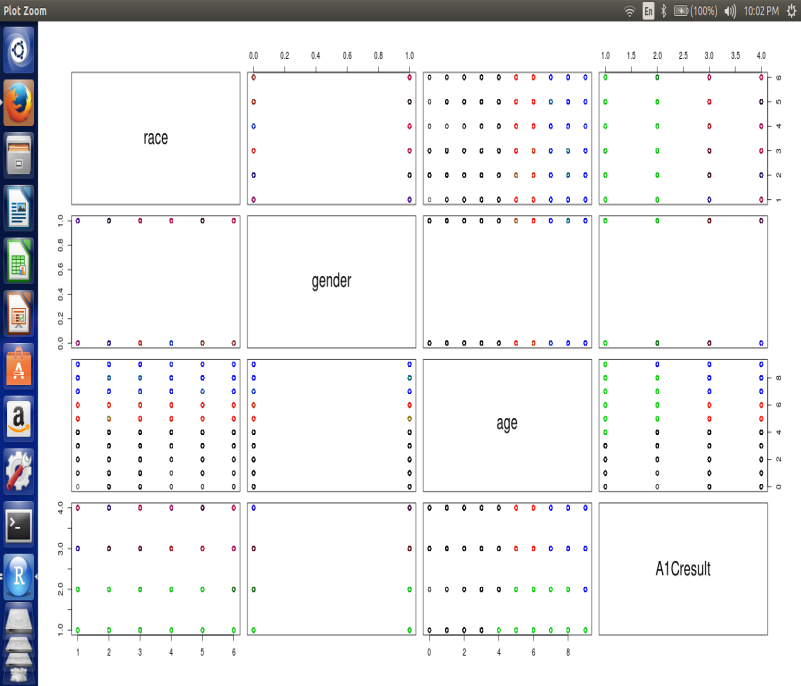
For statistical analysis we selected very commonly programming language R. **R** is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.

The source code for the R software environment is written primarily in C, Fortran, and R.R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. While R has a command line interface, there are several graphical front-ends available. R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made. For computationally intensive tasks, C, C++, and Fortran code can be linked and called at run time. Advanced users can write C, C++, Java, .NET or Python code to manipulate R objects directly. R is highly extensible through the use of user-submitted packages for specific functions or specific areas of study. Due to its S heritage, R has stronger object-oriented programming facilities than most statistical computing languages. Extending R is also eased by its lexical scoping rules. Another strength of R is static graphics, which can produce publication-quality graphs, including mathematical symbols. Dynamic and interactive graphics are available through additional packages. R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hard copy.

1. **Testing and Comparison**

In analytical methods we used clustering by means of programming language R.

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, *k*-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes. The algorithm has a loose relationship to the *k*-nearest neighbour classifier, a popular machine learning technique for classification that is often confused with *k*-means because of the *k* in the name. One can apply the 1-nearest neighbour classifier on the cluster centers obtained by *k*-means to classify new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.



`­ Fig. 1 A k- means clustering on four variables(race,gender,age,HbA1c results)

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression.

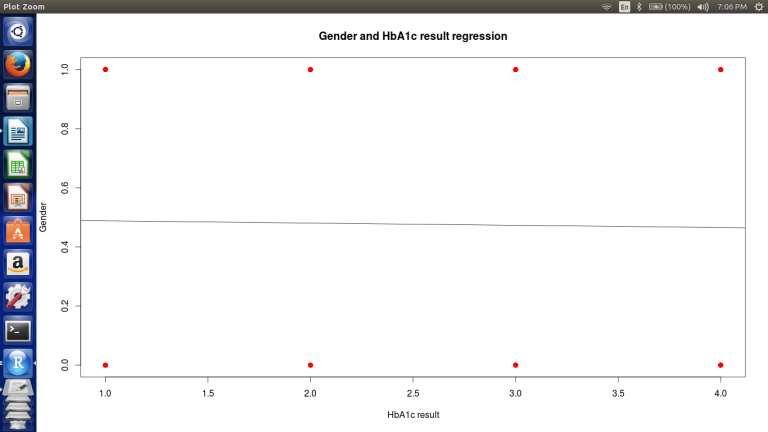
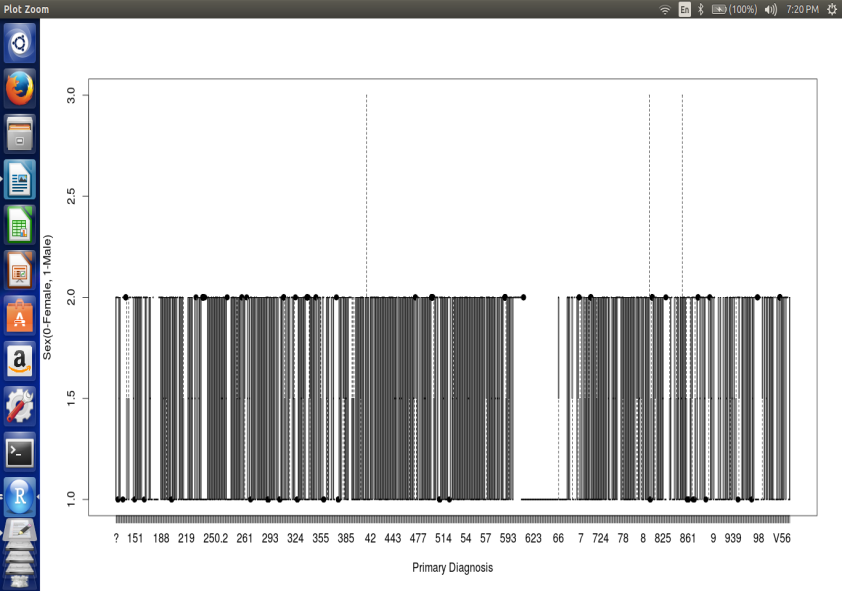


