

IT-Werkzeuge zur Auswertung großer labordiagnostischer Datensätze

Georg Hoffmann

ZUSAMMENFASSUNG

Im April 2010 genehmigte die DGKL ein Forschungs- und Softwareentwicklungsprojekt zum Thema [Data Mining](#) (1). Das Hauptziel war die Erkennung von potenziell interessanten Strukturen und Zusammenhängen in klinischen Datensätzen, ein wichtiges Nebenziel die Normalisierung von Laborwerten auf Mittelwert 0 und Standardabweichung ± 1 .

Nach einem Jahr Entwicklungsarbeit ist ein Projektstatus erreicht, in dem zwei Programme (*Trillium Reader* und *Trillium Explorer*) kostenlos zur Verfügung gestellt werden können. Sie sind unter *MS Excel* auf jedem PC ohne spezielle Systemvoraussetzungen lauffähig. Datensätze können mit dem „Reader“ aus Textdateien eingelesen und mit dem „Explorer“ analysiert werden. Die vorliegende Arbeit konzentriert sich - unter Verzicht auf publizierte Formeln und Algorithmen (2, 3) - auf die Darstellung des Arbeitsablaufs und stellt ein Praxisbeispiel aus der Nierendiagnostik vor, bei dem das Programm einen bislang zu wenig beachteten Zusammenhang zwischen Entzündung und tubulären Schäden aufdeckte.

Der *Reader* kann unter www.trillium.de mit dem Menüpunkt *Software* direkt aus dem Internet geladen werden, der *Explorer* ist für den Einsatz in DGKL-Projekten beim Verfasser abrufbar (hoffmann@trillium.de). Mit dem *Reader* aufbereitete Datensätze können zur Auswertung an die Arbeitsgruppe Bioinformatik der DGKL eingesandt werden.

DANKSAGUNG

Der Autor dankt der Stiftung für Pathobiochemie und Molekulare Diagnostik der DGKL für die Förderung des Forschungsprojekts. Ein weiterer Dank gilt Prof. P. Cramer und Dr. A. Tresch (Genzentrum der Universität München), Prof. W. Hofmann und Dr. von Meyer (Klinikum München) sowie Prof. M. Vogeser (Klinikum Großhadern der Universität München) für die Überlassung der hier gezeigten Datensätze.

Korrespondenzadresse

Prof. Dr. med. Georg Hoffmann
Trillium GmbH, Hauptstraße 12b, 82284 Grafrath, www.trillium.de

PROJEKTbeschreibung

Der Antrag bestand aus zwei Teilen:

- 1) Entwicklung einer Software in zwei Stufen (MS Officeprogramme, Internetprogramm)
- 2) Experimenteller Einsatz für die explorative Auswertung simulierter und echter Daten.

Der Schwerpunkt des Projekts liegt in der fachgerechten Aufbereitung von Laborwerten und anderen klinischen Daten. Dazu gehören die flexible Übernahme aus unterschiedlichen Quellformaten (aktuell .txt, .csv und .xls), die Integration von Referenzintervallen sowie die Transformation der Absolutwerte, das heißt die methodenunabhängige Normalisierung auf eine einheitliche Skala mit dem Referenzmittelwert 0 und der Standardabweichung 1. Da diese Schritte zeitaufwändig und fehlerträchtig sind, wurde vor allem auf die Entwicklung einfach bedienbarer „Assistenten“ für eine intuitive Benutzerführung Wert gelegt (Abb. 1).

