Data Mining in der medizinischen Versorgung und Vorsorge

Werner Dilger, Bianca Loth, Jens Zeidler, Technische Universität Chemnitz
Nanette Schloot, Diabetes-Forschungsinstitut, Düsseldorf

Zusammenfassung

Es werden ein elektronischer Erhebungsbogen zur Erfassung von Daten über Diabetes-Patienten sowie erste Ergebnisse einer Untersuchung von Labordaten aus der Diabetes-Forschung beschrieben. Der Erhebungsbogen bietet erstmals die Möglichkeit zur systematischen und umfassenden Erhebung relevanter Daten, die dann sofort zur elektronischen Bearbeitung zur Verfügung stehen. Da der Bogen bisher nur in Papierform genutzt wird, sind die Daten momentan noch nicht verwendbar. Statt dessen werden Labordaten analysiert. Diese haben die Eigenschaft, dass sie sehr detailliert sind und deshalb die Möglichkeit zu einer vielseitigen Analyse bieten. Sie zielen auf die Früherkennung von Risikopersonen ab und bieten damit die Möglichkeit für präventive Maßnahmen. Diese erscheinen als die wirkungsvollste Möglichkeit zur Bekämpfung der Volkskrankheit Diabetes. Bereits die ersten Analyseversuche haben überraschende Ergebnisse gebracht, die aber durch weitere Untersuchungen zu erhärten sind.

1. Einleitung

Diabetes ist in den Industriestaaten zu einer Volkskrankheit geworden. Sie stellt damit einen bedeutenden Kostenfaktor im Gesundheitswesen dar. Allein in Deutschland hat die Diabetes-Behandlung hat im Jahr 1998 nach einer Schätzung in der CODE-2-Studie ca. 31,4 Mrd. DM gekostet. Es ist bekannt, dass die Krankheit verzögert und Folgeerkrankungen wie Blindheit oder Herzinsuffizienz bzw. Folgemaßnahmen Amputationen aufgehalten oder vermieden werden können, wenn eine regelmäßige Kontrolle und Behandlung durchgeführt wird. Insbesondere gibt es eine Reihe von Screening-Parametern, deren Verlauf Aufschluss über den Zustand der Patienten und die rechtzeitige Einleitung von Maßnahmen erlauben. Vgl. hierzu [Jörgens 98].

Durch Untersuchung der Entwicklung der Screening-Parameterwerte mit Verfahren des Data Mining, speziell der Zeitreihenanalyse, können präzisere oder auch neue Erkenntnisse gewonnen werden. Hier sind vor allem einfach zu erhebende Parameterwerte wichtig, die für ein frühzeitig beginnendes und langfristiges Screening geeignet sind. Bei einer automatisierten Kontrolle der Werte können rechtzeitig Warnmeldungen ausgegeben werden (vgl. entsprechende Überlegungen in der Telemedizin). Mit derartigen präventiven Maßnahmen können beträchtliche Kosteneinsparungen erreicht werden.

In den folgenden Abschnitten wird zunächst die Erfassung von Patientendaten mit Hilfe spezieller Formulare beschrieben. Die systematische Sammlung dieser Daten ist eine wesentliche Voraussetzung für die Analyse einer relevanten Menge von Datensätzen mittels Data Mining-Verfahren. Datensätze dieser Art existieren derzeit noch nicht in größerem Umfang, weil die Voraussetzungen für ihre Erhebung fehlten. Für anderweitig erhobene Datensätze konnte aber bereits eine erste Analyse durchgeführt werden.

2. Die Erfassung von Diabetes-Daten

Am Beispiel von Daten über Diabetes-Erkrankungen wird dargestellt, welche Bedeutung die systematische Erfassung und intelligente Auswertung von Patientendaten für die Versorgung und insbesondere auch für die Vorsorge haben kann. Formulare für die systematische Erfassung wurden in Papierform gemeinsam von Ärzten des Klinikums Chemnitz und des Diabetes-Forschungsinstituts an der Heinrich-Heine-Universität in Düsseldorf entwickelt. Nach dem Vorbild der Papierform wurden an der Technischen Universität Chemnitz und bei der Firma Prudential Systems Software GmbH elektronische Varianten der Formulare erstellt [Gericke et al. 00].