Fig. 2 A linear regression showing relation between gender and HbA1c test.

­In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. This article covers the case of binary dependent variables—that is, where it can take only two values, such as pass/fail, win/lose, alive/dead or healthy/diseased. Cases with more than two categories are referred to as multinomial logistic regression, or, if the multiple categories are ordered, as ordinal logistic regression.

 Fig. 3 A logistic regression showing relation between gender and primary diagnosis test.

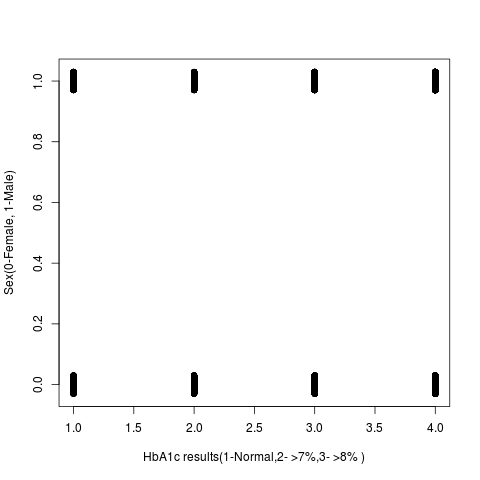


Fig.4 A logistic regression showing relation between gender and HbA1c test.

A chi-squared test, also referred to as **χ²**test (or chi-squaretest), is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true. Chi-squared tests are often constructed from a sum of squared errors, or through the sample variance. Test statistics that follow a chi-squared distribution arise from an assumption of independent normally distributed data, which is valid in many cases due to the central limit theorem. A chi-squared test can then be used to reject the hypothesis that the data are independent.

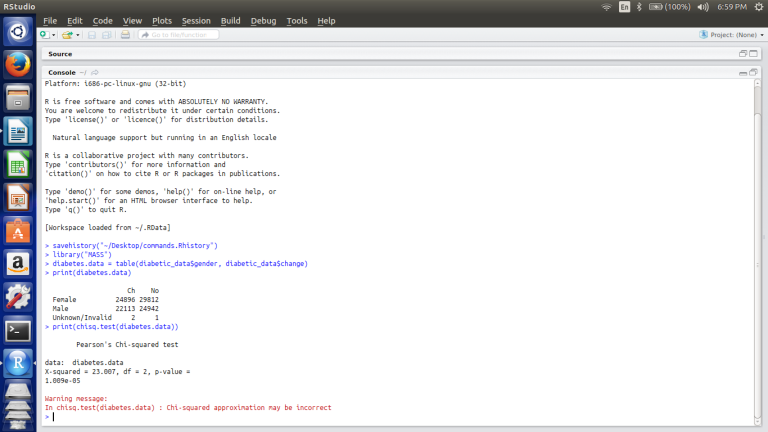


Fig. 5 A chi squared test on two variables gender and change in medication

1. **Results**

The measurement of HbA1c was infrequent, occurring in only 18.4% of encounters where diabetes mellitus was included as an admission diagnosis. Of those in whom the test was ordered, 51.4% were less than 8%. When an HbA1c was not obtained, 42.5% of patients had a medication change during the hospitalization, whereas those providers who ordered the test appear to have been somewhat more responsive as determined by changes in medication.