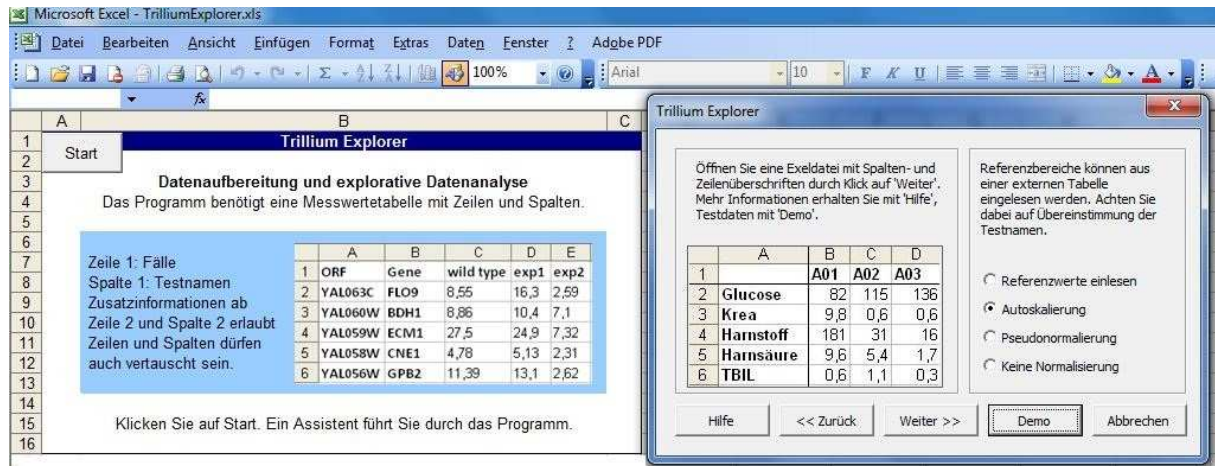


Abb. 1: Aufruf des Programms *Trillium Explorer* aus MS Excel heraus (oben). Nach Klick auf den Startknopf links oben öffnet sich ein Formular im Stil eines MS Windows-Assistenten zur Ablaufsteuerung. Es beinhaltet eine Hilfefunktion und die Simulation von Demo-Daten. Links ist eine typische Datentabelle aus der Molekularbiologie, rechts aus der Labormedizin zu sehen. Beide können in identischer Weise verarbeitet werden.

Die Programmierung erfolgte in MS Excel unter Visual Basic for Applications (VBA). Da das ebenfalls geplante Internetprogramm (s.o.) von den Gutachtern zurückgestellt wurde, liegt das Schwergewicht des vorliegenden Berichts auf der Bedienung des Excelprogramms und der Auswertung klinischer Daten. In Kooperation mit dem Softwarehersteller Neumann & Kindler (www.labcore.de) soll der *Trillium Reader* zu einem eigenständigen Produkt mit HL7-Schnittstellen für LIS und KIS weiterentwickelt werden (4). Interessierte Arbeitsgruppen der DGKL erhalten eine *Community Version* zu Testzwecken kostenlos.

TRILLIUM READER

Mit *MS Excel* kann man Textdateien oft nur mit erheblichem Zusatzaufwand korrekt einlesen. Typische Probleme sind das Überschreiten der maximalen Spaltenzahl von 256, die Deutung von Zahlen mit Dezimalpunkt als Text oder die fehlerhafte Umwandlung von Zahlen- und Textangaben in Datumsformate. Abb. 2 zeigt dies am Beispiel von Genexpressionsprofilen (5): *Excel* deutet „1.82“ als „Jan 82“ und den Zeitpunkt „06-12“ (Minuten) als „06. Dez.“, der *Trillium Reader* liest diese Werte korrekt ein.