Die Erfassung der Daten erfolgt mittels eines Aufnahmebogens und eines Fragebogens für Diabetiker. Sie erlauben eine vollständige und systematische Erfassung aller Patientendaten in formatierter Gestalt. Das gilt sowohl für die Papierfassung als auch für die elektronische Form der Formulare. Die elektronische Version hat darüber hinaus einige Vorteile: Sie erhöht die Flexibilität und den Komfort der Eingabe, sie kann als Eingaben erzwingende Checkliste verwendet werden, es können Plausibilitätstests durchgeführt werden und die Daten sind sofort für die weitere Auswertung verwendbar.

Abbildung 1 zeigt die Startseite des elektronischen Aufnahmebogens, Abbildung 2 eine typische Seite aus dem Bogen und Abbildung 3 eine Seite für die graphische Eingabe von Markierungen.

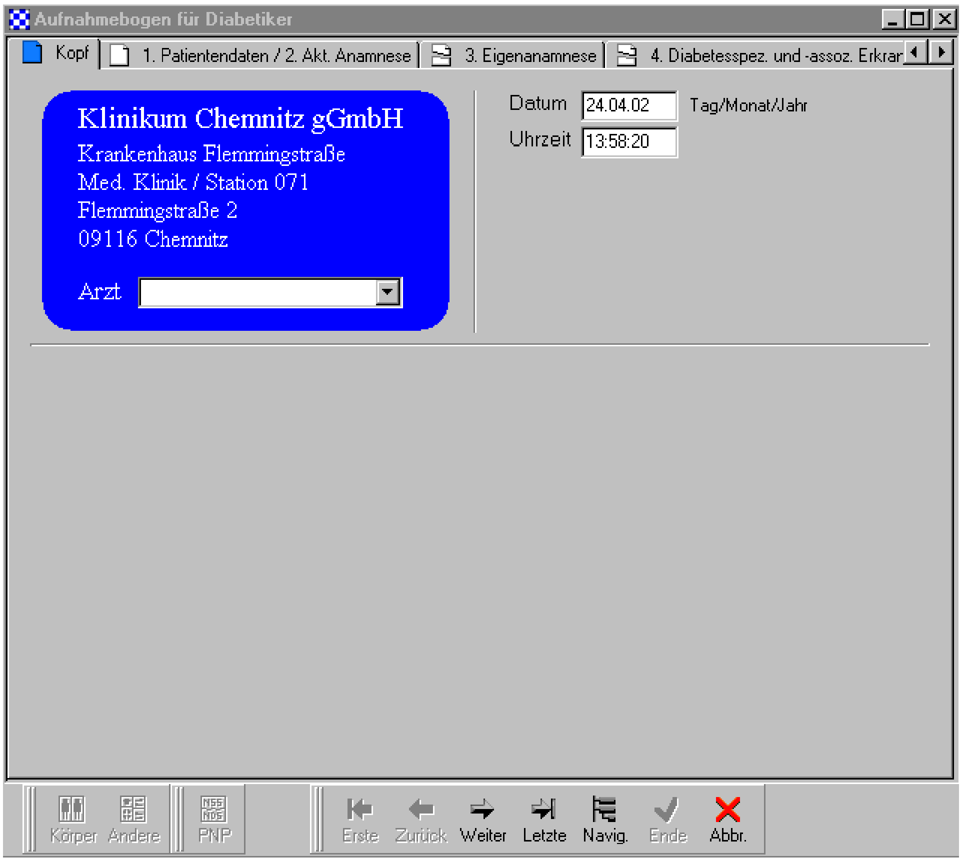

Abbildung 1: Startseite des elektronischen Aufnahmebogens

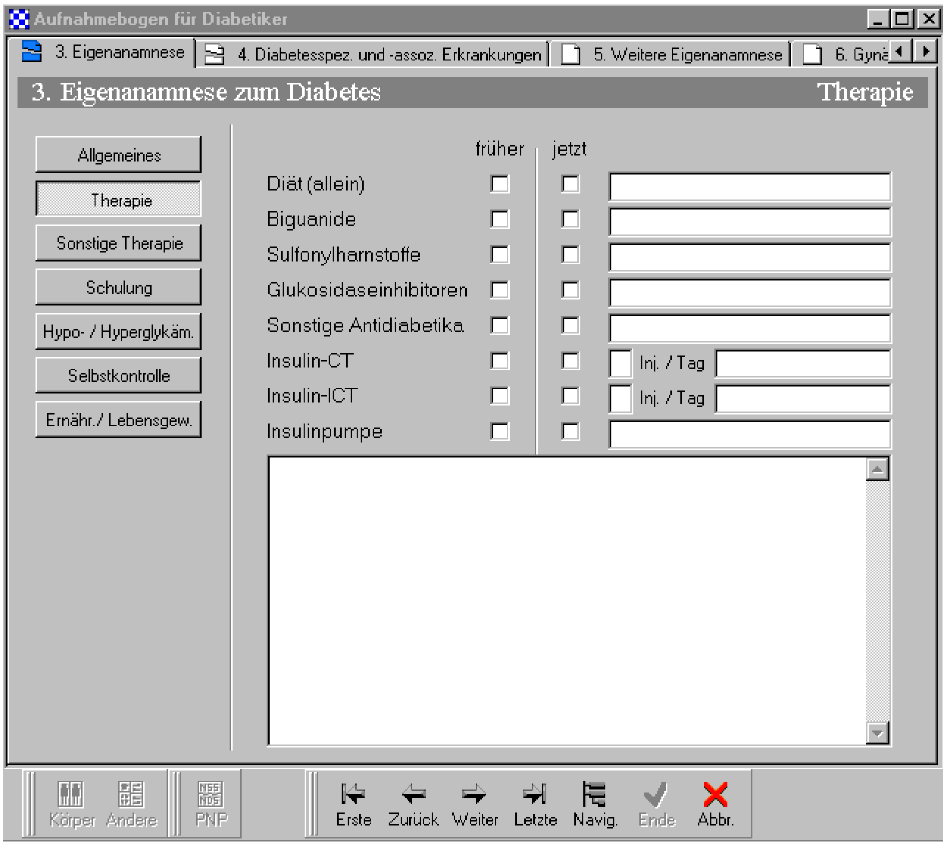

Abbildung 2: Erfassung der Daten zur Eigenanamnese, speziell Therapie

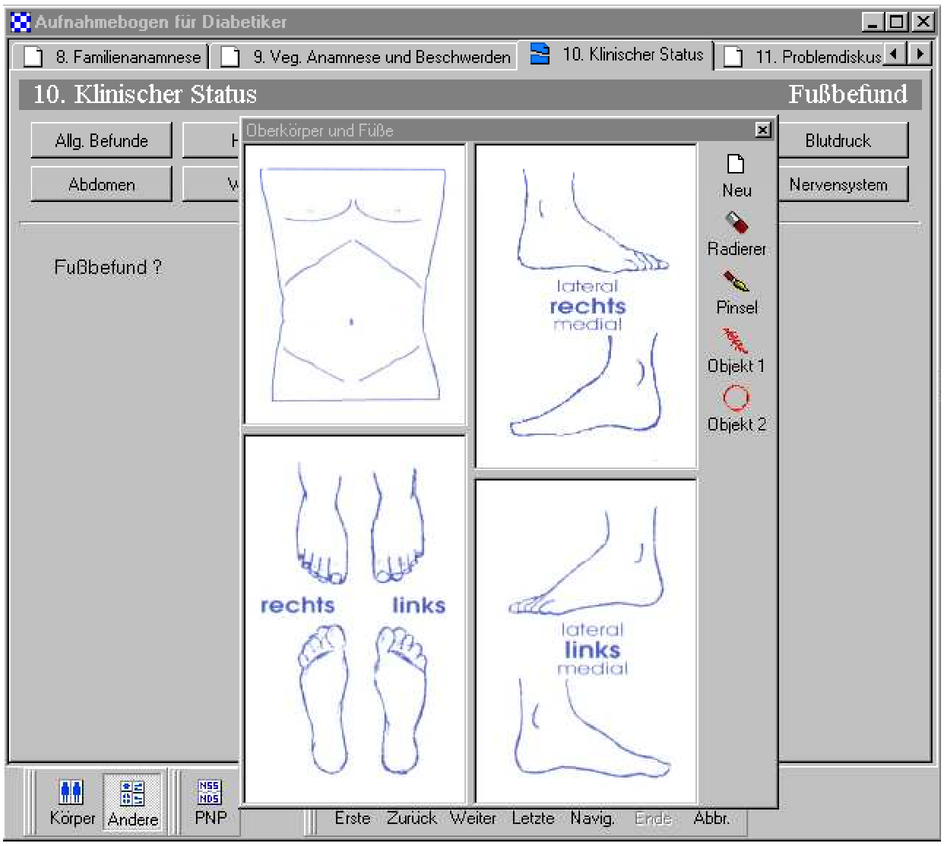

