Lauren Mautner
IRB Protocol
CRC Block 6
November 25, 2013

A Diagnostic Assistant for Multiple Myeloma

A. Study Purpose and Rationale

The incidence of multiple myeloma in the United States is roughly 4 to 5 per 100,000, and the disease represents 1% of all malignancies in the US. (1). The disease is slightly more common in men than in women (1.4:1), and is typically a disease of older adults, with a median age of 66 years at diagnosis and less than 10% of all patients younger than 50 years.

The presenting features at the time of diagnosis of multiple myeloma are non specific and most frequently include anemia, bone pain, increased creatinine, fatigue, generalized weakness, global feeling of ill health, dyspnea, hypercalcemia, and weight loss (2,3). Because of this non-specific presentation, the diagnosis of multiple myeloma may often be delayed. For example, in a subset of 127 Dutch patients who initially presented with anemia and bone pain and were ultimately diagnosed with multiple myeloma between 1991 and 1999, the clinician's initial differential diagnosis in 37% of cases did not include multiple myeloma (4).

Delayed diagnosis of malignancies are especially concerning because, in many cases, there is increased likelihood of disease-free survival when diagnosis is achieved at earlier stages of disease (5). One study which specifically looked at initial presentation of multiple myeloma and time until diagnosis showed that patients whose diagnosis was delayed ≥6 months were more likely to have an increased number of complications, more likely to be diagnosed at a later stage (e.g. Durie-Salmon stage III) and reduced disease-free survival, measured both from onset of symptoms and from time of diagnosis (3).

In a subset of diseases that are often associated with delay in diagnosis, signs/symptoms suggestive of the disease are present in medical notes prior to the diagnosis in 25% of cases (6); this may indicate that more timely diagnosis can be made through the analysis of past clinical notes. Manual review of clinical notes for thousands of patients is unrealistic. The purpose of this study is to investigate whether a machine-learning program can analyze past clinical notes and objective patient data (such as lab values and patient weights) to suggest the diagnosis of multiple myeloma prior to the physician's diagnosis. Implications for the development of a diagnostic assistant for multiple myeloma are more timely diagnosis of the disease, with the possibility of preventing disease related complications and increasing disease free survival, as well as fewer medical resources wasted while investigating other less likely diagnoses.

This study is therefore based on the following hypothesis: by using natural language processing to interpret patient clinical notes as well as objective patient data, a machine-learning program can suggest the diagnosis of multiple myeloma at an earlier time point than the health care provider.

B. Study Design and Statistical Analysis

The first part of this study is to use a machine-learning program to diagnose MM based on clinical notes and objective patient data. Data will be extracted from AIM clinic records in the clinical records data warehouse to identify patients who have already been diagnosed with MM. The clinical notes and objective data from these patients will be collected and run through a machine learning program that will develop an algorithm for diagnosis of MM based on similarities in patients' clinical notes and objective data. The algorithm will then be tested for accuracy against known diagnoses and the time scale against provider diagnoses.

In specific, out of 10,000 AIM patients, 400 have the ICD9 code for the diagnosis of Multiple Myeloma and have been followed in clinic for at least 5 years prior to the diagnosis. These patients as well as 400 age-matched controls (also AIM patients) will be used for machine learning to develop the algorithm. This model will then be applied to see if it can correctly diagnose myeloma. In order to test the accuracy of the model, a separate group of 241 patients with known myeloma will be used. A 1 group chi square analysis will be used on these 241 patients which will be powered to detect a sensitivity of 90% for our model (using an alpha of 0.05 and power of 0.8). We will use the same statistical model to show the sensitivity of the model at different time points prior to the diagnosis being made by the clinician (e.g. 3 months, 6 months, 1 year prior to diagnosis, and so on).

C-F. Study Procedure, Drugs, Medical Devices, and Questionnaires
There are no procedures, drugs, medical devices, or questionnaires that will be used in this study.

G. Study Subjects
There will be no patients directly involved in this study. The clinical data warehouse, which stores patient medical records, will be accessed through data extraction methods. All patient data extracted will be de-identified electronically.

H. Recruitment of Subjects
There are no subjects to recruit.

I. Confidentiality of Study Data
Patient data that is extracted from the data warehouse will be devoid of identifiers as per HIPAA regulations. The de-identified data will then be given randomly assigned "patient codes" so that there will be no relationship between patient MRN and the random code assigned to the patient. Additionally, data will be stored on encrypted hard drives in password-protected files.

J. Potential Conflict of Interest
None

K. Location of Study
This study will be conducted within the Department of Bioinformatics at CUMC under the supervision of Herbert Chase, MD.

L. Potential Risks
There are no potential risks to patients since only extracted and de-identified patient data will be used, and there are no procedures or interventions on actual human subjects for this study.

M. Potential Benefits
The patients in this study will not directly benefit. Future patients may benefit from earlier diagnosis of multiple myeloma. Benefits may include reduced stress, fewer complications from disease, and longer disease free survival.

N. Alternative Therapies
The only alternative approach would be manual review and analysis of clinical notes and objective patient data, which is not feasible.

O. Compensation to Subjects
None

P. Cost to Subjects
None

References:

1) Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. CA Cancer J Clin 2013; 63:11.
2) Kyle RA, Gertz MA, Witzig TE, et al. Review of 1027 patients with newly diagnosed multiple myeloma. Mayo Clin Proc 2003; 78:21.
3) Kariyawasan CC, Hughes DA, Jayatillake MM, Mehta AB Multiple myeloma: causes and consequences of delay in diagnosis. QJM. 2007 Oct;100(10):635-40. Epub 2007 Sep 10.
4) Ong F, Hermans J, Noordijk EM, Wijermans PW, Kluin-Nelemans JC. Presenting signs and symptoms in multiple myeloma: high percentages of stage III among patients without apparent myeloma-associated symptoms. Ann Hematol. 1995;70:149–152.
5) Riwa M, Reid J, Handley C, Grimwood J, Ward S, Turner K, et al. Less haste more speed: factors that prolong the interval from presentation to diagnosis in some cancers. Fam Pract. 2004;21:299–303
6) Feldman, MJ et al. Presence of key findings in the medical record prior to a documented high-risk diagnosis. JAMA 2012; 19(4): 591-596.