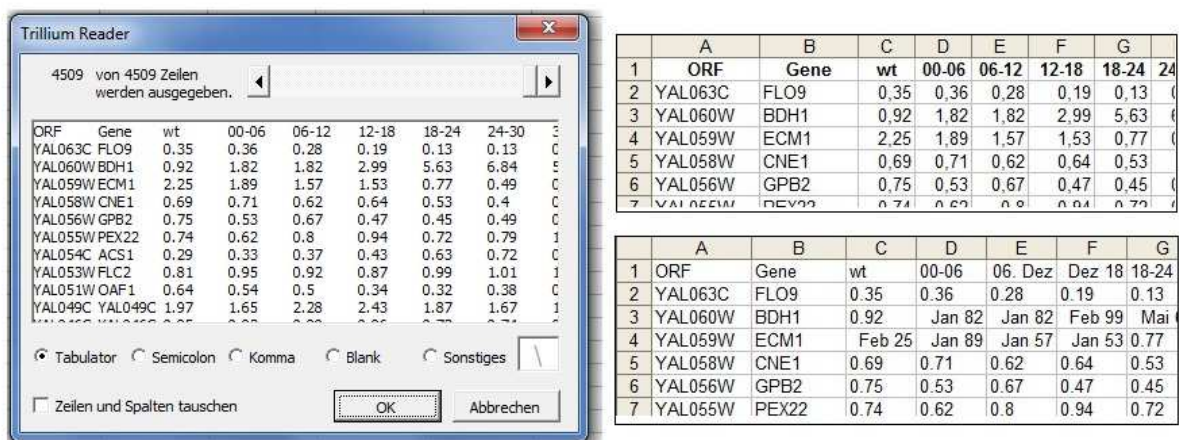


Abb. 2: Der *Trillium Reader* liest Textdateien mit TAB, Semicolon und anderen Trennzeichen (links) in MS Excel ein (rechts oben). Zahlen mit Dezimalkomma wie auch Dezimalpunkt werden korrekt erkannt. Typische Microsoft-Probleme, wie die unerwünschte Umwandlung von Texten in Datumsformate (rechts unten) werden vermieden.

Anstelle der Normalisierung mit Referenzwertetabellen ist es auch möglich, Richtwerte aus den Daten selbst schätzen zu lassen. Je nach Anzahl und Verteilung der Werte verfolgt das Programm dabei selbsttätig drei verschiedene Strategien:

1. Stimmen Median und Mittelwert gut überein, so wird als klassisches Streuungsmaß für symmetrische Verteilungen die Standardabweichung berechnet.
2. Ist die Verteilung links- oder rechts-schief und die Wertezahl ausreichend, dann wird die zweifache Standardabweichung aus dem steileren Kurvenast geschätzt, indem man die Differenz aus Median und dem entsprechenden Quantil (1/6 oder 5/6) bildet.
3. Bei schiefen Verteilungen und geringer Wertezahl wird eine logarithmische Pseudonormalisierung (2, 3) angeboten. Hierfür benötigt man nur einen einzigen Referenzwert (z. B. Median) ohne Streuungsmaß.

Bei allen drei Verfahren entsteht letztlich durch Transformation ein Datensatz mit dem mittleren Referenzwert 0 und einer Schwankungsbreite von etwa -5 bis +5. Entsprechend viele Farben - von dunkelblau bis dunkelrot - werden den einzelnen Feldern zugeordnet. Der „Farbwert“ wird auf zwei Kommastellen genau beim Darüberfahren mit der Maus angezeigt.