Abbildung 3: Erfassung der Daten zum klinischen Status, graphische Eingabe auf den Skizzen möglich

Wie Abbildung 2 zeigt, ist außer formatierter Eingabe durch Anklicken auch Eingabe von Freitext möglich. Durch die graphische Eingabe mit den angegebenen Werkzeugen können bestimmte Stellen an den Skizzen markiert werden, an denen bei einem Patienten auffällige Veränderungen festgestellt wurden. Wenn Ja/Nein-Fragen zu beantworten sind, ist es typisch, dass nur bei positiver Antwort ein zusätzliches Fenster geöffnet wird, in das weitere Angaben einzugeben sind.

3. Probleme bei der Analyse von Patientendaten

Data Mining-Verfahren ermöglichen eine breit gefächerte Auswertung großer Datenmengen nach unterschiedlichen Kriterien. Insbesondere erlauben sie die Erhebung neuen Wissens, das implizit, aber für Menschen nicht erkennbar, in den Daten enthalten ist („Knowledge Discovery"). Typische Anwendungen im medizinischen Bereich sind Untersuchungen zu Zusammenhängen zwischen Diagnosen und Therapien, zwischen Diagnosen und Verweildauern in Krankenhäusern und über die zeitliche Entwicklung von Parameterwerten, die für bestimmte Krankheitsbilder charakteristisch sind.

Durch die Diabetes-Formulare ist es möglich detaillierte Daten in großem Umfang von einzelnen Patienten zu erheben. Sie können außerdem noch durch Labordaten ergänzt werden. Die Datensätze sind hochdimensional und sehr komplex, sie enthalten Werte aller möglichen Typen und der Umfang verschiedener Datensätze kann sehr unterschiedlich sein. Der Grund dafür ist, dass abhängig von der Beantwortung bestimmter Ja-/Nein-Fragen zusätzliche Angaben angefordert werden oder nicht. Wird z.B. die Frage nach Dyslipoproteinanämie positiv beantwortet, dann wird zusätzlich gefragt, ob und wenn ja, welche Medikamente genommen werden und welche Werte bestimmte Blutinhaltsstoffe wie Triglyceride, Cholesterin, LDL-Cholesterin und HDL-Cholesterin mit Zeitangabe der Messung haben. Außerdem kann noch Freitext eingegeben werden. Bei negativer Antwort erscheinen keine entsprechenden Eingabefenster, d.h. Informationen dieser Art entfallen.