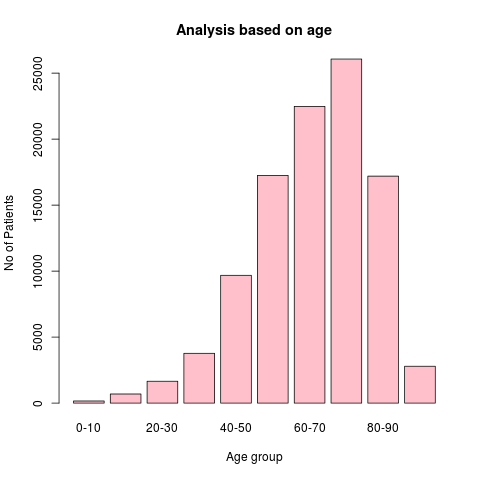


Fig. 6 A bar graph showing distribution of number of patients based on age group in intervals of 10 years

We also analysed the data based on the age groups with the interval of 10 years. We initially made a hypothesis stating that people of age group 60-80 are most affected by diabetes. We used a graphical method to analysis this hypothesis. The bar graph was plotted having age groups on x-axis and number of patients on y-axis.

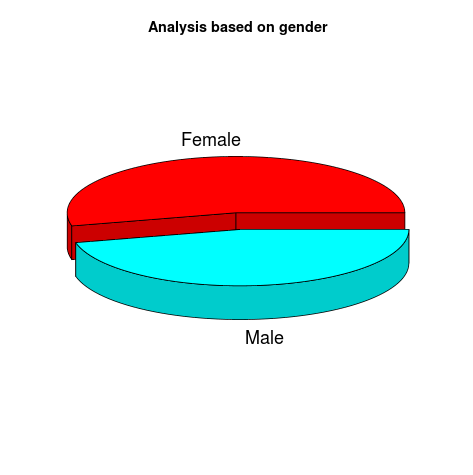


Fig. 7 A 3D pie chart showing distribution based on gender

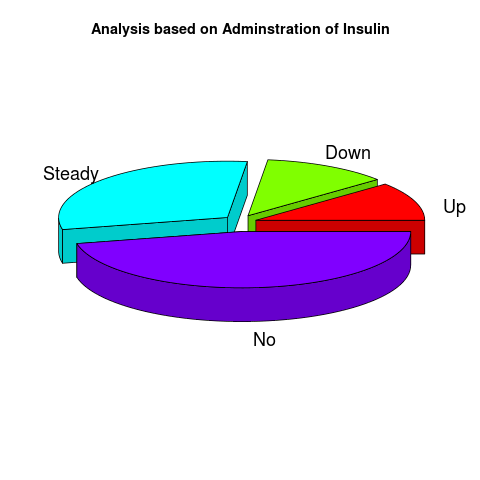


Fig. 8 A 3D pie chart representing analysis of administration of insulin

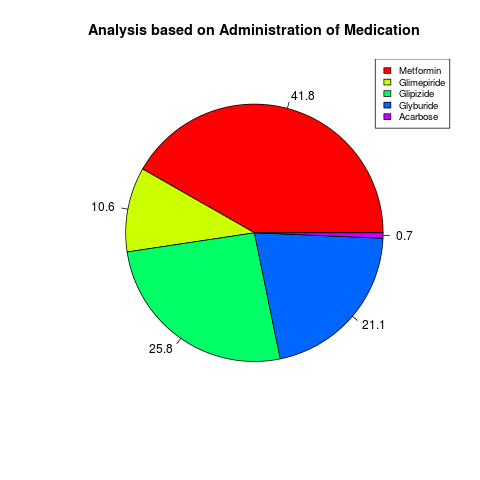


Fig. 9 A pie chart representing the analysis based administration of medication

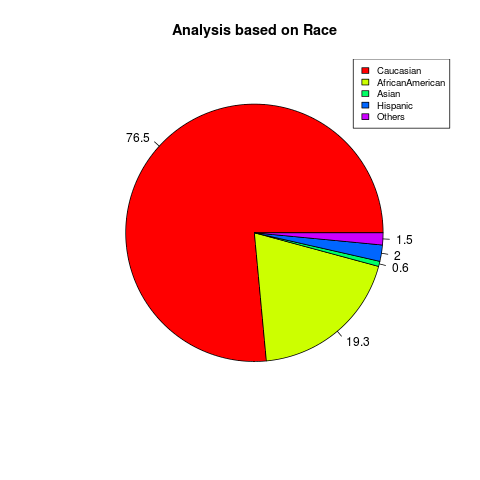


Fig. 10 Analysis based on race

1. **Scope and Future Work**

The future involves the in analyzing the dataset by using the advanced algorithm. It reduces the variations in the prediction and makes the performance faster. It includes removing the deviation, constructing better models with minimum possible deviation in the realistic and predicted data. We recognize that the results from the present analysis represent a preliminary observation with limitations intrinsic to such large health records which in future can be eliminated.