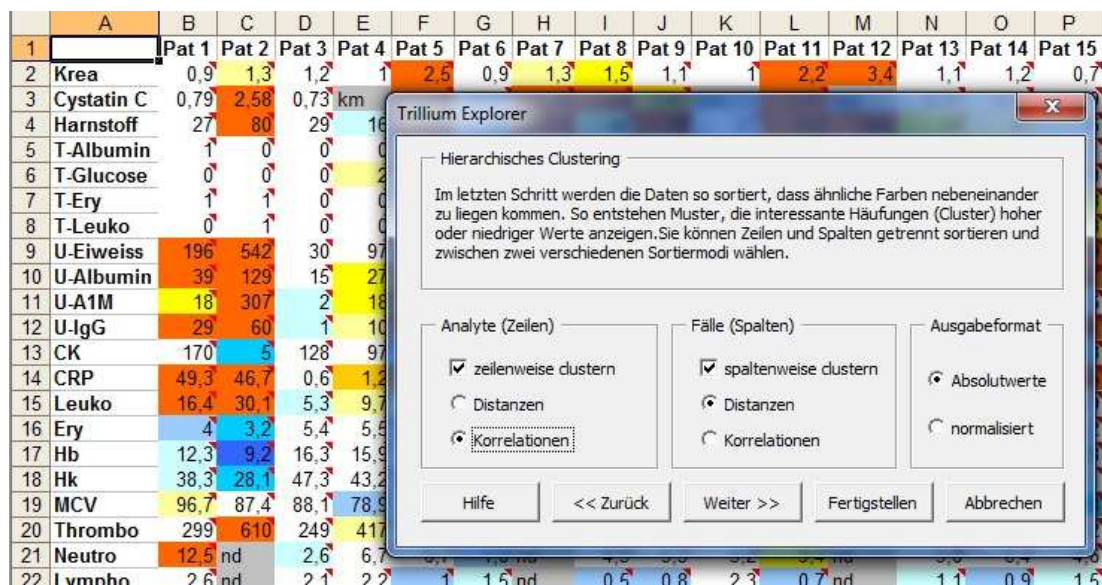


Abb. 4b: Clustering der angefärbten Werte. Der Assistent erlaubt die Auswahl von der Ähnlichkeitsmaße Distanz und Korrelation für Zeilen und Spalten. Im dargestellten Beispiel aus der Nierendiagnostik wird für die Tests ein Korrelationsmaß (Pearson) und für die Patienten ein Distanzmaß (Manhattan) ausgewählt (1, 2).

Nach der Normalisierung erfolgt die eigentliche explorative Analyse. Das Ergebnis ist eine neue Exceltabelle, auf der man zusammenhängende farbige Flächen (*Cluster*) erkennt. Um sowohl den Überblick über den gesamten Datensatz zu erhalten als auch Details beurteilen zu können, bietet das Programm zwei unterschiedliche Ansichten an: eine Detailansicht mit Zahlenwerten und eine Überblicksansicht mit schmalen Spalten, bei der nur die Farben ohne Zahlenwerte zu sehen sind. Das Beispiel in Abb. 5 zeigt beide Ansichten nebeneinander.

Man erkennt in der Übersicht (links) auf einen Blick Datenstrukturen und Zusammenhänge, die mit laborärztlichem Fachwissen aus dem Bereich der Nierendiagnostik in Einklang stehen. So teilt das Programm die Patienten in drei unterschiedliche Schweregrade ein und findet bei mittelschweren und schweren Fällen eine gehäufte Koinzidenz von Niereninsuffizienz und Anämien. Diese ist mit verminderter Erythropoetinbildung in der Niere erklärbar. Ferner erkennt man, dass die Urinproteine bereits in Gruppe 2 erhöht sind, die Serummarker erst in Gruppe 3. Diese höhere Empfindlichkeit wird die Früherkennung von Nephropathien genützt (6).