Bei der Analyse der Datensätze müssen die fehlenden Werte ergänzt werden. Dies geschieht abhängig vom Typ. Bei numerischen Werten kann der Wert 0 sinnvoll sein, wenn das Fehlen eines Werts bedeutet, dass für das betreffende Attribut nichts nachgewiesen werden konnte. Ansonsten, wie auch bei Attributen anderen Typs, kommt der Wert NIL in Frage. Die Behandlung der Freitexte stellt ein besonderes Problem dar. Es ist anzunehmen, dass auch sie verwertbare Informationen enthalten. Um sie nützen zu können müssten sie durch geeignete Verfahren analysiert werden. Dafür kommen Textmining-Verfahren in Frage, jedoch wird diese Aufgabe zunächst zurück gestellt. Ein offenes Problem ist weiterhin die Einbindung von Labordaten, die ebenfalls regelmäßig, aber gesondert erhoben und in einer anderen Datenbank abgelegt werden. Diese Einbindung ist momentan aus organisatorischen Gründen schwer zu realisieren. Schließlich ist noch zu beachten, dass Diabetes-Patienten Langzeitpatienten sind, die immer wieder zur Untersuchung kommen und behandelt werden müssen. Eine wichtige Information steckt daher im zeitlichen Verlauf der erhobenen Daten. Um sie zu erheben sind Verfahren zur Zeitreihenanalyse einzusetzen, wobei aber die üblichen Schwierigkeiten beim Umgang mit medizinischen Daten beachtet werden müssen, vgl. [Tsumoto 99].

Die Auflistung der Probleme zeigt, dass für die adäquate Behandlung der hier vorliegenden oder zu erhebenden Daten ein Bündel von Verfahren verwendet werden muss, eine uniforme Prozedur würde die Möglichkeiten der Analyse einschränken. Zum gegenwärtigen Zeitpunkt werden Daten bereits regelmäßig erhoben, allerdings noch mit der Papierform der Formulare. Die breite Nutzung von Notebooks oder Notepads würde Investitionen erfordern, die derzeit nicht möglich sind. Eine Prozedur zum Einscannen der Daten ist in Vorbereitung so dass die Daten in elektronischer Form bereit gestellt werden können.

4. Analyse von Labordaten

Eine erste Analyse von Daten konnte bereits jetzt durchgeführt werden. Dazu wurden vom Diabetes-Forschungsinstitut Labordaten zur Verfügung gestellt. Die Datensätze sind vollständig anonymisiert, es ist nur die Zugehörigkeit von Patienten zu einer von vier verschiedenen Gruppen angegeben. Jeder Datensatz besteht aus 15 Attributen, die bis auf das ordinale Gruppen-Attribut alle von numerischem Typ sind und die Werte verschiedener Blutinhaltsstoffe wiedergeben. In einer zweiten Variante der Daten wurden neue Attribute durch Division von je zwei Attributwerten gebildet. In dieser Form sind die Daten für die Fachleute aussagekräftiger. Man erhält so insgesamt 183 Attribute. Die Daten sind fast vollständig, so dass eine Vorverarbeitung entfällt. Die Zahl der Datensätze ist klein, nämlich nur 150, so dass statistisch relevante Aussagen nicht möglich sind, jedoch ist die Zahl für erste Untersuchungen der Besonderheiten von Diabetes-Daten ausreichend.

Die erste Analyse der Daten wurde „blind“ durchgeführt im Sinne eines Knowledge Discovery-Prozesses ohne Vorinformationen. Es ging darum, Zusammenhänge zwischen der Gruppenzugehörigkeit und bestimmten Inhaltsstoffen zu entdecken. Zunächst wurde das ursprüngliche Vierklassenproblem in ein Zweiklassenproblem umgewandelt, indem die Gruppe 1 als eine eigene Klasse betrachtet wurde und die übrigen drei Gruppen 2, 3 und 4 zu einer zweiten Klasse zusammengefasst wurden. Es wurden zwei Entscheidungsbäume mit Zielmerkmal *Klasse 1* erstellt Die Erstellung des ersten orientierte sich vorrangig an den Attributen mit der Namensform „IL_##IL_##“. Etwa die Hälfte der numerischen Attribute hat diese Form, so dass anzunehmen war, dass diese Attribute eine wichtige Rolle spielen. Die Erstellung des zweiten orientierte sich an den Attributen mit der Namensform „M...“. Die Wertebereiche dieser Attribute lassen sich in naheliegender Weise in zwei oder drei Teilmengen diskretisieren. Die Entscheidungsbäume wurde mit dem Prudential Discoverer erstellt. Besonders interessant ist der zweite Baum. Er ist in Abbildung 4 dargestellt.