In future, more number of district comparison may takes place and analysis may be done. The future work will reduce the redundant data, improves the performance, effectiveness and accuracy of the search. This extension will give the better result for the performance ratio.

1. **Conclusion**

In conclusion, the decision to obtain a measurement of HbA1c for patients with diabetes mellitus is a useful predictor of readmission rates which may prove valuable in the development of strategies to reduce readmission rates and costs for the care of individuals with diabetes mellitus. For instance, our analysis showed that the profile of readmission differed significantly in patients where Hba1c was checked in the setting of a primary diabetes diagnosis, when compared to those with a primary circulatory disorder. While readmission rates remained the highest for patients with circulatory diagnoses, readmission rates for patients with diabetes appeared to be associated with the decision to test for HbA1c, rather than the values of the HbA1c result.

**References**

1. <http://www.gardenersown.co.uk/education/lectures/R/index.htm>
2. <https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008>
3. G. E. Umpierrez, S. D. Isaacs, N. Bazargan, X. You, L. M. Thaler, and A. E. Kitabchi, “Hyperglycemia: an independent marker of in-hospital mortality in patients with undiagnosed diabetes,” Journal of Clinical Endocrinology and Metabolism, vol. 87, no. 3, pp. 978–982, 2002. View at Publisher · View at Google Scholar · View at Scopus
4. C. S. Levetan, M. Passaro, K. Jablonski, M. Kass, and R. E. Ratner, “Unrecognized diabetes among hospitalized patients,” Diabetes Care, vol. 21, no. 2, pp. 246–249, 1998. View at Google Scholar · View at Scopus
5. S. E. Siegelaar, J. B. L. Hoekstra, and J. H. Devries, “Special considerations for the diabetic patient in the ICU; targets for treatment and risks of hypoglycaemia,” Best Practice and Research: Clinical Endocrinology and Metabolism, vol. 25, no. 5, pp. 825–834, 2011. View at Publisher · View at Google Scholar · View at Scopus
6. A. G. Pittas, R. D. Siegel, and J. Lau, “Insulin therapy for critically ill hospitalized patients: a meta-analysis of randomized controlled trials,” Archives of Internal Medicine, vol. 164, no. 18, pp. 2005–2011, 2004. View at Publisher · View at Google Scholar · View at Scopus
7. A. C. Tricco, N. M. Ivers, J. M. Grimshaw et al., “Effectiveness of quality improvement strategies on the management of diabetes: a systematic review and meta-analysis,” The Lancet, vol. 379, no. 9833, pp. 2252–2261, 2012. View at Publisher · View at Google Scholar
8. M. C. Lansang and G. E. Umpierrez, “Management of inpatient hyperglycemia in noncritically ill patients,” Diabetes Spectrum, vol. 21, no. 4, pp. 248–255, 2008. View at Publisher · View at Google Scholar · View at Scopus
9. R. Vinik and J. Clements, “Management of the hyperglycemic inpatient: tips, tools, and protocols for the clinician,” Hospital Practice, vol. 39, no. 2, pp. 40–46, 2011. View at Google Scholar · View at Scopus
10. K. J. Cios and G. W. Moore, “Uniqueness of medical data mining,” Artificial Intelligence in Medicine, vol. 26, no. 1-2, pp. 1–24, 2002. View at Publisher · View at Google Scholar · View at Scopus
11. A. Frank and A. Asuncion, UCI Machine Learning Repository, University of California, School of Information and Computer Science, 2010.
12. R. M. Bergenstal, J. L. Fahrbach, S. R. Iorga, Y. Fan, and S. A. Foster, “Preadmission glycemic control and changes to diabetes mellitus treatment regimen after hospitalization,” Endocrine Practice, vol. 18, no. 3, pp. 371–375, 2012. View at Publisher · View at Google Scholar
13. D. Baldwin, G. Villanueva, R. McNutt, and S. Bhatnagar, “Eliminating inpatient sliding-scale insulin: a reeducation project with medical house staff,” Diabetes Care, vol. 28, no. 5, pp. 1008–1011, 2005. View at Publisher · View at Google Scholar · View at Scopus
14. H. Anwar, C. M. Fischbacher, G. P. Leese et al., “Assessment of the under-reporting of diabetes in hospital admission data: a study from the Scottish diabetes research network epidemiology group,” Diabetic Medicine, vol. 28, no. 12, pp. 1514–1519, 2011. View at Publisher · View at Google Scholar · View at Scopus