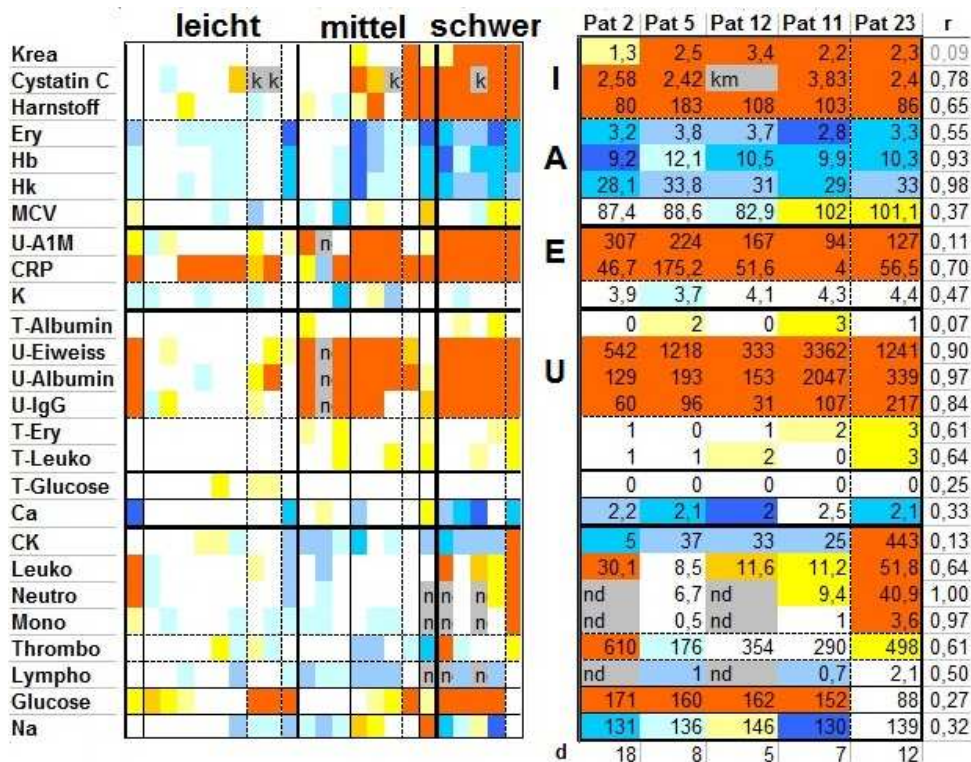


Abb. 5: Ergebnis einer explorativen Analyse aus der Nierendiagnostik (Serummarker, Urinteststreifen, Blutbild, Eiweißdifferenzierung im Urin). Im Gesamtüberblick (links) erkennt man drei Gruppen von Patienten (senkrecht) mit unterschiedlicher Schwere der Erkrankung und vier Gruppen von Markern (waagrecht) für die Insuffizienz (I), Anämie (A), Entzündung (E) und erhöhte Urineiweißausscheidung (U).

In der Detailansicht (rechts) werden zusätzlich die Korrelationskoeffizienten (senkrecht) und Distanzen (waagrecht) der jeweils benachbarten Spalten und Zeilen angegeben. Betrachtet man die Testnamen und zugehörigen Korrelationskoeffizienten genauer, so überrascht, dass das alpha-1-Mikroglobulin im Urin (U-A1M) vom *Explorer* aus der Gruppe der übrigen Urinproteine herausgelöst wurde. Das Programm ordnet diesen Marker dem CRP im Serum zu und ermittelt dafür eine hoch signifikante Korrelation ($p < 0,001$), die der von etablierten Nierenmarkern wie Kreatinin versus Cystatin C nicht nachsteht. Dieser Befund verlangt nach weitergehenden Untersuchungen zur Doppelrolle von A1M bei tubulären Nierenschäden (6) und Entzündungen (7). Es erscheint auf Grund des Data-Mining-Ergebnisses auch lohnend, generell nach Korrelationen zwischen Nieren- und Entzündungsmarkern im Blut und Urin zu fahnden, um entzündlich-toxische Schäden am Tubulussystem besser erklären und früher erkennen zu können.

DISKUSSION UND AUSBLICK

Mit der Programmierung der hier vorgestellten beiden Excel-Programme ist der wesentliche Teil des Projekts nach etwa einjähriger Laufzeit abgeschlossen. Der *Trillium Reader* kann aus dem Internet heruntergeladen werden (www.trillium.de) und befindet sich in der externen Weiterentwicklung, um Daten aus Labor- und Krankenhaus-Informationssystemen via HL7 übernehmen zu können. Der *Trillium Explorer* wird in Details - etwa der Ermittlung vorläufiger Referenzwerte aus Studiendaten (9) - in Kooperation mit der AG Richtwerte der DGKL noch verbessert und ist deshalb vorerst nur per E-Mail anforderbar (hoffmann@trillium.de).

Die Auswertung zahlreicher Datensätze aus klinischen und experimentellen Studien hat den praktischen Nutzen der beiden Werkzeuge bereits bewiesen. Neben dem hier vorgestellten Beispiel aus der Nierendiagnostik fand sich ein überraschender Zusammenhang zwischen niedrigem LDL-C und stark erhöhtem Pro-Insulin (3). Weitere explorative Analysen zur Genexpression (5) und zum Serumpeptidom (10) sind Erfolg versprechend angelaufen.

Die hier vorgestellten Programme eröffnen ein weites Betätigungsfeld für Ergänzungen und Folgeprojekte. So zeigte sich, dass der ursprünglich konzipierte Einsatz für die Analysen von sehr vielen Tests und vergleichsweise wenigen Fällen (siehe Abb. 2) in den bislang eingesandten Datensätzen eher die Ausnahme ist. Häufiger werden Tabellen mit vielen Patienten und vergleichsweise wenigen Tests eingesandt (Abb. 6). Das Vertauschen von Zeilen und Spalten mit entsprechender geänderter Normalisierung ist in Arbeit.