Es fällt auf, dass bereits bei Tiefe 5 fast nur homogene Mengen vorliegen und nur 8 Attribute benötigt wurden, obwohl eine große Zahl von Attributen zur Verfügung steht. Das ist natürlich zum Teil der geringen Zahl der Datensätze geschuldet, aber es deutet auch darauf hin, dass die für die Abspaltung der Gruppe 1 benötigten Attribute einen besonderen Bezug zu dieser Gruppe haben müssen. Bei genauerer Betrachtung des Entscheidungsbaums sieht man, dass bereits auf Ebene 3 mit Hilfe von 3 Attributen mehr als die Hälfte der Beispiele aus Gruppe 1, nämlich 19, aus Gruppe 1 separiert sind. Mit 5 Attributen kann man auf dieser Ebene bereits 25, also fast ¾ der Beispiele separieren.

Der medizinische Hintergrund der Daten ist folgender: Die Probanden der Gruppe 1 sind die gesunden Kontrollpersonen ohne Verwandtschaft zu Diabetikern. In Gruppe 2 sind die Verwandten ersten Grades ohne Diabetes-assoziierte Antikörper ('low risk'). Zu Gruppe 3 gehören die Verwandten ersten Grades mit mindestens 2 Diabetes-assoziierten Antikörpern ('high risk'). Die Gruppe 4 besteht aus den Patienten mit Diabetes mellitus bei Diagnose. Die Ergebnisse der ersten Analyseschritte besagen bei vorsichtiger Interpretation, dass die Konzentration der Zytokine im Blut eine Rolle für die Diabetes-Diagnose spielt, was nicht erwartet worden war.