	A	B	C	D	E	F	G	H	I	J	K
1		Insulin	Pro-Insulin	RR_Sys	Bauchum	BMI	LDLC	HDLC	TG	HbA1c	Glucose
2	P029	3,56	3,08	125	99	30,0	116	71,9	68	5,00	113
3	P030	7,01	1,62	140	105	31,7	165	118,0	239	4,69	78
4	P031	13,72	4,32	130	115	37,5	120	62,6	298	5,09	98
5	P032	10,05	5,09	140	104	32,3	77	50,4	83	4,80	102
6	P033	4,18	0,96	150	98	35,0	73	56,1	49	5,40	94
7	P034	2,46	1,47	160	116	31,8	141	63,9	144	5,19	96
8	P036	11,82	0,61	130	109	30,8	111	90,9	137	5,69	111
9	P040	8,77	1,49	125	99	31,0	173	53,1	125	4,40	95

Abb. 6: Klinische Studie mit Laborwerten und klinischen Daten von rund 500 Patienten (3). Die übliche Darstellung (siehe Abb. 1-5) ist in MS Excel nicht realisierbar, da die Zahl der Spalten auf 256 begrenzt ist.

In der Zukunft ist damit zu rechnen, dass die Labordiagnostik immer häufiger Datensätze mit Hunderten oder Tausenden von Tests pro Fall auszuwerten hat, beispielweise bei der Hochdurchsatzsequenzierung von Genomen oder der Massenspektrometrie von Proteomen und Metabolomen. Immer öfter erhalten wir auch dynamische Daten, die zusätzlich zu den Tests und Patienten die Zeit als dritte Dimension einführen. Hier erweist sich die Darstellung in Excel-Tabellenform als unbefriedigend; dynamische Bildfolgen im gif-Format wären besser geeignet. In der DGKL-Arbeitsgruppe Bioinformatik werden solche Ansätze intensiv verfolgt.

LITERATUR

- 1) Hoffmann G. IT-Werkzeuge zur Auswertung großer labordiagnostischer Datensätze. Antrag zur Förderung eines Forschungs- und Entwicklungsvorhabens an die Stiftung Pathobiochemie und Molekulare Diagnostik der DGKL vom 22. März 2010
- 2) Hoffmann G, Zapatka M, Findeisen P, Wörner S, Martus P, Neumaier M. Data-Mining in klinischen Datensätzen. J Lab Med 2010; 34; 227-33
- 3) Hoffmann G, Zapatka M, Findeisen P, Vogeser M. DGKL-Forschungsprojekte zum Data Mining in der Laboratoriumsmedizin. Klin Chem Mitteilungen 2010; 41: 160-4
- 4) Hoffmann G. Software-Entwicklungsprojekt zum Data Mining in der Medizin – Die Suche nach verborgenen Schätzen. Trillium-Report 2010; 8(4): 244
- 5) Miller C, Schwalb B, Maier K et al. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Molecular Systems Biology 2011;7: 458
- 6) Hofmann W, Edel H, Guder W et al. Harnuntersuchungen zur differenzierten Diagnostik einer Proteinurie. Dtsch Ärztebl 2001; 98(12): A-756-63
- 7) Akerström B. Role of alpha1 microglobulin in immune response and inflammation. Folia cytochem histobiol 1992; 30: 183-6
- 8) MicroArray Quality Control (MAQC) project: www.nature.com/nbt/focus/maqc/
- 9) Arzideh F, Brandhorst G, Gurr E et al. An improved indirect approach for determining reference limits from intra-laboratory data bases exemplified by concentrations of electrolytes. J Lab Med 2009; 33:52-66
- 10) Findeisen P, Sismanidis D, Riedl M et al. Preanalytical impact of sample handling on proteome profiling experiments with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Clinical chemistry 2005;51:2409-11