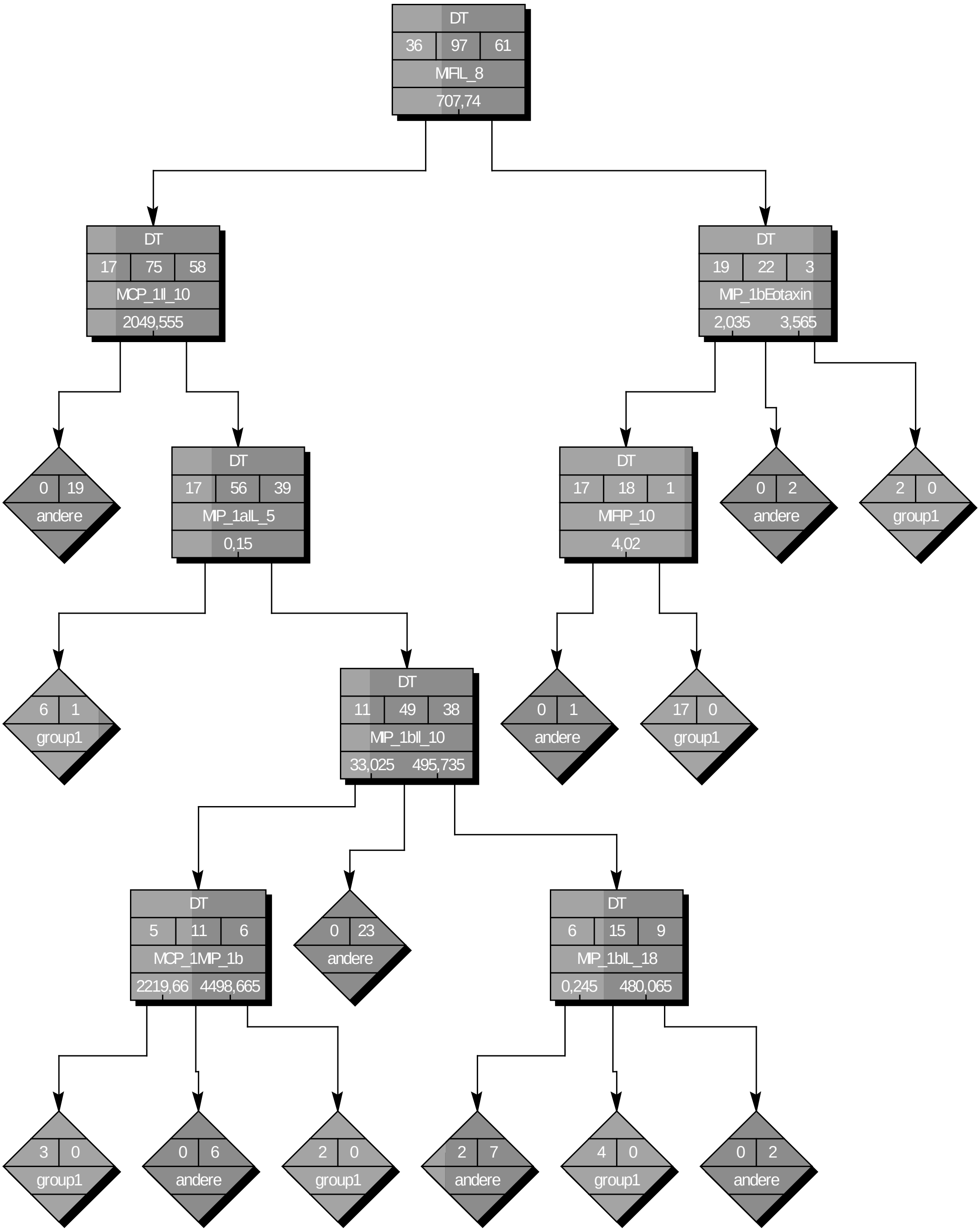

Abbildung 4: Entscheidungsbaum für Labordaten.

5. Weitere Arbeiten

Es ist klar, dass dies erst der Anfang der Untersuchungen ist. Insbesondere an den Labordaten lassen sich noch weitere Aspekte betrachten. Die nächsten Schritte werden sein, dass andere Klassenbildungen vorgenommen werden, etwa erscheint eine Zusammenfassung der Gruppen 1 und 2 sowie der Gruppen 3 und 4 sinnvoll. Die bisherigen Analysen liefen nur mit achsenparallelen Trennungen der Daten. Andere Trennungen, insbesondere solcher höherer Ordnung, sollen untersucht werden. Sie könnten weitere interessante Zusammenhänge aufdecken.

So bald die systematische Erhebung der Daten mit den Formularen angelaufen ist und ein gewisser Bestand erreicht ist, kann auf diesen Daten weiter gearbeitet werden. Es ist dabei zu beachten, dass die routinemäßig erhobenen Labordaten nicht so umfangreich sind wie die hier betrachteten, da deren Ermittlung zum Teil sehr aufwändig ist. Dafür können andere Angaben wichtige Hinweise liefern, insbesondere die über einen längeren Zeitraum hinweg erhobenen Daten.

Der bisher ermittelte Befund bestätigt die Bedeutung der Früherkennung dieser Krankheit. Der Diabetes kann man wirkungsvoll letztlich nur durch präventive Maßnahmen begegnen, deshalb ist die Früherkennung im Rahmen der Vorsorgeuntersuchung wichtig. Wir hoffen, durch intelligente Datenanalyse dazu einen Beitrag leisten zu können.

6. Literatur

[Gericke et al. 00]
Gericke, G., Köhler, G., Moser, W., Grau, C., Bauch, K., Koschinsky, T., Gries, F.A. und die DIADOQ-Gruppe: Funktionelles standardisiertes Dokumentationssystem für die Diabetikerbetreuung (DIADOQ): Anforderungen und Design, *35. Jahrestagung der Deutschen Diabetes-Gesellschaft (DDG)*, München, 2000, *Endocrinologie & Diabetes, Suppl 1/00*, p. 198.

[Jörgens 98]
Jörgens: Vorsorgeuntersuchung zur Früherkennung der Folgeschäden bei Diabetes mellitus. Modellprojekt in Wolfsburg. *Diabetes und Stoffwechsel*, 1998, pp. 25 – 28.

[Tsumoto 99]
Tsumoto, S.: Rule discovery in large time-series medical databases. *Proceedings of PKKD-99*, Prague, 1999, pp. 23 – 31.